Appendix

A License and Intended Use

The VisIT-Bench dataset, along with its various contributions such as instructions, reference outputs, and model ranking annotations, is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). This license applies to all the images we have directly contributed, each of which carries a public license specification in the "public images metadata" field within the dataset sheets. However, the dataset also incorporates images sourced from pre-existing collections. For these images, the original licensing terms are respected and remain applicable.

VisIT-Bench's primary purpose is to function as a dynamic benchmark that continuously evolves and evaluates instruction-following vision-language models. In the current landscape, commercial chatbots are often trained on non-disclosed and non-public datasets, which raises concerns about potential data contamination and inadvertent training on our evaluation data [8]. This risk is further highlighted by recent studies [55, 56]. To mitigate such concerns, we have chosen to withhold the complete VisIT-Bench test set from public disclosure, while still making the images and instructions available for direct download. Researchers, however, can utilize VisIT-Bench to its full potential as a dynamic benchmark by submitting their model predictions for evaluation. We will assess their models using the undisclosed test set, ensuring the ongoing evolution of the benchmark. Moreover, we are open to releasing the test data upon receiving reasonable and justified requests, particularly when additional analysis is necessary, provided that requesters agree to our non-contamination policy which prohibits the use of this data for training commercial chatbots. This approach strikes a balance between the need for robust model evaluation and the mitigation of potential data contamination.

A Dataset Analysis

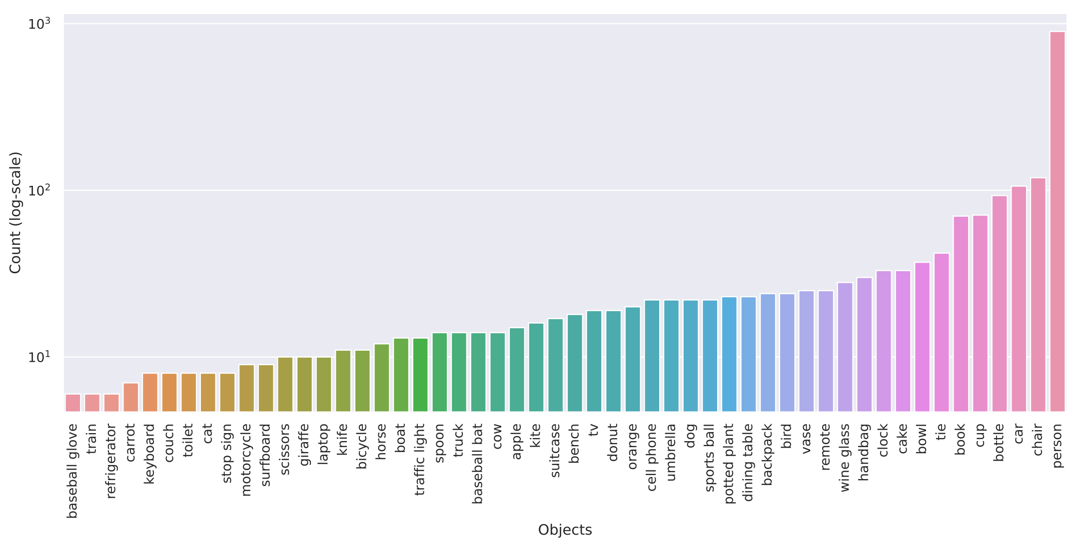

Figure 9: Count of the various COCO objects present in the VisIT-Bench images on a log-scale. The object detection was performed by Yolov5-Large [39]. We observe that the object 'person' occurs most of the time in the dataset images i.e., 875 times.

Figure 10: Most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in the VisIT-Bench instructions.

B Related Work

Multimodal Models for Image-Text Understanding: Recently, the field of machine learning has experienced a rapid proliferation of new models which can perform various image-text tasks [12, 15, 13, 50, 18, 14]. This growth has been driven by several factors, including the emergence of large-scale multimodal datasets (e.g. LAION-5B [57], Multimodal C4 [11]), improved software and hardware frameworks, and advances in modality-specific models such as language models (e.g., [10]). Our work specifically evaluates models which can generate textual outputs, given one or more images, and text. Recent examples of such models include LLaVA [13], mPLUG-Owl [17], InstructBLIP, LLaMA-Adapter, Flamingo [12] and OpenFlamingo [11], PandaGPT [18], and GPT-4 [7] (which reports multimodal capabilities but has not yet seen a release of the multimodal variant).

Instruction Following: "Instruction-following" is an emerging paradigm for training models via language, where instead of being trained to complete only a single, fixed task (such as image classification or captioning), models are trained to follow textual instructions that describe an arbitrary task, with the aim of generalizing to novel instructions. Examples of instruction-following models include Alpaca [5], LLaMA-Adapter [16], Koala [51], InstructBLIP [14], LLaVA [13], and mPLUG-owl [17]. As the downstream capabilities of these models are influenced by the quality of the training dataset, there has also been extensive work on developing instruction-following datasets [38, 58, 59, 13, 60] .

To build powerful these models, two broad approaches have been shown to be effective. One approach focuses on leveraging existing pretrained task-specific tools such as image captioners [15], object detectors [61] and text-to-image generators [62] by either creating multimodal prompt interfaces [42, 63] or by executing LLM-generated programs [64, 65, 66]. The other approach [13, 16, 67, 45, 68, 17, 11] focuses on building a single pretrained model that can follow instructions by supervised finetuning on multimodal vision-language data.

Despite the success of both these approaches on the existing vision-language datasets e.g., VQA, GQA, Image Captioning [21, 52, 20], there is a lack of a high-quality benchmarking dataset for multimodal instruction-following tasks that reliably replicates the way in which humans would interact with multimodal chatbots in the wild. Similar to the image-text models discussed above, many instruction-following models have been released directly as open-source without undergoing peer review or thorough evaluation. As a result, the effectiveness of these models for many tasks is not well-understood.

Benchmarks for Machine Learning: High-quality evaluation datasets have served both to (re)assess, and to accelerate, progress on many machine learning tasks [53]. For example, our work draws particularly from the fields of computer vision and natural language processing, where benchmarking datasets have been critical drivers of progress. On the vision side, datasets such as ImageNet [69] and CIFAR [70] have proven to be critical yardsticks of progress. On the language side, benchmarks such as SQuAD [71], SST [72], GLUE/SuperGLUE [73, 74] and more [75, 76] seen wide use. Recent work has indicated that improvements on these high-quality benchmark datasets is not the result of overfitting, and is a reliable indicator of genuine progress beyond the benchmark data [77, 78, 79, 80].

However, high-quality benchmarking datasets and evaluation methods do not yet exist for multimodal instruction-following. As a result, it is difficult to assess progress in this direction, which both reduces the field's ability to identify true breakthroughs and increases vulnerability to potential pitfalls of evaluation that have hampered progress in other areas of machine learning [53, 54].

C Interfaces for Collecting Human Annotations

In this section, we provide the templates we used to collect human annotations for the instruction generation (Figure 11), the dense caption generation (Figure 12), the model verification (Figure 13 and Figure 14), and the model rating (Figure 15).

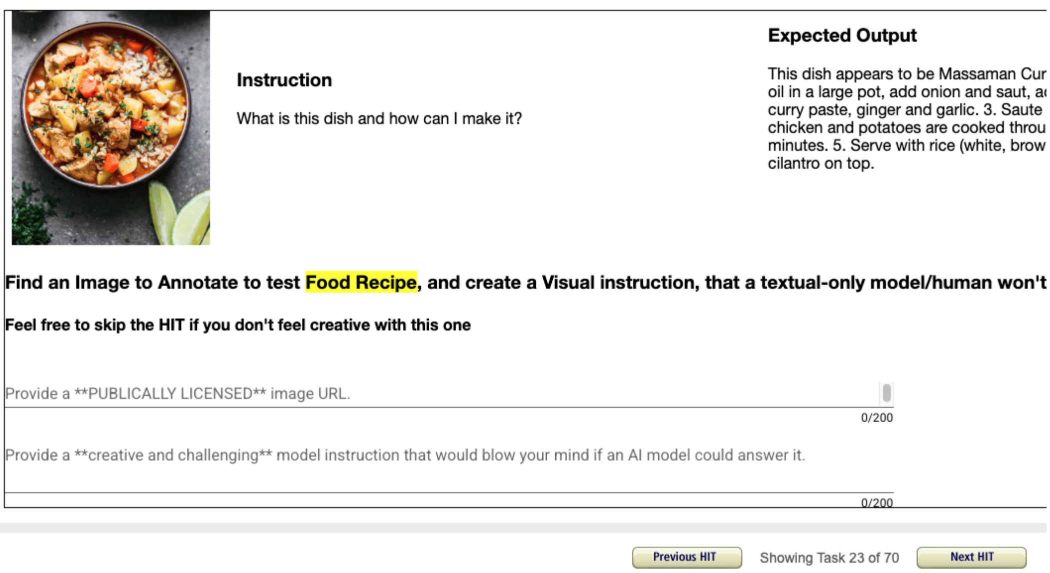

Figure 11: An interface that collects the image and corresponding instruction.

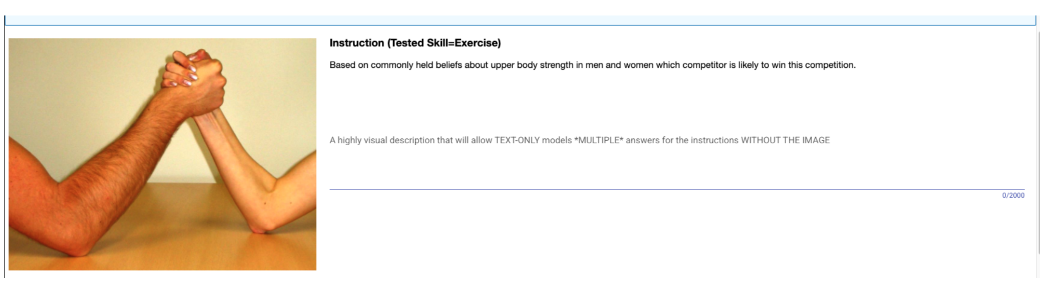

Figure 12: An interface that collects the dense caption given an image-and-instruction pair.

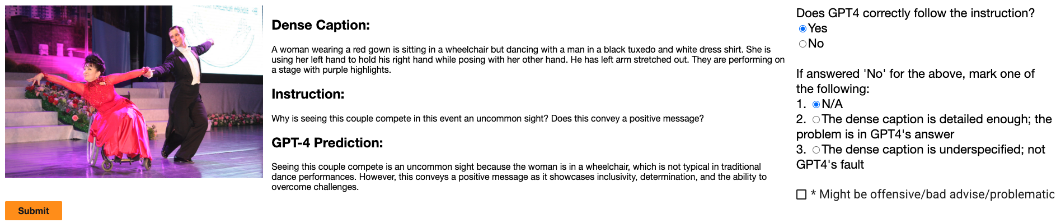

Figure 13: A sample of the model verification where the GPT4 follows the instruction correctly.

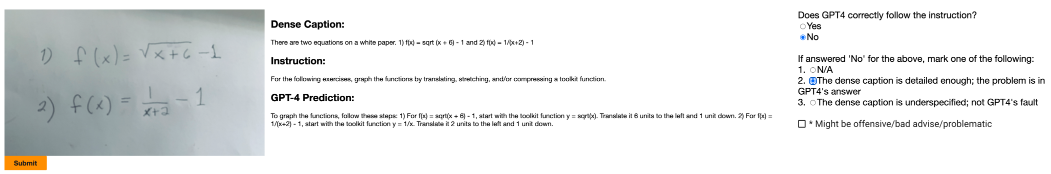

Figure 14: A sample of the model verification where the GPT4 failed to follow the instruction due to its incorrect reasoning.

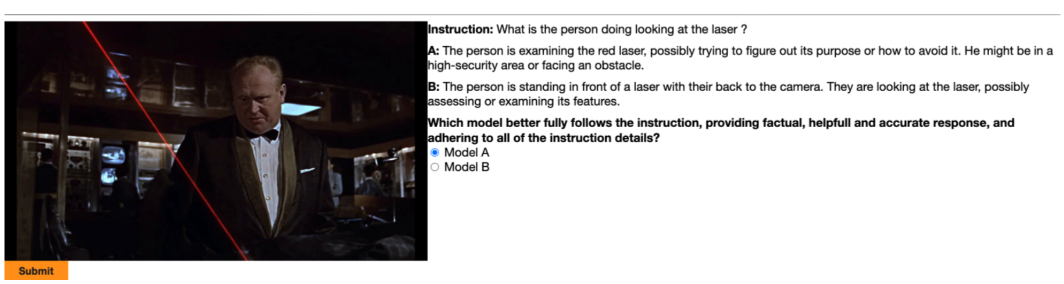

Figure 15: An interface that collects the feedback of the model rating.

D Existing Datasets incorporated in VisIT-Bench

In Table 5, we listed the existing datasets that are incoprated in our VisIT-Bench. Among these datasets, 15 contain a single image in each sample pair, and 10 require reasoning based on multiple images.

Table 5: List of existing datasets in VisIT-Bench, categorized as single and multiple image datasets.

	Dataset	Topic
Single	VQA [21]	Visual Question Answering
	VCR [81]	Cognition-level Visual Understanding
	GD-VCR [82]	Geo-Diverse Commonsense Reasoning
	WHOOPS [40]	What Makes this Image Strange
	Newyork Caption [83]	Humor Understanding
	CLEVR [84]	Visual Question Answering
	Kilogram [85]	Tangrams Identification
	Harmful Memes [86]	Memes Understanding
	ScienceQA [87]	Science Question Answering
	OK-VQA [88]	Outside Knowledge Visual Question Answering
	AOK-VQA [89]	Outside Knowledge Visual Question
	AOK-VQA [89]	Question Generation
	VizWiz [90]	Visual Question Answering
	GQA [52]	Visual Question Answering on Scene Graphs
	TextCaps [91]	Visual Question Answering on Texts
Multiple	Robust Change Captioning [22]	Describing What has Change in a Scene
	NLVR2 [36]	Testing Visual Language Bias
	ImageCoDE [92]	Image Retrieval
	Spot-the-Diff [93]	Identifying Differences
	VASR [94]	Visual Analogies
	WinoGavil [95]	Visual Associations
	IRFL (Metaphor) [96]	Figurative Speech Understanding
	IRFL (Idioms) [96]	Figurative Speech Understanding
	IconQA [97]	Abstract Diagram Understanding
	Pick-a-Pic[98]	Text-to-Image User Preferences

Table 6: List of skills and existing datasets in VisIT-Bench

‘scienceqa’, ‘ocr math’, ‘recognition’, ‘okvqa’, ‘house plan understanding’, ‘nlvr2’, ‘gardening tips’, ‘textcaps’, ‘architectural styles’, ‘dressing sense’, ‘winoground’, ‘food recipe’, ‘paper folding’, ‘whoops’, ‘spot the diff’, ‘wino-gavil’, ‘imagecode’, ‘exercise’, ‘art knowledge’, ‘gqa’, ‘physical knowledge’, ‘contextual knowledge of events’, ‘home renovation’, ‘aokvqa’, ‘animals’, ‘vasr’, ‘counting’, ‘board games’, ‘solving geometry problems’, ‘who to call?’, ‘clevr’, ‘building materials’, ‘hazard identification’, ‘pickapick’, ‘astronomy’, ‘figurative speech explanation’, ‘write a story’, ‘gestures understanding’, ‘newyork’, ‘cultural knowledge’, ‘aokvqg’, ‘traffic sign identification’, ‘pop culture’, ‘fashion products’, ‘harmful memes’, ‘write a poem’, ‘vizwiz’, ‘guesstimate of capacity’, ‘location understanding’, ‘graph reasoning’, ‘vqa’, ‘game playing’, ‘differently abled’, ‘chemical identification’, ‘history knowledge’, ‘climate and weather understanding’, ‘irfl metaphor’, ‘human emotion recognition’, ‘medical’, ‘gd vcr’, ‘vcr’, ‘technical support’, ‘catchy titles’, ‘kilogram’, ‘anagrams’, ‘color’, ‘tour guide’, ‘directions’, ‘irfl idiom’, ‘rcc’

E Elo Rating

For many years, the Elo rating has been popular in ranking players in zero-sum games such as chess [25]. Recently, it has been adopted to rate large language models (LLMs) against each other on the user instructions. In this work, we adopt the same strategy to rank a set of instruction-following vision-language models, that can grow dynamically with further advances in the field.

Given two multimodal chatbots $\mathcal{C}_a$ and $\mathcal{C}_b$ with their absolute Elo rating $\mathcal{R}_a$ and $\mathcal{R}_b$, respectively. Simply put, the probability of $\mathcal{C}_a$ winning over $\mathcal{C}_b$ in a head-to-head battle is given by:

$$P(\mathcal{C}_a \text{ wins over } \mathcal{C}_b) = \frac{1}{1 + 10^{(\mathcal{R}_a - \mathcal{R}_b)/400}} \tag{1}$$

In practice, calculating the Elo rating requires us to set hyperparameters to decide the weightage for each win and loss in a head-to-head battle between two models. In our work, we use the open implementation of Elo for LLMs by FastChat at `https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/elo_analysis.py`.

F GPT-4 Pairwise Evaluation Prompts

The specific prompts we use to extract pairwise judgements from our language model are provided in Table 16 (reference-free version) and Table 17 (reference-backed version). When applied to GPT-4 [7], these prompts usually solicit a definitive pairwise response by the model. But, in some cases, the model either produces a pairwise judgement in an unexpected format, or, refuses to issue a judgement at all. For cases like these, we issue an additional query to ChatGPT to extract an answer (or decide there is no answer) using an additional prompt, given in Table 18. If after this step there is still no definitive pairwise judgment, we call the result a tie.

[UGFt]A head-to-head comparison involves querying GPT-4 two times. Each query has approximately 800 input/250 output tokens (output includes chain of thought; see Supplementary for full prompt/response example). Running a head-to-head comparison between a new model and a randomly sampled existing model across the 700 instances is usually sufficient for a stable Elo estimate. In total, at current GPT-4 prices, the 700 head-to-head comparisons required to assess a new model costs 50-70 dollars.

`system prompt` (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. An instruction: This is a question, an imperative request, or something similar about the image which requires a response.
3. Two responses, response A and response B: These two responses attempt to address the instruction in the context of the image.
Your job is to judge whether response A or response B better. A and B are randomly ordered.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency, specificity, etc.
- You think step-by-step, but ultimately respond with "Response A" or "Response B"

`user` (human authored)

I will describe an image to you, and provide an instruction. Then, I will give you two candidate responses that address the instruction in the context of the image: these will be labelled "Response A" and "Response B". Your job is to first reason step-by-step about which response is best in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of each response, in the end, respond with "Overall, Response X is better." where X is either A or B.

`assistant` (human authored)

Sure, please provide the image context, the instruction, and the two candidate responses, Response A and Response B. Then, I will think step-by-step and provide my ultimate judgement as to which response is better.

`user` (from human-authored template)

OK. Here is the image description, the instruction, and the two response options, Response A and Response B. Image context: A skier is wearing a green jacket, burgundy pants, white shoes, white gloves, and a gray beanie. There are people sitting on the background with their snowboards.

Instruction: Answer the following question based on the provided image: Are the trousers the same color as the shoes?

Response A: no

Response B: No, the trousers are not the same color as the shoes. The trousers are burgundy, while the shoes are white.

Think step-by-step and finish your response with "Overall, Response X is better." where X is either A or B.

`assistant` (machine generated; used to derive the prediction)

Step 1: Identify the key information from the image description.

The skier is wearing:
- Burgundy pants
- White shoes

Step 2: Assess the accuracy of each response in addressing the instruction.

Both Response A and Response B convey that the trousers and shoes are not of the same color.

Step 3: Compare specificity and detail of each response.

Response A is concise and provides a straightforward answer, but lacks detail.
Response B provides a clear, comprehensive answer that includes the color of both the trousers and shoes.

Step 4: Consider the fluency and correctness.

Both responses are fluent and grammatically correct.

Overall, Response B is better.

Figure 16: The prompt we use for reference-free evaluation, along with a sample instance/completion from GPT-4 [7].

`system prompt` (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. An instruction: This is a question, an imparative request, or something similar about the image which requires a response.
3. A reference output: This is a high-quality example output that humans have judged to be an accurate response for the input instruction.
4. Two responses, response A and response B: These two responses attempt to address the instruction in the context of the image.

Your job is to judge whether response A or response B better. A and B are randomly ordered. Some things to remember:
- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- Take the high-quality reference into account when making your judgements, but remember: some instructions are more open-ended than others, so for those cases, a high quality response can differ from the reference.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency, specificity, etc.
- You think step-by-step, but ultimately respond with "Response A" or "Response B"

`user` (human authored)

I will describe an image to you, and provide an instruction. Then, I will provide a reference output which is an example of a high quality output for that instruction in the context of the image. Then, I will give you two candidate responses that address the instruction in the context of the image: these will be labelled "Response A" and "Response B". Your job is to first reason step-by-step about which response is best in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of each response, in the end, respond with "Overall, Response X is better." where X is either A or B.

`assistant` (human authored)

Sure, please provide the image context, the instruction, the reference, and the two candidate responses, Response A and Response B. Then, I will think step-by-step and provide my ultimate judgement as to which response is better.

`user` (from human-authored template)

OK. Here is the image description, the instruction, the high-quality reference, and the two response options, Response A and Response B.
Image context: There is a white plate that has on it a small pizza with many topics, mostly unrecognizable, due to the pie being very well done, charred to the point of almost being burnt, and the crust blackened in many places.

Instruction: Someone told the creator of this item that they are very talented. What is the implied meaning?
High quality reference: The implied meaning is that the person complimenting the creator is being sarcastic, as the pizza is overcooked and almost burnt, which is not a sign of talent.

Response A: When someone says that they are "very talented" in the context of the image, it implies that the person responsible for creating the pizza is skilled in the craft and able to create an appealing and tasty dish. The person praising the pizza's creator recognizes their culinary skills, which can be seen through the combination of ingredients and the overall look of the finished dish. The praise is a form of appreciation for their efforts and the enjoyable result they achieved.
Response B: It is difficult to determine the implied meaning without further context or information about the image. It appears to be a pizza that has been burnt or charred on the surface and has various toppings, including pepperoni, onion, and mushrooms. The image is shown on a white plate or surface. The implied meaning of this image could be that it is a dish that has been cooked badly or that it is a dish with a unique or unusual combination of ingredients.

Think step-by-step and finish your response with "Overall, Response X is better." where X is either A or B.

`assistant` (machine generated; used to derive the prediction)

Step 1: Comparing both responses with the high-quality reference, it's clear that Response A describes the pizza as well-made and appealing, while Response B is more focused on the fact that it is burnt or charred, somewhat aligning it with the reference.
Step 2: Evaluating specificity, Response A defines the creator as skilled, even though the image description contradicts this by stating that the pizza is overcooked. Response B, while not as specific, acknowledges the overcooked nature of the pizza.
Step 3: In terms of fluency, both responses are easily understood and written in coherent language.
Step 4: Considering accuracy, Response A is inaccurate in relation to the context and instruction because it implies a positive meaning when it should be sarcastic. Response B doesn't fully capture the sarcasm but is closer to the truth about the pizza being badly cooked.
Step 5: Understanding relevance, Response A is not entirely relevant due to the inaccuracy in the creator's talent, while Response B, although less direct, is more relevant by mentioning the burnt or charred nature of the pizza.
Overall, Response B is better.

Figure 17: The prompt we use for reference-backed evaluation, along with a sample instance/completion from GPT-4 [7].

`system prompt` (human authored)

You are FinalAnswerExtractionGPT, an expert language model at extracting multiple choice answers from written out from longer explanations. You will be given several sentences describing a thought process which should eventually contain a final answer, either A or B. Your job is to extract the final answer that the explanation arrives at.

Some things to remember:

- Keep your answers short: only output "Final Answer: X" where X is A or B
- If the explanation doesn't mention a final choice explicitly, you can output "Unknown"

...
(`+ 2 in-context examples` (human authored))
...

Example One: GPT-4 gives a definitive answer, but in the wrong format

`user` (GPT-4's raw pairwise judgement)

Step 1: Compare the two responses to the instruction
- The instruction asks if people are required to stop and requests an explanation.

Step 2: Assess Response A
- Response A indicates that people should stop due to the presence of a stop sign. However, it doesn't consider the fact that the sign is pressed against the bus, and the LED lights are off.

Step 3: Assess Response B
- Response B simply says "yes," acknowledging that people should stop, but it doesn't provide the explanation required by the instruction.

Step 4: Compare specificity, fluency, and other factors
- Response A is more specific and fluent but misconstrues the image description.
- Response B is accurate in that it acknowledges the stop sign but lacks an explanation.

Overall, since neither response is fully correct, it is difficult to declare a better response. However, given the fact that Response B acknowledges the prominent stop sign, I would say that Response B is slightly better, but it should also explain how the situation indicates that people should stop.
Please extract the final answer from the above text.

`assistant` (machine generated; used to derive the prediction)

Final Answer: Response B is slightly better, but it should also explain how the situation indicates that people should stop.

Example Two: GPT-4 refuses to answer.

`user` (GPT-4's raw pairwise judgement)

Step 1: Consider the image context, which states that the original thirteen colonies are shown with New Hampshire specifically highlighted in a darker green.

Step 2: Assess the accuracy of Response A, which claims that the colony shown is "Virginia." This is not accurate since the image context indicates New Hampshire being highlighted.

Step 3: Assess the accuracy of Response B, which claims that the colony shown is the Massachusetts Bay Colony. This response is also not accurate, as the image context states New Hampshire is the highlighted colony.

Step 4: In terms of accuracy, both responses are incorrect. Neither response is consistent with the image context.

Step 5: Consider other factors like fluency and specificity. Both responses seem fluent and specific in nature.

After considering these factors, none of the responses are accurate given the information provided in the image context. However, I am unable to provide an "Overall, Response X is better." judgement in this case as both responses are incorrect and neither stands out as being better than the other.

Please extract the final answer from the above text.

`assistant` (machine generated; used to derive the prediction)

Final Answer: Unknown.

Figure 18: The prompt we use to extract answers from ill-formatted GPT-4 [7] responses, along with two example outputs. In one case, GPT-4's pairwise judgement is given, but in a format different than requested. In the other case, GPT-4 "defies" the prompt, and refuses to issue a judgement because both options are bad.