

VisionArena: 230K Real World User-VLM Conversations with Preference Labels

Supplementary Material

8. Acknowledgments

We extend our gratitude to Lianmin Zheng and Ying Sheng for their early discussions and implementation of the vision arena in the LMArena platform. We also thank Tianle Li and Anastasios Angelopoulos for their valuable insights and feedback on both the platform and the manuscript. Most of all, we appreciate the community for their contributions through questions and votes, as this platform and dataset would not be possible without their participation.

9. Interface Details

We implement our interface in Gradio [1]. If a user uploads an image in the first turn, two random VLMs are selected to answer the query. A user can only chat with one image per conversation. As shown in Fig. 10, a user can also select a random image which will select from our preset examples. Note that the user still needs to come up with a query, even for these preset images.

10. Topic Distribution

In Fig. 11 we show the top 20 topic clusters from VisionArena-Chat and WildVision-Chat. The topic clusters are extracted from 50K sampled English conversations from VisionArena-Chat and 37K English conversations (all) from WildVision-Chat. We see that VisionArena-Chat includes more diverse and broad topics especially in the STEM field. Furthermore, WildVision has very specific clusters like "detailed cartoon dog description", "Shock absorber assembly analysis", "rice lead disease identification". Looking at these clusters we see that they contain a large number of duplicate prompts.

11. More Data Stats

We provide further information on language distribution (Fig. 15, Fig. 16), battle counts (Fig. 12, Fig. 14), token count (Fig. 13), turn distribution (Fig. 18, Fig. 19), proportion of refusals (Fig. 20), battle outcome counts (Fig. 17), win rates (Fig. 21), and category overlap (Fig. 24).

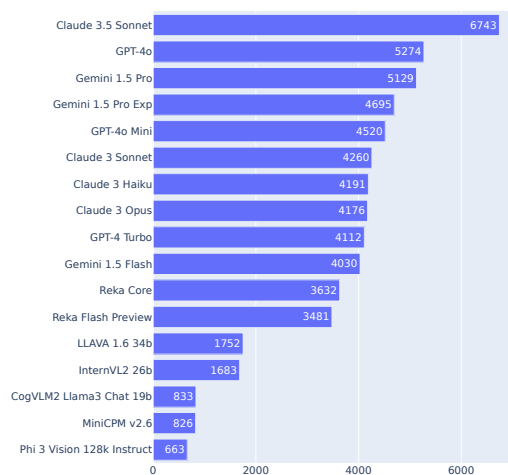


Figure 12. VisionArena-Battle counts per model.

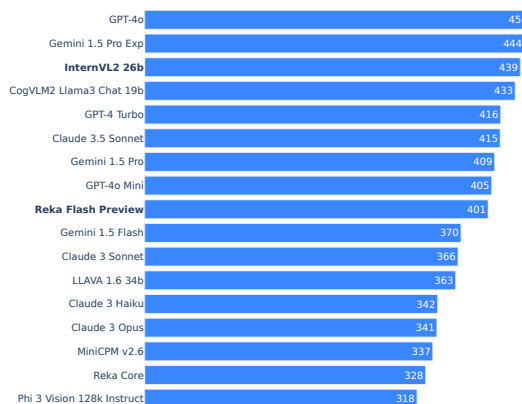


Figure 13. Model token count in VisionArena-Battle. Models in bold see a large decrease in rank when style control is applied.

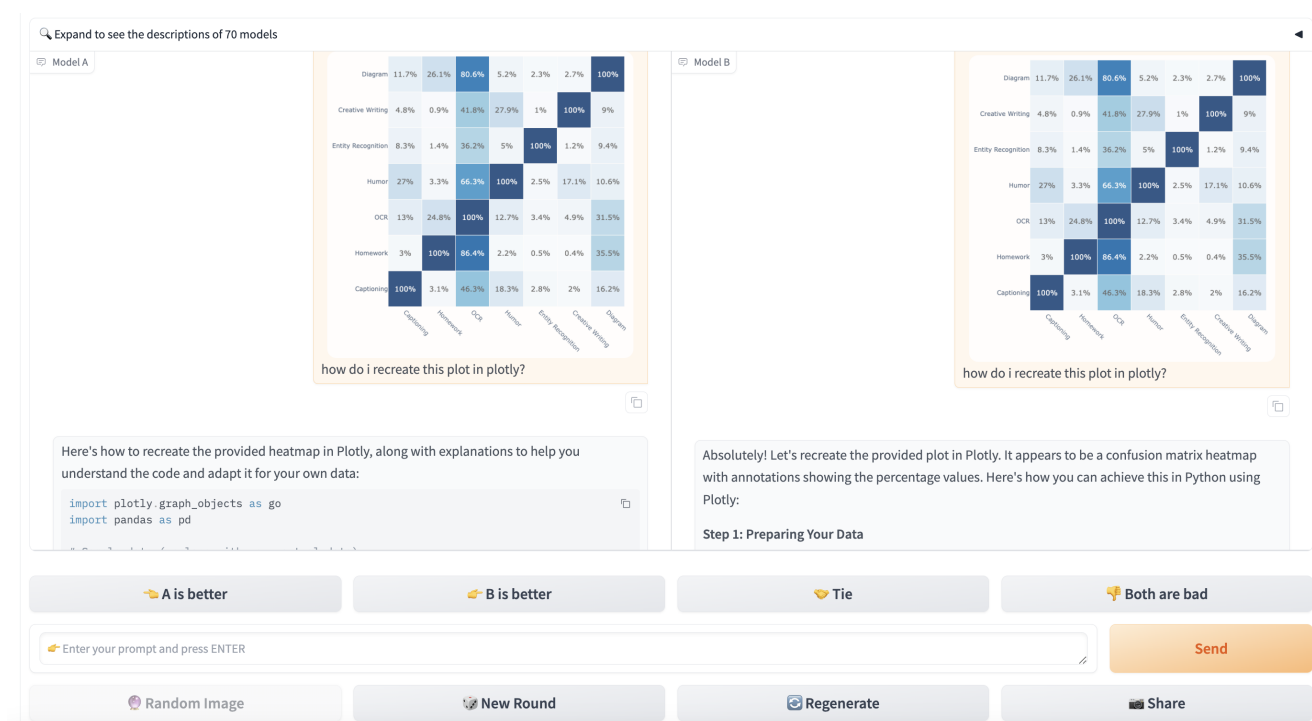


Figure 10. Interface for anonymous side-by-side chat.

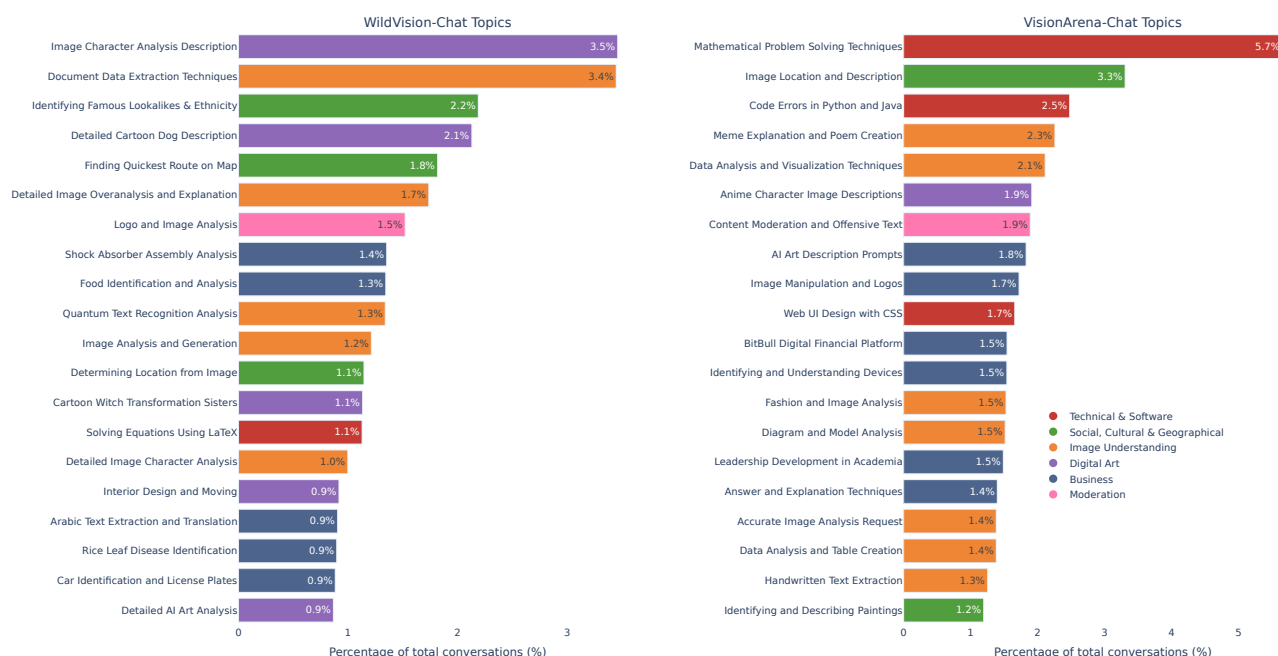


Figure 11. Top 20 topic clusters of VisionArena-Chat compared to WildVision-Chat. VisionArena-Chat includes more diverse and broad topics especially in the STEM field.

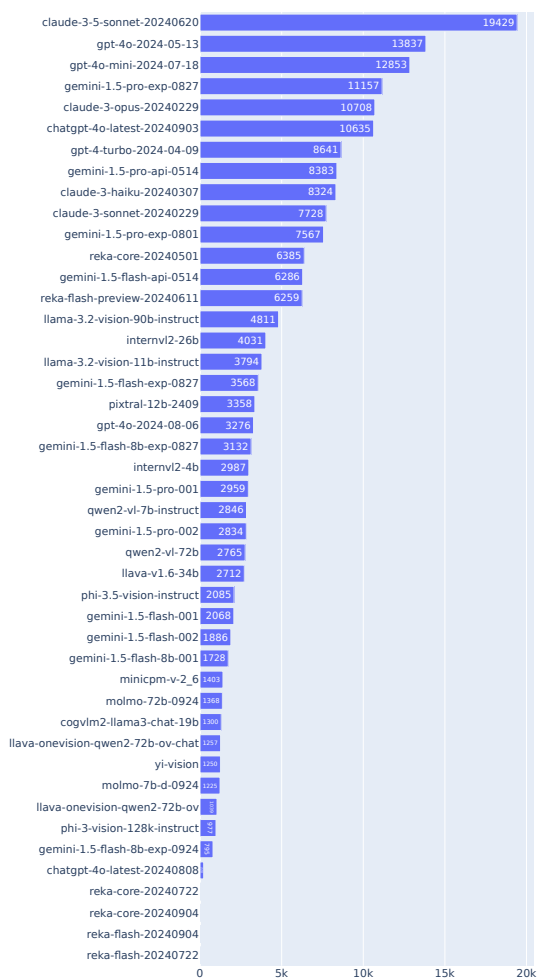


Figure 14. VisionArena-Chat counts per model.

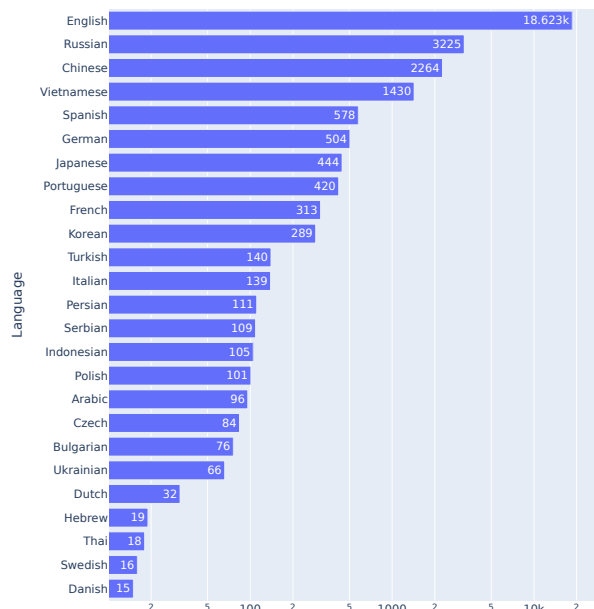


Figure 15. VisionArena-Battle counts for the top 25 languages.

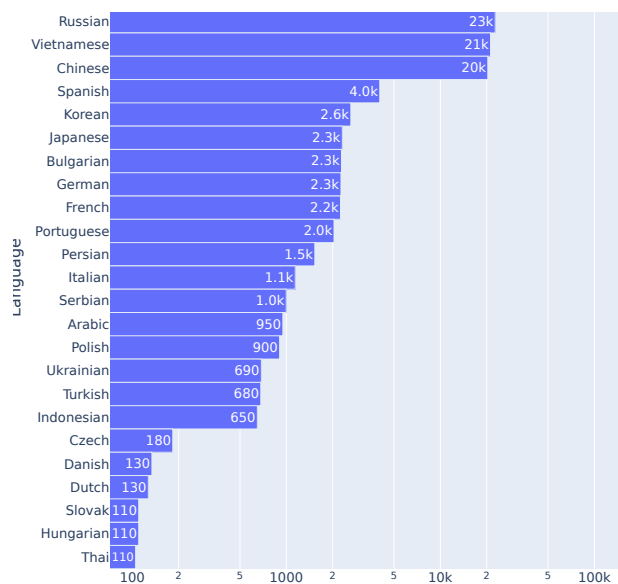


Figure 16. VisionArena-Chat counts for the top 25 languages.

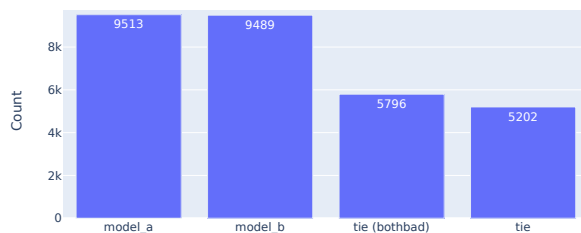


Figure 17. Battle Outcome Counts.

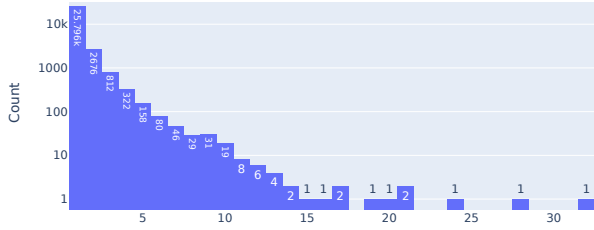


Figure 18. VisionArena-Battle Conversation Turn Distribution

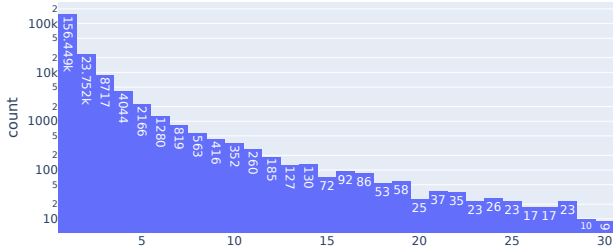


Figure 19. VisionArena-Chat Conversation Turn Distribution

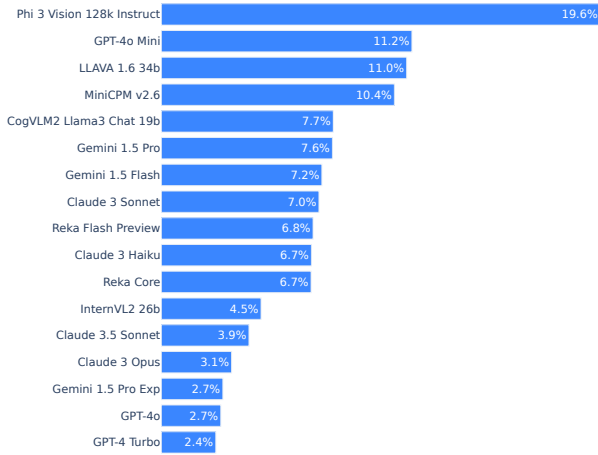


Figure 20. Proportion of Refusals per model.

12. Category Details

Below are the system prompts used to classify user prompts into the categories described in Section 4. We find classifications are more accurate for certain categories when using only the prompt or only the image, as indicated in the prompt titles. We use Gemini 1.5 Flash for classification and show in Table 5 that our classifications achieve high agreement to using SOTA models as category labelers.

OCR System Prompt (image + prompt)

You are tasked with determining if a given VQA question is an optical character recognition (OCR) question. An OCR question requires reading and understanding text in the image to answer. If there is some amount of text in the image and the question requires reading the text in any capacity it should be classified as Optical Character Recognition.

Output your verdict in the following format:
 "<decision>[yes/no]<decision>". Do NOT explain.

Refusal System Prompt (responses only)

You are tasked with determining if any of the given model responses are a refusal to answer. A refusal occurs when the model explicitly declines to answer or complete a task, due to reasons like safety, moderation, or model limitations (e.g. the user asks the model to search the web but it does not have that capability). You will be given the responses of 2 models, A and B and you are to determine if A refuses to answer, B refuses, both refuse, or neither refuse.

Output your verdict in the following format:
 "<decision>[yes/no]<decision>". Do NOT explain.

Captioning System Prompt (prompt only)

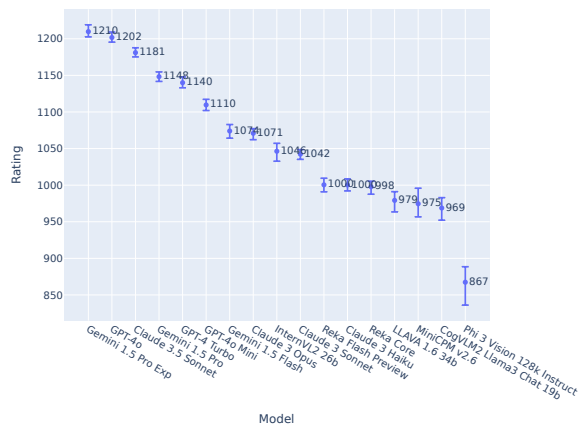
You are tasked with determining if a given VQA question is a captioning question. A captioning question asks for a general, overall description of the entire image. It must be a single, open-ended query that does NOT ask about particular objects, people, or parts of the image, nor require interpretation beyond a broad description of what is visually present. Examples include 'What is happening in this image?', 'Describe this picture.', 'Explain', etc. An example of a non-captioning question is 'Describe what is funny in this picture.' because it asks for a specific interpretation of the image content.

Output your verdict in the following format:
 "<decision>[yes/no]<decision>". Do NOT explain.

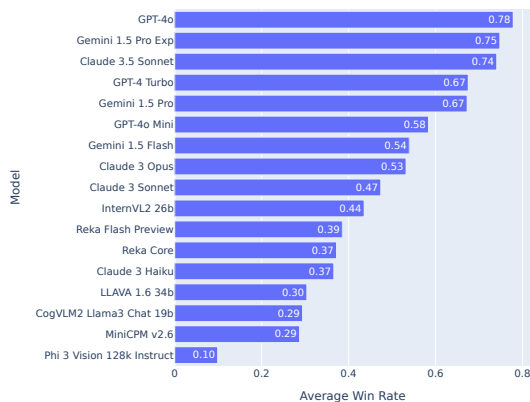
Homework System Prompt (image only)

You are tasked with determining if the given image contains a homework or exam question. A homework or exam question typically contains text with a well-defined question or task which asks for a solution. In addition, many homework and exam questions contain multiple choice, equations, and question numbers. You may also see text referring to showing your work or providing justification. Note that documents such as resumes, business cards, records, or personal notes are NOT considered homework or exam questions; homework and exam questions explicitly ask for a solution or explanation.

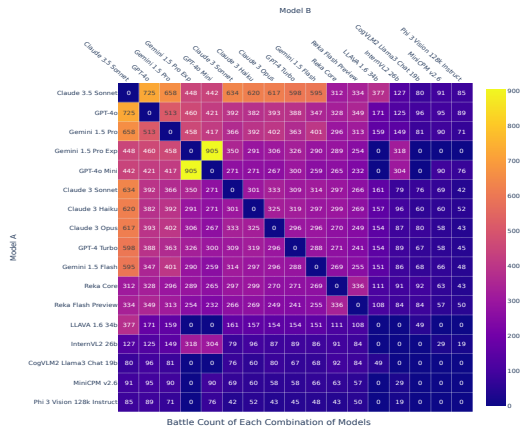
Output your verdict in the following format:
 "<decision>[yes/no]<decision>". Do NOT explain.



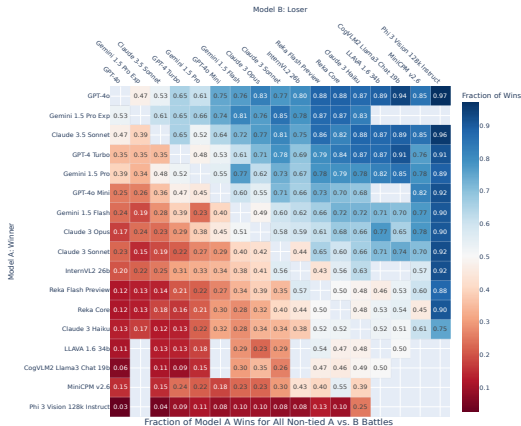
(a) Bootstrap ELO Ratings



(b) Average Win Rate



(c) Battle Counts



(d) Win Fractions

Figure 21. VisionArena-Battle Model Ranking Results.

Category	Uses Image	Labeler	Accuracy	Precision	Recall
Homework	Yes	gemini-1.5-pro-exp-0827	0.987	0.929	0.967
Captioning	No	claude-3-5-sonnet-20240620	0.967	0.938	0.934
Humor	Yes	gemini-1.5-pro-exp-0827	0.925	0.788	0.636
OCR	Yes	gemini-1.5-pro-exp-0827	0.818	0.954	0.769
Entity Recognition	No	claude-3-5-sonnet-20240620	0.952	0.728	0.830
Creative Writing	No	claude-3-5-sonnet-20240620	0.964	0.680	0.810
Diagram	Yes	gemini-1.5-pro-exp-0827	0.961	0.858	0.953

Table 5. Comparing Gemini-1.5-Flash question categorization to larger models. Gemini-1.5-Flash is evaluated against SOTA models on 1000 samples from VisionArena-Chat, using Gemini-1.5-Pro for image-based prompts and Claude-3.5-Sonnet for text-based prompts. Gemini-1.5-Flash achieves high agreement with SOTA models for category classification.

Humor Systems Prompt (image + prompt)

You are tasked with determining if a given VQA question is a humor question. A humor question asks for a humorous or funny response based on the image or asks to understand what is funny about an image. This includes questions that

ask to explain an image which is humorous, such as memes.

Output your verdict in the following format:
 "<decision>[yes/no]<decision>". Do NOT explain.

Captioning	100%	3.1%	46.3%	2.2%	18.3%	2.8%	2%	16.2%
Homework	3%	100%	86.4%	17.9%	2.2%	0.5%	0.4%	35.5%
OCR	13%	24.8%	100%	12.5%	12.7%	3.4%	4.9%	31.5%
Coding	4%	32.9%	80.3%	100%	2.3%	0.8%	1.2%	34.9%
Humor	27%	3.3%	66.3%	1.9%	100%	2.5%	17.1%	10.6%
Entity Recognition	8.3%	1.4%	36.2%	1.3%	5%	100%	1.2%	9.4%
Creative Writing	4.8%	0.9%	41.8%	1.6%	27.9%	1%	100%	9%
Diagram	11.7%	26.1%	80.6%	13.9%	5.2%	2.3%	2.7%	100%

Figure 24. VisionArena-Battle category overlap.

Entity Recognition System Prompt (prompt only)

You are tasked with determining if a given VQA question is an entity recognition question. An entity recognition question asks for the identification of specific objects or people in the image. This does NOT include questions that ask for a general description of the image, questions that only ask for object counts, or questions that only require reading text in the image.

Output your verdict in the following format:
<decision>[yes/no]<decision>. Do NOT explain.

Diagram System Prompt (image only)

You are tasked with determining whether the given image contains a chart, diagram, or figure. Carefully examine the user prompt and consider the following aspects:

1. Does the image contain visual elements such as graphs, flowcharts, tables, method figures, chemical structures, or other visual representations of data or concepts?
2. Does the prompt require interpreting or analyzing the flow of information, relationships between elements, or the structure of the visual representation in the image?
3. Does the prompt require spatial reasoning and understanding the layout or structure of the visual elements?
4. Does the image contain only text, tables, handwriting, or photographs without any visual representations of data or concepts? If so, it is NOT considered a chart or diagram.

Output your verdict in the following format:
"<decision>[yes/no]<decision>". Do NOT explain.

Creative Writing System Prompt (prompt only)

You are tasked with determining whether a given VQA user prompt is asking for creative writing. Creative writing is defined as any form of writing that goes beyond standard professional, journalistic, academic, or technical literature. It typically involves imagination, originality, and expression of

thoughts and emotions. Prompts which only ask to caption the image without any other requests do NOT count as creative writing. Creative writing can include, but is not limited to, the following formats:

- Fiction (e.g., short stories, novels),
- Poetry (e.g., sonnets, free verse),
- Dramatic writing (e.g., screenplays, monologues, scripts),
- Personal essays (focusing on subjective experiences or narrative storytelling),
- Songs and lyrics

Carefully analyze the user prompt and consider whether it primarily requires creative writing. Think about the following aspects:

1. Does the prompt ask for fictional content, speculative scenarios, or the use of imagination to construct narratives?
2. Does it encourage the expression of thoughts, emotions, or personal experiences beyond mere factual reporting or analysis?
3. Is it asking for writing in a specific creative format (e.g., story, poem, script, etc)?
4. Is the primary purpose of the prompt to foster creative expression or originality rather than information delivery, technical documentation, or analytical reasoning?
5. Does the prompt request stylistic or rhetorical elements often associated with creative writing, such as metaphor, imagery, dialogue, etc?
6. Does the prompt expect a response in natural language (e.g., sentences, paragraphs) rather than visual, mathematical, or non-linguistic output?

Output your verdict in the following format:
"<decision>[yes/no]<decision>". Do NOT explain.

13. Contamination with Existing Benchmarks

To ensure that our results from [Sec. 5](#) are not due to training on questions from the test sets, we investigate the rate of benchmark contamination in VisionArena-Chat. Using OpenAI’s text-embedding-small embeddings, we compute the cosine similarity between each VisionArena-Chat question and all benchmark questions, selecting the nearest neighbor with the highest similarity score. We then count the number of cases where this similarity is ≥ 0.8 , indicating minor rephrasings of the same question. [Table 6](#) shows that less than 2% of benchmark questions are seen on VisionArena-Chat.

Dataset	# Matches	% dataset	% VisionArena-Chat
MMMU	47	0.4%	0.02%
MME	0	0.0%	0.0%
HallusionBench	0	0.0%	0.00%
MMMU Pro	23	1.3%	0.01%

Table 6. Proportion of benchmark data in VisionArena-Chat.

14. VisionArena-Bench

VLM-as-a-Judge System Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label: 1. Assistant A is significantly better: $[[A \gg B]]$

2. Assistant A is slightly better: $[[A > B]]$

3. Tie, relatively the same: $[[A = B]]$

4. Assistant B is slightly better: $[[B > A]]$

5. Assistant B is significantly better: $[[B \gg A]]$

Example output: "My final verdict is tie: $[[A = B]]$ ".

Model	Score	95% CI	Token #
gpt-4o-2024-05-13	67.7	(-1.7, 1.8)	316
gemini-1.5-pro-exp-0827	66.2	(-1.8, 1.5)	329
gemini-1.5-flash-exp-0827	60.3	(-1.9, 1.9)	367
claude-3.5-sonnet-20240620	54.5	(-2.1, 1.9)	262
gpt-4-turbo-2024-04-09	50.0	(0.0, 0.0)	258
gemini-1.5-pro-001	45.5	(-1.8, 2.0)	261
gpt-4o-mini-2024-07-18	40.0	(-2.3, 1.9)	224
gemini-1.5-flash-8b-exp-0827	30.6	(-2.3, 1.8)	354
internvl2-26b	23.3	(-2.1, 1.1)	515
gemini-1.5-flash-001	23.0	(-1.1, 1.6)	271
claude-3-opus-20240229	18.9	(-1.9, 1.7)	201
claude-3-sonnet-20240229	18.4	(-1.4, 1.3)	205
reka-core-20240501	15.6	(-1.3, 1.4)	252
llama-3.2-11b-vision-instruct	11.2	(-1.3, 1.1)	457
claude-3-haiku-20240307	9.6	(-1.1, 1.0)	155
internvl2-4b	6.8	(-0.9, 0.8)	421

Table 7. VisionArena-Bench leaderboard (baseline: GPT-4-Turbo)

Model	Score	95% CI	Token #
gemini-1.5-pro-exp-0827	87.6	(-1.1, 1.1)	329
gpt-4o-2024-05-13	86.8	(-1.1, 1.3)	316
claude-3-5-sonnet-20240620	86.3	(-1.3, 1.2)	262
gemini-1.5-flash-exp-0827	83.5	(-1.7, 1.1)	367
gpt-4-turbo-2024-04-09	80.7	(-1.0, 1.6)	258
gemini-1.5-pro-001	75.3	(-1.7, 1.4)	261
gpt-4o-mini-2024-07-18	73.0	(-1.3, 1.4)	224
gemini-1.5-flash-8b-exp-0827	64.7	(-1.5, 2.4)	354
gemini-1.5-flash-001	58.4	(-1.9, 1.5)	271
internvl2-26b	54.2	(-1.7, 1.7)	515
claude-3-opus-20240229	52.0	(-2.0, 1.7)	201
claude-3-sonnet-20240229	50.0	(0.0, 0.0)	205
reka-core-20240501	37.9	(-1.9, 1.7)	252
llama-3.2-11b-vision-instruct	32.6	(-1.7, 1.8)	457
claude-3-haiku-20240307	30.7	(-2.3, 1.6)	155
internvl2-4b	19.6	(-1.9, 1.3)	421

Table 8. VisionArena-Bench leaderboard (baseline: claude-3-sonnet-20240229)

15. Additional model details

Table 9 shows the mapping from the model names used in Section 4 to the exact model versions.

Model Version	Model Name
claude-3-5-sonnet-20240620	Claude 3.5 Sonnet
claude-3-haiku-20240307	Claude 3 Haiku
claude-3-opus-20240229	Claude 3 Opus
claude-3-sonnet-20240229	Claude 3 Sonnet
cogvlm2-llama3-chat-19b	CogVLM2 Llama3 Chat 19b
gemini-1.5-flash-api-0514	Gemini 1.5 Flash
gemini-1.5-pro-api-0514	Gemini 1.5 Pro
gemini-1.5-pro-exp-0801	Gemini 1.5 Pro Exp
gpt-4-turbo-2024-04-09	GPT-4 Turbo
gpt-4o-2024-05-13	GPT-4o
gpt-4o-mini-2024-07-18	GPT-4o Mini
internvl2-26b	InternVL2 26b
llava-v1.6-34b	LLAVA 1.6 34b
minicpm-v-2.6	MiniCPM v2.6
phi-3-vision-128k-instruct	Phi 3 Vision 128k Instruct
reka-core-20240501	Reka Core
reka-flash-preview-20240611	Reka Flash Preview

Table 9. Model Name to exact model version

16. Failure Cases

Hard OCR (Fig. 25). While VLMs perform well at transcribing easily legible text, they struggle with perturbed text (e.g., rotations, blur). Reading such difficult text is essential for real-world applications. We show two failure cases: one with unclear handwriting and another with rotated text.

Model	K-Pop	Sign	Shapes	Triangles	Meme	Map	Shoes	Chess	ARC
gemini-1.5-pro-exp-0827	X	X	X	X	X	X	X	X	X
gpt-4o-2024-05-13	X	O	X	X	X	X	O	X	X
claude-3-5-sonnet-20240620	X	X	X	X	X	X	X	X	X
claude-3-opus-20240229	X	X	X	X	X	X	O	X	X
gpt-4-turbo-2024-04-09	X	X	X	X	X	X	X	X	X
gpt-4o-mini-2024-07-18	X	X	X	X	X	X	X	X	X
gemini-1.5-pro-001	X	X	X	X	X	X	X	X	X
gemini-1.5-flash-8b-exp-0827	X	X	X	X	X	X	X	X	X
gemini-1.5-flash-exp-0827	X	O	O	O	X	X	X	X	X
internvl2-26b	X	X	X	X	X	X	X	X	X
gemini-1.5-flash-001	X	X	O	X	X	X	X	X	X
claude-3-sonnet-20240229	X	X	X	X	X	X	O	X	X
llama-3.2-11b-vision-instruct	X	O	X	O	X	X	X	X	X
claude-3-haiku-20240307	X	X	X	X	X	X	X	X	X
internvl2-4b	X	X	X	X	X	X	X	X	X

Table 10. **Model performance across several hard tasks.** O indicates that the model solves the problem and X indicates that the model fails to solve the problem. 9 out of 16 models fail all questions.

Counting (Fig. 26). Counting is a critical skill for decision-making across education, organization, and daily life. While humans count effortlessly, VLMs still struggle. We provide two examples where the top three models fail: one requiring counting based on shape and color, and another involving intersecting triangles.

Reasoning (Fig. 27, Fig. 28). Reasoning is essential for helping users tackle complex problems. While reasoning remains a challenge for both VLMs and LLMs, we present five unique VLM failure cases.

$$K-Pe^3$$

claude-3-5-sonnet-20240620: The image shows a grid of small squares, with some squares filled in blue to form letters. The blue squares spell out **"THE LOTION"** in a pixelated or dot-matrix style font across the grid.

(a) K-Pop Failure Case

A close-up, low-angle shot of a road sign. The sign is dark with white text that reads "6 Avenue West End". To the right of the text is a red circular logo with a white stylized letter "B" inside. The sign is mounted on a post, and the background is a blurred view of a street with buildings.

claude-3-5-sonnet-20240620: The image shows text that reads "**West 6 Avenue**" when read from right to left. The text appears to be on a street sign or similar type of signage. The lettering is white on a dark blue background, and the image has a slightly blurred or distorted quality to it, possibly due to the angle or lighting conditions when the photo was taken.

(b) Sign Failure Case

Figure 25. **Hard OCR Failure Cases.** These failure cases show that models still fail in cases where the text is perturbed such as rotations or messy hand-writing.

how many of each shape and color

Answer: 3 stars, 6 rectangles, 3 ovals
5 green shapes, 3 orange shapes, 4 blue shapes

gemini-1.5-pro-exp-0827: Here's a count of the shapes and colors:

```

**Shapes:**
***Stars:** 3
***Ovals:** 3
***Rectangles:** 5
**Colors:**
***Light Blue:** 4 (3 stars, 1 rectangle)
***Light Green:** 4 (2 ovals, 2 rectangles)
***Light Orange/Peach:** 2 (1 oval, 1 rectangle)
Let me know if you have any other questions!

```

claude-3-5-sonnet-20240620: Here's a breakdown of the shapes and colors in the image:

```

Shapes:
- Stars: 3 (all light blue)
- Circles/Ovals: 3 (2 green, 1 peach/light orange)
- Rectangles: 5 (3 green, 1 light blue, 1 peach/light orange)

Colors:
- Light Blue: 4 shapes (3 stars, 1 rectangle)
- Green: 5 shapes (2 circles, 3 rectangles)
- Peach/Light Orange: 2 shapes (1 oval, 1 rectangle)

In total, there are 11 shapes in the image.

```

gpt-4o-2024-05-13: In the image, there are various shapes and colors. Here is the count for each shape and color:

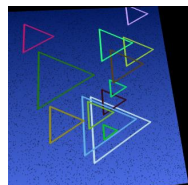
```

**Shapes:**
- Stars: 4
- Ovals: 3
- Rectangles: 6
**Colors:**
- Blue: 5 (4 stars, 1 rectangle)
- Green: 5 (3 ovals, 2 rectangles)
- Orange: 3 (1 oval, 2 rectangles)

```

(a) Shapes Failure Case

How many distinct triangles there are on the image? Intersections between figures dont count.



Answer: 13

gemini-1.5-pro-exp-0827: Let's count the distinct triangles in the image: 1. Top left: 1 (pink) 2. Top right: 1 (light purple) 3. Middle left: 1 (green) 4. Center: - Largest: 1 (light blue) - Medium: 1 (white) - Smallest: 1 (bright green) - Overlapping: 1 (brown) - Tiny: 1 (dark red) 5. Right: 1 (light blue)

Total count: 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 = 9

There are ****9**** distinct triangles in the image.

gpt-4o-2024-05-13: There are **11** distinct triangles in the image.

claude-3-5-sonnet-20240620: In the image, there are **10** distinct triangles. These triangles are of various sizes and colors, scattered across a blue background. The triangles appear in different colors including white, green, pink, yellow, and light blue. Some of the triangles are overlapping, but as per your instruction, we're not counting intersections between figures. Each triangle is considered as a separate, distinct shape regardless of its position or overlap with others.

(b) Triangles Failure Case

Figure 26. **Hard Counting Failure Cases.** These cases show that proprietary models still fail at counting tasks involving shapes even when the format is clear.

Which line connects Trolskaholm to Solköping?



Answer: The Ferry Line

gemini-1.5-pro-exp-0827: The ****Sunrise**** line connects Trolskaholm to Solköping.

It's the yellow/gold line on the map. **You can follow it from Trolskaholm all the way down to Solköping.**

gpt-4o-2024-05-13: The line that connects **Trolskaholm to Solköping** is the **"Midnight" line**, which is represented by the blue color on the map.

claude-3-5-sonnet-20240620: The line that connects Trolskaholm to Solköping is the Bullet line, which is represented by the red line on the map. **This red line runs from Trolskaholm through several stations including Alfheim Central and Singer Park**, before reaching Solköping in the southern part of the map.

(a) Map Failure Case

gemini-1.5-pro-exp-0827: The image is a humorous meme playing on the mathematical concept of square roots.

****The Setup:****

* It shows two pictures of the same cat from behind.

* The cat has a distinctive black marking on its back that vaguely resembles a square root symbol ($\sqrt{}$).

****The Joke:****

* The first image is labeled "Cat".
* The second image is labeled " $\sqrt{\text{Cat}}$ ".

****The Meaning:****

The joke lies in the visual pun. **The cat's marking looks like a square root symbol**, so the second image humorously implies that the cat has been "square rooted" – as if it were a mathematical equation. It's a silly and absurd concept, which is what makes it funny.

gpt-4o-2024-05-13: This image is a humorous visual pun involving a cat.

On the left side, there is a picture of a cat sitting on the ground, viewed from behind. Above this image, the word "Cat" is written.

On the right side, there is a similar picture of the same cat, but the image has been edited to make the cat appear wider and shorter. Above this image, the text " $\sqrt{\text{Cat}}$ " (square root of cat) is written.

The joke is that **the cat on the right looks like it has been "squared" (increased in width and decreased in height), playing on the mathematical concept of squaring and square roots**. The cat on the right is meant to humorously represent the "square root" of the cat on the left.

claude-3-5-sonnet-20240620: This image is a humorous play on words and mathematics, using pictures of a cat to illustrate the concept.

The top of the image shows two text labels:

1. "Cat" on the left

2. " $\sqrt{\text{Cat}}$ " (square root of Cat) on the right

Below these labels are two photos of the same cat, presumably:

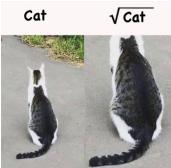
1. On the left, we see a normal-sized cat sitting on a paved surface. It has a black and white coat.

2. On the right, we see what appears to be the same cat, but it looks elongated or stretched vertically. This creates the illusion that the cat is taller or longer than in the left image.

The joke here is that the **"square root" of Cat ($\sqrt{\text{Cat}}$) is visually represented by a stretched version of the cat, as if taking the square root has "elongated" the animal**. It's a clever visual pun combining mathematical notation with a literal interpretation using cat photos.

This type of humor often resonates well on social media platforms, combining familiar concepts (cats and basic math) in an unexpected and visually amusing way.

Explain this image.



Answer: The inside fur of the cat looks like another cat, so the cat's square root is its own cat-looking fur.

(b) Meme Failure Case

The photo shows a boy's new white sneaker and a girl's black leather shoe. Can you describe what happened?



Answer: The girl with the black leather shoe stepped on the boy with the white sneaker.

gemini-1.5-pro-exp-0827: Based on the photo, **it appears the boy with the white sneaker has stepped on the girl's black leather shoe**. The white sneaker is positioned over the black shoe, suggesting that it has come down on top of it. This could have happened accidentally in a crowded space, or perhaps during play or a moment of clumsiness. Without more context, it's impossible to know the exact circumstances, but the positioning of the shoes strongly indicates that the white sneaker has stepped on the black shoe.

gpt-4o-2024-05-13: The photo shows **a black leather shoe stepping on a white sneaker**. It appears that the girl wearing the black leather shoe has accidentally or intentionally stepped on the boy's new white sneaker. This could have resulted in the white sneaker getting scuffed or dirty.

claude-3-5-sonnet-20240620: Based on the image, it appears that the white sneaker and the black leather shoe have come into contact with each other. The white sneaker looks new and clean, while the **black leather shoe seems to be pressing against or stepping on the back part of the white sneaker**. This situation often occurs when someone accidentally steps on the back of another person's shoe while walking closely behind them. It can lead to the heel of the shoe being pushed down or the back of the shoe getting scuffed. The contrast between the pristine white sneaker and the darker leather shoe is quite noticeable in the image. This kind of incident can be frustrating, especially with new shoes, as it might leave marks or damage the shoe's appearance. It's a common mishap that can happen in crowded areas or when people are walking in close proximity to each other.

(c) Shoes Failure Case

Figure 27. **Hard Reasoning Failure Cases (part 1).** The failure cases highlight the model’s inability to connect the visual reasoning with language reasoning. While the vision model identifies an object, it is not able to identify the relationship between them correctly.

I'm playing as white, what should I do in this position?? I'm confused.



Possible Answer: Re7 Rb7 to take the Rook

gemini-1.5-pro-exp-0827: The best move for white in this position is
 Rxe7+.
 Explanation:

1. ****Rxe7+** forces the black king to capture the rook (Kxe7).**
2. ****Qxg7+** follows, checking the black king again.**
3. The black king is forced to move to h8 (Kh8).
4. ****Qxh8#** delivers checkmate.**

This is a forced checkmate sequence that wins the game for white.

Why other moves are not as good:

- * Moving pawns on the queenside doesn't achieve anything significant and allows black to consolidate their position.
- * Developing the knight to f3 doesn't pose an immediate threat and black can easily defend.
- * Other rook moves don't lead to a forced checkmate.

Therefore, Rxe7+ is the only winning move for white in this position.

gpt-4o-2024-05-13: In this position, you have a strong attacking opportunity as White. You can deliver a checkmate in two moves. Here's what you should do:

1. ****Rd8+** (Rook to d8 check)**
 - This move forces the Black king to move to e7 because the rook on d8 is **protected by the knight on c5**.
2. ****Re8#** (Rook to e8 checkmate)**
 - The rook on e8 delivers checkmate as the Black king has no legal moves and cannot escape the check.

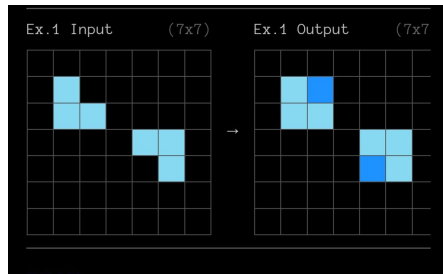
So, the sequence is:

1. Rd8+ Kx7
2. Re8#

This results in a checkmate.

(a) Chess Failure Case

Convert the image into a text based format for both the input and output image, use "[]" to represent each square in a grid, "[b]" to represent light blue grids and "[B]" to represent dark blue grids



Answer:

[] [] [] [] [] [] []	[] [] [] [] [] [] []
[] [b] [] [] [] [] []	[] [b] [B] [] [] [] []
[] [b] [b] [] [] [] []	[] [b] [b] [] [] [] []
[] [] [] [b] [b] [] []	[] [] [] [b] [b] [] []
[] [] [] [] [b] [] []	[] [] [] [] [B] [b] []
[] [] [] [] [] [] []	[] [] [] [] [] [] []
[] [] [] [] [] [] []	[] [] [] [] [] [] []

(b) ARC Failure Case

claude-3-5-sonnet-20240620: In this chess position, you're playing as White and have a few good options to consider. Here are some suggestions:

1. The most immediate threat is **Black's bishop on f6**, which is attacking **your rook on d4**. You should address this threat first.
2. One strong move would be to capture the **black pawn on e7** with your rook: **Rxe7+**. This move puts Black's king in check and removes the defender of the f6 bishop.
3. After Black responds to the check (likely by moving the king to f8), you can then capture the bishop on f6 with your bishop: **Bxf6**.
4. This sequence would result in you winning a pawn and a bishop for the exchange of one of your rooks, which is generally a good trade.
5. Another option, if you want to play more cautiously, is to simply move your rook away from the attack, perhaps to d3 or d2.
6. You could also consider developing your pieces further, such as moving your knight on c5 to a more active square like e6, which would attack Black's rook on b7.

Remember, chess is about long-term strategy as well as immediate tactics. Consider how your move will affect the overall position and try to improve your piece placement while addressing immediate threats.

gemini-1.5-pro-exp-0827

Input:

```
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [b] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [b] [b] [ ] [ ]
[ ] [ ] [ ] [b] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
```

Output:

```
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [B] [ ] [ ] [ ] [ ]
[ ] [b] [b] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [B] [b] [ ] [ ]
[ ] [ ] [ ] [b] [b] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
```

gpt-4o-2024-05-13

Ex.1 Input (7x7) ### Ex.1 Output (7x7)

```
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [b] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [b] [b] [ ] [ ]
[ ] [ ] [ ] [ ] [b] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
```

```
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [B] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [b] [b] [ ] [ ]
[ ] [ ] [ ] [ ] [B] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
```

claude-3-5-sonnet-20240620

Input (7x7):

```
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [b] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [b] [b] [ ] [ ]
[ ] [ ] [ ] [ ] [b] [ ] [ ]
[ ] [ ] [ ] [b] [ ] [b] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
```

Output (7x7):

```
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [b] [B] [ ] [ ] [ ] [ ]
[ ] [b] [b] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [b] [b] [ ] [ ]
[ ] [ ] [ ] [ ] [B] [ ] [ ]
[ ] [ ] [ ] [b] [ ] [b] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ]
```

Figure 28. **Hard Reasoning Failure Cases (part 2).** These failure cases highlight the inability for the model to be able to correctly map out a grid-like structure and the various pieces in it.