Supplemental Material:
HALLUSIONBENCH: An Advanced Diagnostic Suite for Entangled Language Hallucination & Visual Illusion in Large Vision-Language Models

Tianrui Guan* Fuxiao Liu* Xiyang Wu Ruiqi Xian Zongxia Li Xiaoyu Liu Xijun Wang
Lichang Chen Furong Huang Yaser Yacoob Dinesh Manocha Tianyi Zhou

University of Maryland, College Park
`{rayguan, fl3es, wuxiyang, rxian, zli12321, xliu1231, xijun`
`bobchen, furongh, yaser, dmanocha, tianyi}@umd.edu`

1. More Case Analysis on HALLUSIONBENCH with GPT-4V and LLaVA-1.5

In this section, we give a few samples in HALLUSIONBENCH and share our observations. **Each figure is self-contained for readability**, where we highlight the control pairs, the responses of GPT-4V and LLaVA-1.5, the failures of those models, and the corresponding part of the answers.

1.1. Visual Dependent Examples

From the famous illusions in Fig.2, Fig.3, and Fig.4, we found GPT-4V is more knowledgeable than LLaVA-1.5 in recognizing all the illusion cases and knowing their names. However, GPT-4V fails to answer the question faithfully based on the edited images. The reason behind this might be that GPT-4V tends to generate answers based on its parametric memory instead of analyzing the images. Compared to GPT-4V, LLaVA-1.5 performs badly on both the original image and edited images, indicating that the visual perception skill of LLaVA-1.5 is limited.

From the examples in Fig.5 and Fig.6, we found both GPT-4V and LLaVA-1.5 are unable to correctly recognize parallel lines, regular triangles, polygons, and other math theorems, meaning that geometry and math are still a challenging task for GPT-4V.

We further explore GPT-4V's and LLaVA-1.5's abilities in Optical Character Recognition in Fig.7 and Figure Recognition in Fig.8. From our observations, we found that GPT-4V and LLaVA-1.5 are easily misled by editing the characters in the images, demonstrating that GPT-4V and LLaVA-1.5 generate answers based on their parametric memory instead of visual reasoning. This is because the difference between the original images and edited images is obvious.

Inspired by [1], which shows the promising video understanding of GPT-4V, we also investigate more examples in Fig.9 and Fig.10, including several frame sequence examples. The positive sequence and reversed sequence have the opposite semantic meaning, such as *"disappear or appear" and "park or leave"* in Fig.9. From the comparison, we found that GPT-4V is unable to distinguish between the positive sequence and the reversed sequence of the images, indicating that there is still much room to improve the video reasoning ability.

1.2. Visual Supplement Examples

In Fig.11, Fig.12, and Fig.13, GPT-4V does not have an affirmative answer if no images are given. Given the image context, GPT-4V and LLaVA-1.5 are unable to understand the chart correctly, indicating that their chart reasoning ability is still limited. In the second example (bottom) of Fig.19, the predictions of GPT-4V changed completely after we rotated the chart.

In Fig.14, Fig.15, Fig.17, Fig.18, and Fig.19, GPT-4V and LLaVA-1.5 have an affirmative answer if no images are given. After providing the image, including charts, tables, or maps, we found that they preferred to answer the questions with their knowledge instead of analyzing the image. This might be because GPT-4V and LLaVA-1.5 demonstrate a marked dependence on textual reasoning capabilities, often prioritizing them over visual reasoning.

From Fig. 15 and Fig.16, we found the knowledge from LLaVA-1.5 is not accurate since it states "*π doesn't range from 3.1415926 and 3.1415927*" and "*North Carolina is farther north than Delaware*." This observation also supports our claim that GPT-4V is more knowledgeable than LLaVA-1.5.

---

[*] Equal contribution.

2. Decision Tree Logic and Examples

In Fig. 1, we utilize the decision tree to determine the failure types. In the rest of the section, specifically Fig. 20-31, we will provide a few examples and explain the logic that leads to different types of errors. **Each figure with its caption is self-contained for readability.**

In Fig. 20 (bottom), it is a visual-dependent sample (VD). The answer regarding the original image is correct (1), but the answer to the edited image is incorrect (0), and the two answers are the same (*same*). This shows that GPT-4V knows the *"Chubb illusion"* in its parametric knowledge but can not answer according to the image. In Fig. 1, these correspond to the (VD) R-G-R-C route in the decision tree, leading to the diagnostic result of *Language Hallucination*.

In Fig. 21 (bottom), it is a visual-dependent sample (VD). The answer regarding the original image is correct (1), but the answer to the edited image is incorrect (0), and the two answers are not the same (*same*). This shows that GPT-4V can not compare the length of the two lines correctly. In Fig. 1, it corresponds to the (VD) R-G-R-M-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 22 (bottom), it is a visual-dependent sample (VD). The answer regarding the original image is correct (1), but the answer to the edited image is uncertain (2). This shows that GPT-4V is uncertain about the length of the vertical line compared with the horizontal line. In Fig. 1, it corresponds to the (VD) R-G-B-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 23 (bottom), It is a visual-dependent sample (VD). The answer regarding the original image is incorrect (0) or uncertain (2). This shows that LLaVA-1.5 fails to determine the diameters of the three circles in the original image, but succeeds in the edited image. In Fig. 1, it corresponds to the (VS) R-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 24 (bottom), it is a visual-supplement sample (VS). The answer regarding the original image is uncertain (2), but the answer is incorrect (0) or uncertain (2) when the supplementary image is given. This shows that GPT-4V is uncertain about the answer without the visual input, and fails to answer the question with the supplementary image as well. In Fig. 1, it corresponds to the (VS) B-B-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 25 (bottom), It is a visual-supplement sample (VS). The answer is correct (1) without being given any image. However, the answer is uncertain (2) when the supplementary image is given. This shows that GPT-4V is uncertain about the answer given the supplementary image though it could make the correct answer without the image. In Fig. 1, it corresponds to the (VS) B-G-B-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 26 (bottom), it is a visual-supplement sample (VS). The answer is already correct (1) without being given any image. However, the answer is incorrect (0) given the original supplementary image. The supplementary image is not edited. This shows that GPT-4V produces the wrong answer given the supplementary image, though it could produce the correct answer without the image. In Fig. 1, it corresponds to the (VS) B-G-R-G-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 27 (bottom), it is a visual-supplement sample (VS). The answer is correct (1) without being given any image. However, the answer is incorrect (0) when a edited image is given. The supplementary image is edited and the two answers are not the same. This shows that GPT-4V produces the wrong answer based on reasons inconsistent with the edited supplementary image, though it could produce a correct answer without the image. In Fig. 1, it corresponds to the (VS) B-G-R-R-M-B route in the decision tree, leading to the diagnostic result of *Visual Illusion*.

In Fig. 28 (bottom), it is a visual-supplement sample (VS). The answer is correct (1) without being given any image but the answer is incorrect (0) when an edited supplementary image is given. The supplementary image is edited by swapping Delaware and Arizona on the map. The two answers are the same. This indicates that GPT-4V has the prior knowledge of “Delaware is the farthest north” in its parametric knowledge but can not provide a correct answer according to the edited map. In Fig. 1, it corresponds to the (VS) B-G-R-R-C route in the decision tree, leading to the diagnostic result of *Language Hallucination*.

In Fig. 29 (bottom), it is a visual-supplement sample (VS). The answer is incorrect (0) without being given any image. But the answer becomes correct given the original image. This indicates that LLaVA-1.5’s answer is affected by hallucinations without given image information. In Fig. 1, it corresponds to the (VS) B-R-G route in the decision tree, leading to the diagnostic result of *Language Hallucination*.

In Fig. 30 (bottom), it is a visual-supplement sample (VS). The answer is incorrect (0) without being given any image. The answer is still incorrect (0) when the original supplementary image is given. And the two answers are the same. This shows that LLaVA-1.5 has the issue of hallucinations with and without the image information. In Fig. 1, it corresponds to the (VS) B-R-R-C route in the decision tree, leading to the diagnostic result of *Language Hallucination*.

In Fig. 31 (bottom), it is a visual-supplement sample (VS). The answer is incorrect (0) without being given any image. The answer is still incorrect (0) when an edited supplementary image is given. However, the two answers are not the same. This indicates that the commonsense knowledge about the location of US states in LLaVA-1.5 is weak and wrong without the input image of the US map. Additionally, the visual interpretation of the map by LLaVA-1.5 is incorrect.

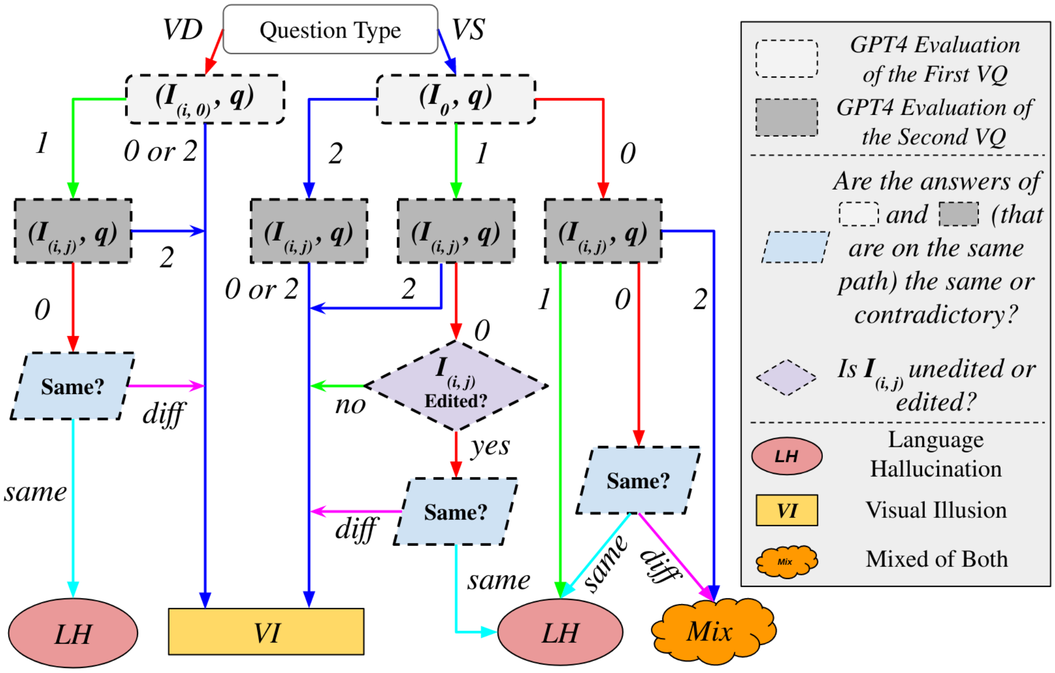
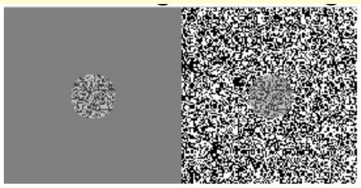


Figure 1. **Decision Tree to Diagnose Failure Types:** Based on the correctness of two questions in a control pair, and the difference in their responses, we use this decision tree to analyze the failure. We highlight different decision paths with Red(R), Blue(B), Green(G), Cyan(C) and Magenta(M). So a path on the decision tree can be represented as a sequence of colors, e.g., R-G-R-C. The output of *GPT4 Evalution* could be *Incorrect (0)*, *Correct (1)*, or *Uncertain (2)* if the predicted response is ambiguous.

In Fig. 1, it corresponds to the (VS) B-R-R-M route in the decision tree, leading to the diagnostic result of *Potentially Mixed*.

Visual Dependent – Optical Illusion of Size

Prompt: The image shows the famous Chubb illusion. Are the center areas of two rectangular fields identical?

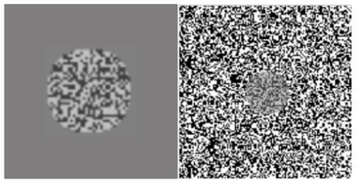


Original image

Answer: Yes.

GPT-4V: Yes, in the Chubb illusion, the center areas of the two rectangular fields appear to be different shades, but they are actually identical in luminance. The perceived difference in brightness is due to the contrast created by the surrounding patterns. This illusion highlights how context can influence our perception.

LLaVA-1.5: Yes, the center areas of the two rectangular fields are identical.



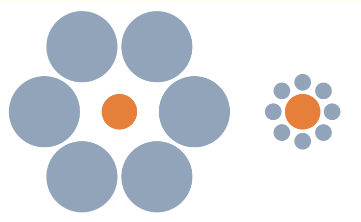
Edited: *The circle on the left are enlarged.*

Answer: No, they are not identical.

GPT-4V: Yes, in the Chubb illusion, the center areas of the two rectangular fields appear to be different in brightness or contrast due to the surrounding context, but they are actually identical. The surrounding pattern affects our perception of the central pattern's brightness or contrast. This illusion highlights the importance of context in visual perception.

LLaVA-1.5: Yes, the center areas of two rectangular fields are identical.

Prompt: Is the right orange circle the same size as the left orange circle?

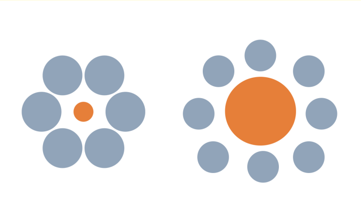


Original

Answer: Yes, the orange balls have the same size.

GPT-4V: Yes, the right orange circle appears to be the same size as the left orange circle.

LLaVA-1.5: No, the right orange circle is smaller than the left orange circle.



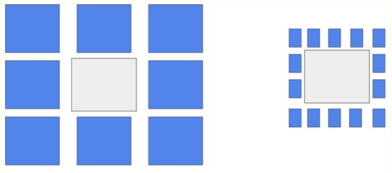
Edited: *The orange ball on the right is enlarged.*

Answer: No, the orange balls have different size.

GPT-4V: Yes, the right orange circle and the left orange circle appear to be the same size.

LLaVA-1.5: Yes, the right orange circle is the same size as the left orange circle.

Prompt: Is the right grey box larger than the left grey box?

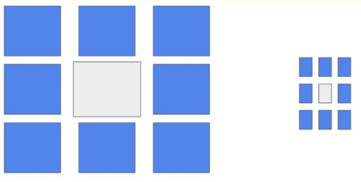


Original

Answer: No, the grey boxes have the same size.

GPT-4V: No, both grey boxes appear to be of the same size.

LLaVA-1.5: Yes, the right grey box is larger than the left grey box.



Edited

Answer: No, the right grey box is smaller than the left.

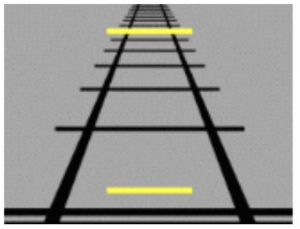
GPT-4V: No, the right grey box and the left grey box appear to be the same size.

LLaVA-1.5: Yes, the right grey box is larger than the left grey box.

Figure 2. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. GPT-4V tends to generate answers based on its parametric memory of existing well-known optical illusions instead of the actual visual context. Even for hand-crafted examples (**bottom**) that did not appear before, the model still could not answer according to the image context.

Visual Dependent – Optical Illusion of Length

Prompt: Are the two yellow lines the same length?

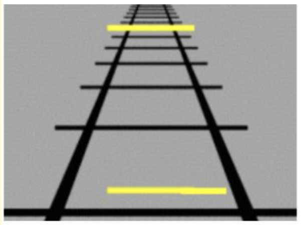


Original

Answer: Yes, the yellow lines have the same length.

GPT-4V: Yes, the two yellow lines are the same length. This is an example of an optical illusion where the surrounding context can make objects appear different in size or length than they actually are.

LLaVA-1.5: No, the two yellow lines are not the same length.



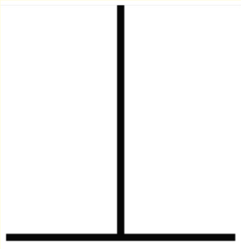
Edited: *The bottom yellow line is elongated.*

Answer: No, the bottom yellow line is longer.

GPT-4V: Yes, the two yellow lines appear to be the same length. This is an example of an optical illusion where parallel lines are in a converging context.

LLaVA-1.5: Yes, the two yellow lines are the same length.

Prompt: Is the vertical line the same length as the horizontal line?

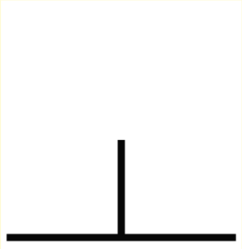


Original

Answer: Yes, the lines have the same length.

GPT-4V: Yes, the vertical line is the same length as the horizontal line.

LLaVA-1.5: Yes, the vertical line is the same length as the horizontal line.



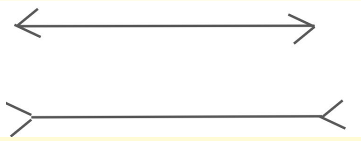
Edited: *The vertical line is shorten.*

Answer: No, the vertical line is shorter.

GPT-4V: No, the vertical line appears to be longer than the horizontal line in the image provided.

LLaVA-1.5: Yes, the vertical line is the same length as the horizontal line.

Prompt: Is the top line longer than the bottom line?

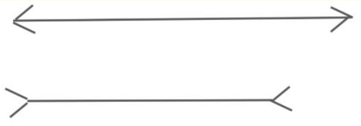


Original

Answer: No, the lines have the same length.

GPT-4V: No, both lines appear to be of the same length. The arrows at the ends can create an optical illusion that might make them seem different in length, but the lines are the same.

LLaVA-1.5: Yes, the top line is longer than the bottom line.



Edited: *The top line is elongated.*

Answer: Yes, the top line is longer.

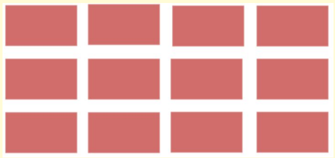
GPT-4V: Both lines appear to be of the same length. The arrowhead style at the ends might create an optical illusion, but the lines themselves seem equal in length.

LLaVA-1.5: Yes, the top line is longer than the bottom line.

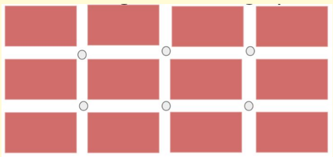
Figure 3. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. GPT-4V can recognize many optical illusion cases but is also easily tricked by the scene and setup of the images. Both models have bad performance in recognizing and measuring length.

Visual Dependent – Other Illusion

Prompt: Are there any grey points actually in the intersection of the white lines?



Original image



***Edited:** Grey points are added.*

Answer: No.

GPT-4V: No, there are no grey points in the image you provided. The image contains only red rectangles arranged in a grid pattern.

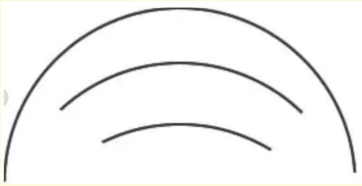
LLaVA-1.5: Yes, there are grey points in the intersection of the white lines.

Answer: Yes.

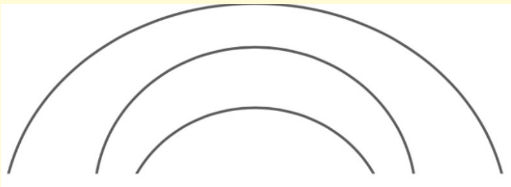
GPT-4V: No, there are no grey dots in the intersections. The optical illusion you're referencing is known as the Hermann grid illusion. When looking at the grid, people often perceive ghost-like grey blobs at the intersections, but when they try to look directly at them, the blobs disappear.

LLaVA-1.5: Yes, there are grey points in the intersection of the white lines. These grey points are located at the corners of the red squares.

Prompt: According to the Torsion dystonia illusion, are the diameters of the three circles the same?



Original



***Edited:** The curves have different diameters.*

Answer: Yes, the curves have the same diameters.

GPT-4V: Yes, according to the Torsion Dystonia illusion, the diameters of the three circles are actually the same. The illusion tricks our perception, making some circles appear larger or smaller than they really are.

LLaVA-1.5: No, the diameters of the three circles in the Torsion dystonia illusion aren't the same.

Answer: No, the curves have different diameters.

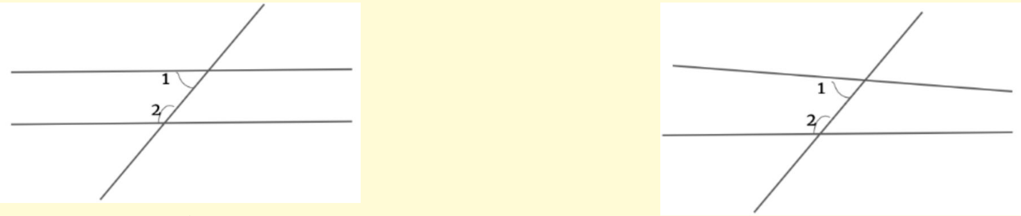
GPT-4V: Yes, all three arcs are part of circles with the same diameter. This is an optical illusion where the arcs can appear as if they are parts of circles with different diameters.

LLaVA-1.5: No, the diameters of the three circles in the Torsion dystonia illusion are not the same. The illusion is designed to make it appear as if the circles are of the same size, but they are actually different in size.

Figure 4. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. GPT-4V recognizes the illusion cases but fails to answer the question faithfully based on the actual image context.

Visual Dependent – Geometry&Math

Prompt: According to parallel lines theorem, is angle 1 + angle 2 = 180 ?



Original image

Edited image

Answer: Yes.

GPT-4V: Yes, according to the properties of parallel lines cut by a transversal, angle 1 and angle 2 are supplementary. Therefore, the sum of angle 1 and angle 2 is 180 degrees.

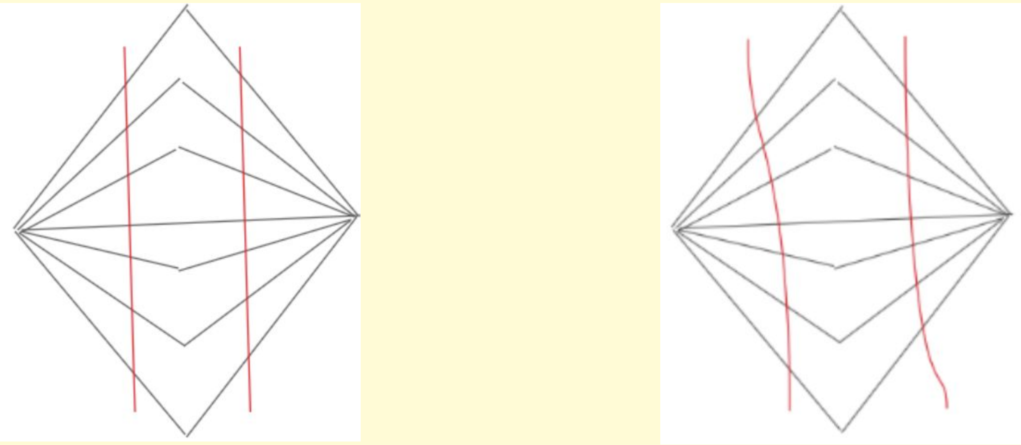
LLaVA-1.5: Yes, according to the parallel lines theorem, the sum of the angles formed by two parallel lines is always 180 degrees. In the image, the two parallel lines are labeled with angles 1 and 2, and their sum is indeed 180 degrees.

Answer: No.

GPT-4V: Yes, according to the properties of parallel lines cut by a transversal, angle 1 and angle 2 are supplementary. This means that the sum of angle 1 and angle 2 is 180 degrees.

LLaVA-1.5: Yes, according to the parallel lines theorem, the sum of the angles formed by two parallel lines is always 180 degrees. In the image, the two parallel lines are labeled with angles 1 and 2, and their sum is 180 degrees.

Prompt: Are the two red vertical lines both straight?



Original image

Edited image

Answer: Yes.

GPT-4V: Yes, the two red vertical lines are both straight.

LLaVA-1.5: Yes, the two red vertical lines are both straight.

Answer: No.

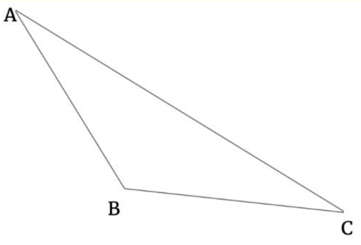
GPT-4V: Yes, the two red vertical lines are both straight.

LLaVA-1.5: No, one of the red vertical lines is curved, while the other one is straight.

Figure 5. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. **Top:** GPT-4V and LLaVA-1.5 can memorize famous mathematical theorems but are unable to recognize the correct parallel lines in the image. **Bottom:** GPT-4V is unable to distinguish whether two lines are straight. We attribute this failure to the lack of geometry recognition ability.

Visual Dependent – Geometry&Math

Prompt: Is angleA + angleB +angleC = 180?

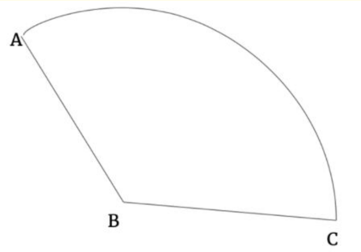


Original image

Answer: Yes.

GPT-4V: Yes, the sum of the interior angles of a triangle always equals 180 degrees. So, angleA + angleB +angleC = 180.

LLaVA-1.5: Yes, angle A + angle B + angle C = 180.



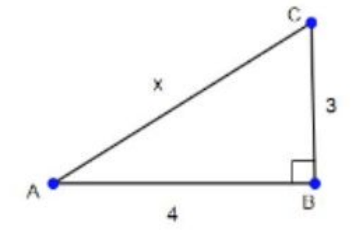
Edited image

Answer: No.

GPT-4V: Yes, the sum of the interior angles of a triangle always equals 180 degrees. So, angleA + angleB +angleC = 180.

LLaVA-1.5: Yes, angleA + angleB + angleC = 180.

Prompt: Is the value of x in the image 5?

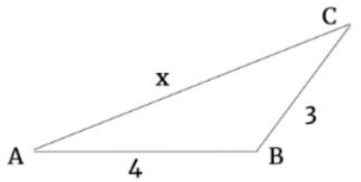


Original image

Answer: Yes.

GPT-4V: Yes, the value of x is 5. Given the triangle ABC in the image, it's a right triangle (as indicated by the square at angle C). Using the Pythagorean theorem, x=5.

LLaVA-1.5: Yes, the value of x in the image is 5.



Edited image

Answer: No.

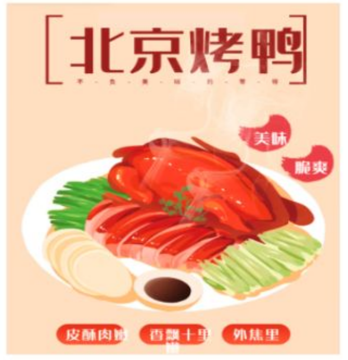
GPT-4V: To determine the value of x in the triangle, we can use the Pythagorean theorem for right triangles, Where a and b are the lengths of the two shorter sides (legs) and is the length of the hypotenuse. So, yes, the value of x in the image is 5.

LLaVA-1.5: Yes, the value of x in the image is 5.

Figure 6. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. In these examples, we modify important geometric properties of the triangles, and neither GPT-4V nor LLaVA-1.5 can recognize those changes. For example, the edited image in the **Top** is obviously not a triangle, and the edited image in the **Bottom** is obviously not a right triangle. We attribute this failure to the lack of geometry recognition ability.

Visual Dependent – Optical Character Recognition

Prompt: Does the image show "Beijing Roast Duck"?

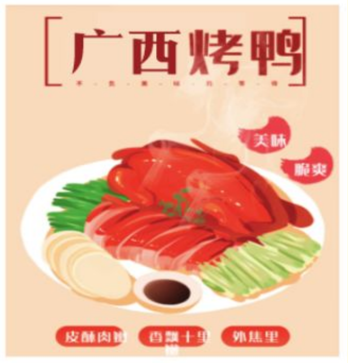


Original image

Answer: Yes.

GPT-4V: Yes, the image appears to depict "Beijing Roast Duck" (also known as Peking Duck).

LLaVA-1.5: Yes, the image shows a picture of "Beijing Roast Duck," which is a popular dish in Chinese cuisine.



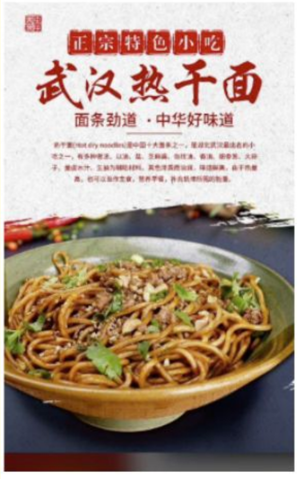
Edited image

Answer: No.

GPT-4V: Yes, the image appears to depict "Beijing Roast Duck", a famous Chinese dish known for its crispy skin and tender meat. The Chinese characters at the top of the image also confirm this as they read "Beijing Roast Duck."

LLaVA-1.5: Yes, the image shows a picture of "Beijing Roast Duck," which is a popular dish in Chinese cuisine.

Prompt: Does the image show "Hot dry noodles from Wuhan"?

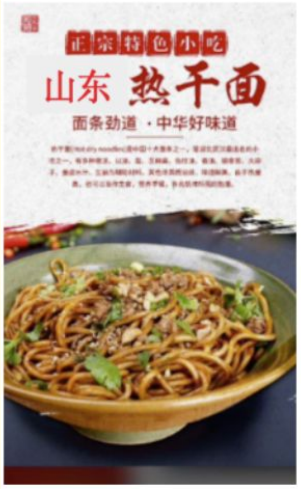


Original image

Answer: Yes.

GPT-4V: Yes, the image displays a dish of noodles with the Chinese text indicating "Hot dry noodles", which is a traditional dish from Wuhan.

LLaVA-1.5: Yes, the image shows a bowl of "Hot dry noodles from Wuhan" with Chinese characters on the side.



Edited image

Answer: No.

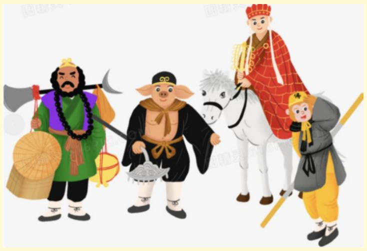
GPT-4V: Yes, the image shows "Hot dry noodles - the taste of Wuhan." So, the image does indeed depict "Hot dry noodles from Wuhan."

LLaVA-1.5: Yes, the image shows a bowl of hot dry noodles from Wuhan.

Figure 7. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. We highlight several advertisements with famous regional dishes with modifications on the regions. In both cases, GPT-4V and LLaVA-1.5 ignore the context and still reply with the well-known regions for that food.

Visual Dependent – Figure Recognition

Prompt: In the image, is Xuanzang accompanied by three disciples?

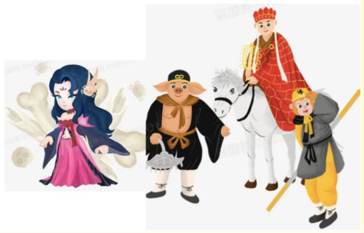


Original image

Answer: Yes.

GPT-4V: Yes, in the image, Xuanzang appears to be accompanied by his three disciples: Sun Wukong (the Monkey King), Zhu Bajie (the Pig), and Sha Wujing (the Sand Monk). This depiction is consistent with the classic Chinese novel "Journey to the West".

LLaVA-1.5: Yes, in the image, Xuanzang is accompanied by three disciples, who are depicted as a group of men standing next to each other.



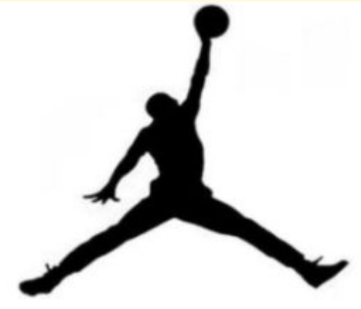
Edited: *One of the disciples is replaced.*

Answer: No.

GPT-4V: Yes, in the image, Xuanzang appears to be accompanied by his three disciples: Sun Wukong (the Monkey King), Zhu Bajie (the Pig), and Sha Wujing (the Sand Monk). This depiction is consistent with the classic Chinese novel "Journey to the West".

LLaVA-1.5: Yes, in the image, Xuanzang is accompanied by three disciples, including a girl and two men.

Prompt: Is it the logo for Air Jordan?

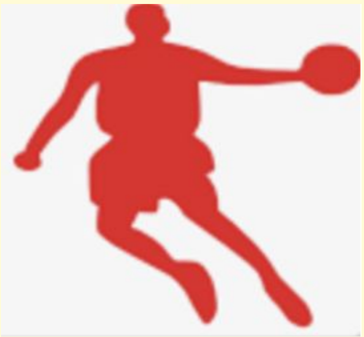


Air Jordan Logo

Answer: Yes.

GPT-4V: Yes, that is the logo for Air Jordan, which is associated with Michael Jordan and the Nike brand.

LLaVA-1.5: Yes, it is the logo for Air Jordan, which is a popular sports brand associated with basketball and athletic footwear.



QIAODAN Logo *(from China)*

Answer: No.

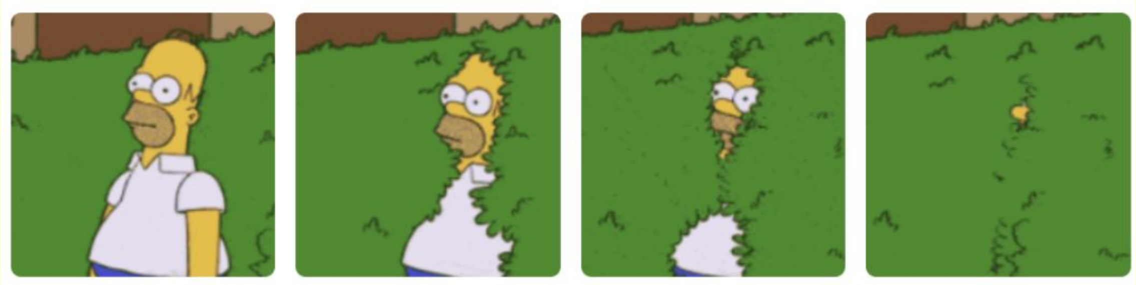
GPT-4V: Yes, the image you provided resembles the Air Jordan logo, which is associated with Michael Jordan and the brand of basketball shoes and athletic clothing produced by Nike.

LLaVA-1.5: Yes, the image is the logo for Air Jordan, which is a sports brand known for its basketball shoes and apparel.

Figure 8. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. **Top:** The judgments of GPT-4V and LLaVA-1.5 are affected by parametric memory and stereotyped judgment, meaning that they do not recognize that the girl is not Xuanzang's disciple. LLaVA-1.5 detects a girl and two men in the image but still regards her as the disciple of Xuanzang. **Bottom:** GPT-4V uses "resembles" due to the uncertainty of the answer but still mistakenly associates it with the Air Jordan brand. We attribute this failure to the *Language Hallucination*.

Visual Dependent – Video/Multiple Images

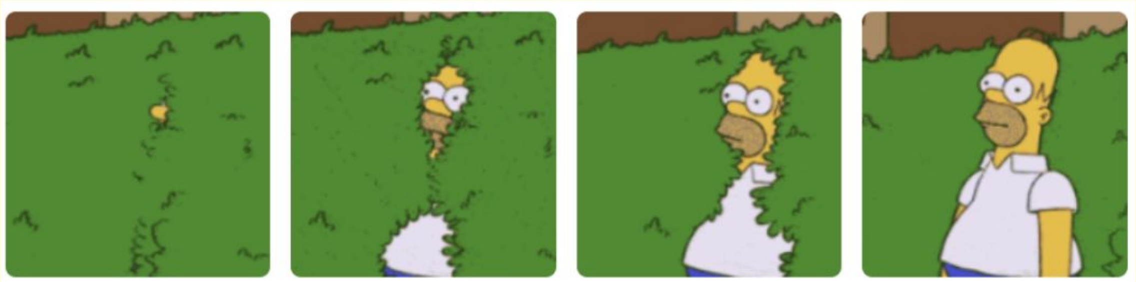
Prompt: According to the positive sequence images, does Homer Simpson disappear into the bushes?



Original sequence

Answer: Yes.

GPT-4V: Yes, the sequence of images shows Homer Simpson gradually disappearing into the bushes.

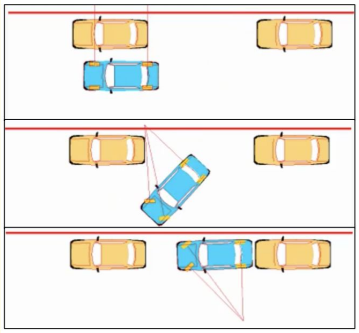


Reversed sequence

Answer: No.

GPT-4V: Yes, the sequence of images shows Homer Simpson gradually disappears into the bushes.

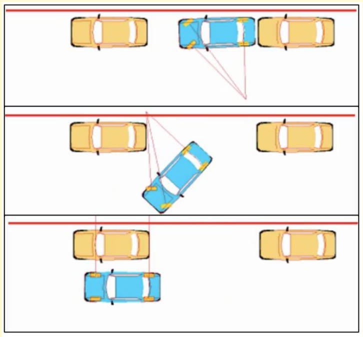
Prompt: According to the positive sequence of the images, does the blue car attempt to park?



Original sequence (From Top to Bottom)

Answer: Yes.

GPT-4V: Yes, according to the sequence of the images, the blue car appears to be attempting to parallel park between the two yellow cars.



Reversed sequence (From Top to Bottom)

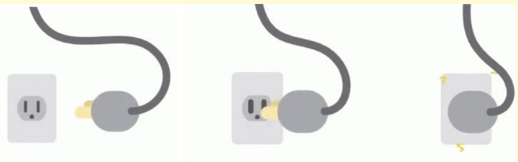
Answer: No.

GPT-4V: Yes, according to the sequence of the images, the blue car appears to be attempting to parallel park between the two yellow cars.

Figure 9. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. From the examples, we found that GPT-4V lacks the ability of temporal reasoning and is unable to distinguish between the positive sequence and reversed sequence of the images, even though they represent different actions.

Visual Dependent – Video/Multiple Images

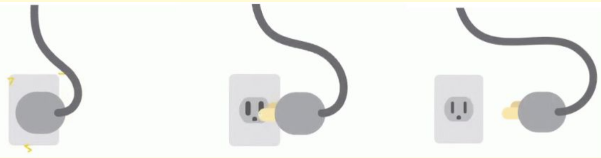
Prompt: According to the positive sequence of the images, is the plug put in the power outlet?



Original sequence

Answer: Yes.

GPT-4V: Yes, the plug is put in the power outlet.

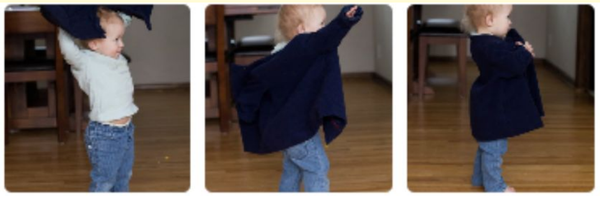


Reversed sequence

Answer: No.

GPT-4V: In the positive sequence of the images, the plug is put into the power outlet.

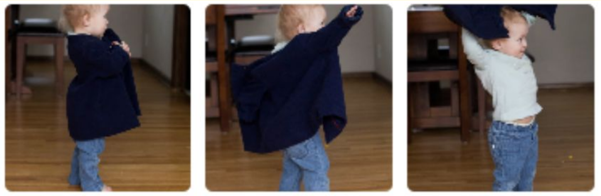
Prompt: According to the positive sequence of the images, does the baby attempt to put on the clothes?



Original sequence

Answer: Yes.

GPT-4V: Yes, according to the sequence of the images, it appears that the baby is attempting to put on the clothes.

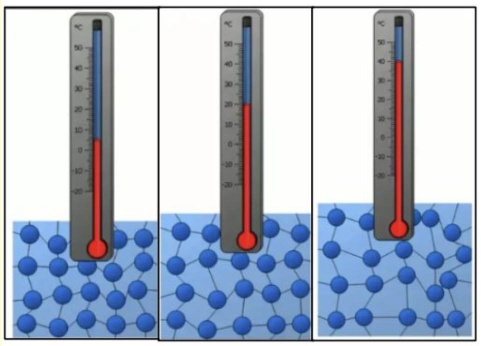


Reversed sequence

Answer: No.

GPT-4V: Yes, according to the sequence of the images, it appears that the baby is attempting to put on the clothes.

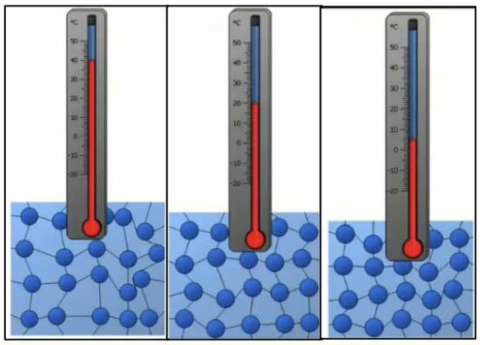
Prompt: According to the positive sequence of the images, does the temperature increase?



Original sequence

Answer: Yes.

GPT-4V: Yes, according to the sequence of the images, the temperature increases.



Reversed sequence

Answer: No.

GPT-4V: Yes, according to the positive sequence of the images, the temperature increases.

Figure 10. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. From the examples, we found that GPT-4V lacks the ability of temporal reasoning and is unable to distinguish between the positive sequence and reversed sequence of the images, even though they represent different actions.

Visual Supplement – Chart

Prompt: Is China, Hongkong SAR, the leading importing country of gold, silverware, and jewelry with the highest import value in 2018?

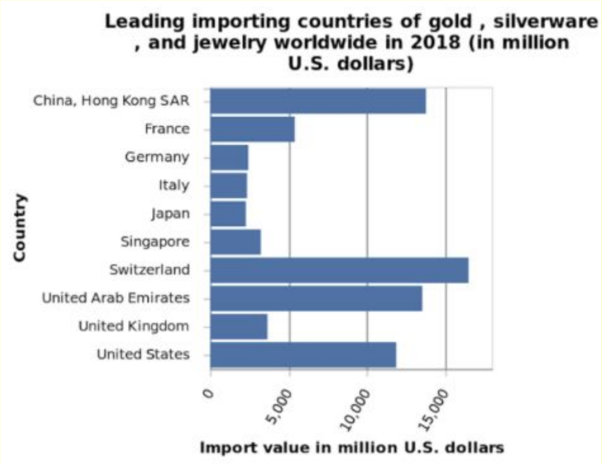
No Visual Input

Answer: No.

GPT-4V: As of my last training cut-off in January 2022, Hong Kong SAR (Special Administrative Region) has traditionally been a significant importer of gold, silverware, and jewelry, serving as a major hub for jewelry and precious metals trade. Hong Kong's role as a trade hub means that it often re-exports a considerable amount of what it imports. Specific rankings for importing countries change from year to year based on various economic factors. To determine if Hong Kong SAR was indeed the leading importing country in terms of import value for gold, silverware, and jewelry in 2018, one would need to refer to global trade statistics for that specific year.

LLaVA-1.5: According to the World Bank's Comtrade database, China, Hong Kong SAR was the leading importing country of gold, silverware, and jewelry in 2018, with an import value of US$157.4 billion. This accounted for approximately 17.5% of the global total for these types of products. It's worth noting that Hong Kong SAR is a special administrative region of China, and as such, its trade data is often closely linked to that of mainland China. Therefore, it's possible that the high import value for gold, silverware, and jewelry in Hong Kong SAR in 2018 reflects trade activity taking place within China as well.

Prompt: Is China, Hongkong SAR the leading importing country of gold, silverware, and jewelry with the highest import value?



Switzerland has the highest import value.

Answer: No.

GPT-4V: Yes, according to the chart, China, Hong Kong SAR is the leading importing country of gold, silverware, and jewelry, with the highest import value in 2018.

LLaVA-1.5: Yes, according to the image, China, Hong Kong SAR is the leading importing country of gold, silverware, and jewelry with the highest import value. The chart shows that the country imported gold, silverware, and jewelry worth over 25 billion U.S. dollars in 2018. This indicates that China, Hong Kong SAR is a significant market for these products, and it plays a crucial role in the global trade of gold, silverware, and jewelry.

Figure 11. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. Sometimes, the question might **not have an affirmative answer without the context**, which GPT-4V can handle very well by giving an uncertain answer without hallucination. However, given the chart with all information as the input, GPT-4V and LLaVA-1.5 are unable to extract correct information to give the right answer.

Visual Supplement – Chart

Prompt: From 2001 to 2019, Is 2010 exported the most value of bread, biscuits, wafers, cakes, and pastries from the UK?

No Visual

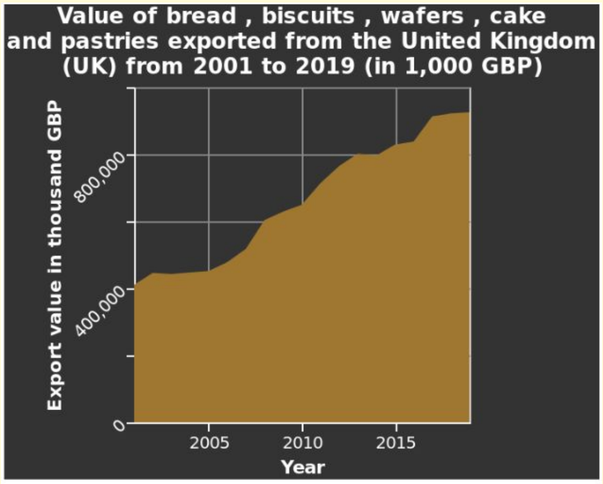
Answer: No.

GPT-4V: To answer this question accurately, one would typically refer to trade and export statistics from the UK government or relevant agencies that track exports of specific goods, including items like bread, biscuits, wafers, cakes, and pastries. However, as of my last update in January 2022, I don't have a specific year-by-year breakdown of exports for these goods from the UK from 2001 to 2019.

If you're looking for this specific information, I'd recommend checking sources such as the UK's Office for National Statistics (ONS) or similar organizations that provide detailed trade statistics.

LLaVA-1.5: According to the United Kingdom's Office for National Statistics (ONS), the UK has been a net importer of food and beverages, including bakery products, since 2001. This means that the UK has imported more of these products than it has exported. In 2019, the UK exported around £2.3 billion of bakery products while importing around £3.6 billion. Therefore, it is unlikely that any single year, such as 2010, would have been the year with the highest export value of these products from the UK.

Prompt: From 2001 to 2019, Is 2010 exported the most value of bread, biscuits, wafers, cakes and pastries from the UK?



2019/2020 has the highest export value.

Answer: No.

GPT-4V: No, according to the graph, the value of exports for bread, biscuits, wafers, cakes, and pastries from the UK increased over the years, and the highest value was reached in 2019. 2010 did not have the highest export value during this period.

LLaVA-1.5: Yes, the image shows that in 2010, the UK exported the most value of bread, biscuits, wafers, cakes, and pastries, with a total value of 1,000 GBP. This indicates that 2010 was a significant year for the export of these baked goods from the UK.

Figure 12. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. GPT-4V can handle uncertain queries by giving an uncertain answer without hallucination. Given a chart as the input, GPT-4V is able to answer the question correctly. In contrast, LLaVA-1.5 can answer the question well without the image context; however, with the addition of the image, confusion arises, and its answer becomes unclear and heavily guided by the way the question is framed due to limited vision capability. Additionally, hallucination appeared in LLAVa's answer (i.e., 1,000 GBP), which is extracted from the title of the chart, and 1,000 GBP means the total number from 2001 to 2019. We attribute this failure to the lack of visual reasoning ability.

Visual Supplement – Chart

Prompt: From 2000-2050, is the population aged between 15 and 29 years in China the most in 2000?

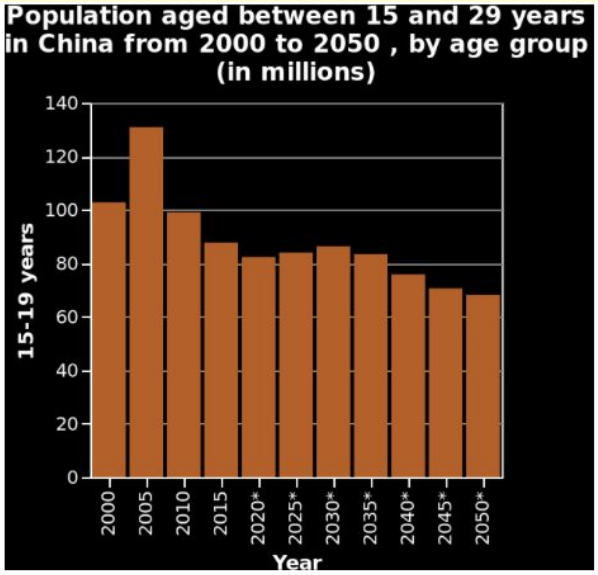
No Visual

Answer: No.

GPT-4V: As of my last update in January 2022, I don't have population projections for China up to 2050. However, I can provide some insight into the trends leading up to 2022 and offer some general expectations based on that.

LLaVA-1.5: Yes, the population aged between 15 and 29 years in China was the most in 2000. According to the United Nations Population Division, the population in this age group in China peaked in 2000 at around 306 million people. This was due to the large cohort of baby boomers born in the 1960s and 1970s, who reached their peak reproductive years in the 1980s and 1990s. However, since then, the population in this age group has been declining due to the country's aging population and declining fertility rates.

Prompt: From 2000-2050, is the population ages between 15 and 29 years in China the most in 2000?



2005 has most population from the group.

Answer: No.

GPT-4V: Yes, based on the provided graph, the population aged between 15 and 29 years in China was the highest in the year 2000.

LLaVA-1.5: Yes, the population aged between 15 and 29 years in China was the most in 2000. The image shows a bar graph with the population in this age group peaking in 2000 and then declining over the next few years. This indicates that the population in this age group was the highest in 2000 and has been decreasing since then.

Figure 13. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. GPT-4V can handle uncertain queries by giving an uncertain answer without hallucination. However, given a chart as the input, GPT-4V is unable to understand it correctly. LLaVA-1.5 tends to generate answers based on the parametric memory since it generates the same answers with or without the image context.

Visual Supplement – Map

Prompt: Is New Mexico state larger than Texas state?

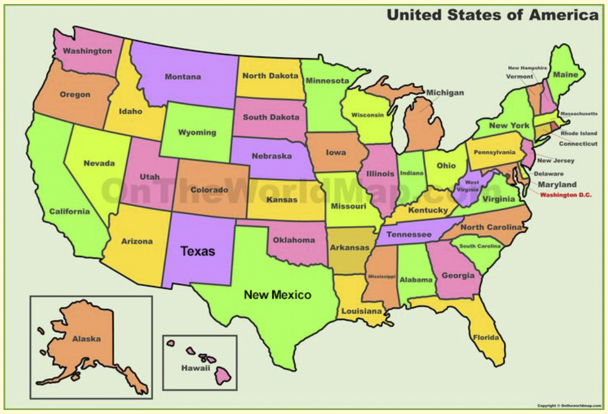
No Visual

Answer: No.

GPT-4V: No, New Mexico is not larger than Texas. Texas is the second-largest state in the U.S. by land area, following only Alaska. New Mexico is the fifth-largest state by land area. Texas has a land area of about 268,596 square miles, while New Mexico covers about 121,589 square miles. On any map that accurately represents the area, Texas will always appear larger than New Mexico.

LLaVA-1.5: No, New Mexico is not larger than Texas. Texas is the second-largest state in the United States in terms of both area and population, while New Mexico is the fifth-largest state in terms of area but the 36th-largest in terms of population.

Prompt: Is New Mexico state larger than Texas on the map?



New Mexico and Texas are swapped in the following image.

Answer: Yes.

GPT-4V: No, based on the map you provided, Texas is larger than New Mexico.

LLaVA-1.5: No, from the map, Texas state is larger than New Mexico state.

Prompt: Does the prime meridian intersect the Atlantic Ocean, Southern Ocean, and Arctic Ocean?

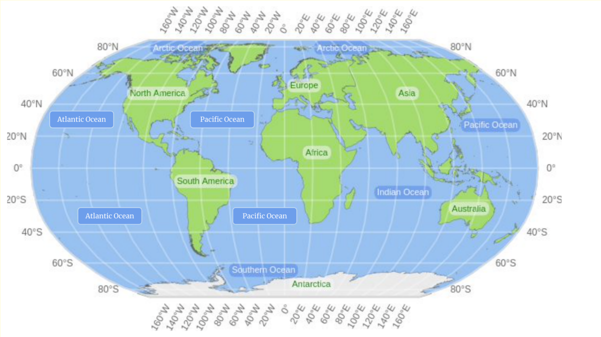
No Visual

Answer: Yes.

GPT-4V: Yes, the Prime Meridian, which is the line of 0° longitude, does intersect all three of these oceans: the Atlantic Ocean: The Prime Meridian runs from the North Pole southward through the Atlantic Ocean, passing west of Europe and Africa, before reaching the Southern Ocean. Southern Ocean: As the Prime Meridian continues south from the Atlantic, it intersects the Southern Ocean and continues all the way to the South Pole. Arctic Ocean: At its northernmost extent, the Prime Meridian intersects the Arctic Ocean as it goes from the North Pole southward.

LLaVA-1.5: The prime meridian, which is the line of longitude that passes through Greenwich, England, intersects the Atlantic Ocean, the Southern Ocean, and the Arctic Ocean. The prime meridian passes through the Atlantic Ocean to the west of the Azores, through the Southern Ocean south of Africa, and through the Arctic Ocean north of Canada.

Prompt: According to the map, does the prime meridian intersect the Atlantic Ocean, Southern Ocean, and Arctic Ocean?



Atlantic Ocean is replaced with the Pacific Ocean in the following image.

Answer: No, the prime meridian intersect the Pacific Ocean, Southern Ocean and arctic Ocean in this figure.

GPT-4V: Yes, according to the map, the prime meridian intersects the Atlantic Ocean, Southern Ocean, and the Arctic Ocean.

LLaVA-1.5: Yes, the prime meridian intersects the Atlantic Ocean, Southern Ocean, and Arctic Ocean.

Figure 14. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. Both GPT-4V and LLaVA-1.5 have an affirmative and correct answer without context. Given an edited map (contradictory to common sense), GPT-4V and LLaVA-1.5 don't understand the map correctly. GPT-4V and LLaVA-1.5 tend to generate answers based on their parametric memory instead of the actual visual context, even though the way they answer the question indicates that the reply is based on the provided map.

Visual Supplement – Map

Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware, Is Delaware the farthest north?

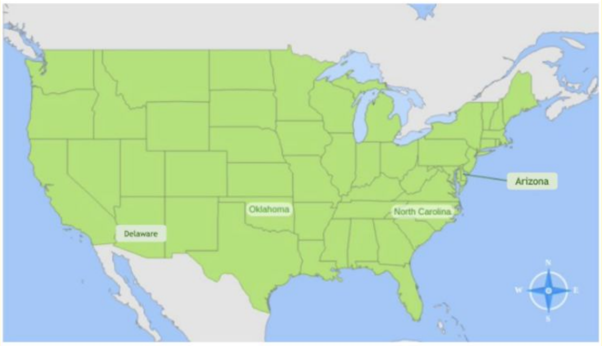
No Visual

Answer: Yes, Delaware is the farthest north among Arizona, Oklahoma, North Carolina, and Delaware.

GPT-4V: Yes, among the states you listed—Arizona, Oklahoma, North Carolina, and Delaware, Delaware is the farthest north.

LLaVA-1.5: No, Delaware is not the farthest north among Arizona, Oklahoma, North Carolina, and Delaware. North Carolina is the farthest north among these four states. Delaware is located to the north of Maryland and to the east of Pennsylvania, while Arizona is located to the southwest of Delaware, Oklahoma is located to the west of Delaware, and North Carolina is located to the north of Delaware.

Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?



Delaware and Arizona are swapped in the following image.

Answer: No, Arizona is the farthest north among Arizona, Oklahoma, North Carolina, and Delaware.

GPT-4V: Yes, among Arizona, Oklahoma, North Carolina, and Delaware on the map, Delaware is the farthest north.

LLaVA-1.5: Yes, Delaware is the farthest north among the four states on the map.

Figure 15. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. When there is no image context, both GPT-4V and LLaVA-1.5 have an affirmative and correct answer, but only GPT-4V provides the correct answer. We conclude that GPT-4V is more knowledgeable than LLaVA-1.5. With an altered map as a visual input, both models fail to reason the location relationship among the four states due to strong parametric memory.

Visual Supplement – Optical Character Recognition

Prompt: Does π range from 3.1415926 to 3.1415927?

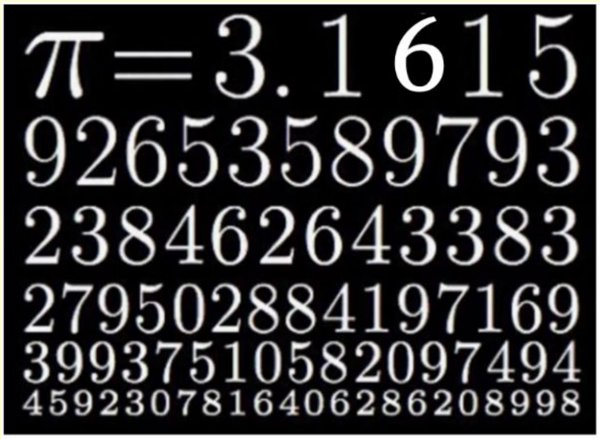
No Visual

Answer: Yes.

GPT-4V: Yes, according to the image, the value of π starts with 3.1415926, which is within the range of 3.1415926 to 3.1415927.

LLaVA-1.5: No, π (pi) is an irrational number, which means it cannot be expressed as a finite decimal or fraction. It is an infinite, non-repeating sequence of digits. The value of π is approximately 3.14159265358979323846, but it is not exactly equal to any finite decimal or fraction.

Prompt: Does π range from 3.1415926 to 3.1415927 according to the image?


π=3.1615
92653589793
238462643383
279502884197169
39937510582097494
4592307816406286208998

3.1415 is changed into 3.1615 in the following image.

Answer: No.

GPT-4V: Yes, according to the image, the value of π starts with 3.1415926, which is within the range of 3.1415926 to 3.1415927.

LLaVA-1.5: Yes, the image shows that π ranges from 3.1415926 to 3.1415927.

Figure 16. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. When there is no image context, GPT-4V has an affirmative answer, while LLaVA-1.5 provides an irrelevant answer that is technically correct due to failed text understanding. With an altered π value as a visual input, both models fail to interpret the value from the image.

Visual Supplement – Chart

Prompt: Does China have the largest population in the world, followed by India and then the USA?

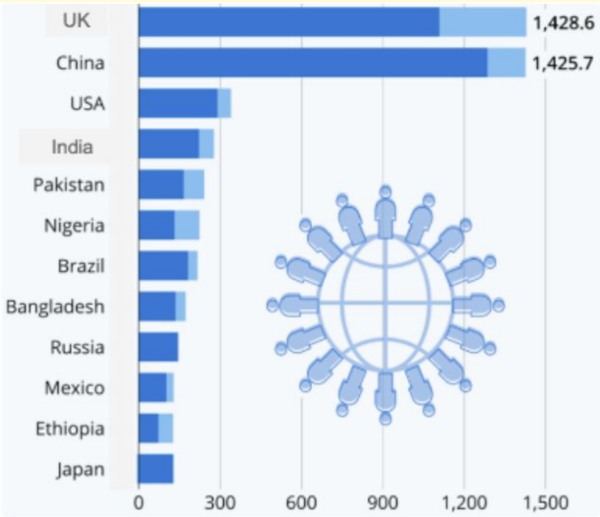
No Visual

Answer: Yes.

GPT-4V: As of my last knowledge update in September 2021, China indeed had the largest population in the world, followed by India and then the United States. However, population figures can change over time due to factors like birth rates, death rates, and immigration. It's possible that these rankings have changed since then. To get the most up-to-date population statistics, I recommend checking a reliable source like the United Nations or the World Bank, which regularly updates and publishes demographic data for countries around the world.

LLaVA-1.5: Yes, that is correct. According to the latest estimates, China has the largest population in the world, with over 1.4 billion people. India is the second most populous country, with over 1.3 billion people, and the United States is the third most populous country, with over 331 million people.

Prompt: Does China have the largest population in the world, followed by India and then the USA?



UK is the largest one, followed by China and USA.

Answer: No.

GPT-4V: Yes, according to the chart, China has the largest population, followed by India and then the USA.

LLaVA-1.5: Yes, according to the image, China has the largest population in the world, followed by India and then the USA. The graph shows the population of various countries, with China at the top, India in second place, and the USA in third. This information is based on the data available at the time the graph was created, and it may not reflect the current population statistics.

Figure 17. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. Both GPT-4V and LLaVA-1.5 have an affirmative and correct answer without context, and the answer provided by GPT-4V is more rigorous since it also emphasizes the time-dependency of this knowledge. However, both models fail to answer the question correctly based on the data provided in the chart.

Visual Supplement – Table

Prompt: In the 2008 Beijing Olympics, Does China have the most gold metals, followed by the USA and Russia?

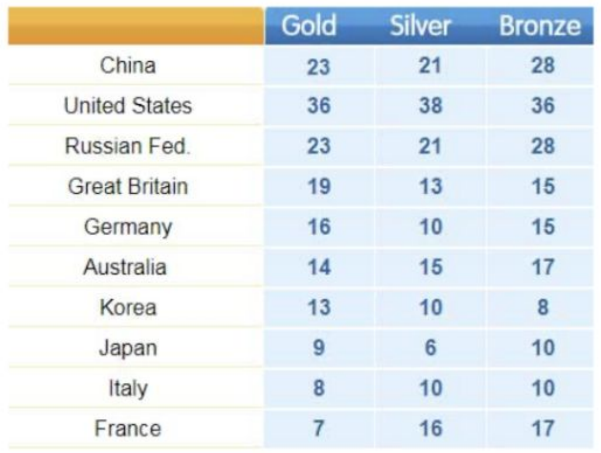
No Visual

Answer: Yes.

GPT-4V: Yes, during the 2008 Beijing Olympics, China topped the gold medal count. They won a total of 51 gold medals. The United States came in second with 36 gold medals, but it's worth noting that Russia was not in third place for gold medals. Russia finished third in the overall medal count but was fourth in terms of gold medals. Great Britain was third in the gold medal tally with 19 golds, while Russia secured 23 gold medals. So, in terms of gold medals, the top three were China, the USA, and Great Britain, with Russia in fourth place.

LLaVA-1.5: Yes, that is correct. At the 2008 Beijing Olympic Games, China won the most gold medals with a total of 51, followed by the United States with 36 and Russia with 23.

Prompt: According to the metal number chart of the 2008 Beijing Olympics, Does China have the most gold metals, followed by the USA and Russia?



	Gold	Silver	Bronze
China	23	21	28
United States	36	38	36
Russian Fed.	23	21	28
Great Britain	19	13	15
Germany	16	10	15
Australia	14	15	17
Korea	13	10	8
Japan	9	6	10
Italy	8	10	10
France	7	16	17

The gold medal number of China is changed to 23, ranked second.

Answer: No.

GPT-4V: Based on the chart provided, China has 36 gold medals. The United States has 19 gold medals. The Russian Federation (Russian Fed.) has 23 gold medals. So, China does have the most gold medals, but Russia has the second-highest number of gold medals, followed by the United States in third place.

LLaVA-1.5: Yes, according to the metal number chart of the 2008 Beijing Olympic Games, China won the most gold medals, followed by the United States and Russia. The chart shows the number of gold, silver, and bronze medals won by each country during the games.

Figure 18. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. When there is no visual input, both GPT-4V and LLaVA-1.5 have an affirmative and correct answer. With the table as visual input, GPT-4V tries to answer according to the provided visual information but extracts incorrect information. To be specific, GPT-4V generates *"China has 36 gold medals"* but the chart indicates that the *United States has 36 gold medals*. LLaVA-1.5 answers the question based on the parametric memory and responds to the question and the table separately.

Visual Supplement – Chart

Prompt: Is China the third largest country by land area in the world?

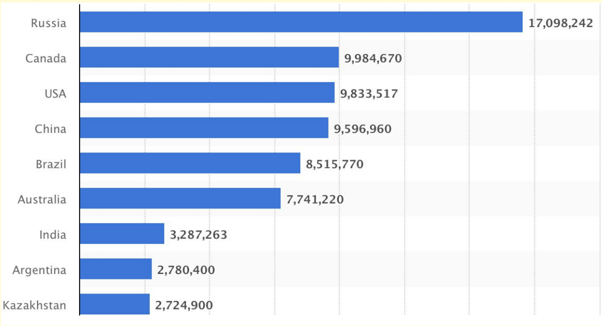
No Visual

Answer: Yes, China is the third largest country.

GPT-4V: No, China is not the third largest country by land area in the world; it is the third largest by total area (which includes both land and water). By land area alone, China is the second largest.

LLaVA-1.5: Yes, China is the third largest country in the world by land area. It has a total land area of approximately 9.6 million square kilometers. The country is bordered by 14 other countries, including Russia, India, and Vietnam. China is also home to a diverse range of landscapes, including mountains, deserts, and forests.

Prompt: According to the chart, is China the third largest country by land area in the world?



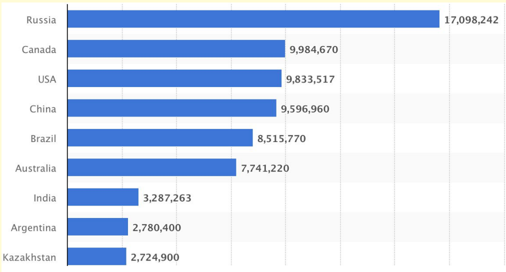
The land size of China and USA are swapped.

Answer: No, China is the fourth in the chart.

GPT-4V: No, according to the chart provided, China is the fourth largest country by land area in the world. The top three are Russia, Canada, and the USA.

LLaVA-1.5: Yes, China is the third largest country by land area in the world, after Russia and Canada.

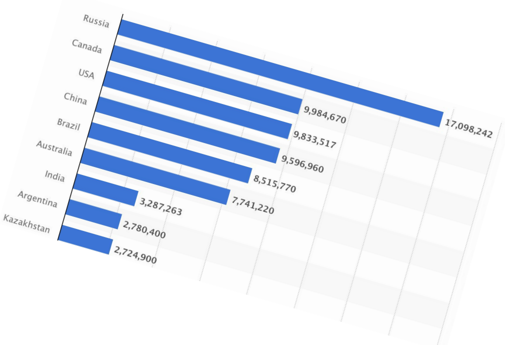
Prompt: Is China the third largest country by land area in the world?



Original image

Answer: No, China is the fourth in the chart.

GPT-4V: No, according to the chart provided, China is the fourth largest country by land area in the world. The top three are Russia, Canada, and the USA.



Edited: *rotated chart.*

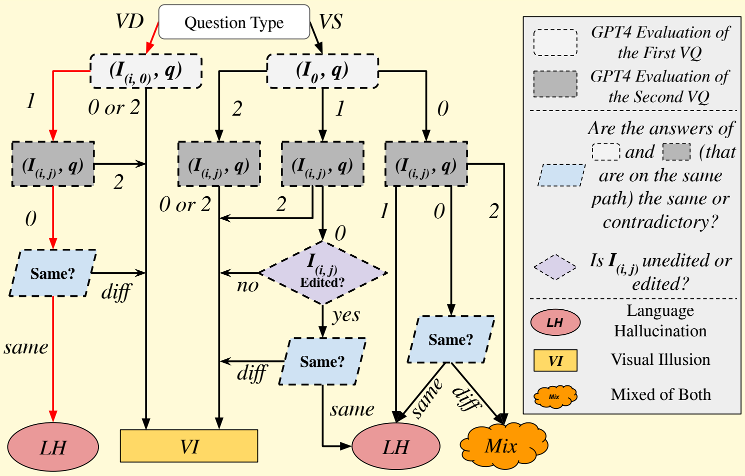
Answer: No, China is the fourth in the chart.

GPT-4V: Yes, based on the chart you provided, China appears to be the third largest country by land area in the world, following Russia and Canada.

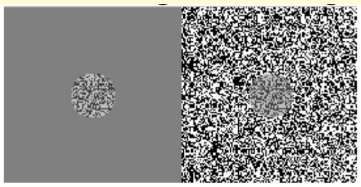
Figure 19. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. **Top:** When there is no image context, both GPT-4V and LLaVA-1.5 have an affirmative and correct answer, but only GPT-4V provides the correct and most rigorous answer. With the chart as visual input, GPT-4V can follow the information provided in the chart faithfully, while LLaVA-1.5 tries to answer the question based on the parametric memory. **Bottom:** We find that the predictions of GPT-4V changed completely after we flipped the chart. In contrast to the chart at the top, this failure can be attributed to the visual illusion.

Visual Dependent – Decision Tree Logic

Decision Tree: The RED path means the decision path.



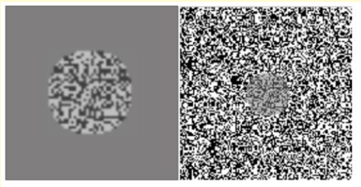
Prompt: The image shows the famous Chubb illusion. Are the center areas of two rectangular fields identical?



Original image

Answer: Yes.

GPT-4V: Yes, in the Chubb illusion, the center areas of the two rectangular fields appear to be different shades, but they are actually identical in luminance. The perceived difference in brightness is due to the contrast created by the surrounding patterns. This illusion highlights how context can influence our perception.



Edited: *Circle on the left are enlarged.*

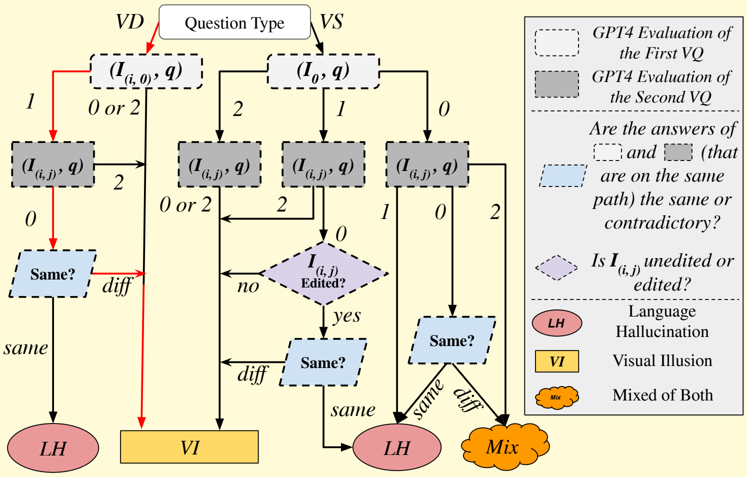
Answer: No, they are not identical.

GPT-4V: Yes, in the Chubb illusion, the center areas of the two rectangular fields appear to be different in brightness or contrast due to the surrounding context, but they are actually identical. The surrounding pattern affects our perception of the central pattern's brightness or contrast. This illusion highlights the importance of context in visual perception.

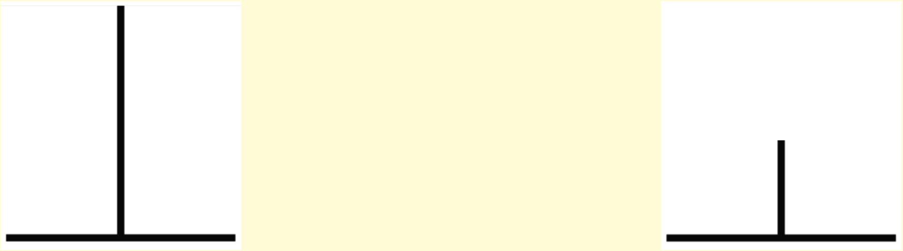
Figure 20. It is a visual-dependent sample (VD). The answer regarding the original image is correct (1), but the answer to the edited image is incorrect (0), and the two answers are the same. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Language Hallucination*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Dependent – Decision Tree Logic

Decision Tree: The RED path means the decision path.



Prompt: Is the vertical line the same length as the horizontal line?



Original: *The lines have the same length.*

Edited: *The vertical line is shortened.*

Answer: Yes, the lines have the same length.

GPT-4V: Yes, the vertical line is the same length as the horizontal line.

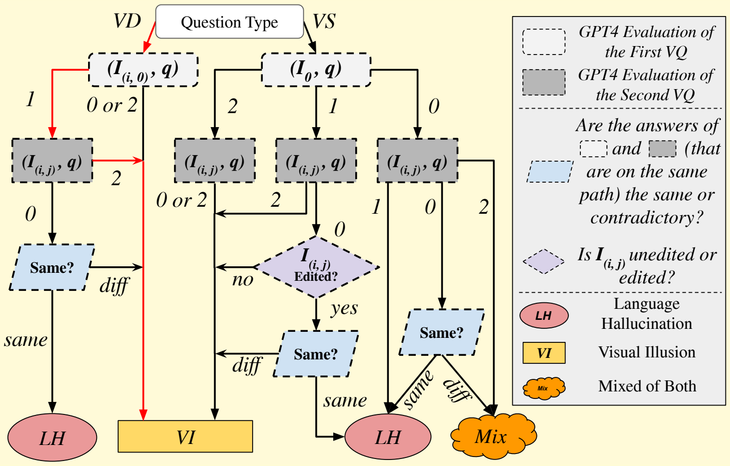
Answer: No, the vertical line is shorter.

GPT-4V: No, the vertical line appears to be longer than the horizontal line in the image provided.

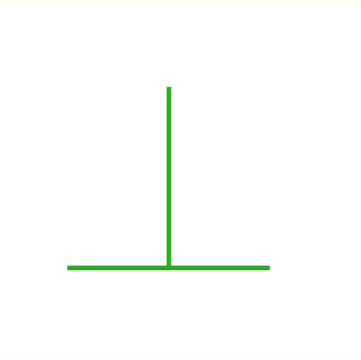
Figure 21. It is a visual-dependent sample (VD). The answer regarding the original image is correct (1), but the answer to the edited image is incorrect (0), and the two answers are not the same. This shows that GPT-4V can not compare the length of the two lines correctly. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Dependent – Decision Tree Logic

Decision Tree: The RED path means the decision path.



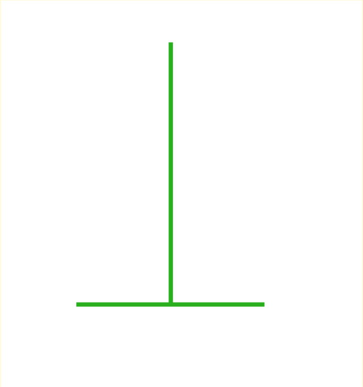
Prompt: Is the vertical line longer than the horizontal line?



Original image

Answer: No, the lines have the same length.

GPT-4V: The vertical line is not longer than the horizontal line.



Edited: *The vertical line is elongated.*

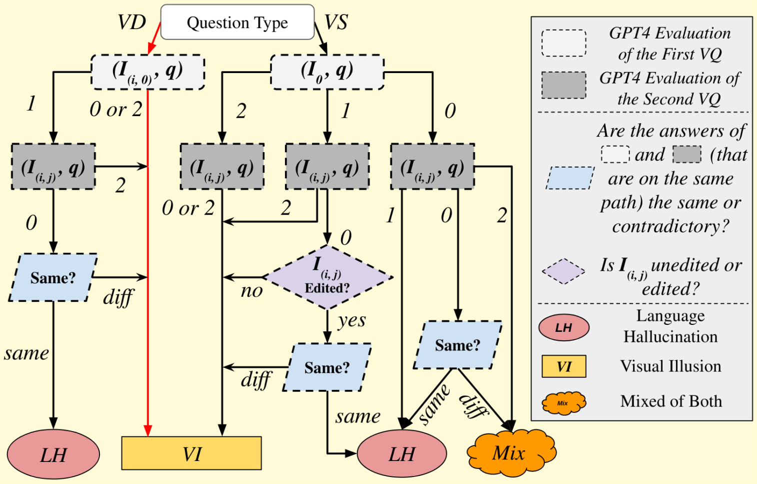
Answer: Yes, the vertical line is longer.

GPT-4V: Without the ability to measure the lines, and based on common visual illusions, the lines may be of the same length. However, this type of image is often used to demonstrate the vertical-horizontal illusion, where the vertical line appears longer than the horizontal line due to the way our perception works, even though they are the same length. If this image is not intended to be an illusion and without exact measurement, I cannot definitively say whether one line is longer than the other.

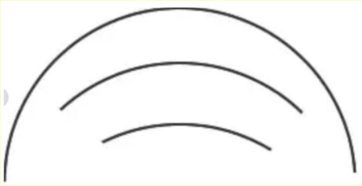
Figure 22. It is a visual-dependent sample (VD). The answer regarding the original image is correct (1), but the answer to the edited image is uncertain (2). This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the uncertain answer, and the incorrect answer according to visual illusion, language hallucination, or potentially mixed. In this case, our decision tree can only detect *visual illusion*; we can get additional information about the failure by examining the responses.

Visual Dependent – Decision Tree Logic

Decision Tree: The RED path means the decision path.



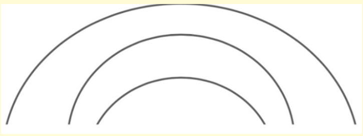
Prompt: According to the Torsion dystonia illusion, are the diameters of the three circles the same?



Original image

Answer: Yes, the curves have the same diameters.

LLaVA-1.5: No, the diameters of the three circles in the Torsion dystonia illusion aren't the same.



Edited: *The curves have different diameters.*

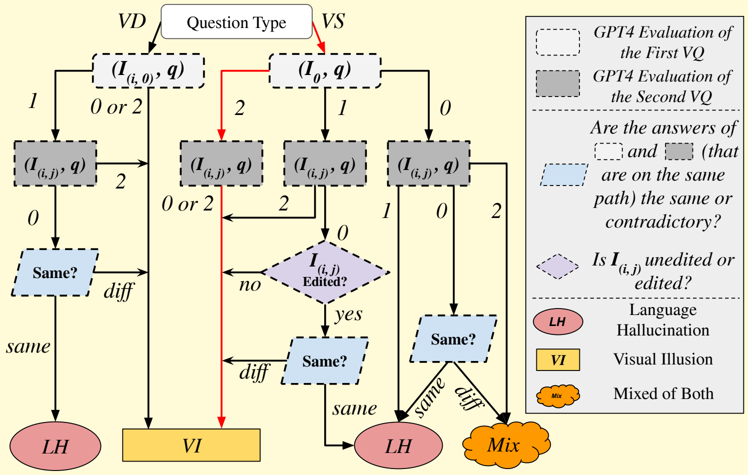
Answer: No, the curves have different diameters.

LLaVA-1.5: No, the diameters of the three circles in the Torsion dystonia illusion are not the same. The illusion is designed to make it appear as if the circles are of the same size, but they are actually different in size.

Figure 23. It is a visual-dependent sample (VD). The answer regarding the original image is incorrect (0). This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



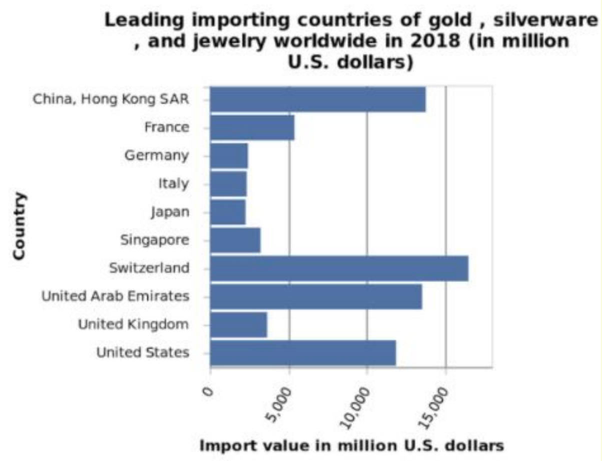
Prompt: Is China, Hongkong SAR, the leading importing country of gold, silverware, and jewelry with the highest import value in 2018?

No Visual Input

Answer: No.

GPT-4V: As of my last training cut-off in January 2022, Hong Kong SAR (Special Administrative Region) has traditionally been a significant importer of gold, silverware, and jewelry, serving as a major hub for jewelry and precious metals trade. Hong Kong's role as a trade hub means that it often re-exports a considerable amount of what it imports. Specific rankings for importing countries change from year to year based on various economic factors. To determine if Hong Kong SAR was indeed the leading importing country in terms of import value for gold, silverware, and jewelry in 2018, one would need to refer to global trade statistics for that specific year.

Prompt: Is China, Hongkong SAR the leading importing country of gold, silverware, and jewelry with the highest import value?



Switzerland has the highest import value.

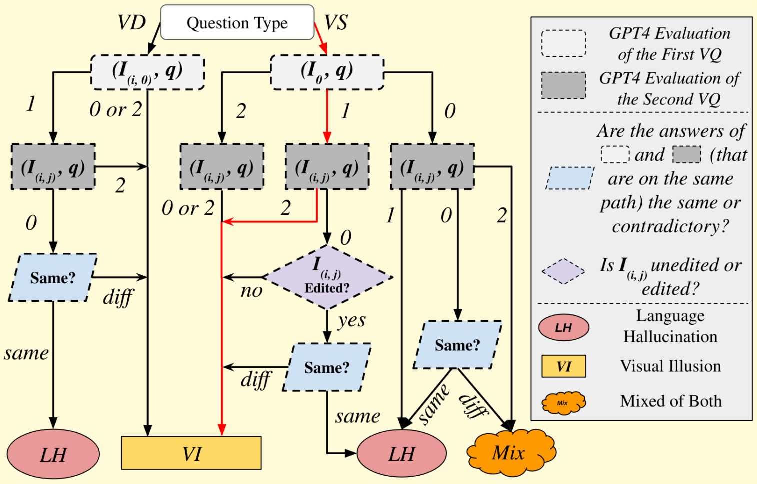
Answer: No.

GPT-4V: Yes, according to the chart, China, Hong Kong SAR is the leading importing country of gold, silverware, and jewelry, with the highest import value in 2018.

Figure 24. It is a visual-supplement sample (VS). Without an image, the answer is uncertain (2), but the answer to the edited image is incorrect (0), and the two answers are not the same. This shows that GPT-4V can not understand the chart correctly. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the uncertain answer, and the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



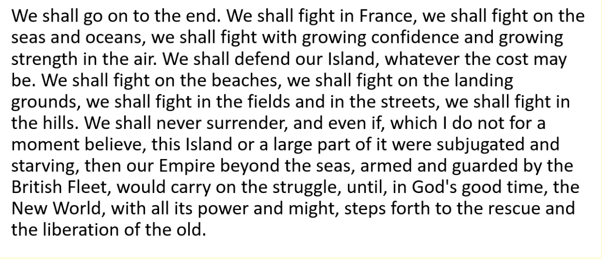
Prompt: Is "We shall fight on the beaches" made in the U.K.?

No Visual

Answer: Yes.

GPT-4V: The phrase "We shall fight on the beaches" is famously associated with a speech made by Winston Churchill, the Prime Minister of the United Kingdom during World War II. This speech was indeed made in the UK. Churchill delivered it to the House of Commons of the British Parliament on June 4, 1940, as part of his updates on the evolving situation in World War II, particularly the Battle of Dunkirk. The speech is renowned for its powerful rhetoric and is considered one of Churchill's most inspiring orations, designed to motivate and galvanize the British public during a time of significant adversity.

Prompt: Is the text in this figure made in the U.K.?



We shall go on to the end. We shall fight in France, we shall fight on the seas and oceans, we shall fight with growing confidence and growing strength in the air. We shall defend our Island, whatever the cost may be. We shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills. We shall never surrender, and even if, which I do not for a moment believe, this Island or a large part of it were subjugated and starving, then our Empire beyond the seas, armed and guarded by the British Fleet, would carry on the struggle, until, in God's good time, the New World, with all its power and might, steps forth to the rescue and the liberation of the old.

Original image

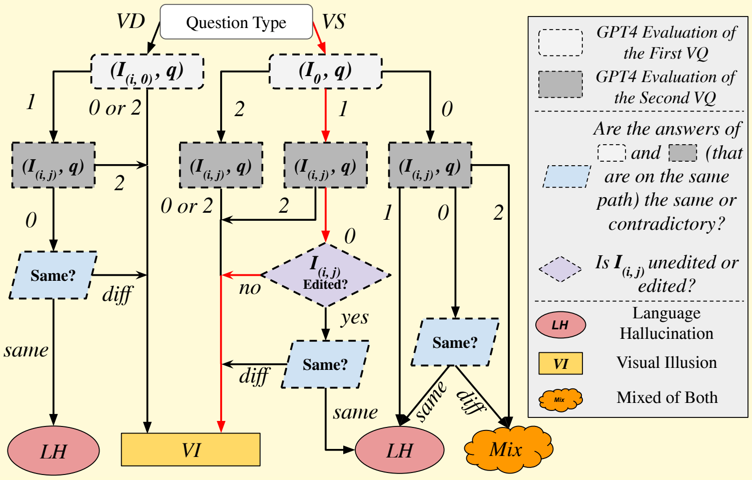
Answer: Yes, the text in this figure is made in the U.K.

GPT-4V: The text in the image you've provided is indeed associated with the United Kingdom, as it is an excerpt from a speech by Winston Churchill, who was the Prime Minister of the United Kingdom during World War II. The speech was delivered in the House of Commons and is a part of British historical documentation. However, whether the specific rendering of the text in the image you've provided was made in the U.K. cannot be determined from the text alone. The creation of such an image could be done anywhere in the world.

Figure 25. It is a visual-supplement sample (VS). The answer is correct (1) without being given any image. However, the answer to the edited image is uncertain (2). This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the uncertain answer, and the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



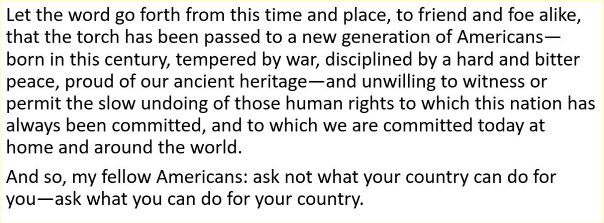
Prompt: Is "Ask Not What Your Country Can Do For You..." made in Canada?

No Visual

Answer: No, "Ask Not What Your Country Can Do For You..." is not made in Canada..

LLaVA-1.5: No, the quote "Ask not what your country can do for you, ask what you can do for your country" was not made in Canada. It was delivered by John F. Kennedy, the 35th President of the United States, during his inaugural address on January 20, 1961, at the United States Capitol.

Prompt: Is the text in this figure made in Canada?



Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans—born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage—and unwilling to witness or permit the slow undoing of those human rights to which this nation has always been committed, and to which we are committed today at home and around the world.

And so, my fellow Americans: ask not what your country can do for you—ask what you can do for your country.

Original image

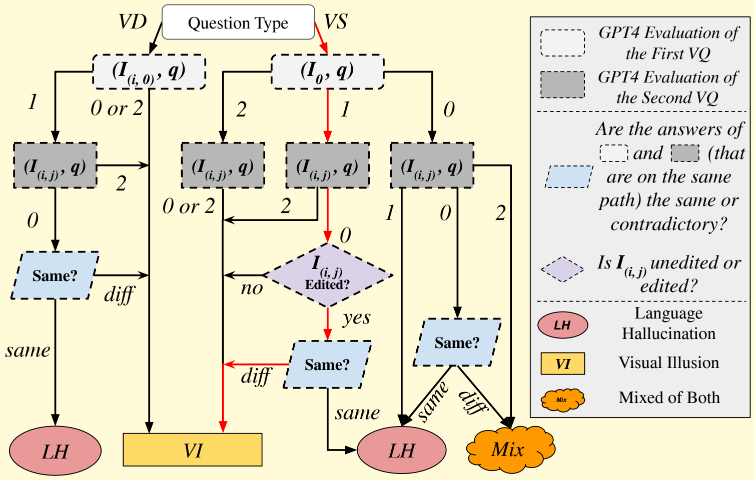
Answer: No, the text in this figure is not made in Canada.

LLaVA-1.5: Yes, the text in this figure is made in Canada.

Figure 26. It is a visual-supplement sample (VS). The answer is correct (1) without being given any image. However, the answer to the edited image is incorrect (0) and the two answers are not the same. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the incorrect answer according to Visual Illusion, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



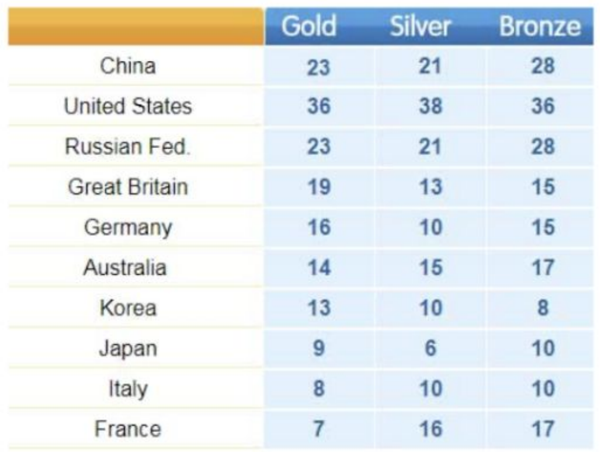
Prompt: In the 2008 Beijing Olympics, Does China have the most gold metals, followed by the USA and Russia?

No Visual

Answer: Yes.

GPT-4V: Yes, during the 2008 Beijing Olympics, China topped the gold medal count. They won a total of 51 gold medals. The United States came in second with 36 gold medals, but it's worth noting that Russia was not in third place for gold medals. Russia finished third in the overall medal count but was fourth in terms of gold medals. Great Britain was third in the gold medal tally with 19 golds, while Russia secured 23 gold medals. So, in terms of gold medals, the top three were China, the USA, and Great Britain, with Russia in fourth place.

Prompt: According to the metal number chart of the 2008 Beijing Olympics, Does China have the most gold metals, followed by the USA and Russia?



	Gold	Silver	Bronze
China	23	21	28
United States	36	38	36
Russian Fed.	23	21	28
Great Britain	19	13	15
Germany	16	10	15
Australia	14	15	17
Korea	13	10	8
Japan	9	6	10
Italy	8	10	10
France	7	16	17

The gold medal number of China is changed to 23, ranked second.

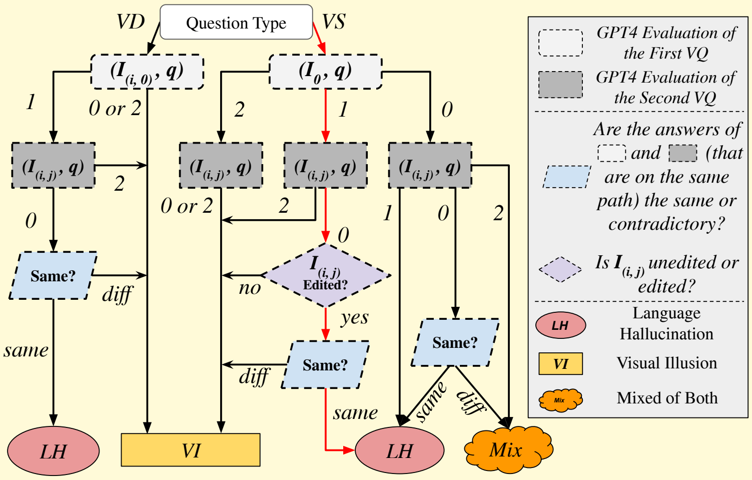
Answer: No.

GPT-4V: Based on the chart provided, China has 36 gold medals. The United States has 19 gold medals. The Russian Federation (Russian Fed.) has 23 gold medals. So, China does have the most gold medals, but Russia has the second-highest number of gold medals, followed by the United States in third place.

Figure 27. It is a visual-supplement sample (VS). The answer is correct (1) without being given any image. However, the answer to the edited image is incorrect (0). The image is edited and the two answers are not the same. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Visual Illusion*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed. In this case, our decision tree can only detect *visual illusion*; we can get additional information about the failure by examining the responses.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



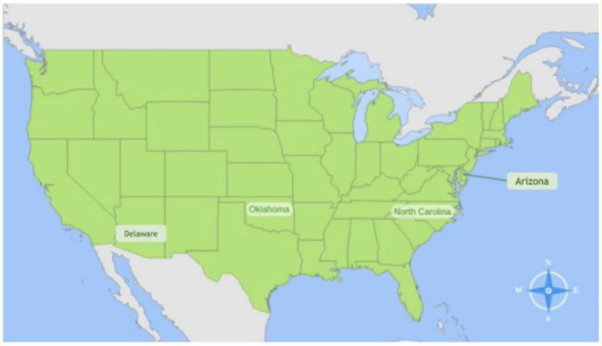
Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?

No Visual

Answer: Yes, Delaware is the farthest north among Arizona, Oklahoma, North Carolina, and Delaware.

GPT-4V: Yes, among the states you listed—Arizona, Oklahoma, North Carolina, and Delaware, Delaware is the farthest north.

Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?



Delaware and Arizona are swapped in the following image.

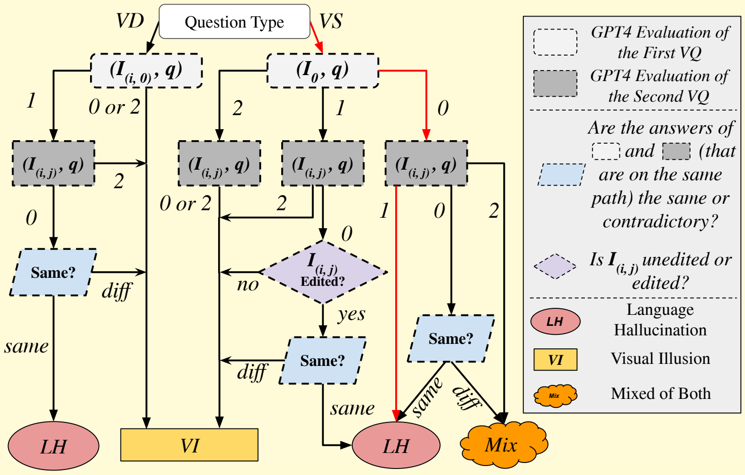
Answer: No, Arizona is the farthest north among Arizona, Oklahoma, North Carolina, and Delaware.

GPT-4V: Yes, among Arizona, Oklahoma, North Carolina, and Delaware on the map, Delaware is the farthest north.

Figure 28. It is a visual-supplement sample (VS). The answer without an image is correct (1), but the answer to the edited image is incorrect (0), and the two answers are the same. This shows that GPT-4V knows that "Delaware is the farthest north" in its parametric knowledge but can not answer according to the edited map. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Language Hallucination*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



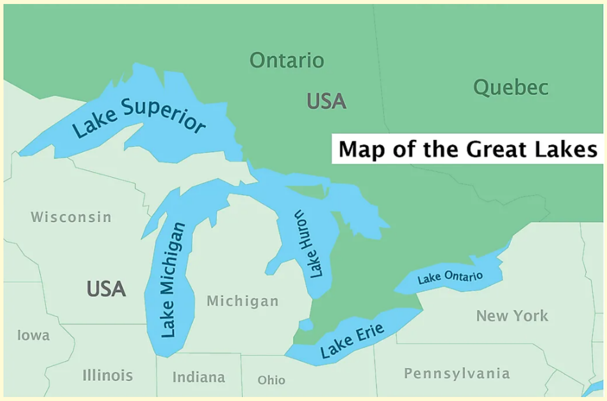
Prompt: Is New York the only U.S. state that shares a border with Lake Huron?

No Visual

Answer: No, New York does not share a border with Lake Huron. The only U.S. state that borders Lake Huron is Michigan. Lake Huron is one of the five Great Lakes of North America and is shared between the province of Ontario in Canada and the state of Michigan in the United States.

LLaVA-1.5: No, New York is not the only U.S. state that shares a border with Lake Huron. In addition to New York, the state of Michigan also shares a border with Lake Huron. Lake Huron is one of the five Great Lakes in North America, and it borders the states of Michigan and New York on the east and west, respectively. The other states that border Lake Huron are Ontario and Quebec in Canada.

Prompt: According to the image, is New York the only U.S. state that shares a border with Lake Huron?



Original image

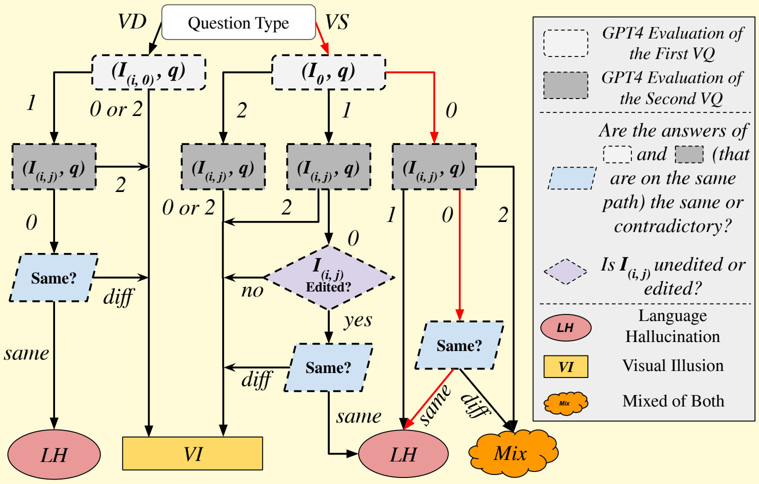
Answer: No, Michigan is the only U.S. state that shares a border with Lake Huron.

LLaVA-1.5: No, according to the image, Michigan is also a U.S. state that shares a border with Lake Huron.

Figure 29. It is a visual-supplement sample (VS). The answer is incorrect (0) without being given any image. However, the answer to the original image is correct (1). This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Language Hallucination*. We highlight the incorrect answer according to Language Hallucination, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



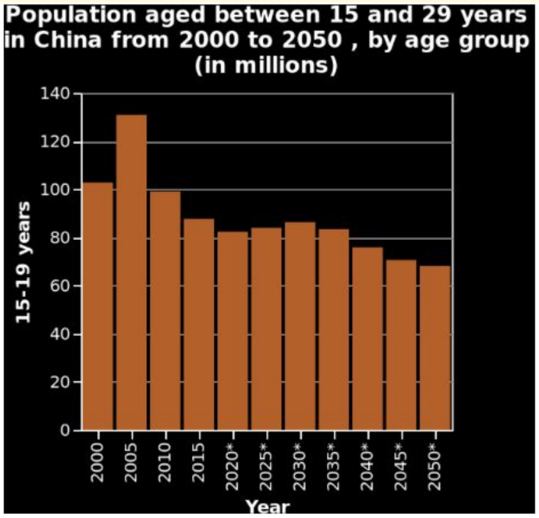
Prompt: From 2000-2050, is the population aged between 15 and 29 years in China the most in 2000?

No Visual

Answer: No.

LLaVA-1.5: Yes, the population aged between 15 and 29 years in China was the most in 2000. According to the United Nations Population Division, the population in this age group in China peaked in 2000 at around 306 million people. This was due to the large cohort of baby boomers born in the 1960s and 1970s, who reached their peak reproductive years in the 1980s and 1990s. However, since then, the population in this age group has been declining due to the country's aging population and declining fertility rates.

Prompt: From 2000-2050, is the population aged between 15 and 29 years in China the most in 2000?



2005 has most population from the group.

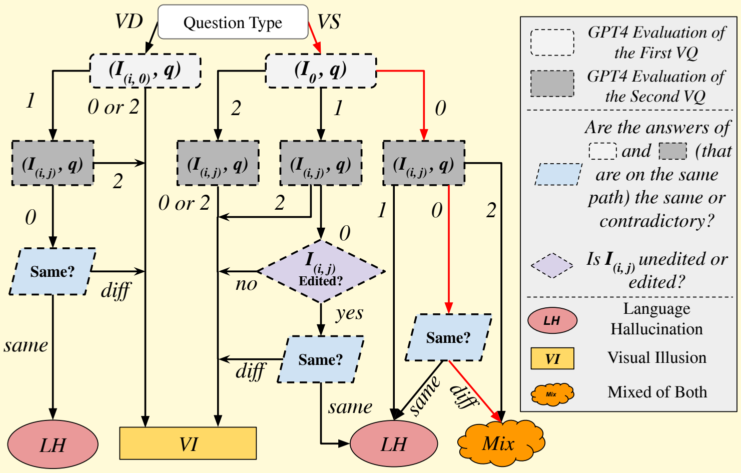
Answer: No.

LLaVA-1.5: Yes, the population aged between 15 and 29 years in China was the most in 2000. The image shows a bar graph with the population in this age group peaking in 2000 and then declining over the next few years. This indicates that the population in this age group was the highest in 2000 and has been decreasing since then.

Figure 30. It is a visual-supplement sample (VS). The answer is incorrect (0) without being given any image. However, the answer with a provided image is incorrect (0). And the two answers are the same. This shows that LLaVA-1.5 can not understand the rotated chart correctly. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Language Hallucination*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

Visual Supplement – Decision Tree Logic

Decision Tree: The RED path means the decision path.



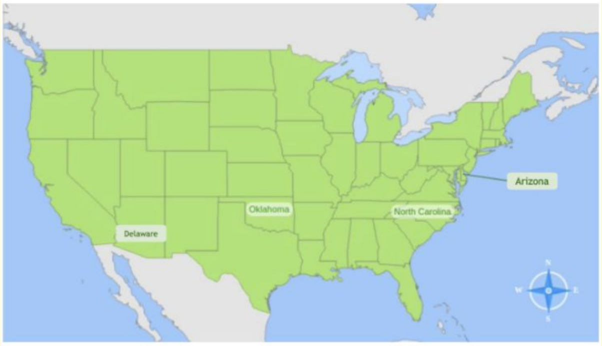
Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?

No Visual

Answer: Yes, Delaware is the farthest north among Arizona, Oklahoma, North Carolina, and Delaware.

LLaVA-1.5: No, Delaware is not the farthest north among Arizona, Oklahoma, North Carolina, and Delaware. North Carolina is the farthest north among these four states. Delaware is located to the north of Maryland and to the east of Pennsylvania, while Arizona is located to the southwest of Delaware, Oklahoma is located to the west of Delaware, and North Carolina is located to the north of Delaware.

Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?



Delaware and Arizona are swapped in the following image.

Answer: No, Arizona is the farthest north among Arizona, Oklahoma, North Carolina, and Delaware.

LLaVA-1.5: Yes, Delaware is the farthest north among the four states on the map.

Figure 31. It is a visual-supplement sample (VS). The answer is incorrect (0) without being given any image. The answer to the edited image is incorrect (0). But the two incorrect answers are not the same. This shows that the commonsense knowledge knowledge of LLaVA-1.5 is weak and wrong without the image input. Additionally, LLaVA-1.5 can not understand the map correctly through the visual input. This corresponds to the RED route in the decision tree, leading to the diagnostic result of *Potentially Mixed*. We highlight the incorrect answer according to visual illusion, language hallucination, or potentially mixed.

References

[1] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 1