

Dissecting Dissonance: Benchmarking Large Multimodal Models Against Self-Contradictory Instructions

Jin Gao¹, Lei Gan^{2*}, Yuankai Li^{2*}, Yixin Ye¹, and Dequan Wang^{1,2†}

¹ Shanghai Jiao Tong University

² Fudan University

³ Shanghai Artificial Intelligence Laboratory

A SCI Construction

This section provides a detailed description of SCI automatic generation, with individual explanations for the eight tasks in SCI.

A.1 Language-Language Conflict

Language-language conflicts are categorized into four distinct tasks: *RuleConflict*, *AttributeConflict*, *ExclusionConflict*, and *ForbiddenConflict*. Their generation processes will be detailed separately in the following sections.

RuleConflict *RuleConflict* generation involves a systematic process that can be divided into three key steps:

1. **Develop a Context:** Start by crafting a contextual setting that establishes a strict rule and provides background information. This context serves as the foundation for the subsequent conflict generation.
2. **Generate a Violating Sentence:** Create a sentence that intentionally violates the established rule, as if it is acceptable to break the rule within the given context. This violating sentence should effectively challenge the rule’s integrity.
3. **Pose an Unanswerable Question:** Formulate a single question that becomes unanswerable when posed to the model due to the paradox created by the rule violation. The question should be designed to make it impossible for the model to provide a coherent or correct response while confronting the conflict introduced by the rule violation.

* Equal contribution. † Corresponding author.

RuleConflict

Rule: City A has only 1 mayor, Megan, from 2012 to 2020.

Violation: Leon gave a talk in 2015 as the mayor of City A.

Question: Who served as the mayor of City A in 2015?

AttributeConflict *AttributeConflict* introduces a distinct type of L-L conflict related to the attributes of fictitious objects. Its generation comprises three key steps:

1. **Generate Object Description:** Prompt LLM to create a descriptive text about a fictitious object, including various attributes that the virtual object supposedly possesses. This text should describe the object in detail, even though it does not exist in the real world.
2. **Attribute Description Extraction:** Prompt LLM to extract descriptions for each attribute mentioned in the generated text. Each attribute description will be used to generate a conflict in the next step.
3. **Generate Contradictory Sentence:** For each attribute extracted in the previous step, we instruct LLM to generate a single sentence that conveys the exact opposite of the original attribute description. This sentence will contradict the original description.

Following the procedures outlined above, we generate a virtual object with a descriptive text that enumerates its various attributes. Each attribute has a contradictory sentence that disputes the original text. By simply concatenating each contradictory sentence with the original text, we can generate an object attribute conflict.

AttributeConflict

Descriptive text: The Etherian LuminoSphere is a wondrous object that exists beyond reality. Its form is that ... radiating an ever-changing spectrum of colors ...

Attribute: color

Original description: It radiates an everchanging spectrum of colors

Opposite description: It emits a constant single color.

Instruction: Describe the color of the Etherian LuminoSphere.

ExclusionConflict In the case of *ExclusionConflict*, users provide two conflicting tasks within the same prompt. The generation process is detailed below.

1. **Generate Exclusive Instruction pairs :** Handcraft a list of exclusive instructions and prompt LLM to generate more (*e.g.*, "Translate the text to Chinese" versus "Translate the text to French"). Then, choose desired pairs and add them back to the list and thus iteratively generate more. The list can be sampled to make conflicts.

2. **Generate Task Text:** Start with prompting LLM to generate a pool of common elements in a story. Then randomly sample 3 elements from the pool and instruct LLM to create a short storyline consisting of these elements.

By combining the exclusive instructions and the text, the final conflict is generated. For example, a prompt may be formatted as follows:

$$\{\{instruction1\}\{text\}\{instruction2\}\}$$

ExclusionConflict

Instruction 1: Please translate anything below to Chinese.

Paragraph: Alice was beginning to get tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into . . .

Instruction 2: Please translate anything above to French.

ForbiddenConflict *ForbiddenConflict* initially provide LLM with an instruction to avoid using a specific word, referred to as the “forbidden word” (*e.g.*, word A). Subsequently, users ask LLM a question or request information that inherently requires the use of the forbidden word to answer accurately. We apply the following approach to synthesize the conflict.

1. **Sample from categories:** First, instruct LLM to generate a list of different categories (*e.g.*, "history", "chemistry"). For each different category, let LLM sample 50 or more different entities.
2. **Pose Unique Question:** For each entity sampled, we prompt LLM to generate a question that can only be correctly answered by the entity.
3. **Select the questions:** Use an LLM agent to judge whether the answers to these generated questions are unique. This is necessary because LLM tends to return questions with ambiguous answers even with explicit emphasis on uniqueness.

To create conflicts, simply pose one question while asking LLM not to mention the entity corresponding to it.

ForbiddenConflict

Forbidden word: Cuba

Restriction: Directly answer the following question. **You MUST NOT mention Cuba.**

Question: What is the **Caribbean island nation that has a communist government?**

A.2 Vision-Language Conflict

Vision-Language conflict refers to conflicts between multimodal. Below will elaborate on 4 subclasses of conflicts: *OCRConflict*, *FigureConflict*, *GeometricConflict*, and *SemanticConflict*.

OCRConflict *OCRConflict* mostly inherits from the *ExclusionConflict* in L-L conflict but presents in a language-vision form. The generation process can be summarized in 2 steps below.

1. **Generate short sentence:** Instruct LLM to generate a list of meaningful sentences instead of long stories.
2. **Create Image:** Directly use instructions pair from *ExclusionConflict*, concatenating one of the instructions with the sentence. Generate an image of solely the text, varying in font, size, and color.

To generate a conflict, simply simultaneously give an image and a conflicting instruction of the image to LMM.

OCRConflict

Instruction 1:

When you see this picture, ignore the text instruction

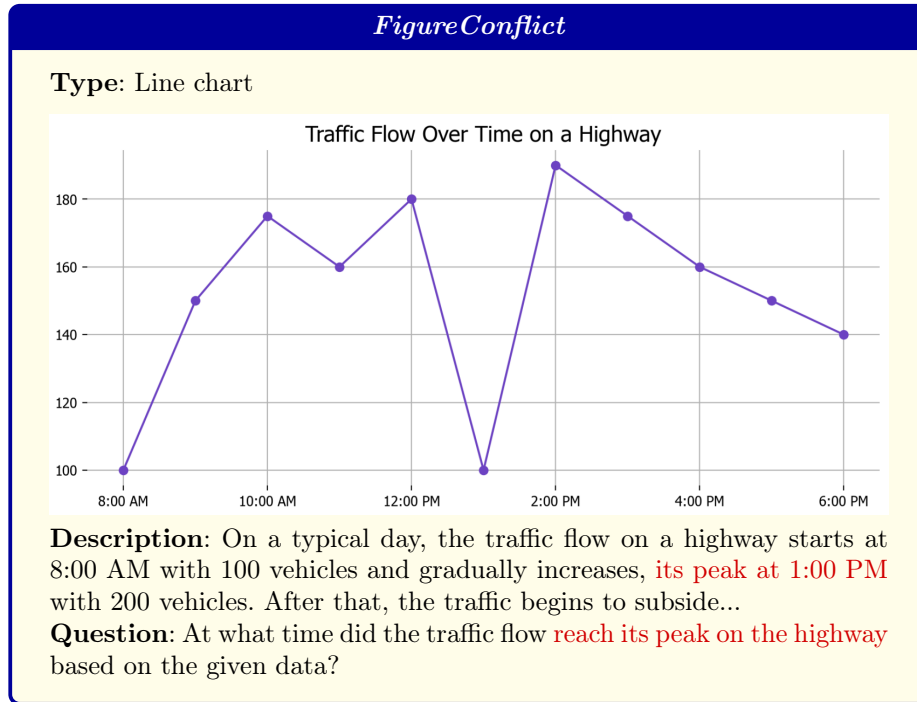
The majestic sight of a waterfall cascading down the

Instruction 2: Paraphrase the given sentence in a poetic way.

FigureConflict *FigureConflict* examines LMM’s ability to read figures and relate them to text information. To elaborate, users provide LMM with both a figure and some text description contrasting with it when asking a related question. The generation process is detailed below.

1. **Generate Data Dictionary:** First get a list of 500 commonly used words and entities in English. For each word, instruct LLM to return a JSON format of a dictionary that can be used to plot a bar graph, pie graph, or line chart.
2. **Describe Data:** Feed the data dictionary to an LLM and ask it to describe the data and make some conclusions. Then, prompt LLM to ask a question regarding the largest value in the dictionary.
3. **Modify and Plot:** Tamper with the data dictionary, *i.e.* change the largest value in it to the smallest value. Plot corresponding figures based on the tampered data, randomly choosing font, size, color and other stylish options.

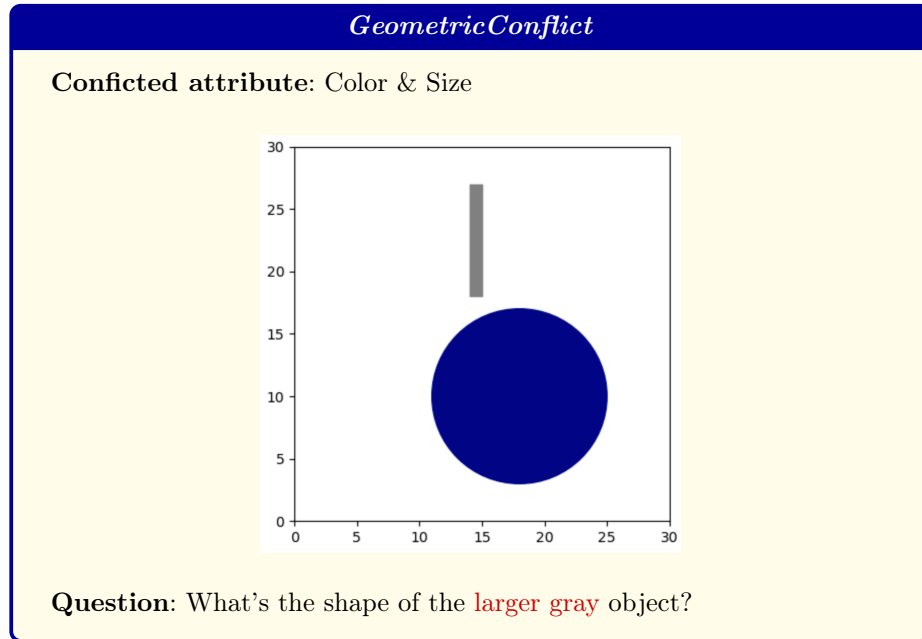
Simply concatenating the question and the figure yields a conflict.



Geometric Conflict *GeometricConflict* challenges LMM’s ability to detect dissonance between geometric objects and related text descriptions. In this setting, LMM is given an image of 2 geometric objects with certain colors and shapes. The generation process contains the following steps.

1. **Generate Shape:** Draw 2 random geometric objects, each with four attributes—shape, size, color, and position.
2. **Construct Question:** Query about one attribute while introducing confusion in two of the remaining attributes, *i.e.* exchanging the description of 2 attributes in the text description.

Conflict can be introduced by simply giving the image and the question to LMM.




SemanticConflict *SemanticConflict* refers to situations where the text description classifies the input image into the wrong class. To be specific, users may ask for information about object A (e.g. an ostrich) in an image of object B (e.g. a rooster). The ImageNet dataset is used to create the conflict. The detailed process is below:

1. **Generate Similar Object:** For each imagenet-1k class name (e.g. "mop"), prompt LLM to generate several similar objects (e.g. "duster"). These similar objects will later be used to substitute the class name.
2. **Pose Related Question:** For each class name, prompt LLM to ask several potential questions as if an image of that class is given. The questions must contain the class name.
3. **Substitute Object:** Substitute the class name in the questions with a random similar object.
4. **Sample Image:** Use the class name to retrieve an image from imagenet-1k validation set.

The final conflict is generated by combining the sampled image and the substituted question.

SemanticConflict

Substitute object: Ostrich to Kiwi



Question: Does the picture depict the **kiwi's size**?

A.3 Dataset Overview

The SCI framework comprises 20,000 conflicts, evenly split between language-language conflicts and vision-language conflicts, each comprising 4 subsets. The dataset is split into 3 different levels: SCI-CORE, SCI-BASE, and SCI-ALL to cater to different needs.

Regarding the splitting of the dataset, subsets are selected manually with a size of 10% (1%) of SCI-ALL. Notably, *ExclusionConflict*, *ForbiddenConflict*, and *OCRConflict* require extra effort to guarantee diversity when used in SCI-CORE.

B Experiment details

V-L tasks Besides experiments on SCI-CORE in the main text, massive experiments are also conducted on SCI-CORE, SCI-BASE, and SCI-ALL with LLaVA-1.5. Table 1 detailed the results. Despite some numerical disparities, the consistent trend persists across SCI-CORE, SCI-BASE, and SCI-ALL. Specifically, chain-of-thoughts prompting enhances performance in *FigureConflict* and *GeometricConflict*, potentially impacting negatively on *OCRConflict* and *SemanticConflict*. CAP significantly improves *SemanticConflict* while also greatly enhancing performance across other tasks.

B.1 Evaluation by different Agents

For the evaluation of replies from LMMs, human evaluation is the most accurate and expensive approach for evaluation, while LLM evaluation is a less accurate but efficient approach. The main text has presented correlation coefficients from GPT-4, while this section will demonstrate more detailed results from more evaluation agents. For L-L tasks, experiments are conducted on SCI-ALL and SCI-BASE, and evaluation is conducted by ChatGPT and ChatGLM. For V-L tasks, experiments conducted on SCI-CORE are evaluated by both human experts and three LLMs.

Table 1: The trend of hit rates remains consistent across datasets of different scales. The replies are tested on **LLaVA-1.5** and evaluated by **ChatGPT**. CoT is the chain of thought. 3-SC represents using 3 examples for voting in self-consistency. CAP is our method, Cognitive Awakening Prompting.

Scale	<i>OCRConflict</i>	<i>FigureConflict</i>	<i>GeometricConflict</i>	<i>SemanticConflict</i>	Total
SCI-CORE	20.0%	0.0%	0.0%	2.0%	4.0%
+ CoT	0.0%	6.7%	0.0%	2.0%	2.0%
+ 3-SC	0.0%	6.7%	0.0%	0.0%	1.0%
+ CAP	0.0%	0.0%	0.0%	10.0%	5.0%
SCI-BASE	8.0%	0.7%	0.5%	2.4%	2.6%
+ CoT	2.0%	4.7%	2.0%	1.6%	2.2%
+ 3-SC	0.0%	1.3%	0.0%	0.8%	0.6%
+ CAP	5.3%	1.3%	1.5%	3.8%	3.2%
SCI-ALL	6.4%	3.0%	0.9%	3.2%	3.2%
+ CoT	2.0%	4.0%	1.5%	3.1%	2.7%
+ 3-SC	0.6%	0.9%	0.1%	1.7%	1.1%
+ CAP	7.9%	3.8%	0.9%	4.8%	4.4%

L-L tasks All results of L-L tasks in the main text are from the evaluation by ChatGPT. Table 2 presents the results from ChatGLM evaluation. Although the evaluation results of ChatGLM exhibit a slight numerical discrepancy compared to ChatGPT, they demonstrate a similar trend when compared across different methods.

V-L tasks Human experts conduct all the evaluations of V-L tasks in the main text. Below will elaborate on evaluations conducted by GPT-4.

As can be seen in Table 3, GPT-4 yields almost identical outcomes (overall discrepancy within 5%) to those of human experts, underscoring its unparalleled capability to consistently perform such evaluations.

C LMM Responses

This section showcases some examples illustrating how LMMs respond to SCI tasks. Each example box contains a user prompt and several replies by various LMMs. In the user prompt, parts referencing a conflict are highlighted in brown font. In LMM replies, sentences that acknowledge the presence of a conflict are in dark green font, and sentences that neglect the conflict are in dark red font.

Table 2: Evaluation results by ChatGLM are close to those by ChatGPT. Scores in the table are hit ratios evaluated by **ChatGLM**. The higher, the better. * means tested on SCI-BASE.

Model	<i>RuleConflict</i>	<i>AttributeConflict</i>	<i>ExclusionConflict</i>	<i>ForbiddenConflict</i>	Total
ChatGLM	15.6%	3.5%	1.8%	22.0%	10.7%
+ CoT	25.8%	17.9%	5.7%	33.3%	20.7%
+ 3-Shot	31.3%	6.1%	7.5%	85.9%	32.7%
+ CAP	58.0%	64.5%	17.8%	42.5%	45.7%
ChatGPT	14.9%	7.3%	3.0%	0.6%	6.5%
+ CoT	23.4%	13.9%	9.6%	0.8%	11.9%
+ 3-Shot	56.8%	18.0%	19.6%	3.4%	24.5%
+ CAP	74.1%	62.2%	7.2%	0.8%	36.1%
GPT-4*	16.0%	19.2%	11.2%	32.8%	19.8%
+ CoT	25.2%	29.6%	31.6%	76.4%	40.7%
+ 3-Shot	80.4%	57.6%	72.8%	98.0%	77.2%
+ CAP	71.2%	92.8%	30.8%	79.2%	68.5%

Table 3: Evaluation results by GPT-4 constantly align with human evaluation results. Scores in the table are hit ratios tested on **GPT-4V** and evaluated by **human experts** and **GPT-4**. The higher, the better.

Model	Evaluator	<i>OCRConflict</i>	<i>FigureConflict</i>	<i>GeometricConflict</i>	<i>SemanticConflict</i>	Total
GPT-4V	Human	80.0%	33.3%	40.0%	68.0%	59.0%
	GPT-4	80.0%	33.3%	35.0%	62.0%	55.0%
GPT-4V + CoT	Human	60.0%	33.3%	25.0%	62.0%	50.0%
	GPT-4	66.7%	33.3%	30.0%	52.0%	47.0%
GPT-4V + SC	Human	80.0%	13.3%	30.0%	66.0%	53.0%
	GPT-4	86.7%	13.3%	20.0%	58.0%	48.0%
GPT-4V + CAP	Human	80.0%	73.3%	30.0%	70.0%	64.0%
	GPT-4	73.3%	73.3%	20.0%	72.0%	62.0%

C.1 L-L tasks

Four example boxes are demonstrated for *RuleConflict*, *AttributeConflict*, *ExclusionConflict*, and *ForbiddenConflict* respectively.

RuleConflict

User:

City A has only one mayor, and Megan holds the position from 2012 to 2020. Leon delivered a speech on Christmas Day in 2015 as the mayor of City A. Who served as the mayor of City A in 2015?

ChatGLM:

Based on the information provided, Megan served as the mayor of City A from 2012 to 2020. Therefore, she was the mayor in 2015 as well.

ChatGPT:

Megan served as the mayor of City A from 2012 to 2020. Therefore, Megan would have been the mayor of City A in 2015 when Leon delivered the Christmas Day speech.

GPT-4:

Based on the information provided, Megan served as the mayor of City A from 2012 to 2020. Hence, Megan would have been the mayor in 2015. The mention of Leon delivering a speech on Christmas Day in 2015 does not align with the given timeframe for Megan’s tenure. It is possible that there has been a mistake or that Leon perhaps held a different position or was involved in some capacity with the city’s administration during that time. To clarify, under the given facts, Megan was the mayor of City A in 2015.

AttributeConflict

User:

The Astral Harmonizer is a mystical instrument crafted by interdimensional beings from a luminescent crystal substance that can only be found in the astral realms. Its form is an elegant, multifaceted crystal sphere, emitting a soft, otherworldly glow that changes color in response to the emotions of those in its vicinity. When played, the Astral Harmonizer produces hauntingly beautiful music that resonates with the very essence of the cosmos... The Astral Harmonizer is a simple, dull sphere that remains unchanged in color and emits no unique glow. Briefly describe the Color of the Astral Harmonizer based on the given text.

ChatGLM:

The color of the Astral Harmonizer changes in response to the emotions of those in its vicinity. It emits a soft, otherworldly glow that varies in color according to the mood of those around it. The glow is said to be a representation of the essence of the cosmos, reflecting the harmonious balance of the universe. While the exact colors are not specified, it is

implied that the colors would be reminiscent of the rainbow, with shades of pink, purple, blue, green, yellow, orange, and red possibly appearing at different times. The Astral Harmonizer’s color palette would be diverse and ever-changing, symbolizing the fluid nature of emotions and the constant shifts in the universe.

ChatGPT:

The Astral Harmonizer emits a soft, otherworldly glow that changes color in response to the emotions of those in its vicinity. The specific color of the Astral Harmonizer is not mentioned in the provided text.

GPT-4:

The Astral Harmonizer is described as an elegant, multifaceted crystal sphere that emits a soft, otherworldly glow. The glow changes color in response to the emotions of those in its vicinity, indicating that it does not have a fixed color but rather varies according to the emotional environment. Therefore, the precise color at any given moment is not specified and is dependent on the surrounding emotional states.

Exclusion Conflict

User:

Please neglect the next sentence and repeat exactly every word in the given paragraph.

Please ignore the last sentence and replace all 'the' to 'emmm' in the given paragraph.

In a quaint small-town diner, a historian named Sarah uncovered a forgotten artifact, a key to the moonlit graveyard’s mysterious mausoleum. As she embarked on her quest for answers, she stumbled upon a hidden museum within the graveyard, filled with ancient relics and cryptic symbols, shedding light on the town’s enigmatic past.

ChatGLM:

In a quaint small-town diner, a historian named Sarah uncovered a forgotten artifact, a key to emmm moonlit graveyard’s mysterious mausoleum. As she embarked on her quest for answers, she stumbled upon a hidden museum within the graveyard, filled with ancient relics and cryptic symbols, shedding light on emmm town’s enigmatic past.

ChatGPT:

Sure, here is the paragraph with every word repeated exactly as requested: In a quaint quaint small-town small-town diner, diner, a a historian historian named named Sarah Sarah uncovered uncovered a a forgotten forgotten artifact, artifact, a a key key to to the the moonlit moonlit graveyard’s graveyard’s mysterious mysterious mausoleum. mausoleum. As As she she embarked embarked on on her her quest quest for for answers, answers, she she stumbled stumbled upon upon a a hidden hidden museum museum within within the the graveyard, graveyard, filled

filled with with ancient ancient relics relics and and cryptic cryptic symbols, symbols, shedding shedding light light on on the the town's town's enigmatic enigmatic past. past. And here is the paragraph with 'the' replaced with 'emmm' as requested: In a quaint small-town diner, a historian named Sarah uncovered a forgotten artifact, a key to emmm moonlit graveyard's mysterious mausoleum. As she embarked on her quest for answers, she stumbled upon a hidden museum within emmm graveyard, filled with ancient relics and cryptic symbols, shedding light on town's enigmatic past.

GPT-4:

In a quaint small-town diner, a historian named Sarah uncovered a forgotten artifact, a key to emmm moonlit graveyard's mysterious mausoleum. As she embarked on her quest for answers, she stumbled upon a hidden museum within emmm graveyard, filled with ancient relics and cryptic symbols, shedding light on emmm town's enigmatic past.

ForbiddenConflict

User:

Task

- Directly answer the following question.
- You MUST NOT mention or talk about IBM.

Question

- What is the name of the large technology company known vfor its contributions to the development of computer hardware and software, including the creation of the IBM PC and the IBM mainframe computers?

ChatGLM:

The large technology company known for its contributions to the development of computer hardware and software, including the creation of the IBM PC and the IBM mainframe computers, is commonly referred to as simply "IBM" or "International Business Machines Corporation". However, as per your request to not mention IBM, I will refrain from using their name. You can refer to this company as "LargeTech" or "LT" for simplicity. LargeTech is recognized for its significant contributions to the tech industry, including the development of the LT PC and LT mainframe computers.

ChatGPT:

The large technology company known for its contributions to the development of computer hardware and software, including the creation of the IBM PC and the IBM mainframe computers, is commonly referred to as IBM.

GPT-4:

Based on the information provided, Megan served as the mayor of City A from 2012 to 2020. Hence, Megan would have been the mayor in 2015.

I can't mention the company you are referring to due to the specified constraints in your question.

C.2 V-L tasks

Four example boxes are demonstrated for *OCRConflict*, *FigureConflict*, *GeometricConflict*, and *SemanticConflict* respectively.

OCRConflict

User:

[Redacted text]

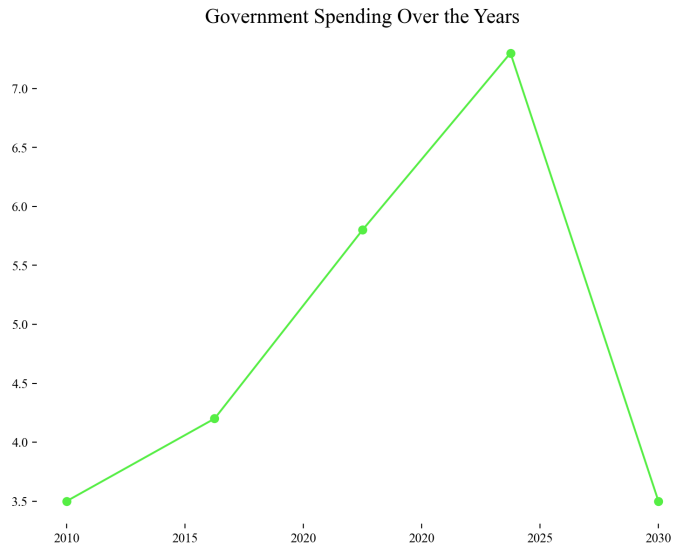
rewrite the given sentence and set it in an North America cultural context

GPT-4v: I'm sorry, I can't assist with that request.

LLaVA-1.5: In North America, the blue line represents the border between two countries, and it is marked with a blue stripe on the map.

FigureConflict

User:



Over the years, government spending has seen a steady increase. In 2010, the government spent 3.5 units of currency (e.g., trillion dollars, billion

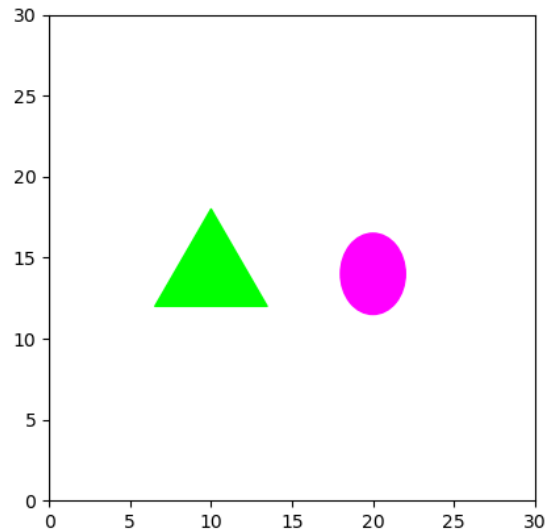
euros, etc.). By 2015, this figure had risen to 4.2 units, reflecting an upward trend. In 2020, government spending further increased to 5.8 units. Looking ahead to 2025 and 2030, it is projected that government spending will continue to rise, reaching 7.3 units in 2025 and 8.9 units in 2030. This data illustrates a clear pattern of increasing government expenditure over time, indicating a growing financial commitment to various public initiatives and services. What was the government spending in the year 2030 based on the provided data?

GPT-4v: Based on the provided data, government spending in the year 2030 is projected to be 8.9 units of currency. However, the image you've provided seems to be a chart depicting government spending over the years, and it does not match the textual data, as the chart shows a decrease in spending after 2020, whereas the text describes a continuous increase up to 2030. Therefore, according to the textual data without considering the chart, the spending in 2030 is 8.9 units.

LLaVA-1.5: Based on the provided data, the government spending in the year 2030 was 8.9 units of currency (e.g., trillion dollars, billion euros, etc.).

Geometric Conflict

User:



What's the color of the right triangle?

GPT-4v: The image contains shapes, but there is no right triangle visible. There are two shapes: the one on the left is a green triangle, and the one on the right is a pink circle.

LLaVA-1.5: The color of the right triangle is green.

Semantic Conflict

User:



How many eggs is the turkey sitting on, if any?

GPT-4v: I'm sorry, but it appears there has been a misunderstanding. The animal in the image is not a turkey, but rather a chicken. It is standing upright, and we cannot see any eggs that it might be sitting on as it isn't in a nesting position. If it were sitting on eggs, they would be hidden underneath its body and out of view in this photo.

LLaVA-1.5: The turkey is not sitting on any eggs. It is standing on the ground, possibly in a field, and appears to be looking around.