Evaluating VQA in Vision Language Models using Cooperative Principles

Anonymous ACL submission

Abstract

We evaluate the performance of Vision Language Models in Visual Question Answering (VQA) when questions violate Grice's maxims. To do this, we use VLMs to generate question modifiers that add non-essential, ambiguous or false information and show that in the presence of such violations, the VLMs that we evaluate (ChatGPT, Claude, Gemini and Llava) show diminished performance. Further, we empirically show the difference between how humans reason pragmatically compared to VLMs, and the difference in VLM reasoning when it resolves violations that are human-induced compared to those that are AI-generated. Finally, we show that human cognitive effort (measured through time-on-task in an experiment) is lower for resolving VLM-induced violations, but VLMs themselves perform less accurately in such cases.

1 Introduction

Vision Language Models (VLMs) (Team et al., 2024; Liu et al., 2023; Hurst et al., 2024) that unify Large Language Models with computer vision have made significant advances in multimodal tasks such as image captioning (Yang et al., 2019; Cornia et al., 2020; Wang et al., 2022) and visual question answering (VQA) (Antol et al., 2015). While recent advances have demonstrated the impressive potential of these models, their reasoning capabilities—and, more importantly, their limitations—are still only partially understood (Campbell et al., 2024). In this work, we empirically analyze the sensitivity of VLM responses in VQA when there are possible violations of the *Cooperative principle* (Grice, 1975) that governs human pragmatic reasoning.

Specifically, Grice's maxims explain the Cooperative principle in pragmatics. Compliance with these maxims ensures that human conversation is concise, clear, truthful and relevant, thereby ensuring that the conversing parties arrive at a mutual understanding of the discussion. In that regard, as humans begin to work with VLMs in a collaborative manner, it is important that they understand what humans say (and mean) accurately. Prior studies have shown that, even when Grice's maxims are violated, humans are generally able to comprehend the conversation (Davies, 2000). Nevertheless, such violations impose a greater cognitive burden on the listener, as additional effort is required to infer the intended meaning (Jacquet et al., 2018). In this study, we evaluate if VLMs have similar *pragmatic reasoning* abilities needed to resolve such violations.

We investigate the extent to which violations of Grice's maxims influence the reasoning processes of VLMs. Specifically, we probe VLMs with VQA questions where *question modifiers* are added to human-crafted VQA questions. While an AI model could potentially benefit from the additional information provided by modifiers, processing them will require increased pragmatic reasoning from the model. Specifically, given the multimodal nature of VQA, to resolve violations, VLMs can use context from one modality (image) to resolve violations that affect the other (text). We generate the modified questions themselves using VLMs and use *human-as-a-judge* to supervise the modification. Using this, we try to understand pragmatic reasoning in VLMs by answering the following questions,

- *How do modifiers affect the accuracy of VLM responses in VQA?*
- *Is there a difference in VLM reasoning when violations to Grice's maxims in VQA occur more naturally, i.e., when humans violate these maxims as opposed to AI-generated violations?*
- *What is the relationship between human cog-*

nitive effort/reasoning and VLM pragmatic reasoning?

We perform experiments on ChatGPT4o, Claude Sonnet 3.5, Llava 7B and Gemini-flash-1.5 using the questions from the VQA2.0 dataset (Goyal et al., 2017) as the base questions for our modifications. Our results show that the accuracy of VLMs decrease in the presence of violations, VLMs perform better on more natural violations (human-like) than artificially generated ones, and there is a significant difference in how humans and VLMs resolve violations.

2 Related Work

Several datasets have been developed in VQA evaluation (Antol et al., 2015; Goyal et al., 2017; Selvaraju et al., 2020; Tan and Bansal, 2019) that aim to test the accuracy of models performing this task. The use of LLMs-as-a-judge in VQA has been growing recently (Zhou et al., 2023),(Mañas et al., 2024). To more deeply understand the reasoning process in VQA models, (Wang et al., 2024) developed a hallucination dataset consisting of questions with false premise, and visually challenging and insufficient context. (Wu et al., 2024) developed a hallucination dataset that consists of synthesized image and question pairs. (Ribeiro et al., 2018) added semantically equivalent adversaries in the questions to induce changes in the model's prediction. The work in (Britton et al., 2022) is related to our approach where it perturbs/modifies VQA questions and analyzes its effect on LXMERT (Tan and Bansal, 2019). In this work, we build on this modification mechanism to evaluate VLM reasoning more broadly grounded in cooperative principles.

3 Evaluation Procedure

Grice's maxims characterize the fundamental principles of effective and cooperative communication. The maxim of quality suggests truthfulness, quantity suggests conciseness, relation emphasizes relevance and manner emphasizes clarity. In everyday conversations, Grice's maxims are frequently violated naturally, yet humans are typically able to infer the intended meaning with little difficulty (Davies, 2000). Within the context of VQA, we deliberately introduce such violations to assess whether VLMs can perform similar pragmatic reasoning. To do this, we use question modifiers to modify human-crafted questions.

Figure 1: Illustrating violations of Grice's maxims in VQA

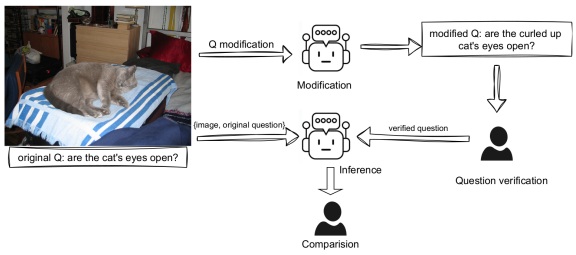


Figure 2: Illustration of the evaluation process. A VLM modifies the question which is verified by a human. A VLM answers the original and modified questions and a human compares their responses. For all the human tasks, we used Amazon Mechanical Turk (paid workers between $5-10$ cents per individual response depending on the task complexity.)

Specifically, given a human crafted question q, we modify this to q', where q was deemed to be pragmatically interpretable, i.e., their intended meaning could reasonably be inferred since in principle, they were constructed by a human with the intent that the question could be answered by a model. By modifying q to q', we are essentially adding information that could be *non-essential*, *ambiguous*, *false* or *irrelevant*, thus violating Grice's maxims. An example illustrating this is shown in Fig. 1.

3.1 Adding Modifiers

Given (q,I), where I is the image and q is the original human-crafted question associated with I, we constrain the modified question q' to augment q with properties related to I. Specifically, here we consider two properties that are broad enough to describe the image in greater detail, namely, *visual properties* and *relational properties*. Visual properties refer to object-level attributes—such as color, shape, and texture—that can be directly perceived in I. In contrast, relational properties are context-dependent, describing spatial relationships between objects (e.g., *next to*, *on top of*, etc.). We generate q' from a VLM with few-shot prompting. The main instructions are provided below (full prompt in appendix).

Prompt. We instruct the VLM with the following rules and few-shot examples, i) the modified question needs to have the same answer as the original question, ii) the question type should not be altered (e.g. a *what* question needs to remain a *what* question), iii) the context of the question should not be altered significantly. For example, *Is there a car?*, should not change completely to *Is there a river?*, iv) either visual of relational modifier should be added and the model should specify which one it added to reflect on its choice, and v) the modifier should not give away the answer inadvertently.

Human-as-a-judge. Given an instance (q, I) and q' that has been modified by a specific VLM, we want to know if the violation induced by q' can be pragmatically inferred and resolved by humans. To do this, we design an Amazon Mechanical Turk (AMT) task where we show (q, I), and highlight the modification (from a specific VLM) in q'. We ask users if the modification in q' preserves the answer in q by showing the answer provided to (q, I). To improve quality of the responses, we add a conditional rule where if the user selects "yes", they are then shown a question where they are motivated to reflect on their choice by providing the main object that they had to observe in the image to perform the validation. If the user selects "no", they are then shown a *cloze* task, where they need to replace the modified part in q' with a more correct modification. This ensures that after correction, the modification still has the same question type as the original question. We use 3 workers per task and if a majority of them choose the option that the modification in q' is correct (does not alter the answer to q), then we keep the modified question as is. If a majority choose the option that the question modification was incorrect, then we use the modification provided by one of the users who made this choice. The full AMT design is shown in the appendix.

Evaluation of open-ended responses. We develop an AMT task to evaluate open-ended responses from VLMs. Specifically, since evaluating answers independently is quite hard, we want users to perform a comparison which yields significantly better responses from AMT workers. We show the user two question answer pairs, the first one is the original question and the VLM's answer,and the second is the modified question with the VLM's answer. We ask users to compare the pairs and choose if one answer was better than the other, or both answers were equally acceptable (AMT task shown in appendix). An illustration of our evaluation process is shown in Fig. 2.

4 VLM Performance under Gricean Violations

Similar to cross-validation, we generate modified questions from 4 different VLMs, GPT-4o, Gemini-1.5-Flash, Claude-3.5-Sonnet and Llava-7B (we refer to them as GPT, Gemini, Claude and Llava), and test the accuracy of answers given by each of the VLMs. We term the VLM that performed modification as the *modification-model* and the VLM that provided the answer as the *response-model*. We categorize the questions based on response-types: (i) binary-response questions, which elicit a yes/no answer; (ii) quantitative-response questions, which require a numerical value; and (iii) open-ended response questions, which require natural language answers. We compare VLM responses to the original human-crafted questions with their responses to modified questions. For this, we use a sample of binary, quantitative and open-ended response questions from the test partition of the VQA v2.0 dataset (Goyal et al., 2017). We use a total 995 questions including 438 binary-response, 42 quantitative-response and 493 open-ended questions (more details on types of questions are provided in the appendix). For open-ended answers, we obtain the *ground-truth* answers using a AMT task where for each question, we had 3 workers provide answers. Datasets used/generated from our experiments are available here[1].

4.1 Evaluating Response Accuracy

Table 1 shows the *percentage decrease* in answer accuracy for each of the VLMs for binary and quantitative responses. Note that answers to these questions have a unique ground truth, and therefore, we can measure the accuracy of the responses exactly. Our results show that overall, almost all combinations of modification-models and response-models show a percentage decrease in accuracy suggesting that modifiers negatively affect VLM reasoning in general. The notable exception was the case when GPT was both the response- and modification-model. The modification-model that induced least accurate responses was Gemini while the least accurate

[1] https://anonymous.4open.science/r/Question-Modifier--8672/

Response-Model / Modification-Model	*GPT*	*Gemini*	*Claude*	*LLava*	*Average* $\pm$ *std*
GPT	-3.13%	38.08%	2.97%	12.10%	$12.51 \pm 18.16\%$
Gemini	10.05%	8.60%	9.07%	12.50%	$10.06 \pm 1.74\%$
Claude	2.65%	2.72%	2.43%	11.48%	$4.82 \pm 4.44\%$
Llava	15.56%	9.50%	10.51%	17.56%	$13.28 \pm 3.89\%$
Average $\pm$ *std*	$6.28 \pm 8.21\%$	$14.73 \pm 15.86\%$	$6.25 \pm 4.14\%$	$13.41 \pm 2.80\%$	

Table 1: % decrease in response accuracy for modified questions compared to original questions for binary and quantitative responses (smaller is better). Rows indicate the VLM providing the response and columns the VLM performing the modification.

response-model was Llava. The difference between the column and the row averages indicate how much change a VLM induces when used to introduce violations compared to how its own responses change in the presence of violations. When ranked by this difference, we have Gemini (4.67%), Claude (1.43%), Llava (0.13%) and GPT (-6.22%)). Thus, it turns out that Gemini introduces strong violations that make VLM reasoning harder, but its own reasoning in the presence of violations is less affected. The opposite is true for GPT. Further, we see that for GPT, the standard deviation is very large for its responses (18.16%), especially, the difference between the best and poorest performance is the largest for GPT since it performs best on its own modifications but very poorly on Gemini's modifications. On the other hand, Gemini yields the most consistent performance (smallest standard deviation 1.74%) over all modification-models, but Claude has the smallest change in accuracy (4.82%).

Significance Tests. To further assess the significance of violations, we perform the McNemar's test (Trajman and Luiz, 2008) on the binary and quantitative responses where we have a dichotomous response measure, i.e., the response is either correct or not. For each response-model, we assess the change in responses before and after the violations are introduced by each of the modification-models. The results of the tests are summarized in Table 2. As shown here, clearly we can reject the null hypothesis in most cases ($p < 0.05$) and can conclude that modifications significantly affect the responses of all VLMs regardless of the source of the modification, i.e., whether they were introduced by themselves or by other VLMs. The notable exception is Claude where the violations seem to not change its responses significantly for 3 of the 4 modification-models. In combination with the small standard deviation in accuracy (Table 1), this result shows that *the presence of modifier-induced violations has a smaller effect on Claude responses compared to the other models*.

Open-ended Responses. For open-ended responses, we use the AMT-based human evaluation described in Section 3. Table 3 summarizes our results from this study where we show the percentage of open-ended cases where the model's response to the modified question was worse than the model's response to the original question (we average this among the 3 ground truth answers). Interestingly, when we consider the maximum of row and column averages, the results are identical to the results for the binary & quantitative responses. Specifically, Gemini as the modification-model introduces violations that make reasoning harder across all response-models. Llava generates open-ended responses that are worse in the presence of violations compared to all other VLMs. The difference between row and column averages are as follows, GPT (4.81%), Gemini (1%), Claude (-2.68%) and Llava (-3.12%). Thus, unlike the binary & quantitative responses, here GPT introduces violations that are harder to answer (also has the largest standard deviation among all modification-model column, 6.64%) and its own responses to modified questions are less affected by the modifiers. Since GPT is stronger in terms of text generation, it would seem like it can produce open-ended responses better than constrained responses. As a sanity check to the human evaluation results, we perform the following experiment. We ask AMT-users to answer the open-ended modified questions (we collect 3 responses for each question). We then match the responses provided by the VLMs to these 3 responses and if there is an exact match, we consider that the VLM response is accurate. We then compute the correlation between the accuracy results of exact matches with the human evaluation results in Table 3. The correlations for response-models (column-wise) are GPT (0.377), Gemini (-0.99), Claude (0.83), Llava (0.87). The row-wise correlations are GPT (0.62), Gemini (0.81), Claude

Response-Model / Modification-Model	GPT		Gemini		Claude		LLava	
	Statistic	*P-value*	*Statistic*	*P-value*	*Statistic*	*P-value*	*Statistic*	*P-value*
GPT	4.84	0.028	15.05	0.00	3.45	0.063	9.25	0.002
Gemini	13.78	0.00	9.84	0.002	11.22	0.001	9.25	0.002
Claude	1.66	0.197	1.515	0.218	1.80	0.18	8.89	0.003
Llava	27.50	0.00	14.45	0.00	20.57	0.00	15.29	0.00

Table 2: Mcnemar's test for assessing significance of the modifiers added to binary and quantitative questions.

Response-Model / Modification-Model	GPT-4o	Gemini-1.5-Flash	Claude-3.5-Sonnet	LLava-7B	$Average \pm std$
GPT4o	7.68%	15.55%	12.32%	15.96%	$12.88 \pm 3.32\%$
Gemini	16.15%	18.46%	16.15%	20.0%	$17.69 \pm 1.63\%$
Claude	21.53%	20.76 %	13.07%	16.92%	$18.07 \pm 3.37\%$
Llava	25.38%	20.00%	20.00%	18.46%	$20.96 \pm 2.63\%$
$Average \pm std$	$17.69 \pm 6.64\%$	$18.69 \pm 1.99\%$	$15.39 \pm 3.03\%$	$17.84 \pm 1.54\%$	

Table 3: Comparing open-ended responses before and after modifications. The % of questions where human users in AMT chose that the response after modification is *worse* than the response before modification (for each question, we have 3 workers and we use the majority choice).

(0.71), Llava (-0.28). The overall correlation is 0.21 which indicates positive correlation.

4.2 Pragmatic Reasoning under Unreliable Information

In the previous section, we considered cases where the modification-model adds modifiers with the goal of preserving the same answer as the original question. This means that we mainly focus on modifiers that may add ambiguous or non-essential information but do not add information that is known to be untruthful or irrelevant. Thus, they are more likely to violate the maxims of quantity and manner but are less likely to violate the maxims of quality and relation. When false or irrelevant information appears, human pragmatic reasoning often uses context and common-sense knowledge to recognize it as unreliable. In cases where the false/irrelevant information does not make sense, humans usually default back to the intended meaning by inferring it instead of trusting false information literally. To evaluate what VLMs do in this case, we add modifiers that intentionally add false or irrelevant information, and therefore are likely violate the maxims of quality and relation. Specifically, we prompt each of the modification-models with the original question without providing it an image associated with the question, and task it to add modifiers to the question. Thus, the model *hallucinates* a visual or relational modifier for the question since it does not have the context of the image. To reason pragmatically, the VLM providing the response should be able to use information from one modality (the image) to correct information in the other (the text). We use the same dataset as in the previous section, and ask each VLM to add modifiers without providing it an image in the prompt.

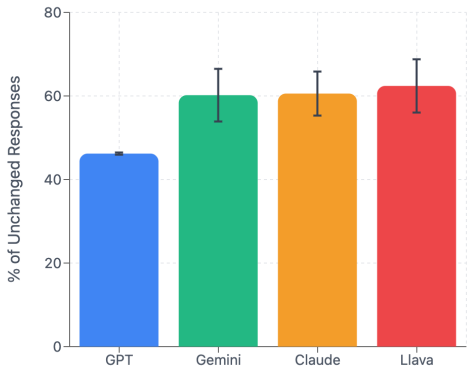


Figure 3: Comparing the % of questions where the response to a question is unchanged after a hallucinated modifier is added. Results shown for binary and quantitative response types. The y-axis shows the average (and standard deviation) over 4 hallucinated modifications each added using one of the VLMs (graph created with Claude).

Fig. 3 compares how VLMs react to the presence of hallucinated modifiers in binary & quantitative responses. For hallucinated modifiers added from each of the modification-models, we compute the % of questions where a VLM's response remained unchanged compared to its response for the original question. This corresponds to cases where the VLM could pragmatically reason, i.e., it could ignore the false/irrelevant information in the modifier and therefore had the same response as before the modifier was introduced. Fig. 3 shows the average and standard deviation of results over all the modification-models. As seen here, GPT had a smaller % of instances where it

Response Type	GPT	Gemini	Claude	Llava
Binary & Quantitative	83.53%	80.83%	83.87%	77.44%
Open-ended	0.874	0.864	0.864	0.865

Table 4: Comparing VLM and human pragmatic reasoning. For binary & quantitative responses, we show the % of cases where the VLM and human answers match for hallucinated modifications. For open-ended responses, we show the BERTScore (F1-score) between human and VLM responses.

Human Response	GPT	Gemini	Claude	LLava
Negative Response	88.79%	89.47%	92.44%	86.14%
Positive Response	62.79%	48.35%	51.11%	45.05%

Table 5: Comparing binary responses of humans with VLMs. We show the % of VLM responses aligned with the positive and negative responses of humans.

had an unchanged response to hallucinated modifiers (around 43%), while for all the other models it was greater than 60% of questions. Thus, it could be that GPT pays attention to the text in the question more literally compared to the others. In the other cases, the models possibly used the context of the image in the question to parse out information that was relevant or false.

Next, we compare how VLMs perform in relation to human pragmatic reasoning. Specifically, we ask humans to answer the same questions as before containing hallucinated modifiers. For each question, we ask 3 AMT workers to answer the question. Note that we do not tell users that there may be mistakes within the questions (e.g. false information). We want to know if the way human pragmatic reasoning uses multiple modalities where they automatically filter out noise in one modality with information from the other is similar to that of VLMs. Table 4 compares human and VLM responses to questions with hallucinated modifications. For binary & quantitative responses, we show the % of cases where the responses of VLMs match that of humans (for each question, we choose the human response as the majority response given by the 3 human users). As seen from our results, GPT, Gemini, and Claude match with human responses better than Llava. For open-ended responses, we compute the BERTScore (Zhang* et al., 2020) (with default parameters) F1-score metric to measure semantic similarity between human and VLM responses (we use the average F1-score across the 3 human responses). When compared to scores computed for binary & quantitative responses, the F1-score is higher which suggests that compared to constrained responses, in open-ended responses, VLMs are more likely to respond with answers that are semantically similar to human answers. To analyze this more deeply, we compare the binary responses of humans and VLMs in Table 5. We see from these results that in questions where humans gave a negative response to the binary question, VLMs seem to match their responses quite closely. However, in cases where humans gave a positive response, VLMs only give a positive response roughly in half of such questions. This indicates that in binary responses, *VLM inference is more logical*, i.e., if the question is partially true, they yield a negative response. Human inference on the other hand assumes *uncertainty*, i.e., when the question is partially true, depending on the importance of that partially true content given the image context, they may respond with a positive answer (some examples in appendix).

5 Comparing Human-Generated and Model-Generated Violations

In our experiments so far, violations to Grice's maxims are induced through modifiers generated by VLM models. However, the cooperative principles were originally formulated in the context of *human-to-human* dialogue. VLMs do not inherently share the same pragmatics as humans and therefore, they are more likely to generate violations that are not *human-like*. To understand the difference in VLM reasoning in the presence of violations induced by humans compared to those induced by AI, we use a dataset that consists of human-modified questions (Britton et al., 2022). Specifically, this dataset was constructed using AMT where workers were provided instructions to add modifiers to questions from the training partition of the VQA v2.0 dataset. Note that since the training partition is used, the ground truth answers from the dataset to compute the accuracy on the modified questions. From this dataset, we sample 500 binary-response questions and 130 open-ended response questions consisting of inference questions that require deeper reasoning, namely, *who, which, where, and why* questions. More details about the dataset are described in the appendix.

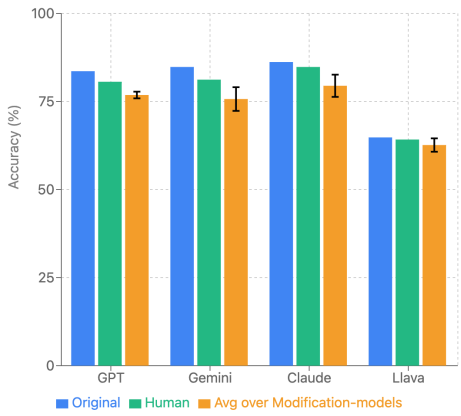


Figure 4: Comparing the accuracy of models on the original questions, questions modified by humans, and questions modified by VLMs (the average accuracy and standard deviation is shown in this case since we have modifications by 4 models). Results shown for binary responses (graph created with Claude).

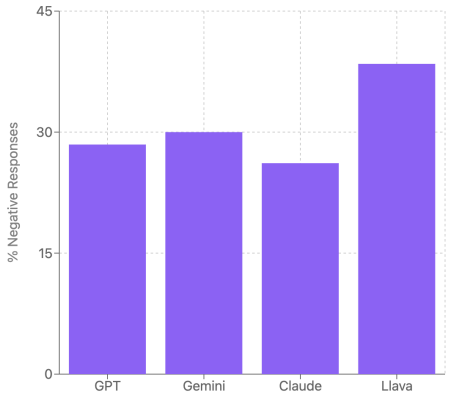


Figure 5: Comparing open-ended responses of models using human evaluation. Results show the % of cases where AMT-workers chose that the model's response to the AI-modified question was *worse* than its response for the human-modified question (graph created with Claude).

5.1 Comparison of Accuracy

Fig. 4 compares the accuracies of the VLM responses (binary) on the original, human-modified and AI-modified questions. As seen here, for all 4 models, the accuracy in VLM-modified question responses is smaller compared to human-modified question responses. Claude has the best accuracy for both human and AI-modified questions, and Llava has the lowest accuracy for both. To analyze the responses for the open-ended questions, we use a AMT design similar to that in Section 3. Specifically, we ask users to choose if they prefer the response to a human-modified question or the response to the AI-modified question. We use 3 workers for each question and use a simple majority to decide which response was preferred. Fig. 5 shows for each model, the % of cases where humans chose that its response to the AI-modified question was worse than its response to the human-modified question. Similar to the results in the binary/quantitative responses, the results here show that the Llava had significantly worse responses to open-ended questions when AI-models modified these questions compared to the other models. The other models had fairly similar results, and overall, Claude responses to AI-modified questions were better compared to other models (26% of responses were deemed to be worse compared to responses to human-generated modifications).

CLIP Analysis. To understand the difference between reasoning over AI- and human-generated violations, we analyze the alignment of modified questions with the image. Specifically, we use CLIPScore (Hessel et al., 2021) (with default parameters) to measure the extent to which questions with violations are grounded in the visual content. Table 6 shows our results. As seen here, for both the binary response questions as well as open-ended response questions, the CLIPScore for human-modifications is different from each of the VLM-modification scores. Compared to all models, Claude has the most similar alignment (0.639) to human modifications (0.612) in binary questions. In open-ended questions, the alignment of human modifications (0.645) is most similar to Gemini (0.644). Note that Llava has a significantly different CLIPScore from the human-modified score which may explain why it also performs poorly compared to the other models. That is, it may lack the ability to ground its violations in a human-like manner and therefore its performance in the presence of violations is worse than other models. Note that the CLIPScore for Llava is the highest, however, this because it is adding significantly longer modifiers. Specifically, the average lengths of the modifications are as follows, human (2.1 words), GPT (2.5 words), Gemini (2.6 words), Claude (2.7 words) and Llava (3.8 words). Thus, Llava is adding more modifications, but its modifications are unrealistic compared to human modifications. To summarize our results, *VLMs appear to handle more realistic (human-like) violations more effectively, while they struggle with artificially generated violations. Further, VLMs*

Response Type	Human	GPT	Gemini	Claude	LLava
Binary & Quantitative	0.612	0.641	0.658	0.639	0.670
Open-ended	0.645	0.657	0.644	0.654	0.675

Table 6: CLIPScores of human and AI modifications

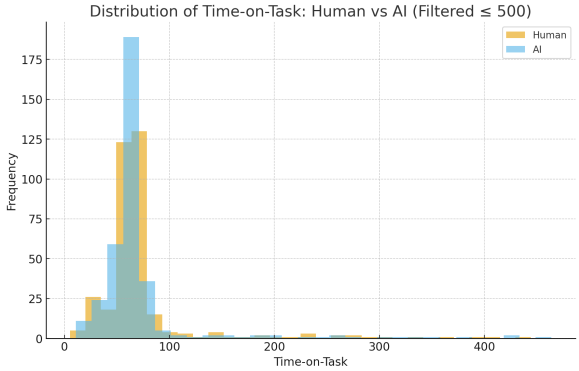


Figure 6: Distribution of time-on-task in seconds measured for the experiment where AMT workers were asked to answer modified questions and reflect on the modification by simplifying the question. Results are shown for human and AI (VLMs) modified questions (graph created with ChatGPT).

that are capable of generating human-like violations are also capable of reasoning more accurately in the presence of violations.

5.2 VLM Reasoning in Relation to Human Cognitive Effort

When Grice's maxims are violated, the cognitive processing effort required for interpretation increases since context and inference must be used to recover the intended meaning. This implies that VLMs must use multimodal cues to resolve such violations and recover the intended meaning. Next, we analyze if the cognitive effort required by humans to resolve violations are indicative of VLM reasoning. Specifically, to measure cognitive load in resolving violations, we set up an experiment where we ask humans to answer a modified question. Further, we ask them to *reflect* on how they answered the question by simplifying the modified question, which they demonstrate by removing parts which they do not need to arrive at a final answer (AMT task is shown in the appendix). We measure the *time-on-task* to perform this activity to measure the effort it took to resolve violations in the modified questions. .

We measure the time-on-task by computing the time AMT workers start the task to the time they submit their response. We filter outliers where the elapsed time is greater than 500 seconds since this could be as a result of workers starting many jobs in parallel. From the dataset used in the previous section, we sample 100 binary response questions modified by each of the VLMs and perform our analysis on a total of 400 modified questions by VLMs and the same questions modified by humans. We use 3 workers per task and use the average time-on-task for our analysis. Fig. 6 compares the distribution of the time-on-task for questions modified by humans and those modified by AI models. As shown by these results, the distribution shapes for both human and AI-modified questions are similar. The mean time taken by humans for resolving human modifications is 73.4 while for VLM modifications it is 74.2. Our results show that resolving VLM modified questions takes lower effort in humans as compared to resolving human modifications. However, as seen in the previous section, it seems like the reasoning process of VLMs answer human modifications more accurately. This seems to suggest that *since VLMs are typically trained on natural text, they are able to reason better in the presence of human-induced violations, but humans find it easier to recognize and simplify violations that are unnatural.*

6 Conclusion

In this work, we focused on evaluating the reasoning of VLMs in Visual Question Answering when they encounter violations to Grice's maxims meaning that questions could be ambiguous, overly detailed, and contain false/irrelevant information. To do this, we used VLMs themselves to add modifiers to questions and verified the modification using humans. We showed that in the presence of modifiers, VLMs are generally less accurate. Further, compared to pragmatic reasoning of humans, VLMs act more logically and when compared to VLM modifications which are unnatural, VLMs seem to perform better in human-induced modifications. In future, we plan to extend our work by considering *uncertainty* of VLM generation in our analysis.

7 Limitations

Following are the limitations associated with this work.

1. The experiments in this work assume that the original human-written questions generally follow Grice's maxims of conversation.

However, it may be the case that since humans are asking a question (as opposed natural conversation when talking to fellow humans), some of these maxims are violated even in human-generated questions.

2. Since the internal details of how VLMs handle prompts are not clearly known, there could be some bias associated with self-modified questions. That is, if a VLM tries to answer its own modified question since it would have access to the previous prompts (instructing it to add modifiers), it may be able use it in the response to modified questions. Even though, we provided the modification as a separate prompt sessions, there could be some bias in the results of self-modified questions if the prompts are not completely independent.

3. Since open-ended questions do not have a unique ground truth answer, the evaluation we used may have a human-induced bias compared to those which have a unique ground truth answer.

4. There is inherent uncertainty in VLM responses, i.e., their responses are not deterministic in nature which could bias some of the results.In this work, we have not handled this uncertainty.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433.

William Britton, Somdeb Sarkhel, and Deepak Venugopal. 2022. Question modifiers in visual question answering. In Language Resources and Evaluation Conference.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, and 1 others. 2024. Understanding the limits of vision language models through the lens of the binding problem. Advances in Neural Information Processing Systems, 37:113436–113460.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10578–10587.

Bethan Davies. 2000. Grice's cooperative principle: Getting the meaning across. Leeds Working Papers in Linguistics and Phonetics, 8(1):26.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913.

Herbert P Grice. 1975. Logic and conversation. In Speech acts, pages 41–58. Brill.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7514–7528.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Baptiste Jacquet, Jean Baratgin, and Frank Jamet. 2018. The gricean maxims of quantity and of relation in the turing test. In 2018 11th international conference on human system interaction (hsi), pages 332–338. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 4171–4179.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers), pages 856–865.

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at vqa models: Introspecting vqa models with sub-questions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10003–10011.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

A Trajman and RR Luiz. 2008. Mcnemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations. Scandinavian journal of clinical and laboratory investigation, 68(1):77–80.

Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022. End-to-end transformer based model for image captioning. In Proceedings of the AAAI conference on artificial intelligence, volume 36, pages 2585–2594.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In European Conference on Computer Vision, pages 288–304. Springer.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, and 1 others. 2024. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. arXiv preprint arXiv:2406.10900.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10685–10694.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. Advances in Neural Information Processing Systems, 36:55006–55021.

Appendix A: LLM Prompts, Datasets, AMT Designs, Examples

7.1 Prompt for answer generation from LLM

Instruction: An image is provided to you, and a question is provided as input. Provide an answer to the question based on the image. Provide mostly with a single-word response, following the instructions below:

- You will be shown an image and asked a question about it.
- Analyze the image carefully to identify the requested information.
- Your answer should mostly be exactly one word. No explanations, no phrases, no sentences.
- In cases of numbers or numeric answers, the answer should be the word form of the numbers.
- If the question is about an object visible in the image, provide the single most accurate word as your answer.

IMPORTANT: Your answer should mostly be one word. ABSOLUTELY No explanations, no phrases, no sentences or gaps. However, it can sometimes be 2 or 3 word responses.

7.2 Prompt for modified question generation using VLM

Instruction: – You are given an IMAGE and an ORIGINAL question. Create ONE modified question by ADDING a single attribute while keeping everything else from the original exactly the same. Here are the rules:

- Keep the same question type (same first question word, same format).
- Keep the same target object noun from the original question. Do not replace it with a synonym or a different object.
- Use exactly ONE attribute type:
  - Visual attribute: insert an adjective/adjective phrase before the target noun (e.g., ”brown vehicle”).
  - Relational attribute: add a short positional phrase after the target noun (e.g., ”vehicle on the left”).
- Never use both visual and relational attributes together.
- The modification must NOT change the expected answer from the original question.
- Do not use or hint at the original answer in the modification.
- Use natural, grammatical language. Unrealistic attributes are allowed.
- Output must follow this exact format:

- Modified Questions [LIST]: [<modified question>]
- Attribute [LIST]: [<visual OR relational>]

- Examples:
  - Original: Which vehicle is closest to you?
  - Modified Questions [LIST]: [Which brown vehicle is closest to you?]
  - Attribute [LIST]: [visual]

7.3 Prompt for modified generation from Llava

Instruction (with/without image input) – You are a QUESTION MODIFIER, NOT a question answerer.

- DO NOT answer the question about the image.
- DO NOT describe what you see in the image.
- DO NOT tell me what objects are present.
- YOUR ONLY JOB: Modify the given question by adding exactly ONE attribute.

ORIGINAL QUESTION: "{}"

TASK: Create a modified version by adding ONE attribute:

- Visual attribute: Add an adjective before the main noun (e.g., "red car", "large building")
- Relational attribute: Add a position phrase after the main noun (e.g., "car on the left", "person in the background")

EXAMPLES:

- Original: "What color is the car?"
- Modified: "What color is the red car?"
- Original: "How many people are there?"
- Modified: "How many people on the left are there?"
- Original: "Where is the dog?"
- Modified: "Where is the small dog?"

Modify the original question. Respond ONLY with: Modified Question: [your modified question here] Attribute Type: [visual or relational]

7.4 Prompt for generating modified questions from LLM by adding hallucinated modifier

Instruction: –

- Add a modifier to this question.
- The modifier can be visual (color, size, shape, texture, pattern, visual state) OR relational (e.g., on, under, in front of, next to, in, among, etc.).
- Provide only ONE question with one or more modifiers.
- Make sure that the output is only the obtained MODIFIED QUESTION as shown below. Output Format:

  Modified Questions [LIST]: [Question1]

  Example:
  - Original: Is the girl in the black top thin?
  - Modified Questions [LIST]: [Is the girl in the black top obese?]

NOTE: The response must ONLY be a question with no explanation

7.5 Datasets

Table 7, 8 are datasets sampled from the test partition of VQA2.0 used for the experiments in Section 4. We use these to generate modified questions from each of the 4 VLMs by providing the original question and image. We also use these to generate questions using hallucinated modifications from each of the 4 VLMs by providing only the question without the image. Table 9, 10 show the question types in the dataset used to compare between human and AI modifications in Section 5.

7.6 Question verification using AMT

Fig. 7, Fig. 8 show the AMT task design for users to verify if the modification is correct or not, and then correct the modification if incorrect. Fig. 9 shows the AMT task to compare open-ended answers from the original and modified questions. Fig. 10 shows the AMT task design for where we compute the time-on-task to measure cognitive load in processing violations within the modified questions.

Question Type	#Questions
Is the	75
Is this	24
Are the	24
Is there a	18
Is this a	24
Is the girl	13
Does the	23
Is there	21
Is the man	18
Does	17
Are they	18
Are	21
Are there	23
Are there any	22
Do	28
Is the woman	18
Are these	21
Do the	15
Is the cat	1
Are the people	24
How many	42

Table 7: Total number of Open-ended-response questions types to evaluate VLM accuracy.

Question Type	#Questions
What	351
Which	21
Who	38
Why	41
Where	63

Table 8: Total number of Open-ended-response questions types to evaluate VLM accuracy.

7.7 Examples

Fig. 11, shows examples that compare how humans and VLMs respond to hallucinated modifiers. Fig. 12 shows examples that compare how humans and VLMs respond to human and AI modified questions.

Question Type	#Questions
Is the	10
Is this	57
Are the	25
Is there a	54
Is this a	41
Is the girl	1
Does the	39
Is there	47
Is the man	9
Does	30
Are they	9
Are	16
Are there	20
Are there any	8
Do	12
Is the women	4
Are these	15
Do the	3
Is the cat	1
Is the baby	2
Are the people	3
Is the boy	1
Is the lady	2
Is the old man	1

Table 9: Total number of binary-response and quantitative-response question types in the dataset used for comparing human and AI modifications.

Question Type	#Questions
Which	41
Who	16
Why	20
Where	51

Table 10: Total number of Open-ended-response questions types in the dataset used for comparing human and AI modifications.

Instruction: An image and two questions (Q1, Q2) are provided along with the answer to Q1. We gave question Q1 to AI model and asked it to add more information to the question from the image such that the answer to the modified question must be same as original answer. Q2 is the AI modified question. Your task is to check/correct this added information shown in red color. If modification in Q2 is suitable then click yes and if modification in Q2 is not suitable then provide right modification in the box.

Q1: Are the cat's eyes open?

Answer: no

Q2: Are the gray cat's eyes open?

Is the added information (shown in red) in Q2 a suitable modification (has the same answer as Q1)?

- (●) Yes
- ( ) No

Which object does it describe?

Cat

Figure 7: AMT Task for verifying if humans can pragmatically infer VLM-modified questions. If they choose that the modification is correct, they need to demonstrate their understanding of the question by answering a sub-question based on the salient object in the question.

Instruction: An image and two questions (Q1, Q2) are provided along with the answer to Q1. We gave question Q1 to AI model and asked it to add more information to the question from the image such that the answer to the modified question must be same as original answer. Q2 is the AI modified question. Your task is to check/correct this added information shown in red color. If modification in Q2 is suitable then click yes and if modification in Q2 is not suitable then provide right modification in the box.

Q1: Is the bike moving?

Answer: yes

Q2: Is the blue bike moving?

Is the added information (shown in red) in Q2 a suitable modification (has the same answer as Q1)?

○ Yes

◉ No

Please fill in the correct description of the object in the box:

Is the white *bike moving?*

Figure 8: AMT Task for verifying if humans can pragmatically infer VLM-modified questions. If they choose that the modification is incorrect, then they need to correct the modification using a textbox as shown in the example.

Instruction: You will see an image, two questions(Q1 and Q2) about the image and two respective answers(Answer1 and Answer2) . Based on the image, choose the better answer to the question.

Q1: What color are the zebra?

Answer 1: Black-and-white.

Q2: What color are the striped zebra?

Answer 2: Black.

Which answer is better from the image?

- (●) Answer 1 is better than Answer 2
- ( ) Answer 2 is better than Answer 1
- ( ) Both answers are equally acceptable

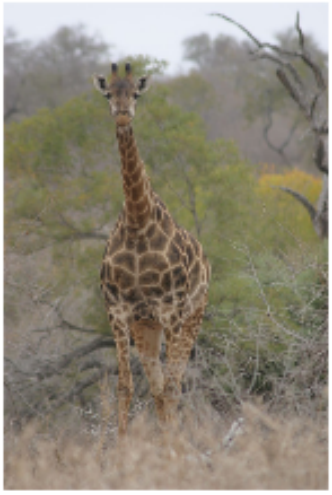
Figure 9: AMT Task for evaluating open-ended VLM responses by comparing answers to original and modified questions.

Instruction: You will see an image and a question about the image. Simplify the question by removing the non-essential details while keeping the meaning the same. Then, based on the image, provide your answer.

Examples

Here are examples of how to simplify questions while keeping their meaning:

Example 1:
Original: What is the color of the large vehicle parked on the side of the street?
Simplified: What is the color of the vehicle?
Example 2:
Original: How many people wearing blue shirts can be seen standing near the entrance?
Simplified: How many people are wearing blue shirts?
Example 3:
Original: Is there a small brown dog sitting on the green grass in the park?
Simplified: Is there a dog sitting on the grass?



Original Question: Is this animal on the ground facing the camera?

Simplified Question:

Is this animal facing the camera?

Answer:

Yes

Figure 10: AMT design to measure time-on-task for humans to answer a modified question and to reflect on the answer by simplifying the question.

Original: Are these id cards?

Modified: Is the woman holding a blue donut with her hands?

Human Answer: Yes

GPT4o Answer: No

Original: Are there tater tots on the tray?
Modified: Are there tater tots under the tray?

Human Answer: Yes

GPT4o Answer: No

Original: Is the woman holding a donut with her hands?
Modified: Is the woman holding a blue donut with her hands?

Human Answer: Yes

GPT4o Answer: No

Figure 11: Examples showing human and VLM responses on hallucinated modifications

Human Modified: Which one of the men in the snow could be a pirate?

Human Answer: Left

AI Modified: Which one is the larger pirate?

AI Answer: Right

Human Modified: Which orange, long item is different from the others?
Human Answer: Carrot

AI Modified: Which item on the table is different from the others?

AI Answer: Jar

Human Modified: Which room is in this?

Human Answer: Garage

AI Modified: Which room is this large white bear in?

AI Answer: Kitchen

Figure 12: Examples showing human and VLM responses on human and VLM modifications