MMMT-IF: A Challenging Multimodal Multi-Turn Instruction Following Benchmark

Anonymous ACL submission

Abstract

Evaluating instruction following capabilities in multimodal, multi-turn dialogue presents significant challenges, particularly when multiple instructions are distributed throughout the conversation. Current evaluation approaches often rely either on time-intensive human ratings or LLM-based judges, which we show have systematic bias toward responses from their own model family. We address these challenges by introducing MMMT-IF, a benchmark that augments image-based question-answering with global answer format instructions distributed between conversation turns. All instructions are verifiable through code execution, enabling objective evaluation. To measure performance, we introduce the Programmatic Instruction Following (PIF) metric, which quantifies the fraction of correctly followed instructions during reasoning tasks. This metric shows 60% correlation with human ratings, validating its reliability. Evaluation of leading models (Gemini 1.5 Pro, GPT-4o, and Claude 3.5 Sonnet) reveals significant performance degradation as conversations progress, with average PIF scores dropping from 0.81 at turn 1 to 0.64 at turn 20. Model performance deteriorates significantly when testing for consistency; when generating four responses per turn, GPT-4o and Gemini successfully follow all instructions only 11% of the time. Notably, when instructions are appended to the conversation end rather than distributed throughout, PIF scores improve by 22.3 points on average, indicating that retrieving multiple instructions from different parts of the input context, rather than instruction following itself, is the major challenge. The MMMT-IF dataset and metric computation code will be open-sourced.

1 Introduction

Despite the significant success of Large Foundation Models (LFMs) (Gemini et al., 2024; Open AI, 2024; Anthropic, 2024), instruction following is still a challenging task (Zhou et al., 2023a). This challenge becomes more pronounced when there are multiple instructions spread out over several turns in a chat setting between a user and a LFM, where the model needs to reason over various turns of the conversation. While there are several instruction following evaluation datasets, for example (Zhou et al., 2023a; Zhang et al., 2024), these evaluations are usually single-turn and most often use text input. Another key challenge is developing objective evaluation criteria for instruction following. In collecting human annotated reference answers for our evaluation dataset, annotators reported that, at each answer turn, rewriting the answer to follow all given instructions took 10 minutes on average, highlighting that human evaluation is time intensive. Recent developments have suggested using LLMs as judges of answer quality, but we found that there was a bias in the LLM judge to favor responses coming from the same model.

A new development has been to create tasks where model answers can be programmatically checked, in the domains of coding (Yang et al., 2023), data science (Huang et al., 2022), and text (Dong et al., 2025), ensuring an objective evaluation. Among these, (Dong et al., 2025) also focus on instruction following, but only in the single-turn, single modality setting. Current chat use cases are often multimodal and multi-turn, showing the need for objective instruction following evaluation datasets in this domain.

To address these limitations, we propose an instruction following benchmark, MMMT-IF, along with new metrics for multimodal multi-turn dialogue. Our proposal extends the MMDU evaluation dataset (Liu et al., 2024d), a multimodal, multi-turn chat task with independent question turns. An overview of our evaluation benchmark and key metric is shown in Figure 1. The MMMT-IF extension augments the MMDU by adding code verifiable format instructions between dialogue turns. Each

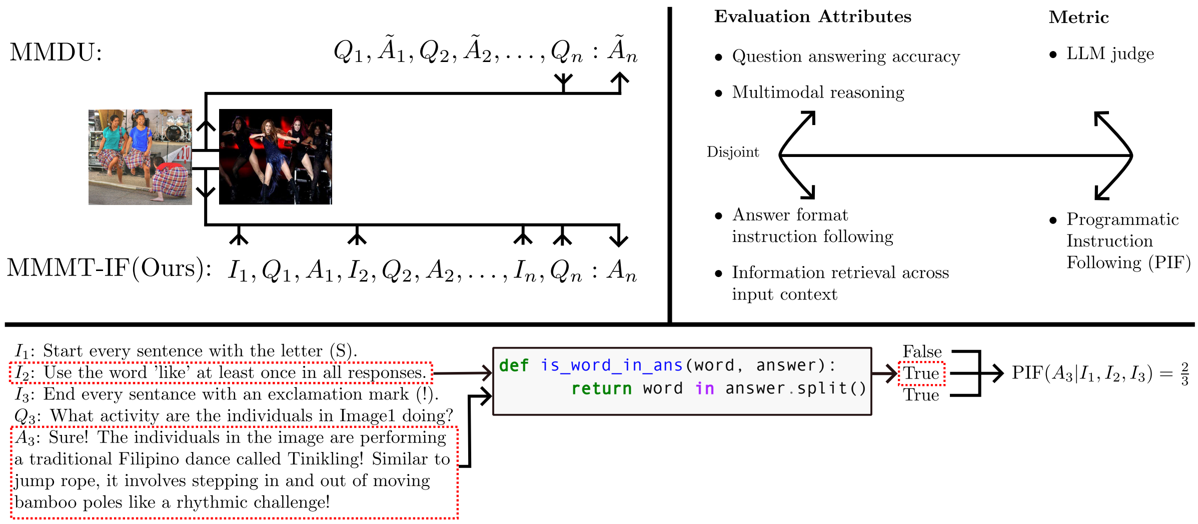

Figure 1: **Upper left**: The MMMT-IF evaluation benchmark is created by adding global instructions before each question turn in the MMDU benchmark. **Upper right**: The MMMT-IF benchmark focus on instruction following and information retrieval over multiple question turns and is evaluated with the programmatic instruction following (PIF) metric, this is disjoint from the MMDU benchmark which focuses on Q&A accuracy and multimodal reasoning, evaluated with an LLM judge. **Lower:** To compute the PIF metric at the current turn, Python functions are used to programmatically check the fraction of the given instructions that are followed.

instruction constrains the response format in a way that can be checked by code execution. Instructions accumulate throughout the conversation, with each new instruction adding to rather than replacing previous ones. This challenges the models as the task requires long context reasoning and retrieval of the instructions from different chat turns, creating a dataset that not only measures single-turn instruction following performance, but also how well a model can follow multiple instructions given throughout a conversation, a common chat use case. The task is not particularly challenging for human raters, who follow on average 94% of given instructions at each turn when writing reference answers for the MMMT-IF evaluation dataset.

We develop two metrics to measure instruction following capabilities: Programmatic Instruction Following (PIF), the fraction of given instructions in the chat that are followed at a certain turn, and (PIF-N-K), to stress test the ability of the models to consistently generate responses that follow all the instructions. To compute the PIF-N-K metric at a turn, we generate N responses, and the PIF-N-K metric is the fraction of the responses where at least K of the response candidates at a given turn follow all instructions, i.e., has PIF metric of 1. We conducted a human study to rate the instruction following capabilities at each turn, and found out that annotators' ratings have a correlation of 60% with the proposed PIF metric on the full MMMT-IF evaluation dataset. This highlights the relevance of the metric. We show that the evaluation suite is challenging for the models with evaluate it on: Gemini 1.5 Pro (Gemini et al., 2024), Claude 3.5 Sonnet (Anthropic, 2024) and GPT-4o (Open AI, 2024), with a significant loss in performance both over multiple turns and over multiple given instructions, as measured by the PIF metric. The average PIF across the models at turn 1 is high at 0.81, while at turn 20, it declines to 0.64. We develop a more nuanced measure by comparing empirical distributions of PIF scores at each question turn. Interestingly, Sonnet 3.5's PIF scores are consistently higher than Gemini's - not just on average, but across the entire distribution of performance outcomes. This means that regardless of how much one values higher PIF scores, Sonnet 3.5 would be preferred over Gemini 1.5 Pro at every turn with respect to the PIF metric.

A similar pattern is seen when conditioned on the number of given instructions. Conditional on having given 6 instructions, the best model in our benchmark, Sonnet 3.5 has a PIF score of 0.74, and Gemini 1.5 Pro has a PIF score of only 0.4. This is in stark contrast to the PIF metric conditional on 1 instruction given, where Gemini 1.5 Pro has an average PIF score of 0.68 and Sonnet 3.5 has an average PIF score of 0.97 on the evaluation dataset.

For the PIF-4-K metric, the PIF-4-4 metric is only 11% for both Gemini 1.5 Pro and GPT-4o, and 28% for Claude 3.5 Sonnet, showing that all models fail to robustly follow all given instructions correctly.

We show that a significant part of the challenge with the evaluation set is not following the instructions, but rather retrieving the instructions from the model context and then reasoning over the instructions. When all instructions are added in the end of the model input context in addition to the model context, the average PIF increased 22.3 points across all models, with Gemini 1.5 Pro improving from 0.473 to 0.739, GPT-4o from 0.647 to 0.856, and Sonnet from 0.771 to 0.974, highlighting that in addition to following the instructions, retrieving the instructions from the input model context remains challenging. This shows similarities with tasks such as multiple needles in a haystack, where the needles are instructions that needs to be reasoned over. Furthermore, our most challenging metric, the PIF-4-4 metric, showed an average improvement of 27 points, from an average of 0.16 across all models to an average of 0.43 when all given instructions were added in the end of the input model context.

To summarize, our main contributions in this work are:

1. We propose a methodology to extend multimodal multi-turn chat datasets to measure answer format instruction following, implemented on the MMDU dataset.

2. Two metrics, PIF and PIF-N-K, to measure, through code execution, the effectiveness for models to follow instruction, as well as their robustness in correctly following all given instructions.

3. We uncover a significant PIF performance degradation for all the models (Gemini 1.5 Pro, GPT-4o and Claude 3.5 Sonnet) as the number of given instructions increases.

4. We show that the main difficulty is not following the given instructions, but rather retrieving the instructions from the input model context and reasoning over them.

2 Dataset

This section describes the MMMT-IF evaluation dataset, as well as the human data we collect to create reference answers and preference ratings.

2.1 Instruction Following Extension

The extension is visualized in Figure 1. The extension adds answer format constraint instructions in between questions in the dialogue from the MMDU benchmark. All instructions are chosen so that the correctness of a response can be verified through code execution, enabling an unbiased and automated evaluation of instruction adherence. The instructions are global within a chat, meaning that all instructions from previous turns needs to be followed for future turns. Each instruction is chosen from separate categories (for example, one category dictates the start character for answer sentences, and another category dictates the end character for answer sentences). All the categories are independent from each other. Each category has either 2 or 3 instruction options. Before each question, with probability $1 - \frac{\#\text{ Instruction given so far}}{6}$ another instruction is added, uniformly at random chosen from a category (in total there are 6 categories) not yet added, hence dialogues includes a maximum of 6 instructions. As a result, most chats will receive 6 instructions between turns 6 and 10. Given an average chat length of 14, this means that 6 instructions will be the most common number received across all turns, as shown in Figure 2. This increases the task's difficulty, as turns with more instructions are harder to satisfy completely. A partial set of all the instructions are in Figure 1. I_1 corresponds to category 'sentence start character', I_2 corresponds to category 'favorite word', and I_3 corresponds to category 'sentence end character'. The full set of all instructions and instruction categories are available in Appendix D.

Note that the extension makes the task also require more long context abilities in the models, as instructions needs to be retrieved from multiple parts of the input model context. We view it as a strength of the work that we are able to reuse the base questions and images from a previous benchmark, as we test model capabilities disjoint from the original benchmark, and use different evaluation metrics. Our method is a general way to extend a Q&A benchmark to test instruction following. Table 1 shows several statistics about the properties of the MMMT-IF evaluation set. We describe the details we used to filter the dataset in Appendix D. Most of the 71 conversations are at least 10 turns, and none are more than 20 turns. The full distribution of conversation lengths are in Appendix D.

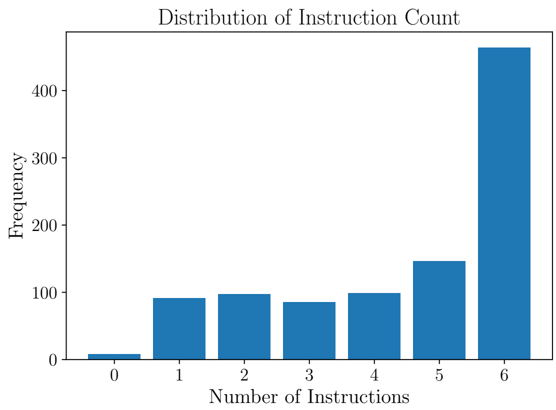

Figure 2: For all 990 turns, the distribution of the number of instructions that were given so far in the chat.

Table 1: Descriptive Statistics of the MMMT-IF dataset

Quantity	Value
# Chat turns	990
# Chats	71
# Images per chat	2 - 5
# Turns per chat	1 - 20
# Instructions per chat	1 - 6

2.2 Human written reference labels

We collect human labels for a reference response that both answers the questions correctly and follows all the constraints from the given instructions. In addition, the human annotators were asked to rate the answer accuracy from 1 to 10, the instruction following accuracy from 1 to 10 and give a pairwise preference score between each of the models (Gemini 1.5 Pro, GPT-4o, and Claude 3.5 Sonnet) in our evaluation set. The full set of instructions given to the human annotators is in the Appendix G.

3 Evaluation Metrics

This section introduces the PIF and PIF-N-K metrics, and provides a rationale for their use.

3.1 Programmatic Instruction Following Metric

Given model input context X (containing the input images, previous instructions, previous questions, and previous answers), and model response Y, we can define the sample PIF metric for that response to be

$$\mathrm{PIF}(X, Y) = \frac{\text{\# Instructions in } X \text{ followed in response } Y}{\text{\# Instructions in input context } X},$$

where we use 'instruction' to refer to the constraints we give on the answer format. Note that the PIF considers whether the response follows all given instructions in previous turns, not just the instruction given at the current turn. The PIF metric does not take into account if the question was answered correctly, but rather, it focuses on if the instructions given to constrain the answer were followed. For our evaluation set, we have $M = 71$ chats, and chat $i \in \{1, \ldots, M\}$ have N_i turns. This gives us our evaluation set: $D = \{(X_{i,j}, Y_{i,j})\}_{i=1,j=1}^{M,N_i}$, where $X_{i,j}$ is the input model context for chat i at turn j, and $Y_{i,j}$ is the model response for chat i at turn j. We define the corpus level (mean) PIF score as

$$\mathrm{PIF}(D) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{1}{N_i} \mathrm{PIF}(X_{i,j}, Y_{i,j}).$$

The corpus level Programmatic Instruction Following Score conditioned on turn j, is given by

$$\mathrm{PIF}(D|\text{turn} = j) = \mathrm{PIF}(\{(X_{i,j}, Y_{i,j})\}_{i=1}^{M}),$$

where chats with less than j turns are excluded. It will be clear from the context whether we refer to the corpus or sample PIF metric.

The PIF metric captures the following aspects:

1. The ability for a model to retrieve several pieces of information from different parts of an input text context and reason over them
2. The ability for the model to follow objective instructions

Of these, we think the most important is the first, as this is a very common scenario for real use-cases, and it's a feature that single-turn based metrics are not capturing well.

3.2 Consistency Metric, PIF-N-K

In addition to having a high average score, we want models to consistently produce the same high quality results. We propose a metric to capture this intuition; for each turn N responses are sampled, and PIF-N-K will then denote the fraction of samples where at least K samples have PIF score 1.

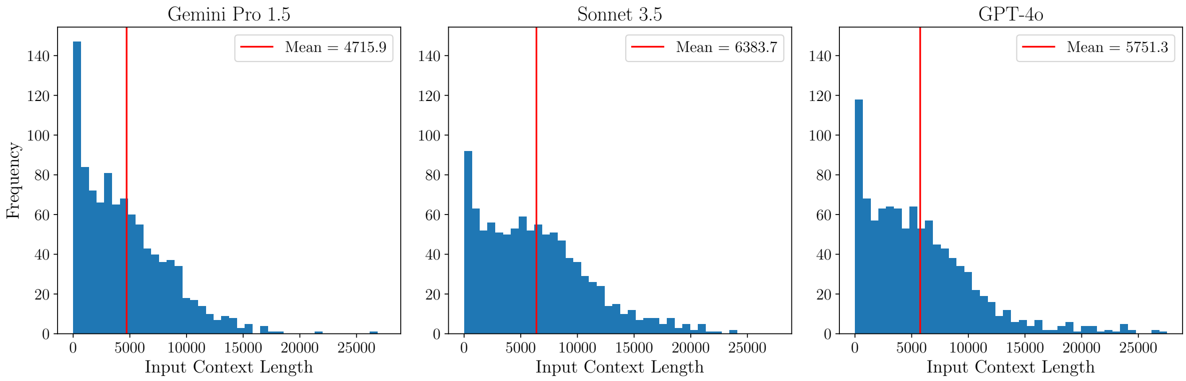

Figure 3: The distribution of the input context lengths for Gemini 1.5 Pro, Claude 3.5 Sonnet and GPT-4o, in the evaluation dataset, along with the mean input context length in characters.

Thus the sample level PIF-N-K, for input model context X, and sampled responses $Y^1, \dots Y^N$ is

$$\text{PIF-N-K}(X, Y^1, \dots, Y^N) =$$

$$\begin{cases} 1, & \text{if } \sum_{i=1}^{N} \mathbf{1}_{\text{PIF(X,Y}^\text{i})=1} \geq K \\ 0, & \text{otherwise} \end{cases}$$

The intuition is that we want to measure how consistently the models can follow all the instructions correctly. We overload notation and define the corpus level (mean) PIF-N-K for a dataset with L turns, $D = \{X_i, Y_i^1, \dots, Y_i^N\}_{i=1}^L$ as

$$\text{PIF-N-K}(D) =$$

$$\frac{1}{L}\sum_{i=1}^{L} \text{PIF-N-K}(X_i, Y_i^1, \dots, Y_i^N).$$

With this definition it holds that, for any dataset D,

$$\text{PIF-N-i}(D) \leq \text{PIF-N-j}(D)$$

when $i > j$.

4 Evaluated Models

This section describes the models evaluated, and provides an analysis of the answer lengths of the models.

4.1 Model Endpoints

We access Gemini 1.5 Pro (abbreviated as Gemini) through the Vertex AI API, using the following model version: 'Gemini-1.5-pro-preview-0514'. We access Claude 3.5 Sonnet (abbreviated as Sonnet) through the Anthropic Vertex API, with the model version 'claude-3-5-sonnet@20240620'. We access GPT-4o from the OpenAI API with the model version 'gpt-4o-2024-05-13'. The hyperparameters for all models are the default settings. The default temperature for all models is 1. The safety filters for all models are the default settings. We don't see questions that are marked as unsafe with the default setting for the models.

4.2 Context Lengths

Figure 3 shows that the mean input context length for Gemini 1.5 Pro is the smallest, as the input context is made up from the questions and model outputs in the previous turns, and the average output generated is shortest by Gemini 1.5 Pro. This does not take into account the images that are inputted at the beginning of each chat. It also shows that the average input context is rather long, thus requiring long context reasoning.

5 Evaluation Results

The section describes the results from the evaluation experiments, starting with results for the PIF metric, then considering similarities with the needle in a haystack experiment, results for the PIF-N-K metric, before finally considering human evaluation results, and inherent biases with a popular alternative approach using LLM judges to measure instruction following performance.

5.1 PIF Metric

Figure 4 shows the PIF conditional on question turn. We note that the PIF metric decreases with the question turn. The 95% confidence bounds for the PIF metric are done on a per-turn basis, using a Bernoulli confidence interval approximation. This gives conservative confidence bounds as the Bernoulli distribution is the distribution that for

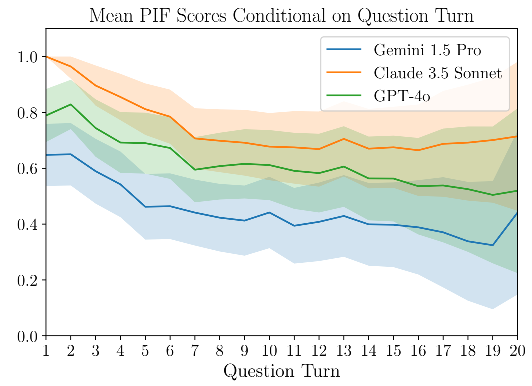

Figure 4: The mean PIF metric conditioned on the question turn with 95% confidence intervals. For a fixed turn i, the mean is taken across all chats at with at least i turns.

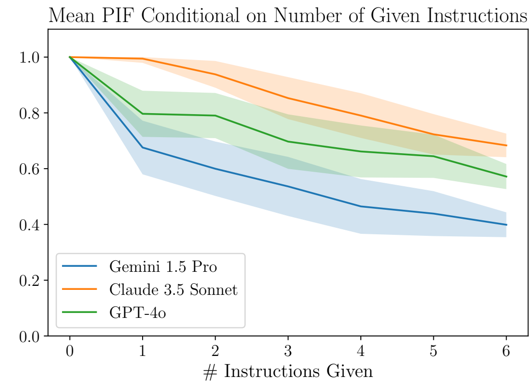

Figure 5: The mean PIF metric conditioned on the number of instructions given in the chat so far. The metric defaults to 1 if no instruction has been given.

a given mean maximizes the variance among all distributions on [0,1].

From Figure 5 we see that the scores decrease with the number of given instructions, as it's harder for the models to follow multiple instructions at the same time. Also note that Gemini 1.5 Pro has a significantly lower score for high number of instructions compared with Sonnet and GPT-4o, highlighting an area for improvement. Finally, note that the programmatic instruction following metrics is automatically evaluated by code execution, which increases the reliability of the shown results. The 95% confidence intervals are computed with a Bernoulli approximation.

Figure 6 shows the empirical cumulative distribution function for the PIF metric. The interpretation of the left graph in Figure 6 is that at turn 2, the programmatic instruction following score can be 0, 0.5, or 1. For Gemini 1.5 Pro, it's 0 with probability 18%, while for GPT-4o it's zero with probability around 10%. The probability that the programmatic instruction following score is less than 1 (i.e., 0.5 or 0) is around 35% for GPT-4o, 52% for Gemini and 10% for Sonnet. Not only is the average PIF score better for Sonnet at each question turn, it's also true that $P(\text{PIF}_{\text{Sonnet}}(X,Y) > x|\text{turn} = i) \geq P(\text{PIF}_{\text{Gemini}}(X,Y) \geq x|\text{turn} = i)$ for all $x \in [0,1]$, and for all turns i, for any model input context X and model response Y in the evaluation set at turn i, and P is the empirical measure from all samples in the MMMT-IF evaluation set.

5.2 Extension of Needles In a Haystack

The Needle in a Haystack test involves embedding a random statement ("needle") within a long context ("haystack") and prompting an LLM to retrieve it. Our experimental setup has several similarities and differences with a needle in a haystack experiment. In our setting, the complex reasoning across the needles (given instructions) is important, in addition to the retrieval of the needles. To understand the impact of where in the input model context the instructions are located, we run the following ablation: In addition to having instructions throughout the input context, we add all given instructions at the end of the input model context. Table 2 shows the results, where we see that the corpus level PIF increased 22.3 points on average across all models, highlighting that in addition to following the instructions, retrieving the instructions from the input model context remains challenging. This suggests a practical method to improve instruction following capabilities in multi-turn chat: find all the instructions and add them to the end of the input model context.

The first row of Table 2 also highlights statistically significant differences in the corpus level (all 990 turns) PIF metric between the evaluated models. We see that the programmatic instruction following score is best for Sonnet, and Gemini has the weakest performance. Using the (non parametric) Wilcoxon Signed Rank test, we reject the hypotheses $H_0 : P(\text{PIF}_{\text{Gemini}} > \text{PIF}_{\text{Sonnet}}) >= 0.5$ with p-value smaller than 10^{-5}. Using the Wilcoxon Signed Rank test, we also reject the hypotheses $H_0 : P(\text{PIF}_{\text{Gemini}} > \text{PIF}_{\text{GPT-4o}}) >= 0.5$ with p-value smaller than 10^{-5}. The difference between the models for the mean programmatic instruction following metric is significant.

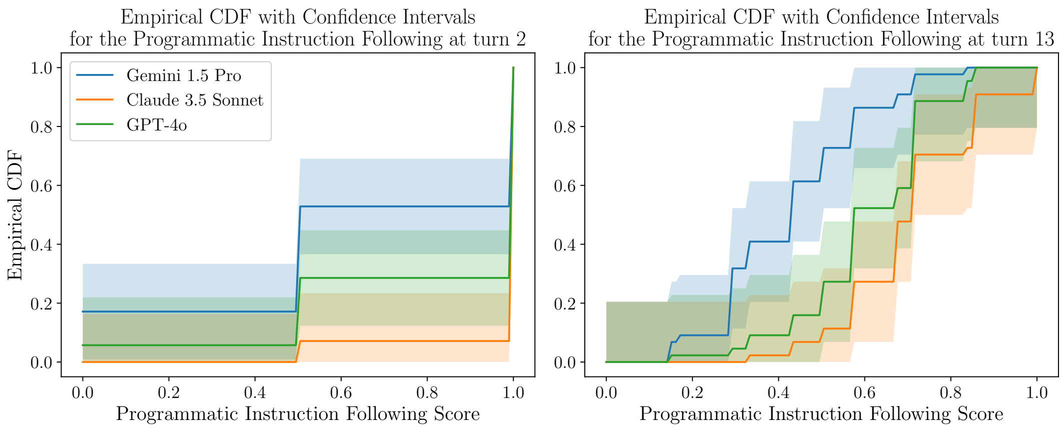

Figure 6: The empirical CDF of the PIF metric conditional conditional on question turn 2 (left) and 13 (right), with confidence intervals. Here, a lower CDF value for a given PIF score is better.

Table 2: Mean PIF metric on the MMMT-IF evaluation dataset.

Metric	Gemini 1.5 Pro	GPT-4o	Sonnet 3.5
Programmatic Instruction Following (PIF)	0.473	0.647	0.771
PIF with all instructions added at end of input prompt	0.739	0.856	0.974

5.3 PIF-N-K **Metric**

We now consider the results for the PIF-N-K, measuring the robustness for following all given instructions correctly. In our experiments we set $N = 4$. Figure 7 shows the results. As expected PIF-4-4, meaning the fraction of turns where all $N = 4$ sampled turn answer candidates got all the instructions correct is quite low, for both Gemini and GPT-4o it's 11%, highlighting that this is a very challenging metric with significant headroom for model improvement. However, note that also for Sonnet 3.5, the model with the strongest performance, the metric rapidly becomes more challenging as we move from PIF-4-1 to PIF-4-4. This points to a significant robustness issue with the models we have studied in this work, as if the model always had the same percentage of instructions followed in its responses, we would not see a decrease in the PIF-N-K metric.

5.4 Human Evaluation

As described in Section 2.2 we collect human evaluations of instruction following, chat accuracy and pairwise preferences. In Figure 8 with human evaluations, we observe that Gemini underperforms GPT-4o and Sonnet, and Sonnet and GPT-4o are broadly similar. We also find that the correlation between the human instruction rating and the PIF metric, the results are shown in Table 3. We note that the average correlation across all models is high, 0.60, indicating the usefulness of the PIF metric to capture the instruction following of the models. In Table 4 we see that the average human evaluation score for accuracy is highest for GPT-4o, highlighting that while PIF score is an important metric, there are several aspects of model performance the metric does not cover.

Table 3: Correlation between the PIF metric and the human rated instruction following metric for 990 samples from human raters.

Overall Correlation	Gemini 1.5 Pro	GPT-4 o	Sonnet 3.5
0.60	0.44	0.68	0.63

5.4.1 How hard is the task for human raters?

Starting with a reference answer from the original MMDU dataset, human raters were instructed to rewrite the responses to both be correct and to follow all the given instructions. The human raters had access to the LLM model responses, the orig-

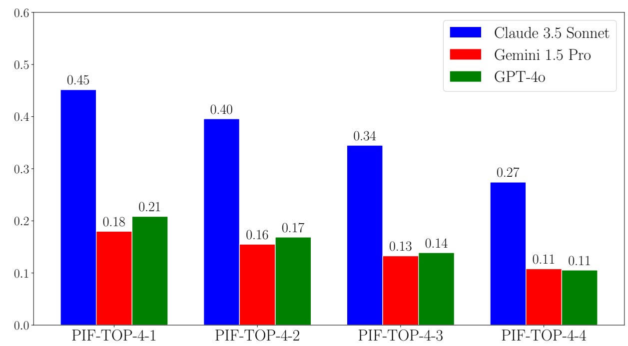

Figure 7: The corpus level PIF-4-j metrics for $j \in \{1, 2, 3, 4\}$.

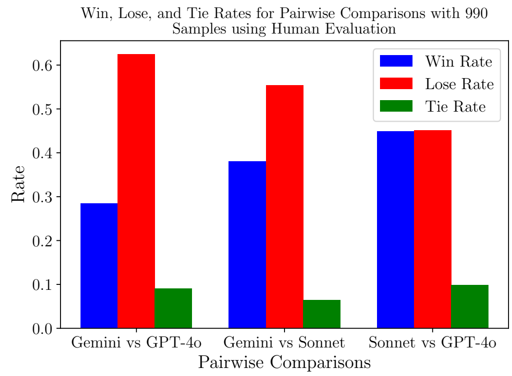

Figure 8: Win, Tie, and Loss rate using human preference rankings for Gemini vs GPT-4o, Gemini vs Sonnet, and Sonnet vs GPT-4o.

inal reference answer for the MMDU dataset, as well as a list of all instructions given in the chat, so they did not have to look at the chat history to find the instructions. The raters reported that it took on average 10 minutes to write the answer and reported that the hardest instructions to satisfy where the constraints on the sentence start word and the constraints on the sentence lengths. The programmatic instruction following scores for the human raters have an average of 0.94, significantly higher than both Gemini and GPT-4o with all instructions in the end of the input context, but actually lower than Sonnet 3.5 in the setting with all instructions added at the end of the input model context, at a mean PIF score of 0.97. This highlights that while the task is challenging, the human raters are able to complete it with great proficiency, indicating that there is headroom for models to improve. The raters reported that having access to the model answers helped speed up the rewriting process by giving inspiration to ways to follows the given constraints.

Table 4: Gemini vs GPT-4o as auto-rater vs human evaluation on the MMMT-IF dataset.

Judge	**Gemini 1.5 Pro**	**GPT-4o**	**Human**
Avg Accuracy			
Gemini-1.5 Pro	6.95	7.36	6.04
GPT-4o	7.07	7.82	6.70
Sonnet 3.5	6.92	7.44	6.33
Avg Instruction Following			
Gemini-1.5 Pro	7.61	8.33	3.80
GPT-4o	7.65	9.06	4.41
Sonnet 3.5	7.81	9.01	5.32

5.5 Auto-Rater Bias

To understand the reliability of using LLM-judges as auto-raters of response quality, we use the an extension of the LLM-based judge as in (Liu et al., 2024d), with different models as the LLM judge. We focus on instruction following and answer accuracy, with the same prompt format for the LFMs as for the human evaluators, described in Appendix G. From Table 4 we observe that with GPT-4o as judge, GPT-4o performs better, and with Gemini as a judge, Gemini has a better performance. The Gemini based auto-rater gets the relative order of the instruction following correct (relative to human evaluation) whereas the GPT-4o judge ranks the GPT-4o as having better instruction following than Sonnet. In addition, we see that the human rater scores are in general more conservative. For the accuracy ratings, the GPT-4o judge has the same relative ranking as the human raters, which the Gemini judge does not. This underscores the need for using objective evaluation criteria.

6 Conclusion

In this work we proposed the MMMT-IF LFM instruction following evaluation set for multimodal, multi-turn dialogue, along with several metrics verifiable by code execution that are highly correlated with human evaluations. We show that all evaluated models have a strongly degrading performance with the number of conversation turns. Our analysis shows that the main difficulty of the task lies not within the instruction following, but rather to retrieve the instructions from different parts of the input context and then reason over them.

7 Limitations

One limitation of our paper is that only a subset of all possible instructions are suitable to be used in our evaluation set. Critically, only instructions that can be confirmed by code execution are used, this often limits it to constraints on the answer format. Many subjective instruction types can't be used. Another limitation is that the benchmark is focused only on English language, limiting cross-lingual insights. Due to resource constraints (each turn takes 10 minutes on average for the annotators), we only used one human annotation per chat turn, with more resources it would have been useful to consider inter-annotator agreement. Finally, the PIF metric only captures binary compliance for each instruction, while sometimes an instruction is almost followed but still get zero score.

8 Ethical Considerations

We declare that all authors of this paper acknowledge the ACL Code of Ethics and honor the code of conduct. We believe our benchmark to be accessible to researchers with low resources, as our main PIF metric relies solely on executing short Python code, rather than using human evaluation or an LLM based evaluation through an API. While our PIF metric shows good correlation with human judgment (0.60), we acknowledge a significant risk that future language models may be trained on this benchmark, potentially compromising its effectiveness as an evaluation tool even if human correlation remains high. We have thoroughly scanned the created evaluation benchmark for offensive content and personally identifying information, none of which were present. The annotators were based in the United States and paid fair wages. We commit to open-sourcing the MMMT-IF dataset and metric computation code under the Apache-2.0 License.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2024. Introducing the next generation of claude.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, and Maarten de Rijke. 2024c. The SIFo benchmark: Investigating the sequential instruction following ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1691–1706, Miami, Florida, USA. Association for Computational Linguistics.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2024. InstructEval: Towards holistic evaluation of instruction-tuned large language models. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 35–64, St. Julian's, Malta. Association for Computational Linguistics.

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2025. Self-play with execution feedback: Improving instruction-following capabilities of large language models. In *The Thirteenth International Conference on Learning Representations*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: a simulation framework for methods that learn from human feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Gemini et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. In *Findings* *of the Association for Computational Linguistics: EMNLP 2024*, pages 10864–10882, Miami, Florida, USA. Association for Computational Linguistics.

Junjie Huang, Chenglong Wang, Jipeng Zhang, Cong Yan, Haotian Cui, Jeevana Priya Inala, Colin Clement, and Nan Duan. 2022. Execution-based evaluation for data science code generation models. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 28–36, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. 2024b. Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024d. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Open AI. 2024. Hello gpt-4o.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.

Ondrej Skopek, Rahul Aralikatte, Sian Gooding, and Victor Carbune. 2023. Towards better evaluation of instruction-following: A case-study in summarization. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 221–237, Singapore. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9750, Bangkok, Thailand. Association for Computational Linguistics.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2024. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *Preprint*, arXiv:2406.11230.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. FOFO: A benchmark to evaluate LLMs' format-following capability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, Bangkok, Thailand. Association for Computational Linguistics.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2025. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1877–1893.

Jianhao Yan, Yun Luo, and Yue Zhang. 2024. RefuteBench: Evaluating refuting instruction-following for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13775–13791, Bangkok, Thailand. Association for Computational Linguistics.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023. Intercode: Standardizing and benchmarking interactive coding with execution feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 23826–23854. Curran Associates, Inc.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Tao Zhang, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Tao Zhang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, Bin Cui, Wentao Zhang, and Zenan Zhou. 2024. Cfbench: A comprehensive constraints-following benchmark for llms. *Preprint*, arXiv:2408.01122.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023a. Instruction-following evaluation for large language models. *ArXiv*, abs/2311.07911.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023b. Controlled text generation with natural language instructions. *Preprint*, arXiv:2304.14293.

A Overview of Appendix

The appendix provides supplementary information across six main sections:

- B. Literature Review: Related work in instruction following evaluation, programmatic instruction following, multimodal evaluation datasets, long context retrieval, and LLM judges.
- C. Additional Experiments: Detailed performance analysis on specific instruction types, model response lengths, and correlation between PIF metric and human accuracy scores.
- D. Additional Dataset Details: Dataset question characteristics, filtering criteria used to create the final dataset and complete set of code verifiable instructions used in the evaluation.
- E. Example Chat: A partial conversation with corresponding images demonstrating the interaction format and instruction following evaluation.
- F. Error Analysis: An examination of common failure modes and error patterns observed across different models.
- G. Human Annotator Instructions: Complete guidelines and rubrics provided to human annotators for dataset creation and evaluation.

B Literature Review

B.1 Instruction Following Evaluation

There are several instruction following evaluation benchmarks focusing on instructions related to answer constraints on a Q&A task, (Xia et al., 2024; Zhou et al., 2023a; Zhang et al., 2024; Tam et al., 2024; He et al., 2024). Compared with these works, we focus on multiple instructions spread out over a long context, testing not only instruction following but also retrieval and complex reasoning from the input context. There have been many other instruction following evaluation sets (Chen et al., 2024c; Zhou et al., 2023b; Adlakha et al., 2024; Sun et al., 2024; Yan et al., 2024; Jiang et al., 2024; Chia et al., 2024; Skopek et al., 2023; Qin et al., 2024), but their focus in not on multiple instructions spread out in the input context for multimodal multi-turn chat.

B.2 Programmatic Instruction Following

Several previous papers use program execution to determine instruction following capability, for code (Yang et al., 2023), Data science (Huang et al., 2022), and text (Dong et al., 2025). Our work is most related to (Dong et al., 2025), but we fix a set of instructions, and instead of a single instruction use case, we focus on multiple instructions, over multiple turns of multimodal question answering.

B.3 Multimodal Evaluation Datasets

There have been several benchmarks suggested for multimodal models, for the multi-turn chat use case (Liu et al., 2024d,b). However, while the datasets are multi-turn, the chat turns can be independently answered, thus making it less relevant for long context models. By introducing given at several locations throughout the chat, we introduce long range dependencies in the data needed to answer questions. Other work for evaluating multimodal models include (Yue et al., 2024; Liu et al., 2024c; Srinivasan et al., 2021; Yu et al., 2024; Xu et al., 2025; Chen et al., 2024b; Wang et al., 2024). None of these focus on multi-turn instruction following.

B.4 Long context Retrieval

There have been several works focusing on the effect of long input context on model performance on downstream tasks, including (Liu et al., 2024a; Levy et al., 2024; An et al., 2024). Similar to the Lost-in-the-middle paper (Liu et al., 2024a), our paper examines the effect of where in the input context information is located. The results in (Levy et al., 2024) are also complementary, as both observe performance degradation with input context length increases. Our evaluation set can also be viewed as a task similar to multiple needles in the haystack (a task where several tokens needs to be retrieved from a long input context), where each needle is an instruction that the model needs to reason over.

B.5 LLM judges

There have been several previous works on using LLMs to judge quality of other LLM responses, including (Dubois et al., 2023; Zheng et al., 2023; Chen et al., 2024a; Dubois et al., 2024; Zeng et al., 2024; Liu et al., 2024d). While these work mostly focus on using LLM judges, we focus on some of the potential drawbacks due to a bias in which a model LLM judge tend to rates higher answers from models within the same model family.

C Additional Experiments

C.1 Performance on Specific Instructions

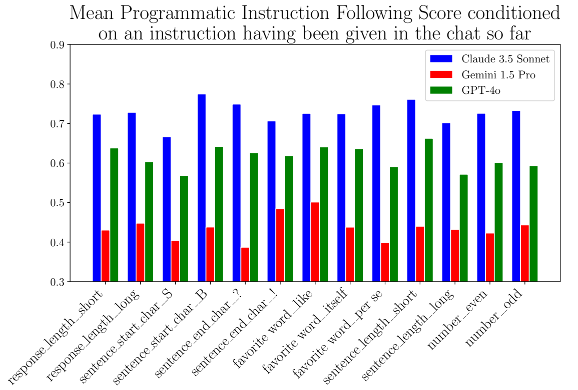

Figure 9: the mean conditional programmatic instruction following score conditioned on an instruction having been given in the chat.

In Figure 9 the PIF score conditional on an instruction having been given is shown. We note that Gemini 1.5 Pro has a hard time following an instruction to end sentences with a question mark, and GPT-4o has some issues with following instructions related to outputting even or odd numbers in its responses. The definition of the categories are presented in Table 5.

C.2 Analysis of Dataset Questions

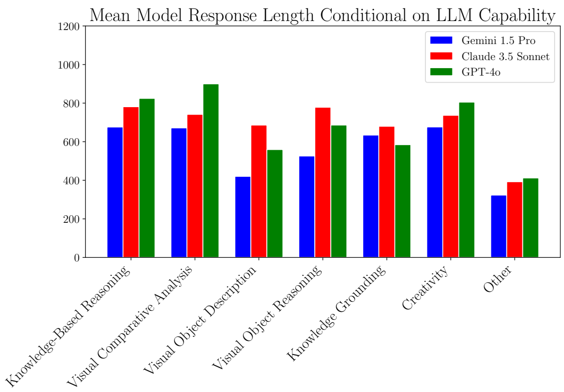

Figure 10: The mean answer length conditional on the LLM capability the question most closely targets.

In Figure 10, we show the average response length of conditioned on the LLM capability the question targets. We see that questions classified as Creativity and Visual Comparative Analysis have longer average answer lengths compared with those classified as visual object description.

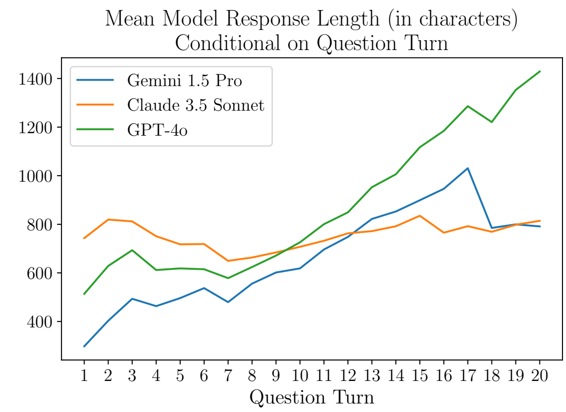

Figure 11: Mean response length (in characters) conditional on question turn in the MMMT-IF evalution set.

C.3 PIF **Metric and Human Accuracy Scores**

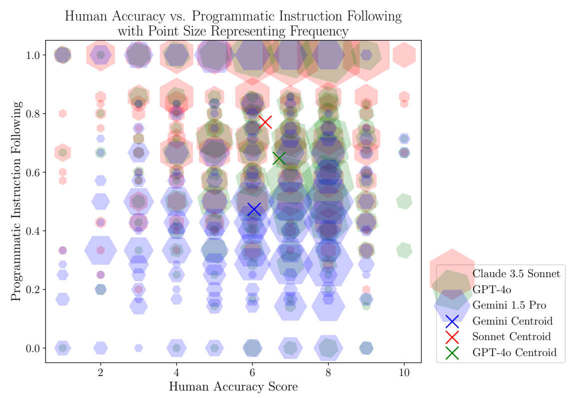

Figure 12: Scatter plot for Gemini 1.5 Pro, GPT-4o and Claude 3.5 Sonnet responses with PIF scores on the y axis and human accuracy scores on the x axis. The size of the points is proportional to the number of samples with the same PIF score and human accuracy score.

While the PIF score is an important metric for instruction following, it's also important to answer the image based questions in the MMMT-IF dataset correctly, not only following the answer constraints. Figure 12 shows a scatter plot with PIF score on the y axis and human accuracy score for each turn on the x axis. It's desirable to both have high accuracy score and high PIF score, but this is relatively uncommon as shown in the Figure, highlighting the challenge of the task. In the Figure the cluser centroids are also shown. Note that GPT-4o responses have the highest average human accuracy scores and Claude 3.5 Sonnet have the highest average PIF scores. Also note that the Sonnet responses have more robustly high PIF score, and the GPT-4o responses have more robustly high human accuracy scores.

D Additional Dataset Details

D.1 Analysis of Dataset Questions

In Figure 13 we display the model capabilities targeted in each question turn, where the classification is done by GPT-4o. We manually reviewed the classifications to ensure they were aligned with human categorizations. Rather then focus on the LLM capability, Figure 14 shows the distribution of the questions topics in the dataset, classified by GPT-4o. Many questions are related to flowers, plants, architecture, food and vehicles.

D.2 Data Filtering

The initial evaluation set, had a total of 1342 turns, from 98 chats, the data was filtered down to 990 turns, with 71 full chats, based on the following criteria:

1. Removing chats where some image is corrupted: 23 chats.
2. Removing chats with more than 5 images: 3 chats.
3. Removing chats containing skipped turns due to model error or content filters: 1 chat.
4. Truncating chats to have a maximum length of 20 turns.

D.3 Full Instructions

The full set of instructions is given in Table 5.

E Example Chat

In Table 6 we show an example chat and Figure 15 shows the images corresponding to the chat. In Table 6, the column "All instructions" shows all the instructions given in previous and current chat turns. However, the model will not be directly given the data in this column, but rather needs to find the instructions in the input context, where each instructions appear before questions in the user turns. The PIF column shows the PIF score for the considered chat turn.

F Error Analysis

In this section we explore a few chats and the errors made. While the average performance of the PIF metric for human raters was at 0.94, the lowest

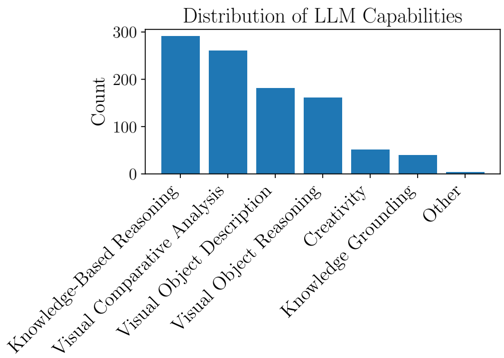

Figure 13: The distribution of LLM capabilities that the questions in the dataset targets.

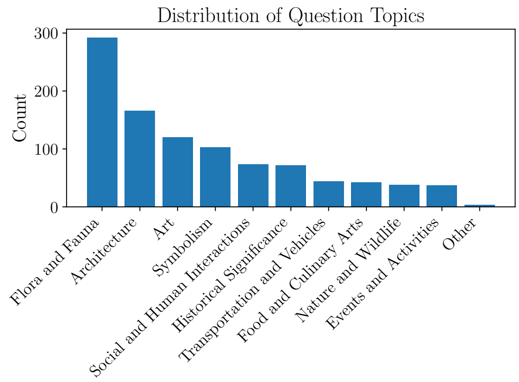

Figure 14: The questions topics, as rated by an GPT-4o with access to the associated images and verified by a human for each turn in the MMMT-IF eval set.

Table 5: Full set of instructions in the MMMT-IF dataset.

Name	Instruction
Response length short	Instruction: Make all the following responses no more than 4 sentences.
Response length long	Instruction: Make all the following responses at least 5 sentences.
Sentence start char S	Instruction: Start every sentence with the letter (S).
Sentence start char B	Instruction: Start every sentence with the letter (B).
Sentence end char ?	Instruction: End every sentence with a question mark (?).
Sentence end char !	Instruction: End every sentence with an exclamation mark (!).
Favorite word like	Instruction: Use the word 'like' at least once in all future responses.
Favorite word itself	Instruction: Use the word 'itself' at least once in all future responses.
Favorite word per se	Instruction: Use the word 'per se' at least once in all future responses.
Sentence length short	Instruction: Only use responses to questions where each sentence in the response is at most 18 words in all future responses.
Sentence length long	Instruction: Only use responses to questions where each sentence in the response is at least 18 words in all future responses.
Number even	Instruction: Include at least one even number bigger than 5 in each of your responses.
Number odd	Instruction: Include at least one odd number bigger than 5 in each of your responses.

observed PIF score for a chat turn was 0.4, the chat turn is shown in Table 7. Note that the word 'like' is not in the response, and that sentence 7 is both longer than 18 words and start with the letter T, hence the PIF score of 0.4. Notice that the answer looks broadly correct and it requires a careful view to spot the errors.

Gemini repeats the same answer One error pattern noticed for Gemini 1.5 Pro was that it's repeating the same answer to multiple questions, see row 1-4 in Table 8, which are consecutive turns in a conversation.

Gemini gives up In Table 9, we show a response where Gemini-1.5 Pro claims the task is impossible, when it's not, and hence fails to adhere to any of the six given instructions.

GPT-4o ignores number instruction In Table 10, we show several chats where GPT-4o ignores instructions related to numbers.

G Human Annotator Instructions

In this section the instructions given to human annotators are presented. There were 8 human annotators working on separate partitions (split by chat) of the MMMT-IF dataset. In Figure 16 we show the instructions (lightly edited to remove details on how to operate the software application used for the task) given to human annotators.

Table 6: Several turns in one of the chats in the MMMT-IF benchmark.

All instructions	Turn	Question	Gemini Response	PIF
Instruction: Include at least one even number in the range bigger than 5 in each of your responses.	1	What activity are the individuals in the Image1 engaged in, which is characterized by rhythmic and fast-paced footwork?	The individuals in Image 1 are performing Tinikling, a traditional Philippine folk dance characterized by rhythmic and fast-paced footwork, often using bamboo poles. Dancers need a lot of agility to avoid getting their feet caught between the poles, especially as the pace increases to 10 or 12 beats per second.	1 (1/1)
⋮	⋮	⋮	⋮	⋮
Instruction: Include at least one even number in the range bigger than 5 in each of your responses. Instruction: End every sentence with a exclamation mark (!). Instruction: Use the word 'itself' at least once in all future responses. Instruction: Only use responses to question where each sentence in the response is at most 18 words in all future responses. Instruction: Make all the following responses no more than 4 sentences. Instruction: Start every sentence with the letter (S).	15	Identify the possible significance of the props or lack thereof in the settings of Image1 and Image2.	Stage props can enhance the narrative or symbolism of a performance. Simple props, like the bamboo poles in Image 1, become integral to the dance itself. Conversely, the absence of props in Image 2 might emphasize the dancers' movements and stage presence. Such choices reflect deliberate artistic decisions in conveying a performance's message.	0.5 (3/6)

(a) Image 1 in example chat.

(b) Image 2 in example chat.

Figure 15: Images for the example chat.

Table 7: The chat turn with the lowest PIF score for the human rater.

All instructions	Question	Human Rater Answer	PIF
Instruction: Include at least one even number in the range bigger than 5 in each of your responses. Instruction: Only use responses to question where each sentence in the response is at most 18 words in all future responses. Instruction: Start every sentence with the letter (B). Instruction: Use the word 'like' at least once in all future responses. Instruction: End every sentence with a exclamation mark (!).	What could be the symbolic meanings of the food items in Image1 and Image2 in the context of their respective culinary cultures?	Both food items in Image1 and Image2 convey different storylines and sentiments within their culinary cultures! Being round and sweet, doughnuts, as depicted in Image1, often symbolize indulgence and reward! Because of their round shape, they can also denote eternity and completeness! By comparison, spaghetti and tomato sauce shown in picture 2, hold deep-rooted symbolism in Italian culture! Being central to many meals, spaghetti symbolizes unity and family cohesion for more than 6 centuries! By making pasta from scratch or preparing a family recipe the second dish speaks to continuity and heritage! The tomato sauce signifies prosperity and abundance, often representative of the lush Italian gardens and the healthful Mediterranean diet! Beyond its flavor, tomato sauce symbolizes prosperity, reflecting Italy's lush gardens, as well as a healthy Mediterranean diet! Both pasta and tomato sauce embody Italy's identity and cultural pride with connection to land, tradition, and family!	0.4 (2/5)

Table 8: A chat where Gemini is repeating the same answer.

All instructions	Question	Answer	PIF
Instruction: Start every sentence with the letter (S).	What type of environment is depicted in the Image1?	Swamps are the typical environment in image 1.	1 (1/1)
Instruction: Start every sentence with the letter (S).	Based on the Image1, how are the cattail plants predominantly arranged or positioned in their natural habitat?	Swamps are the typical environment in image 1. Spatially, they appear to be clustered together in dense stands.	1 (1/1)
Instruction: Start every sentence with the letter (S). Instruction: Only use responses to question where each sentence in the response is at least 18 words in all future responses.	Considering the Image1, what practical uses do the cattail leaves serve in various cultures?	Swamps are the typical environment in image 1. Spatially, they appear to be clustered together in dense stands. Serving a multitude of purposes, the leaves of the cattail plant have been ingeniously incorporated into diverse aspects of life in various cultures throughout history. Specifically, the leaves are recognized for their strength and pliability, making them ideal for weaving into mats, baskets, and even temporary shelters, providing both comfort and protection from the elements.	0.5 (1/2)
Instruction: Start every sentence with the letter (S). Instruction: Only use responses to question where each sentence in the response is at least 18 words in all future responses. Instruction: Include at least one even number in the range bigger than 5 in each of your responses.	What type of flowers are depicted in the Image2, and what might be their historical significance in floral arrangements?	Swamps are the typical environment in image 1. Spatially, they appear to be clustered together in dense stands. Serving a multitude of purposes, the leaves of the cattail plant have been ingeniously incorporated into diverse aspects of life in various cultures throughout history. [...]. The exquisite flowers depicted in Image 2 are Cattleya orchids, renowned for their vibrant hues and intricate blooms, captivating hearts for centuries. [...]	0.33 (1/3)

Table 9: Gemini-1.5 Pro response claim the task is impossible, when it's solvable.

All instructions	Question	Gemini Answer	PIF
Instruction: Only use responses to question where each sentence in the response is at most 18 words in all future responses. Instruction: Include at least one even number in the range bigger than 5 in each of your responses. Instruction: Make all the following responses at least 5 sentences. Instruction: Use the word 'per se' at least once in all future responses. Instruction: End every sentence with a question mark (?). Instruction: Start every sentence with the letter (S).	How does the structure in Image2 reflect the importance of the arts within the context of its location?	Sorry, but I cannot fulfill this request. Combining the existing instructions – to end every sentence with a question mark and to begin every sentence with the letter "S" – creates an impossible grammatical structure for a coherent response.	0 (0/6)

Table 10: GPT-4o ignores instructions related to numbers.

All instructions	Question	GPT-4o Answer	PIF
Instruction: Include at least one odd number bigger than 5 in each of your responses.	Describe the patterns and textures you can observe in the Image3.	In the third image, the pattern and texture are indicative of delicate lacework. The lace features intricate floral designs, with a repeated motif showcasing varying flower shapes and sizes. The detailed, openwork design creates a sense of depth and complexity, with elements of symmetry that contribute to the overall aesthetic. This type of lace pattern may require numerous stitches and precision, reflecting skilled craftsmanship.	0 (0/6)
Instruction: Include at least one even number in the range bigger than 5 in each of your responses.	What activity are the individuals in the Image1 engaged in, which is characterized by rhythmic and fast-paced footwork?	In Image 1, the individuals appear to be engaged in a traditional dance involving rhythmic and fast-paced footwork, likely Tinikling, a popular folk dance from the Philippines that involves stepping deftly between bamboo poles. The dance is known for its vibrant and energetic movements, often done in groups of 2 or more participants.	0 (0/1)
Instruction: Include at least one odd number bigger than 5 in each of your responses.	What type of environment is depicted in Image1?	In Image 1, the environment depicted appears to be a natural setting with tall grassy vegetation, which suggests that this environment is likely a forested or grassland area.	0 (0/1)

Annotation Guidelines

For each sample (total 71), several images (up to 5) are shown and multiple questions and instructions (average 14 questions and 6 instructions be sample) related to the images are asked based on this. Answers from several models are provided for each question.

Part 1: Each question has a reference answer that should be rewritten to ensure it follows the given instructions in addition to correctly answering the question about the images. All given instructions in the chat so far will be given for convenience. True facts can be added in order to fulfill the instructions. You will have access to answers from the three models when rewriting, which can be used for inspiration.

Part 2: For each of the three models and for each question, individual ratings for instruction following and answer accuracy should also be evaluated on a scale from 1 to 10 (see next paragraph).

For each question, 3 model answers will be compared, Gemini 1.5 Pro, GPT-4o, and Sonnet 3.5. The second part is to provide a side-by-side preference ranking between Gemini 1.5 Pro and GPT-4o, Gemini and Sonnet, as well as between Sonnet and GPT-4o. For each comparison, write down the winner model's name ("Gemini", "GTP-4o", or "Sonnet"), or "tie" (use sparingly, only when strictly necessary). The comparison should be based on the similarity to the rewritten reference answer from Part 1.

Metrics

Answer Accuracy

- Scores 1-2 when the answer is significantly inconsistent with the question or contains obvious errors.
- Scores 3-4 when the answer is partially correct but contains some errors or is incomplete, significantly worse accuracy compared to the rewritten reference answer.
- Scores 5-6 when the answer is basically correct but lacks details or is not sufficiently detailed, the accuracy is worse than the reference answer.
- Scores 7-8 when the answer is accurate and detailed, fully corresponding to the question, on par with the reference answer.
- Scores 9-10 when the answer is not only accurate and detailed but also provides additional useful information, exceeding the rewritten reference answer.

Instruction Following

- Scores 1-2 when the answer is completely ignoring most or all of the previously given instructions.
- Scores 3-4 when several of the instructions are followed but some are not followed, significantly worse than the rewritten reference answer.
- Scores 5-6 when most of the instructions are correctly followed, but there are some errors, worse quality than the rewritten reference answer.
- Scores 7-8 when all instructions except perhaps 1 is followed in a good way, on par with the rewritten reference answer.
- Scores 9-10 when all instructions are followed in a clear and insightful way, exceeding the rewritten reference answer.

Figure 16: instructions (lightly edited) given to human annotators when creating reference answers, model scores, and pairwise preferences for the MMMT-IF benchmark.