

Image



Query

What's the model of the white car with the spare tire in the front right?

Reference

To the right front is a white Toyota Prado SUV.

LLM's Answer

The white car with the spare tire in the front right is a Toyota Land Cruiser Prado.

1. Task Instruction

You are an assistant skilled at evaluating text quality. Please act as an impartial judge and assess the quality of an AI assistant's response to a user's query. Since the primary category of the responses you are evaluating is **Vehicle Model Recognition** and the secondary category is {sub_category}, you need to evaluate the responses from the following dimensions: {dimensions}.

2. Type-specific evaluation dimensions

- **"Vehicle Model Recognition"**: [{"Factuality", 3}, {"User Satisfaction", 3}, {"Visual Location", 3}, {"Clarity", 1}, {"Completeness", 2}],
- **"Object Recognition"**: [{"Factuality", 3}, {"User Satisfaction", 3}, {"Visual Location", 3}, {"Clarity", 1}, {"Completeness", 2}],

3. Evaluation Procedure

- **Safety Check**: Verify that the AI assistant's response is safe and harmless. If the response is found to be unsafe or harmful in any way, it will be given a score of 0.
- **Positional Accuracy Check**: If the query involves specific positions or orientations, check whether the AI's answer correctly describes the location or object. Incorrect descriptions will result in a lower score.
- **Answer Comparison**: Compare the AI's response to a reference answer. Identify any discrepancies and provide explanations for the shortcomings.
- **Multi-Dimensional Scoring**: Assess the response across various dimensions. Each dimension will be scored on a scale of 1-10.
- **Overall Score Calculation**: Combine the scores from different dimensions, applying a weighted average based on the importance of each dimension, to compute the overall score.
- **Strict Score Adjustment**: Adjust the overall score according to specific rules to ensure a stringent evaluation. This step is critical to maintain the integrity and accuracy of the scoring process.

4. Grading Rules

- | | |
|--------------------------------|----------------------------------|
| 1-2: irrelevant or fatal error | 3-4: minor error but low quality |
| 5-6: medium quality | 7-8: comparable with reference |
| 9-10: better than reference | |

Evaluation and Explanation

1. ****Factual Accuracy****:

- The AI assistant's response matches the reference answer, identifying the white car in the right front as a Toyota Land Cruiser Prado.
- The image clearly shows that the white car in the right front is indeed a Toyota Land Cruiser Prado with a spare tire.
- Therefore, the AI assistant's response is accurate in terms of factual correctness.
- Score: 10

2. ****Meeting User Needs****:

- The user's question asked for the model of the white car with a spare tire on the right front.
- The AI assistant's response directly identifies the car model, matching the reference answer, thus fully meeting the user's needs.
- Score: 10

3. ****Visual Positioning Ability****:

- The user specifically mentioned "right front" in the question. The AI assistant correctly identified the white car in the right front in the image and accurately described its model.
- Score: 10

4. ****Clarity****:

- The AI assistant's response is concise and clear, using straightforward language and structure, making it easy for the user to understand.
- Score: 10

5. ****Completeness****:

- The AI assistant's response provides sufficient information, i.e., the car model, fully meeting the user's needs without omitting important aspects.
- Score: 10

Comprehensive Score Calculation

- Factual Accuracy: $10 * 3 = 30$
- Meeting User Needs: $10 * 3 = 30$
- Visual Positioning Ability: $10 * 3 = 30$
- Clarity: $10 * 1 = 10$
- Completeness: $10 * 2 = 20$

Comprehensive Score = $(30 + 30 + 30 + 10 + 20) / (3 + 3 + 3 + 1 + 2) = 120 / 12 = 10$

Final Score

{"Factual Accuracy": [10, 3], "Meeting User Needs": [10, 3], "Visual Positioning Ability": [10, 3], "Clarity": [10, 1], "Completeness": [10, 2], "Comprehensive Score": 10}

LLM Judgement