

Woodpecker: Hallucination Correction for Multimodal Large Language Models

Shukang Yin^{1†}, Chaoyou Fu^{*†}, Sirui Zhao^{1*†}, Tong Xu^{1*}, Hao Wang¹,
Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun & Enhong Chen^{1*}

¹*School of Data Science, State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China, Hefei 230026, China*

Appendix A Prompt Templates

In this part, we list our prompt templates for instructing LLM to fulfill various tasks, including key concept extraction, question formulation, hallucination correction, and GPT-4V-aided evaluation.

Appendix A.1 Key Concept Extraction

The template is listed in Table A1.

<p>System message</p> <p>You are a language assistant that helps to extract information from given sentences.</p> <hr/> <p>Prompt</p> <p>Given a sentence, extract the existent entities within the sentence for me.</p> <p>Extract the common objects and summarize them as general categories without repetition, merge essentially similar objects.</p> <p>Avoid extracting abstract or non-specific entities. Only extract concrete, certainly existent objects that fall in general categories and are described in a certain tone in the sentence.</p> <p>Extract entity in the singular form. Output all the extracted types of items in one line and separate each object type with a period. If there is nothing to output, then output a single "None".</p> <p>Examples:</p> <p>{In-context examples}</p> <p>Sentence:</p> <p>{Input sentence}</p> <p>Output:</p>

Table A1 Template for prompting LLM to perform key concept extraction. {In-context examples} are in-context examples used to better instruct the LLM to fulfill the task, and {Input sentence} is the input from which the key concept are extracted.

Appendix A.2 Question Formulation

The template is listed in Table A2.

* Corresponding author (email: bradyfu24@gmail.com, sirui@mail.ustc.edu.cn, tongxu@ustc.edu.cn, cheneh@ustc.edu.cn)

†† Shukang Yin, Chaoyou Fu and Sirui Zhao have the same contribution to this work.

System message

You are a language assistant that helps to ask questions about a sentence.

Prompt

Given a sentence, extract the existent entities within the sentence for me.

Given a sentence and some entities connected by periods, you are required to ask some relevant questions about the specified entities involved in the sentence, so that the questions can help to verify the factuality of the sentence.

Questions may involve basic attributes such as colors and actions mentioned in the sentence. Do not ask questions involving object counts or the existence of objects.

When asking questions about attributes, try to ask simple questions that only involve one entity.

Ask questions that can be easily decided visually. Do not ask questions that require complex reasoning.

Do not ask semantically similar questions. Do not ask questions only about scenes or places.

Use “where” type questions to query the position information of the involved entities.

Do not ask questions about uncertain or conjecture parts of the sentence, for example, the parts described with “maybe” or “likely”, etc.

It is no need to cover all the specified entities. If there is no question to ask, simply output a “None”.

When asking questions, do not assume the claims in the description as true in advance. Only ask questions relevant to the information in the sentence.

Only ask questions about common, specific and concrete entities. The entities involved in the questions are limited to the range within the given entities.

Output only one question in each line. For each line, first output the question, then a single “&”, and finally entities involved in the question, still connected by periods if multiple entities are involved.

Examples:

{In-context examples}

Sentence:

{Input sentence}

Entities:

{Input entities}

Questions:

Table A2 Prompt template for question formulation. {In-context examples} are in-context examples. {Input sentence} and {Input entities} are the inputs, where the latter comes from the step of key concept extraction.

Appendix A.3 Hallucination Correction

The template is listed in Table A3.

Appendix A.4 GPT-4V-aided Evaluation

The template is listed in Table A4.

Appendix B GPT-4V-aided Evaluation Case

To offer a straightforward and intuitive understanding, we list some evaluation cases in Fig. B1, Fig. B2, and Fig. B3, where “Assistant 1” and “Assistant 2” in the evaluation reason part correspond to “MLLM” and “MLLM w/Woodpecker”,

System message

You are a language assistant that helps to refine a passage according to instructions.

Prompt

Given a passage and some supplementary information, you are required to correct and output the refined passage in a fluent and natural style, following these rules:

1. The supplementary information may include some of the following parts:

“Count” information that specifies how many instances of a certain kind of entity exist, and their associated bounding boxes;

“Specific” information that describes attribute information specific to each entity instance, including bounding boxes, colors, etc. The information is arranged in the form of “entity 1: [bbox]” info of this entity. Note that the entity in “Specific” information corresponds to that in the “Count” information.

“Overall” information that may involve information about multiple entity objects.

2. Try to retain the original sentence with minimal changes.

3. The number of entitie instances should match the number in the “Count” information. Also correct the number counts if the number stated in the original sentence does not match the counting information.

4. If the original sentence is already correct, then just keep it. If you need to rewrite the original sentence, when rewriting, try to modify the original sentence as little as possible based on the original sentence, and use the supplementary information as guidance to correct or enrich the original sentence.

5. In the refined passage, when describing entities mentioned in the “Specific” supplementary information, add their associated bounding boxes in parentheses right after them, in the form of “entity([bbox])”. If multiple entities of the same kind are mentioned, then separate the box with “;”, in the form of “entity([bbox1];[bbox2])”

Examples:

{In-context examples}

Supplementary information:

{Input information}

Passage:

{Input passage}

Refined passage:

Table A3 Prompt template for hallucination correction. {In-context examples} are in-context examples. {Input information} is the formatted knowledge base, and {Input passage} is the original response to be corrected.

respectively. GPT-4V gives not only respective scores for responses but also reasons for the judgment.

Appendix C More examples

In this part, we list some examples and make a corresponding qualitative analysis. As shown in Fig. C1, our method formulates diverse questions and covers various aspects, ranging from existence (Is there a cup on the table?), attribute (What color is the plate?) to positional relationships (What is on the left side of the plate?), compositional relationships (What is the riding?), and background information besides foreground objects (What can be seen in the background?).

Prompt

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Note that the descriptions may be accompanied by bounding boxes, indicating the position of objects in the image, which are represented as $[x1, y1, x2, y2]$ with floating numbers ranging from 0 to 1. These values correspond to the top left $x1$, top left $y1$, bottom right $x2$, and bottom right $y2$.

Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:

1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.

2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.

Please output a single line for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

[Assistant 1]

{Response 1}

[End of Assistant 1]

[Assistant 2]

{Response 2}

[End of Assistant 2]

Output format:

Accuracy:

Scores of the two answers:

Reason:

Detailedness:

Scores of the two answers:

Reason:

Table A4 Prompt template for GPT-4V-aided evaluation. {Response 1} and {Response 2} are the original responses and the corrected ones, respectively.

GPT-4V-Aided Evaluation



MLLM:

A plate with a sandwich, broccoli, and cheese on it, along with a fork on the side.

MLLM w/ Woodpecker:

The key elements in this picture are a plate([0.001, 0.014, 0.998, 0.999]) with a sandwich([0.429, 0.148, 0.835, 0.858]) and broccoli([0.182, 0.286, 0.381, 0.543];[0.021, 0.342, 0.091, 0.434];[0.374, 0.353, 0.438, 0.463];[0.042, 0.364, 0.164, 0.629];[0.134, 0.428, 0.279, 0.66];[0.114, 0.145, 0.241, 0.272];[0.217, 0.139, 0.459, 0.357]) on it, along with a fork([0.0, 0.658, 0.58, 0.873]) on the side.

GPT Score:

Accuracy:

79

Reason:

Both assistants correctly identified the main elements in the image: a plate, sandwich, and broccoli. However, Assistant 1 described there being cheese, which seems to be a hallucination as there's no distinct evidence of cheese outside of the broccoli preparation. Assistant 2 didn't mention cheese separately and also provided bounding boxes which seem accurate, indicating specific positions of the objects in the image.

Detailedness:

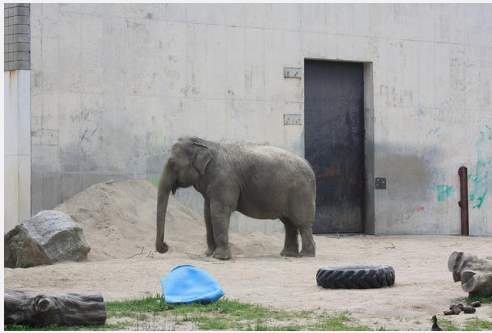
610

Reason:

While Assistant 1 provided a general overview of the image, Assistant 2 was more detailed in its response, specifying the locations of the plate, sandwich, broccoli, and fork using bounding boxes. The additional details provided by Assistant 2, such as multiple bounding boxes for different broccoli pieces, add richness to the description.

Figure B1 The first example for the GPT-4V-aided Evaluation.

GPT-4V-aided Evaluation



MLLM:

An elephant is walking on a dirt ground with a blue bag and a tire nearby.

MLLM w/Woodpecker:

In this snapshot, an elephant([0.306, 0.4, 0.636, 0.783]) is walking on a dirt ground. There is a blue bag([0.319, 0.798, 0.449, 0.919]) nearby the elephant, and there are two tires([0.634, 0.797, 0.795, 0.873];[0.0, 0.874, 0.21, 0.987]) nearby as well.

GPT Score:

Accuracy:

8 10

Reason:

Both assistants have accurately described the presence of an elephant on a dirt ground with a blue bag and tires nearby. However, Assistant 1 mentioned a single tire, while the image shows two tires. Assistant 2 correctly identified the number of tires and their positions.

Detailedness:

6 9

Reason:

While both descriptions provide necessary details about the elephant, blue bag, and tires, Assistant 2 goes a step further by providing bounding boxes for the elephant, blue bag, and both tires, offering a more detailed description of the image content. Assistant 1's description is less detailed in comparison.

Figure B2 Example for the GPT-4V-aided evaluation.

GPT-4V-Aided Evaluation



MLLM:

The image shows a construction site with a wooden structure being built. The room appears to be in the process of being framed, with wooden studs and insulation visible. There are several pieces of furniture, such as a bed and a couch, placed in the room. Additionally, there are two people present in the scene, one near the left edge and the other near the right edge, possibly working on the construction or observing the progress.

MLLM w/ Woodpecker:

The photo is about a construction site([0.007, 0.011, 0.993, 0.987]) with a wooden structure([0.716, 0.008, 0.996, 0.769]) being built. The room appears to be in the process of being framed, with insulation([0.005, 0.531, 0.545, 0.996]) visible.

GPT Score:

Accuracy:

49

Reason:

Assistant 1 mentioned furniture such as a bed and a couch, which are not visible in the image. It also described two people being present in the scene, but there are no people in the provided image. Therefore, the response contains several hallucinations. Assistant 2 provided a more accurate description, aligning with the image content. It also added bounding boxes to indicate specific portions of the image, improving its accuracy score.

Detailedness:

67

Reason:

Assistant 1 provided a detailed description, but some of the details were hallucinations, like the furniture and the two people. Despite this, it described the setting more elaborately as a construction site and the room being framed. Assistant 2, while accurate, was less detailed in its description. It did not provide as much context as Assistant 1, but it did use bounding boxes to indicate portions of the image.

Figure B3 The third example for the GPT-4V-aided Evaluation.



Model output:

The image displays a delicious and well-prepared meal served on a dining table with an orange plate. The sandwich on the plate is filled with meat and is accompanied by a bowl of broccoli, which complements the meal. The plate is situated in the middle of the table, with the broccoli bowl on its left side and the sandwich occupying the majority of the plate space. A fork and a knife are placed on the table, ready for the diner to enjoy their meal. The table also features a cup, which might hold a drink or a side, although it is not clear in the image. Overall, the image portrays a delightful and appetizing dining experience.

Questions:

- Q: What color is the plate?
- Q: What is serving on the dining table?
- Q: What is on the plate?
- Q: What does the sandwich contain?

- Q: What is accompanied by the bowl of broccoli?
- Q: What is on the left side of the plate?
- Q: What occupies the majority of the plate space?
- Q: Is there a cup on the table?

(a)



Model output:

The image features a young man in a black shirt and jeans riding his skateboard on a city street. He is wearing a backpack and appears to be performing a trick. The scene is lively with several other people in the background, some of whom are closer to the foreground while others are more distant. Several cars can be seen parked or driving down the street, and there is a traffic light in the scene, indicating that the area is a bustling urban area. In addition, there are a few traffic cones scattered throughout the scene.

Questions:

- Q: What color is the man's shirt?
- Q: What is the man riding?
- Q: What is the man wearing?
- Q: What can be seen in the background?

- Q: Where is the man riding his skateboard?
- Q: Is the man performing a trick?
- Q: Who is closer to the foreground?
- Q: Who is more distant?

(b)

Figure C1 Qualitative examples of our methods, showing the questions asked in the Question Formulation stage.