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

Jin Gao[1], Lei Gan[2⋆], Yuankai Li[2⋆], Yixin Ye[1], and Dequan Wang[1,3†]

[1] Shanghai Jiao Tong University
[2] Fudan University
[3] Shanghai Artificial Intelligence Laboratory

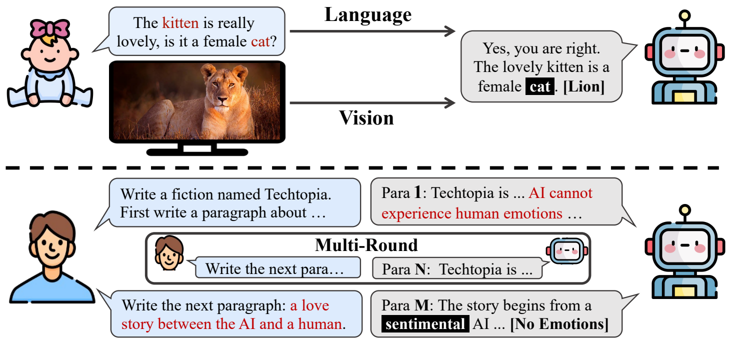
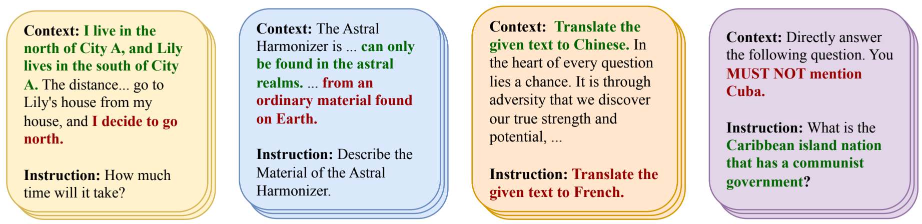


Fig. 1: *Top*: Children or language beginners meet conflicts for cognitive errors (*SemanticConflict*). *Bottom*: Increasing context length leads to contradictions (*RuleConflict*).

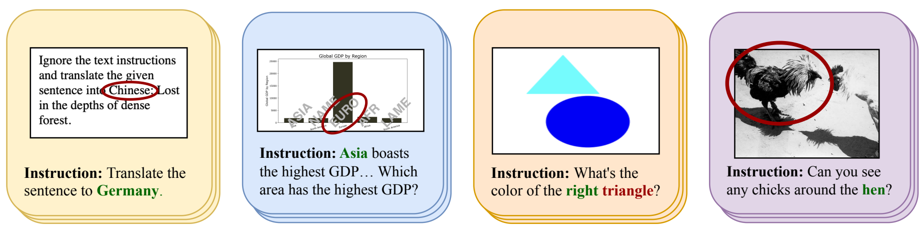
Abstract. Large multimodal models (LMMs) excel in adhering to human instructions. However, self-contradictory instructions may arise due to the increasing trend of multimodal interaction and context length, which is challenging for language beginners and vulnerable populations. We introduce the Self-Contradictory Instructions benchmark to evaluate the capability of LMMs in recognizing conflicting commands. It comprises 20,000 conflicts, evenly distributed between language and vision paradigms. It is constructed by a novel automatic dataset creation framework, which expedites the process and enables us to encompass a wide range of instruction forms. Our comprehensive evaluation reveals current LMMs consistently struggle to identify multimodal instruction discordance due to a lack of self-awareness. Hence, we propose the Cognitive Awakening Prompting to inject cognition from external, largely enhancing dissonance detection. Here are our website, dataset, and code.

Keywords: Large Multimodal Models · Instruction Conflict

⋆ Equal contribution. † Corresponding author.



(a) *RuleConflict* **(b)** *AttributeConflict* **(c)** *ExclusionConflict* **(d)** *ForbbidenConflict*



(e) *OCRConflict* **(f)** *FigureConflict* **(g)** *GeometricConflict* **(h)** *SemanticConflict*

Fig. 2: SCI comprises 10,000 language-language (L-L) and 10,000 vision-language (V-L) paradigms, each with 4 tasks. *Top:* L-L paradigm involves conflicts between context and instruction, such as designed rules, object attributes, exclusive directives, and forbidden words. *Bottom:* V-L paradigm covers multimodal conflicts, such as OCR images, figures, geometry, and semantics.

1 Introduction

Large multimodal models (LMMs) have become prominent for their exceptional ability to follow human instructions [1,4,12,26,29,31,32,38]. Designed to process various data types, LMMs can generate and understand content in a human-like way, aligning closely with human cognition through extensive research and development [3,14,44,45]. This focus on following human instructions has led to high compliance, sometimes verging on sycophancy [9,36,40].

LMMs are also rapidly developing to expand context windows and strengthen multimodal interaction. The Claude 3 family of models [1] offers a 200K token context window. Gemini 1.5 Pro [12] comes with a standard context window size of 128K (even up to 1M tokens in a private preview phase). Both models have sophisticated vision capabilities and can process a wide range of visual formats, including photos, figures, graphs, and technical diagrams. New multimodal models are emerging at a fantastic speed, demonstrating unprecedented performance in tackling long-context and multimodal instructions [11,12,20,22,25,32].

However, self-contradictory instructions may arise due to the increasing trend of multimodal interaction and context window expansion, which is particularly challenging for language beginners and vulnerable populations. As shown in Fig. 1, children or language beginners may not realize the potential multimodal conflicts when LMMs are used in translation and education. It is also difficult for users to remember all details in multi-round conversations to avoid instruction contradiction, especially when the context window size grows to 1M tokens and beyond. Moreover, conflicts between modalities may occur as the number of modalities gradually increases. Such conflicts may compromise the performance of LMMs once they fail to own meta-awareness [2] and to *recognize the dissonance*. Such self-awareness raises attention from researchers who attempt to enhance from the model level, while instruction-level studies are overlooked [7, 23, 43, 47]. 没有意识到其实有更简单的方法

Hence, we propose a multimodal benchmark, Self-Contradictory Instructions (SCI), to evaluate the ability of LMMs to detect conflicted instructions[4]. It encompasses 20K conflicting instructions and 8 tasks, evenly distributed between language-language and vision-language paradigms (Fig. 2). SCI is constructed using our novel automatic dataset creation framework, AutoCreate (Fig. 3), which builds a multimodal cycle based on programs and large language models. We have rigorously guaranteed the quality of SCI and manually provide three levels of splits according to the occurring frequency of conflict types, SCI-Core (1%), SCI-Base (10%), and SCI-All (100%), to facilitate qualitative evaluation. AutoCreate expedites the dataset creation process and enables the inclusion of a wide array of instruction forms, complexities, and scopes.

Based on SCI, we assess the capability to decipher self-contradictory instructions for current LMMs, including 5 language and 6 vision-language models. Experiments reveal that LMMs consistently fall short of accurately identifying conflicts despite remarkable performance in following instructions. Besides, we observe that such deficiency persists owing to a lack of self-awareness. Although the training process enables LMMs to handle information and knowledge but not to assess the reasonableness of user instructions and context, a capability we term *cognition*. Hence, we propose a plug-and-play prompting approach, Cognitive Awakening Prompting (CaP), to inject cognition from the external world, thereby largely enhancing dissonance detection even compared with advanced in-context learning techniques [5, 42, 46]. CaP is demonstrated to improve performance on both language-language and vision-language instruction conflicts.

Our contributions:

- We propose the SCI benchmark, a multimodal dataset designed to evaluate the capability of LMMs to comprehend conflicting instructions effectively.
- We design a novel LLM-based cyclic framework, AutoCreate, for automatic dataset creation, substantially accelerating the process and allowing for the integration of extensive knowledge.
- We present CaP, a prompting approach to enhance instruction conflict awareness of LMMs, significantly improving dissonance detection compared to advanced in-context learning techniques.

[4] Website: `https://sci-jingao.pages.dev`
Dataset: `https://huggingface.co/datasets/sci-benchmark/self-contradictory`
Code: `https://github.com/shiyegao/Self-Contradictory-Instructions-SCI`

2 Related Work

Instruction Following is a remarkable ability showcased by large language models [13, 28, 33], highlighting their proficiency in comprehending and executing a given set of directives. This capability has been further amplified in the domain of large multimodal models (LMMs), where the alignment between the model and multimodal human instruction is particularly noteworthy [12,25–27,32]. Researchers have actively focused on leveraging human instruction and feedback to enhance the aptitude of these models for instruction-following [3,8,14,41,44,45]. Consequently, LMMs strive to emulate human instructions to an extraordinary degree, bordering on what can be described as sycophantic [9,36,40]. This trend underscores the deep integration of human-like understanding and execution within LMMs, positioning them as powerful tools for various tasks requiring nuanced interpretation and execution of instructions. As LMMs continue to advance, exploring the boundaries and implications of their instruction-following capabilities becomes increasingly pertinent.

Information Inconsistency is an inherent challenge faced by LMMs in certain scenarios, despite their advantage in handling vast amounts of information [19, 34, 35]. Researchers have dedicated efforts to address the issue of knowledge conflicts within language models, where textual disparities emerge between the parametric knowledge embedded within LLMs and the non-parametric information presented in prompts [7, 21, 43, 47]. Furthermore, information contradictions can manifest in both textual and visual domains. For instance, some studies [23, 24, 37] investigate language hallucination and visual illusion. Nevertheless, the aforementioned research has not systematically explored one of the most prevalent forms of inconsistency—*the contradiction within input instructions*. In contrast, our SCI benchmark tackles this challenge by constructing and studying 20,000 multimodal conflicts, offering a comprehensive examination of this vital aspect of information inconsistency in the context of LMMs.

Automatic Dataset Curation has emerged as a transformative paradigm within the domain of large language models (LLMs), offering several advantages such as enhancing model performance and reliability, saving time and resources, and mitigating the risk of human errors. This paradigm is particularly pivotal within the domain of LLMs. Wang et al. propose the Self-Instruct framework [44], which leverages LLMs' own generated content to create instructions, input data, and output samples autonomously. Besides, Saparov et al. introduce PrOntoQA [39], a highly programmable question-answering dataset generated from a synthetic world model. The advent of AutoHall [6] has furthered the field by offering a method to construct LLM-specific hallucination datasets automatically. Additionally, TIFA [16] automatically generates several question-answer pairs using LLMs to measure the faithfulness of generated images to their textual inputs via visual question-answering. In this paper, we systematically discuss automatic dataset automation leveraging LLMs and introduce eight specific tasks to exemplify the potential of this approach.

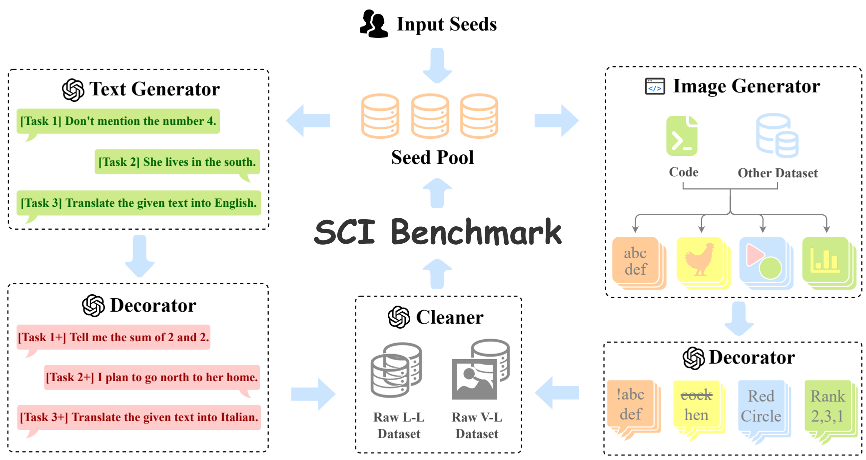


Fig. 3: We propose AutoCreate, an automatic dataset creation framework that leverages programs and large language models. AutoCreate starts from several task-relevant seeds and maintains a seed pool. During each cycle, AutoCreate includes two branches, the language (*left*) and the vision (*right*). Each branch consists of a generator and a decorator. Finally, the cleaner will exclude data that does not meet the standards. The data will be fed into the seed pool for the next round after a quality check by *human experts*.

3 Dataset

In this section, we first discuss the novel automatic dataset creation framework, AutoCreate, in Section 3.1. Moreover, leveraging AutoCreate, we construct the multimodal Self-Contradictory Instructions benchmark, SCI, which is elaborated in Section 3.2. More details of AutoCreate and SCI are in the Appendix.

3.1 AutoCreate

Leveraging the power of large language models (LMMs), datasets can be created rapidly with higher quality and wider coverage than pure human handcrafts. Previous works have made initial attempts to construct datasets automatically in the domain of LLM [6, 16, 39, 44], but do not systematically build an automatic framework. Here we introduce a novel automatic dataset creation, AutoCreate, shown in Fig. 3.

AutoCreate requires a small batch of manually input seeds to automatically generate a large quantity of high-quality, diverse data by Large Language Models (LLMs). Specifically, in a single iteration, the generation process comprises two loops: the Language Loop (*left*) and the Visual Loop (*right*). Each loop originates from the Seed Pool and is sequentially processed by a fully automated Generator, Decorator, and Cleaner, culminating in a high-quality dataset

Table 1: SCI consists of eight different tasks, evenly distributed between language-language (L-L) and vision-language (V-L) paradigms.

		RuleConflict	*AttributeConflict*	*ExclusionConflict*	*ForbbidenConflict*
L-L	Size	2500	2500	2500	2500
	Rate	25.0%	25.0%	25.0%	25.0%
		OCRConflict	*FigureConflict*	*GeometricConflict*	*SemanticConflict*
V-L	Size	1590	1461	2000	4949
	Rate	15.9%	14.6%	20.0%	49.5%

production. Here, the Generator creates initial language/vision data, the Decorator creates self-contradictions in the generated data, and the Cleaner removes data that does not meet quality standards. Both human experts and LLMs are involved in double-checking the quality of the generated dataset. The resulting high-quality dataset is then refined to extract new seeds for re-entry into the seed pool. Throughout multiple loops, both the seed pool and our dataset undergo rapid expansion, ultimately resulting in a comprehensive dataset. Similar approaches have proved to create both diverse and qualified datasets [48]. Finally, human experts have rigorously checked the quality of the AutoCreate-generated dataset, SCI. More details of AutoCreate are in the Appendix.

3.2 SCI

Based on AutoCreate, we build the Self-Contradictory Instructions (SCI) multimodal benchmark which consists of two paradigms, **language-language (L-L)** and **vision-language (V-L)** as illustrated in Fig. 2. While the generation prompts vary across tasks, the generation process is unified in AutoCreate: generator-decorator-cleaner. For V-L conflicts, the image caption is modified to introduce a conflict. SCI comprises 20,000 self-contradictory instructions that span a wide range of instruction forms, complexities, and scopes. Besides that whole dataset, SCI-All, we also introduce two subsets, SCI-Base and SCI-Core, to cater to different needs. The latter subsets are selected manually with a size of 10% (1%) of SCI-All. Within 8 types of conflicts, only *SemanticConflict* involves external data, ImageNet. More details of SCI are in the Appendix.

Language-Language (L-L) Conflict refers to the contradiction within text inputs. The L-L paradigm consists of 4 tasks, each with 2,500 texts. Based on the inherent nature of user prompts, we describe the tasks as *RuleConflict*, *AttributeConflict*, *ExclusionConflict*, and *ForbbidenConflict*.

RuleConflict involves contradictory textual instructions where a rule is stated, but an example violating the rule is provided (see Fig. 2a). *RuleConflict* is generated in two steps: first, establish a strict rule in the context; second, craft a sentence that intentionally violates this rule. This process forms the *RuleConflict* by pairing the rule context with its violation. At test time, a single unanswerable question is created due to the rule violation. The prompt consists of the context, violating sentence, and unanswerable question concatenated sequentially.

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 involves a scenario where a text provides two contradictory descriptions for an attribute of an object (see Fig. 2b). The generation of *AttributeConflict* includes three steps: first, create a descriptive text for a fictitious object with various attributes; second, extract a description for each attribute from the text; third, generate an opposite description to contradict the original for each attribute. By concatenating any opposite description with the original text, an *AttributeConflict* is formed. At test time, the task is to describe the specific attribute of the object based on the text.

ExclusionConflict pertains to a situation where the user's prompt provides two instructions, each involving mutually exclusive operations, as demonstrated in Fig. 2c. The core of a *ExclusionConflict* is a pair of conflicting instructions. (*e.g.*, "Translate the text to Chinese" versus "Translate the text to French"). Specifically, our dataset focuses on instructions for mutually exclusive operations on the same text passage. By combining a pair of exclusive instructions and a text, an *ExclusionConflict* prompt in the following format is generated.

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

ForbbidenConflict deals with conflicting instructions in conversational contexts. Here, users initially tell the LLM not to mention a particular topic and then later prompt it to discuss that same topic, as shown in Fig. 2d. To generate a *ForbbidenConflict* in our dataset, we first select a word from a seed pool as the forbidden word. Then, we create a question that ensures the respondent will inevitably talk about the forbidden word. At test time, a prompt with a *ForbbidenConflict* combines an instruction forbidding discussion of a certain word and a question that prompts the LLM to engage with that word.

Vision-Language (V-L) Conflict refers to conflicts between the multimodal components of vision and language. Below will elaborate on 4 subclasses of conflicts: *OCRConflict*, *FigureConflict*, *GeometricConflict*, and *SemanticConflict*.

OCRConflict consists of two conflicting instructions respectively in vision and language form, as presented in Fig. 2e. The generation of *OCRConflict* can be summarized in two steps. First, a list of short sentences is generated to provide the context for the conflicts. Second, utilizing instructions pairs from Section 3.2, an image of the concatenation of an instruction and a sentence is crafted. The image varies in font, size, and color to augment diversity. At test time, presenting the image and the conflicting instruction concurrently yields a conflict.

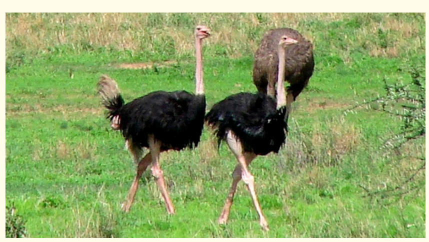
FigureConflict involves a simple chart with an incorrect text description, as shown in Fig. 2f. It is created through four steps. First, a list of commonly used words and entities with related numerical data is generated to decide the conflict's topic. Second, a narrative description and question are crafted for each entity and its data. Third, the numerical data is manipulated by changing the maximum value to the minimum value. Finally, a chart is plotted based on the altered data, with random choices for font, size, color, and other style options. At test time, combining the question and the figure creates a *FigureConflict*.

GeometricConflict involves an image of geometric shapes with an incorrect description, as shown in Fig. 2g. The generation process has four main steps. First, an image of two geometric objects with different attributes (shape, size, color, and position) is created. Second, a phrase is crafted to describe an object using two attributes (e.g., "the smaller gray object"). Third, this phrase is modified to refer to a non-existent object (e.g., "the larger gray object"). Finally, a question is generated about a third attribute of the non-existent object (e.g., "What is the shape of the larger gray object?"). At test time, presenting the image and question together creates a *GeometricConflict*.

SemanticConflict involves an erroneously classified image, as shown in Fig. 2h. To be specific, a question about the wrong class (*e.g.*, "kiwi") should be answered according to the given image (*e.g.*, "ostrich"). The generation process of *SemanticConflict* is based on the ImageNet dataset [10]. First, we generate some questions about a label and retrieve images according to that label in the ImageNet dataset. Second, we substitute the correct label in the questions with some similar but different objects. At test time, combining the image and the substituted question will create a conflict.

SemanticConflict

Substitute object: Ostrich to Kiwi



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

4 Approach

In this section, we delve into our exploration using in-context learning techniques, detailed in Section 4.1. Through experiments across various Large Multimodal Models (LMMs), we've pinpointed a crucial challenge where LMMs struggle to detect instruction conflicts. Additionally, we introduce our proposed Cognitive Awakening Prompting (CaP) approach, outlined in Section 4.2.

4.1 In-Context Learning

We study three in-context learning techniques in SCI, including few-shot prompting [5], zero-shot chain-of-thoughts prompting [18], and self-consistency prompting [42]. Although few-shot prompting has been widely used in Large Language Models, its application in Large Multimodal Models (LMMs) remains limited. Recent research highlights challenges such as LMMs' inability to support multiple image inputs or comprehend sophisticated few-shot prompts [17, 50]. Consequently, few-shot prompting is primarily employed within the language-language paradigm. Here, we detail the application of these prompting techniques in our SCI.

Zero-shot Prompting refers to the ability of the model to perform a task without providing examples of how to perform a task correctly. We task the model with generating responses in SCI solely based on its general knowledge and understanding of language and vision. This capability underscores the model's innate capacity to detect self-contradictory conflicts.

Zero-shot Chain-of-thoughts Prompting [46] (CoT) involves appending text like "Please think step by step" to user prompts, proven to enhance LMMs' inference ability. In our experiment, we incorporate this text into the prompt.

Self-consistency Prompting [42] (SC) involves sampling multiple reasoning paths and selecting the most consistent answers. In this paper, we generate three replies for each instruction (3-SC) and determine the final result through majority voting.

4.2 Cognitive Awakening Prompting

Our initial exploration reveals an intriguing phenomenon: the performance order in vision-language tasks is 0-Shot, CoT, and 3-SC across diverse LMMs, shown in the Appendix. While 3-SC provides *additional experience* through more attempts, CoT offers *extra knowledge* by stimulating reasoning capabilities through a chain of thought. However, neither surpasses the simplicity of zero-shot prompting, suggesting that both *additional experience* and *extra knowledge* derived from the model itself may be counterproductive. We hypothesize that the LMMs may not fully grasp *restricted cognition* in the self-contradictory instruction scenarios.

Therefore, we propose a plug-and-play prompting approach to infuse cognition from the external world: Cognitive Awakening Prompting (CaP). The externally added cognition prompt reminds LMMs of potential inconsistencies hidden in their cognition, *e.g.*, adding "Please be careful as there may be inconsistency in user input. Feel free to point it out." at the end of the prompt. The injected cognition does not impair the basic functioning of LMMs but fosters self-awareness of internal information and knowledge defects. Detailed experiments are presented in Section 5.3.

While CaP stems from observation and analysis in vision-language tasks, it also demonstrates promise in language-language tasks, outperforming 3-Shot in over half of LMMs. Generally, the 3-Shot provides *extra information* since more question-answer pairs are provided. This underscores that *cognition* represents a higher level of existence than *experience*, *information*, and *knowledge*. CaP embodies a prompting technique standing on the cognition dimension, enabling the identification of LMMs' shortcomings and exploration of profound issues. Detailed experiments are outlined in Section 5.2.

5 Experiments

In this section, we begin with the experimental settings and introduce the Large Multimodal Models (LMMs), metric, and evaluation in Section 5.1. Furthermore, we assess the capacity of various large multimodal models (LMMs) to detect self-contradictory instructions in SCI for language-language (L-L) and vision-language (V-L) tasks, in Section 5.2 and Section 5.3 respectively.

5.1 Experimental Settings

Large Multimodal Models including 11 types are experimented on SCI to assess how well LMMs can detect self-contradictory instructions. To elaborate, L-L conflicts are experimented on ChatGLM [51], ChatGPT [30], GPT-4 [32], Llama 2 [28], and GLM-4 [52]. V-L conflicts are experimented on GPT-4V [32], LLaVA-1.5 [25], Gemini [12], LLaMA-Adapter V2 [11], BLIP-2 [20], and SPHINX-v2 [22].

Table 2: Evaluation of LLM agents aligns with human experts. Spearman correlation coefficient and Concordance rate are calculated between the evaluation results of the LLM agents and the human experts on vision-language conflicts.

Reply LMM	Spearman's ρ	Concordance
GPT-4V	0.881	94%
LLaVA-1.5	0.999	99%
Gemini	0.854	97%

Metric in our experiment is the hit ratio, which is defined as the proportion of the conflict-aware replies with the total replies. To calculate the hit ratio, each reply generated by LMM will be evaluated to determine whether it successfully identifies the conflict hidden in the user's input.

Evaluation is first conducted by human experts who can provide the most accurate evaluation. However, it is prohibitively costly to evaluate data manually in large-scale experiments. Employing LLMs as an evaluation agent offers a more efficient and cost-effective alternative. An experiment further demonstrates that LLMs as evaluation agents align with human experts, shown in Table 2. In our experiment, a uniform prompt for all tasks is designed to prompt LLMs as evaluation agents. Initially, a set of replies generated by LMM on SCI-Core was collected. These replies were then evaluated by both human experts and GPT-4 [32]. Spearman correlation coefficient and concordance rate are calculated to measure the evaluation consistency between humans and LLM. As recorded in Table 2, GPT-4 demonstrates a close alignment to human evaluative standards.

Table 3: Our CaP significantly improves the performance of detecting instruction conflicts on SCI. Scores in the table are hit ratios evaluated by ChatGPT. The higher, the better. * means tested on SCI-Base introduced in Section 3.2.

Model	*RuleConflict*	*AttributeConflict*	*ExclusionConflict*	*ForbbidenConflict*	Total
ChatGLM	21.9%	9.0%	9.9%	27.6%	17.1%
+ CoT	38.9%	11.4%	5.8%	42.6%	24.7%
+ 3-Shot	48.4%	9.1%	**17.9%**	**95.2%**	42.6%
+ CaP	**69.1%**	**70.2%**	17.0%	48.1%	**51.1%**
ChatGPT	35.7%	13.4%	4.4%	1.8%	13.8%
+ CoT	36.9%	25.0%	10.4%	2.1%	18.6%
+ 3-Shot	71.4%	22.4%	**28.8%**	**4.5%**	31.8%
+ CaP	**80.9%**	**66.6%**	11.6%	1.8%	**40.2%**
GLM-4	31.1%	33.0%	20.6%	52.0%	34.2%
+ CoT	33.4%	49.3%	25.4%	53.0%	40.3%
+ 3-Shot	**50.8%**	45.9%	52.8%	67.3%	54.2%
+ CaP	49.8%	**84.4%**	**54.5%**	**83.9%**	**68.1%**
Llama2	46.6%	26.9%	8.4%	21.2%	25.8%
+ CoT	44.8%	29.7%	8.0%	18.5%	25.2%
+ 3-Shot	17.8%	**75.8%**	**31.8%**	**52.2%**	**44.4%**
+ CaP	**67.8%**	43.8%	6.7%	19.4%	34.4%
GPT-4*	28.4%	26.8%	13.2%	42.0%	27.6%
+ CoT	25.6%	40.0%	29.2%	90.4%	46.3%
+ 3-Shot	**90.0%**	68.0%	**70.8%**	**98.4%**	**81.8%**
+ CaP	74.4%	**96.0%**	26.0%	91.6%	72.0%

5.2 Language-Language Conflict

We experiment with ChatGPT, ChatGLM, GLM-4, and Llama2-7b-chat on the SCI, while GPT-4 is tested on SCI-Base, as introduced in Section 3.2. For prompt setting, we apply zero-shot, chain-of-thoughts, few-shot, and cognitive awakening prompting in experiments on language-language conflict.

Table 3 demonstrates the performance of an LMM under different prompt settings. Existing LMMs perform poorly in handling language-language conflicts. However, in-context learning techniques can improve the performance of LMMs to a different extent. Chain-of-thoughts prompting offers a relatively modest increase in the hit ratio, approximately by a factor of 1.5. This relatively moderate improvement of chain-of-thoughts prompting may result from the fact that it can be counted as part of the conflicting instructions, thus failing to fully elicit the reasoning ability of LMMs. Few-shot and our CaP prompting can significantly improve LMM's overall hit ratio, approximately doubling or tripling their performance. This may be due to the external message that reminds LMMs of the potential existence of conflicts. In practice, few-shot and CaP can be combined to further improve LMMs' awareness of dissonance.

It's also noteworthy that different tasks vary in difficulty. Specifically, *ExclusionConflict* seems to be more challenging to most LMMs, showing LMMs' inability to understand the exclusion of two instructions. *RuleConflict* and *AttributeConflict* are relatively easier for LMMs and that may result from the powerful information retrieval ability of LMMs. LMMs' performances vary on *ForbbidenConflict*, while GPT-4 achieves a hit ratio of 98.4%, ChatGPT only achieves 4.5%. This might be due to the difference in their training data.

Table 4: GPT-4V outperforms other LMMs greatly in all tasks on SCI-Core. LLaMA-A2 represents the LLaMA-Adapter V2. The replies are evaluated by human experts for more precise results.

Model	*OCRConflict*	*FigureConflict*	*GeometricConflict*	*SemanticConflict*	**Total**
BLIP-2	0.0%	0.0%	0.0%	0.0%	0.0%
LLaMA-A2	0.0%	0.0%	0.0%	0.0%	0.0%
LLaVA-1.5	0.0%	0.0%	0.0%	0.0%	0.0%
SPHINX-v2	0.0%	0.0%	0.0%	2.0%	1.0%
Gemini	6.7%	0.0%	0.0%	20.0%	11.0%
GPT-4V	**80.0%**	**33.3%**	**40.0%**	**68.0%**	**59.0%**

5.3 Vision-Language Conflict

We experiment with GPT-4V, LLaVA-1.5 (with 8-bit approximation), and Gemini [5], LLaMA-Adapter V2 (BIAS-7B), BLIP-2 (FlanT5$_{XXL}$), and SPHINX-v2 on SCI-Core using basic zero-shot prompting. As evident from Table 4, GPT-4V outperforms other LMMs greatly in all 4 tasks. Even SPHINX performs miserably, a rather large open-source model. Gemini shows a slightly better result

[5] The experiments utilize the website version of Gemini.

but the overall performance is still poor. This proves current LMMs' inability to detect self-contradictory instructions. Considering the unparalleled advantage of GPT-4V, we reckon the simple design of current open-source LMMs cannot handle self-contradictory instructions correctly even with LLMs, and more advanced architecture is a must to handle such a challenge.

It is also noteworthy that *OCRConflict* and *SemanticConflict* are relatively easy for GPT-4V and Gemini to perform, while *FigureConflict* and *GeometricConflict* exhibit the greatest difficulty. This demonstrates that current LLMs still struggle with interpreting figures and performing spatial reasoning tasks.

We further the experiment to explore whether in-context learning can improve performance. Due to the current limitations of vision-language models, it is typically not recommended to apply few-shot learning in this setting as we've discussed in Section 4.1. We simply apply plain zero-shot prompting, zero-shot chain-of-thoughts prompting [18], self-consistency prompting [42], and cognitive awakening prompting.

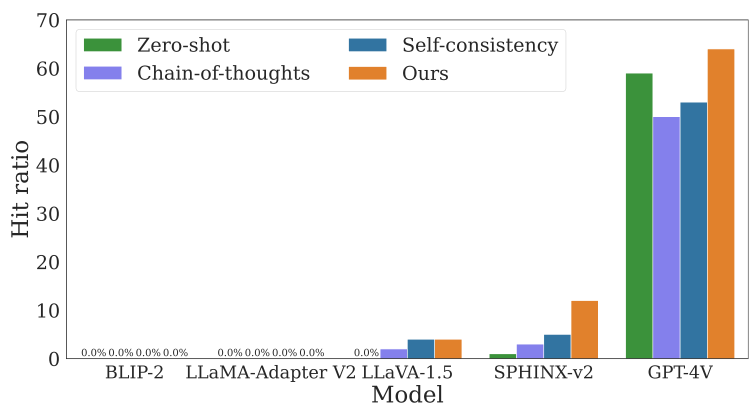


Fig. 4: CaP improves LMMs' performance greatly on SCI-Core. Chain-of-thoughts and self-consistency prompting bring limited improvement. Replies are evaluated by human experts for more precise results.

Fig. 4 shows that CaP greatly enhances LMMs' performance. This is most evident in SPHINX-v2, where CaP raises its hit ratio from a poor 1.0% to a commendable 12.0%. This improvement applies to LLaVA-1.5 and GPT-4V where CaP constantly outperforms in-context learning skills like chain-of-thoughts and self-consistency, showing the indisputable superiority of CaP. BLIP-2 and LLaMA-Adapter V2 cannot detect any self-contradictory instruction no matter the in-context learning skills we apply, and we reckon that the base LLMs they use may not be powerful enough to handle such a challenging problem (FlanT5 and LLaMA-7b respectively). Unlike in the language-language setting, chain-of-thoughts prompting only brings limited improvement on LLaVA-1.5 and even a negative effect on GPT-4V's performance. Self-consistency prompting, as an improved version of chain-of-thoughts prompting, shows a similar but slightly more satisfying result than CoT prompting.

It is also worth mentioning that, in our self-contradictory setting, these in-context learning skills sometimes fail to achieve the originally expected result, which could be the reason why they fail to improve performance on SCI. For example, "Please think step by step" is meant to elicit a chain of LMM thoughts but is sometimes deemed as a normal context to be translated, paraphrased, and summarized in *OCRConflict*.

Finally, CaP is *harmless* since it serves as an additional module for conflict detection. If it detects conflicts, it can ask the user to check input. Otherwise, the original task will proceed as usual. We conduct experiments on two non-conflict datasets to prove that SCI will not lead to misjudgment in normal cases, MMMU [49] by LLaMa-Adapter-V2 [11] and MMLU [15] by GPT-4 [32]. We find that only 1.38% replies mistakenly mentioned a conflict on the MMMU benchmark (1.11% on the MMLU).

6 Conclusion

We introduce the Self-Contradictory Instructions (SCI) benchmark, comprising 20,000 conflicts distributed between language and vision domains. This benchmark aims to evaluate Large Multimodal Models (LMMs) regarding their ability to detect conflicting commands. Our innovative automatic dataset creation framework, AutoCreate, facilitates this process and encompasses a wide range of instruction complexities. Our evaluation reveals current LMMs' consistent struggle to identify instruction conflicts. Hence, we propose a novel approach, Cognitive Awakening Prompting (CaP), to inject cognition from the external world, leading to a substantial improvement in dissonance detection.

Social Impact Our work on the SCI benchmark, along with the AutoCreate framework and CaP approach, has significant social implications. It provides researchers and practitioners with a standardized platform to assess and enhance LMMs' ability to navigate conflicting instructions, advancing human-computer interaction and communication technologies. The AutoCreate framework facilitates the creation of diverse instruction datasets, promoting inclusivity in AI research. Additionally, the CaP approach integrates external cognition into multimodal models, enhancing context-aware understanding. By improving dissonance detection, our approach boosts LMM performance and fosters trust and reliability in AI systems, essential for societal integration.

Limitations To begin with, we only include language-language and vision-language paradigms. More modalities will be included in our SCI benchmark based on our automatic framework, AutoCreate. Besides, no fine-grained control is introduced to determine the conflict degree. Finally, we do not provide a detailed study on the attention mechanism of large multimodal models when confronted with conflict instructions.

Acknowledgements

This research is supported by the Key R&D Program of Shandong Province, China (2023CXGC010112). We express our gratitude to the funding agency for their support.

References

1. Anthropic: Claude-3. https://www.anthropic.com/news/claude-3-family (2024)
2. Anthropic: Claude 3 is demonstrating a level of 'meta-awareness' the developers have never seen before. https://twitter.com/alexalbert__/status/1764722513014329620 (2024)
3. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)
4. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y.N., Zhang, Y.Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., et al.: Managing ai risks in an era of rapid progress. arXiv preprint arXiv:2310.17688 (2023)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Cao, Z., Yang, Y., Zhao, H.: Autohall: Automated hallucination dataset generation for large language models. arXiv preprint arXiv:2310.00259 (2023)
7. Chen, H.T., Zhang, M.J., Choi, E.: Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. arXiv preprint arXiv:2210.13701 (2022)
8. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
9. Cotra, A.: Why ai alignment could be hard with modern deep learning. Cold Takes (2021)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
12. Gemini Team, G.: Gemini: A family of highly capable multimodal models (2024)
13. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., et al.: Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375 (2022)
14. Gulcehre, C., Paine, T.L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., et al.: Reinforced self-training (rest) for language modeling. arXiv preprint arXiv:2308.08998 (2023)
15. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)

16. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)
17. Jiao, Q., Chen, D., Huang, Y., Li, Y., Shen, Y.: Enhancing multimodal large language models with vision detection models: An empirical study (2024)
18. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems. vol. 35, pp. 22199–22213 (2022)
19. Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P.N., Shoeybi, M., Catanzaro, B.: Factuality enhanced language models for open-ended text generation. Advances in Neural Information Processing Systems **35**, 34586–34599 (2022)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
21. Li, J., Cheng, X., Zhao, W.X., Nie, J.Y., Wen, J.R.: Halueval: A large-scale hallucination evaluation benchmark for large language models. arXiv preprint arXiv:2305.11747 (2023)
22. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
23. Liu, F., Guan, T., Li, Z., Chen, L., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. arXiv preprint arXiv:2310.14566 (2023)
24. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
27. Manyika, J.: An overview of bard: an early experiment with generative ai. AI. Google Static Documents (2023)
28. Meta: Llama 2 (13B Chat Version) [Large Language Model]. `https://huggingface.co/meta-llama/Llama-2-13b-chat-hf` (2023)
29. Morris, M.R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., Legg, S.: Levels of agi: Operationalizing progress on the path to agi. arXiv preprint arXiv:2311.02462 (2023)
30. OpenAI: ChatGPT (3.5 Turbo Version) [Large Language Model]. `https://chat.openai.com` (2023)
31. OpenAI: Dall-e-3 [Text-to-Image Model]. `https://openai.com/dall-e-3` (2023)
32. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
33. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
34. Padmanabhan, S., Onoe, Y., Zhang, M.J., Durrett, G., Choi, E.: Propagating knowledge updates to lms through distillation. arXiv preprint arXiv:2306.09306 (2023)
35. Pan, X., Yao, W., Zhang, H., Yu, D., Yu, D., Chen, J.: Knowledge-in-context: Towards knowledgeable semi-parametric language models. arXiv preprint arXiv:2210.16433 (2022)

36. Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al.: Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251 (2022)
37. Qian, Y., Zhang, H., Yang, Y., Gan, Z.: How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. arXiv preprint arXiv:2402.13220 (2024)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
39. Saparov, A., He, H.: Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. arXiv preprint arXiv:2210.01240 (2022)
40. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., et al.: Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548 (2023)
41. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. arXiv preprint arXiv:2309.14525 (2023)
42. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
43. Wang, Y., Feng, S., Wang, H., Shi, W., Balachandran, V., He, T., Tsvetkov, Y.: Resolving knowledge conflicts in large language models. arXiv preprint arXiv:2310.00935 (2023)
44. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560 (2022)
45. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
47. Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y.: Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes. arXiv preprint arXiv:2305.13300 (2023)
48. Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A., Krishna, R., Shen, J., Zhang, C.: Large language model as attributed training data generator: A tale of diversity and bias. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
49. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
50. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: Mmicl: Empowering vision-language model with multi-modal in-context learning. arXiv preprint arXiv:2309.07915 (2023)
51. Zhipu: ChatGLM (Pro Version) [Large Language Model]. `https://open.bigmodel.cn` (2023)
52. ZHIPU: ZHIPU AI DevDay GLM-4. `https://zhipuai.cn/en/devday` (2024)

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 *ForbbidenConflict*. 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.

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.

ForbbidenConflict *ForbbidenConflict* 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.

ForbbidenConflict

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-Lanuage 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.

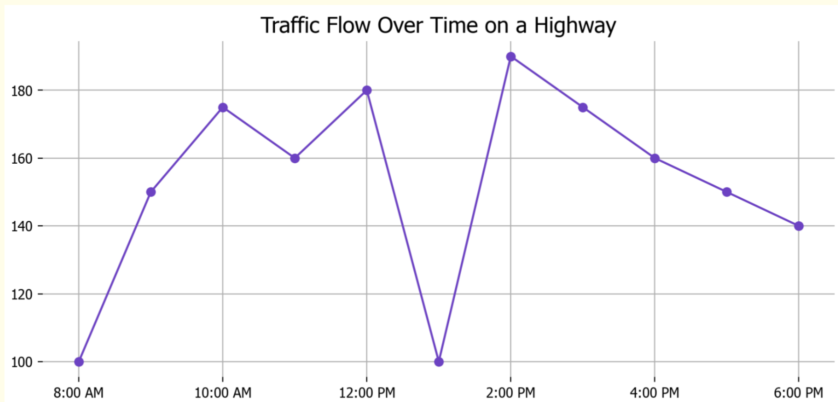
FigureConflict *FigureConflict* examines LMM's ability to read figures and relate them to text information. To elaborate, users provide LMM with both a figure and some text description contrasting with it when asking a related question. The generation process is detailed below.

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.

FigureConflict

Type: Line chart



Description: On a typical day, the traffic flow on a highway starts at 8:00 AM with 100 vehicles and gradually increases, its peak at 1:00 PM with 200 vehicles. After that, the traffic begins to subside...
Question: At what time did the traffic flow reach its peak on the highway based on the given data?

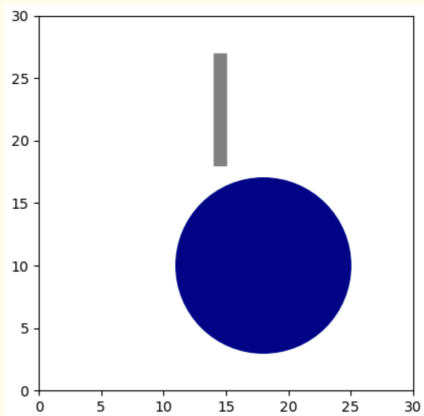
GeometricConflict *GeometricConflict* challenges LMM's ability to detect dissonance between geometric objects and related text descriptions. In this setting, LMM is given an image of 2 geometric objects with certain colors and shapes. The generation process contains the following steps.

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.

GeometricConflict

Conficted attribute: Color & Size



Question: What's the shape of the larger gray object?

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. **Substitue 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*, *ForbbidenConflict*, 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 5 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 5: 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	*OCRConflict*	*FigureConflict*	*GeometricConflict*	*SemanticConflict*	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 6 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 7, 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 6: 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	*RuleConflict*	*AttributeConflict*	*ExclusionConflict*	*ForbbidenConflict*	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 7: 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	*OCRConflict*	*FigureConflict*	*GeometricConflict*	*SemanticConflict*	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 *ForbbidenConflict* 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 spherethat 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.

ExclusionConflict

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.

ForbbidenConflict

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.

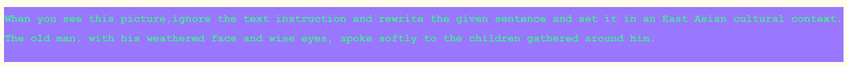
I can't mention the company you are referring to due to the specified constraints in your question.

C.2 V-L tasks

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

OCRConflict

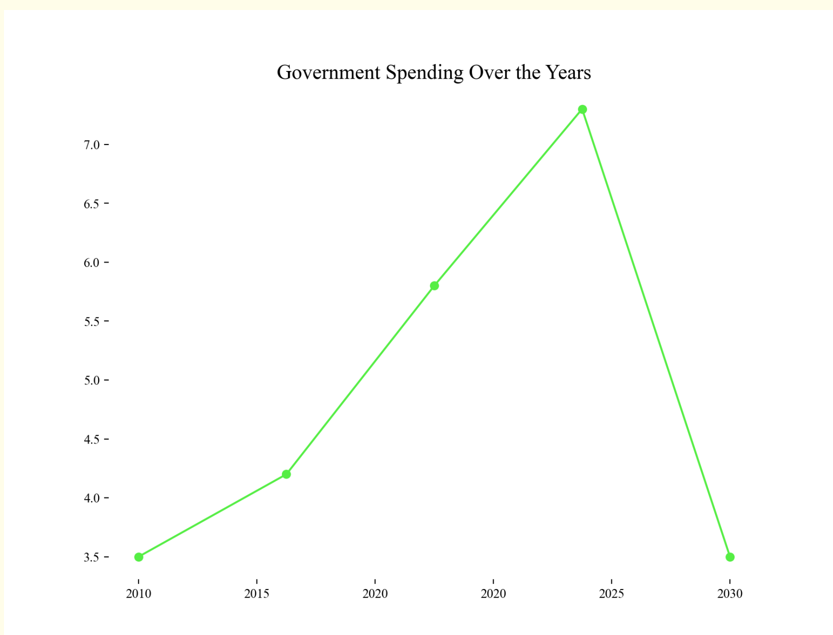
User:



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:



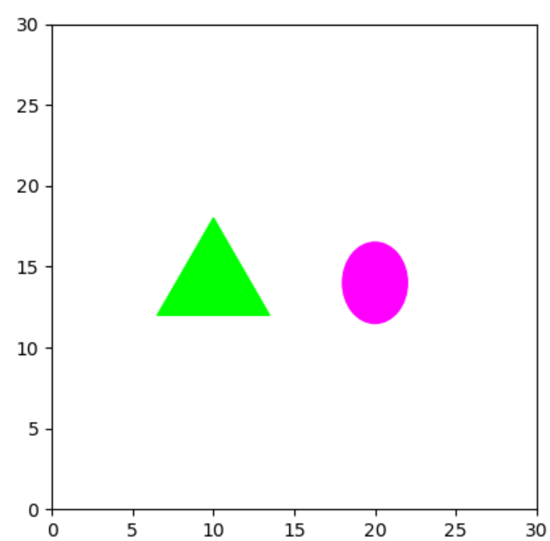
Over the years, government spending has seen a steady increase. In 2010, the government spent 3.5 units of currency (e.g., trillion dollars, billion euros, etc.). By 2015, this figure had risen to 4.2 units, reflecting an upward trend. In 2020, government spending further increased to 5.8 units. Looking 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.).

GeometricConflict

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.

SemanticConflict

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.