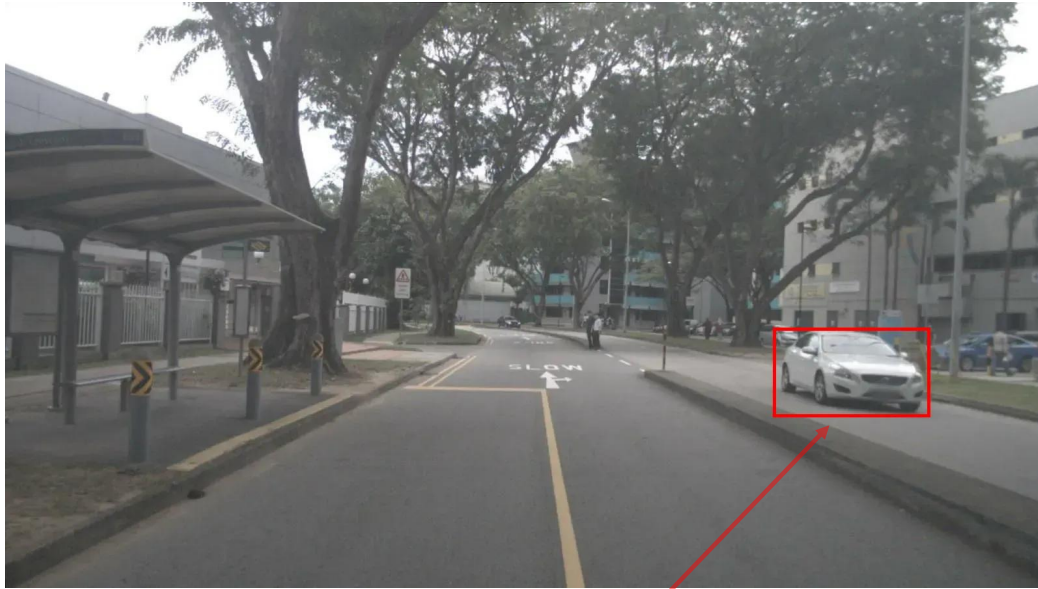


1 **SURDS: Benchmarking Spatial Understanding and Reasoning**  
2 **in Driving Scenarios with Vision Language Models**  
3 **Supplementary Material**

4 **1 Example Illustrating the Reasoning Process**



**Question:**

**Task Description:**

The primary goal of this task is to accurately identify and provide the coordinates of a specified object within a given image. Your task is to analyze the image, locate the object, and return its position in the form of coordinates [x, y].

Question: **Where is the white car located in the image?**

First, detect and list the target objects mentioned in the question from the image. Output their bounding boxes under the <location> tag. Next, reason carefully and step-by-step under the <think> tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the <answer> tag. Use the following format: <location>[[Detected target object]: [xmin, ymin, xmax, ymax], ...]</location> <think>[Step-by-step reasoning with attention to detail and potential error checks]</think><answer>[Final answer]</answer>

**Answer:**

<location>white car: [1179, 500, 1420, 637]</location>

<think>

Step 1: The task is to identify the coordinates of the white car in the image.

Step 2: **The image shows a white car on the right side of the road, close to the curb.**

Step 3: The white car is the object of interest, and its position should be identified.

Step 4: **Estimating the coordinates, the white car appears to be around the middle-right of the image. The x-coordinate is approximately 1320, and the y-coordinate is around 551.**

Step 5: Verifying the solution, the coordinates seem plausible as they place the car near the middle-right of the image.

Step 6: No alternative scenarios are considered as the car is clearly visible and identifiable.

Step 7: The final answer is [1320, 551].

Step 8: The assumptions made are based on the visual cues and the car's position relative to the image boundaries.

</think>

<answer>[1320, 551]</answer> ✓

Figure 1: Example of Pixel Location Estimation task.



**Question:**

Task Description:

The primary goal of this task is to identify the direction that the specified object is facing in the given image. **The camera in the image is facing North**, and you need to analyze the object's orientation based on this reference.

Question: **Which direction is the white car facing in the image?**

Options: \n- North\n- South\n- East\n- West

First, detect and list the target objects mentioned in the question from the image. Output their bounding boxes under the <location> tag. Next, reason carefully and step-by-step under the <think> tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the <answer> tag. Use the following format: <location>[[Detected target object]: [xmin, ymin, xmax, ymax], ...]</location> <think>[Step-by-step reasoning with attention to detail and potential error checks]</think><answer>[Final answer]</answer>

**Answer:**

<location>white car: [789, 519, 895, 601]</location>

<think>

Step 1: Establish a Reference Direction

**The camera is facing North, so this is our baseline direction.**

Step 2: Determine Object Position Relative to Camera

The white car is positioned in front of the camera on the road, **moving towards the camera.**

Step 3: Analyze Object Orientation Based on Reference and Position

**Since the car is moving towards the camera and the camera is facing North, the car must be facing South.**

Step 4: Consider Additional Contextual Clues

There are no visible contextual clues such as road markings or traffic signs that would suggest a different orientation for the car.

Step 5: Verify Consistency with Given Options

The derived orientation of the car facing South is consistent with the given options.

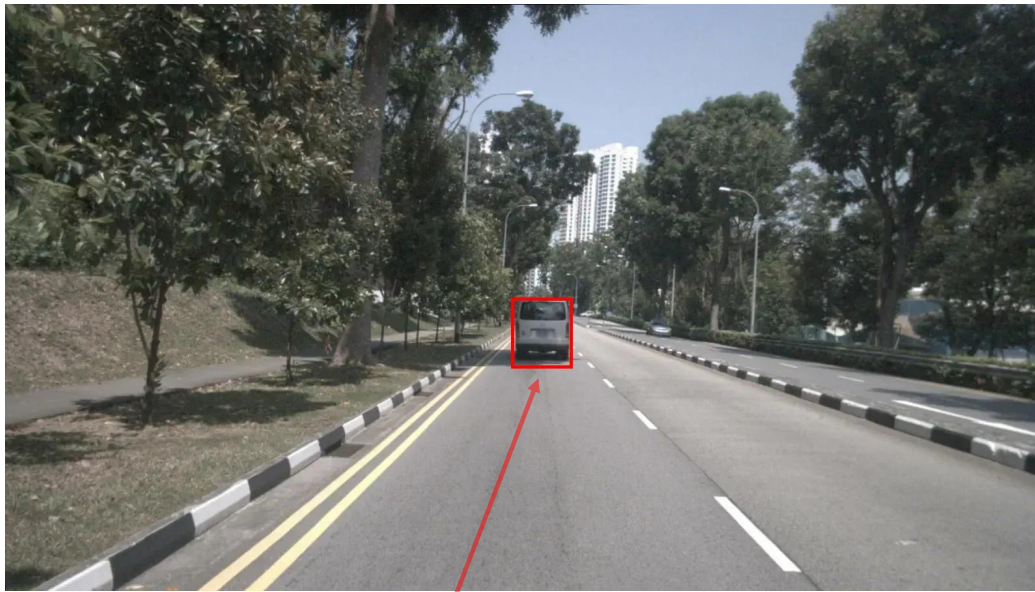
Step 6: Formulate a Clear Final Answer

After verifying the analysis, the final answer is clear.

</think>

<answer>**South**</answer> 

Figure 2: Example of Yaw Angle Determination task.



#### Question:

##### Task Description:

The primary goal of this task is to estimate the vertical distance of the specified object in the image from the camera, which is positioned at the origin. You need to analyze the image and choose the correct range of distance from the camera based on the visual cues provided.

Question: **How far is the vertical distance of the white car in the picture from the camera?**

Options:- Between 15 meters and 18 meters - Between 19 meters and 23 meters - Between 24 meters and 31 meters

First, detect and list the target objects mentioned in the question from the image. Output their bounding boxes under the <location> tag. Next, reason carefully and step-by-step under the <think> tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the <answer> tag. Use the following format: <location>[[Detected target object]: [xmin, ymin, xmax, ymax], ...]</location> <think>[Step-by-step reasoning with attention to detail and potential error checks]</think><answer>[Final answer]</answer>

#### Answer:

<location>white car: [777, 450, 871, 558]</location>

<think>

Step 1: Understand the Context and Question

The question asks for the vertical distance of the white car from the camera, which is positioned at the origin. The options provided are ranges of distances in meters.

Step 2: Identify Visual Cues and Reference Points

The image shows a road with trees on both sides and a white car in the distance. The car appears to be a standard-sized vehicle, and the road markings can be used as reference points.

Step 3: Apply Principles of Perspective and Scaling

The white car appears smaller due to perspective, indicating it is farther away. The road markings can help estimate the distance. The car is positioned near the midpoint of the road, which suggests it is not extremely far away.

Step 4: Consider Environmental and Contextual Clues

The environment is a wide road with trees and a clear sky. The car is driving away from the camera, and the road appears to be a typical urban or suburban road.

Step 5: Make Educated Guesses Based on Reasonable Assumptions

**Assuming the car is a standard-sized vehicle and using the road markings as a reference, the car appears to be between 24 meters and 31 meters away from the camera.**

Step 6: Verify and Refine Estimates

Cross-checking the estimate with the given options, the range of 24 meters to 31 meters seems reasonable based on the visual cues and the typical dimensions of a standard-sized car.

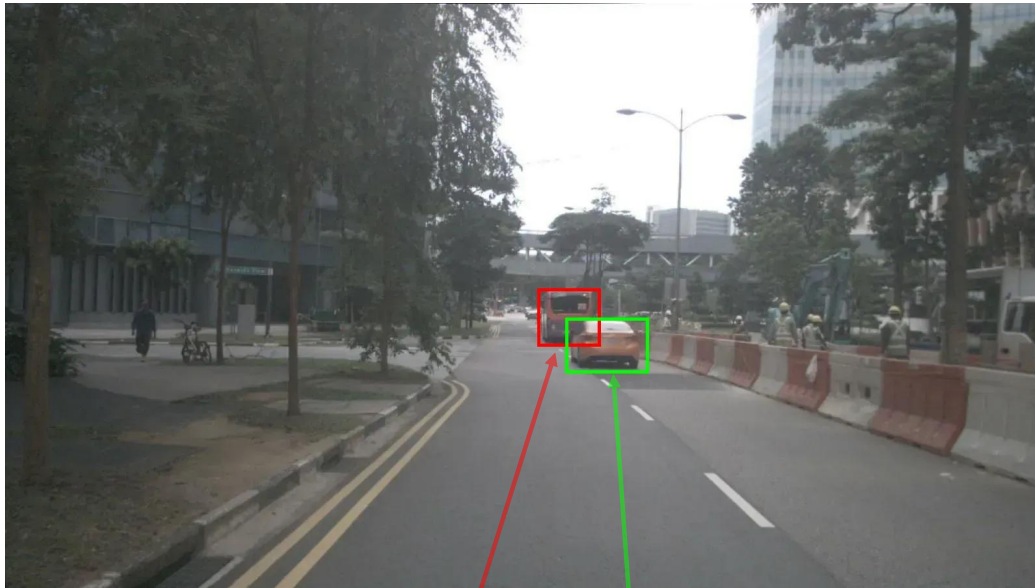
Step 7: Select the Most Appropriate Option

Based on the analysis, the most appropriate option is "Between 24 meters and 31 meters."

</think>

<answer>**Between 24 meters and 31 meters**</answer> ☒

Figure 3: Example of Depth Range Determination task.



#### Question:

##### Task Description:

The primary goal of this task is to determine which of the two objects is closer to the camera that captured the image below. You need to assess the relative distance between the two objects based on the camera's perspective.

Question: **Which object, the orange rigid bus or the orange car, is closer to the camera?**

Options: - The orange rigid bus - The orange car - Almost the same

First, detect and list the target objects mentioned in the question from the image. Output their bounding boxes under the <location> tag. Next, reason carefully and step-by-step under the <think> tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the <answer> tag. Use the following format: <location>[[Detected target object]: [xmin, ymin, xmax, ymax], ...]</location> <think>[Step-by-step reasoning with attention to detail and potential error checks]</think><answer>[Final answer]</answer>

#### Answer:

<location>orange rigid bus: [818, 438, 915, 527], orange car: [861, 483, 989, 568]</location>  
<think>

1. **Identify Scene and Objects**: The scene is a road with an orange rigid bus and an orange car visible. **The bus is farther down the road, and the car is closer to the camera.**
2. **Consider Size and Perspective**: **The bus appears smaller compared to the car. Since objects closer to the camera appear larger, this suggests the car is closer to the camera.**
3. **Evaluate Clarity and Detail**: The car appears more detailed and sharper compared to the bus, which is further away and less detailed.
4. **Analyze Positioning and Overlap**: **The car is positioned closer to the camera and partially overlaps the bus, indicating it is nearer to the camera's viewpoint.**
5. **Examine Depth Cues**: **Linear perspective shows that the road narrows into the distance, and the bus is closer to this narrowing point, suggesting it is farther away.**
6. **Consider Lighting and Shadows**: The lighting and shadows on both objects are consistent with the car being closer to the camera.
7. **Visualize Camera Perspective**: Visualizing from the camera's perspective confirms that the car is closer, as it is larger, more detailed, and positioned closer in the frame.
8. **Cross-Verify Observations**: All observations consistently support the car being closer to the camera.
9. **Select the Best Answer**: Based on the analysis, the most appropriate answer is \"The orange car.\"

</think>

<answer>**The orange car**</answer> ☒

Figure 4: Example of Distance Estimation task.



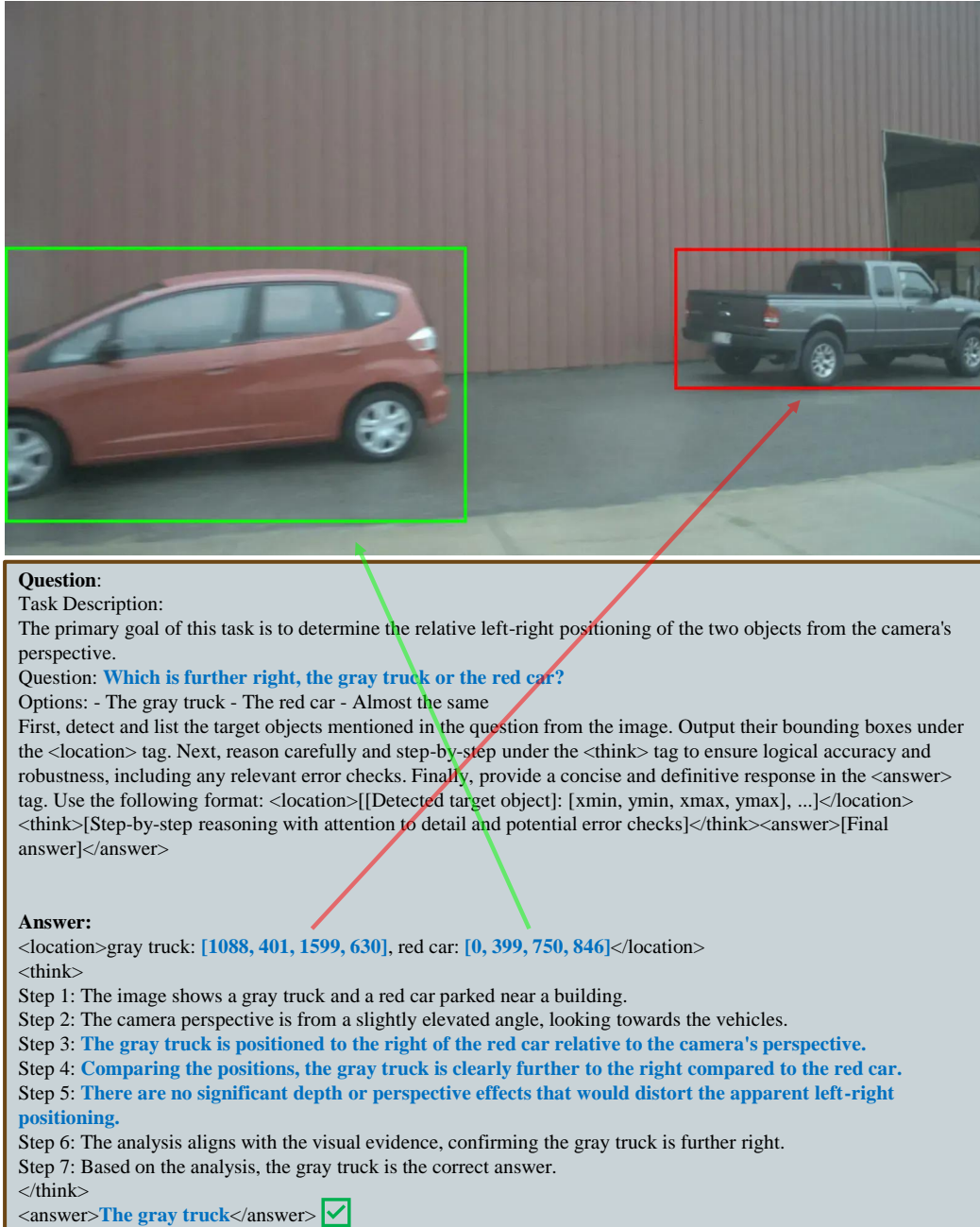
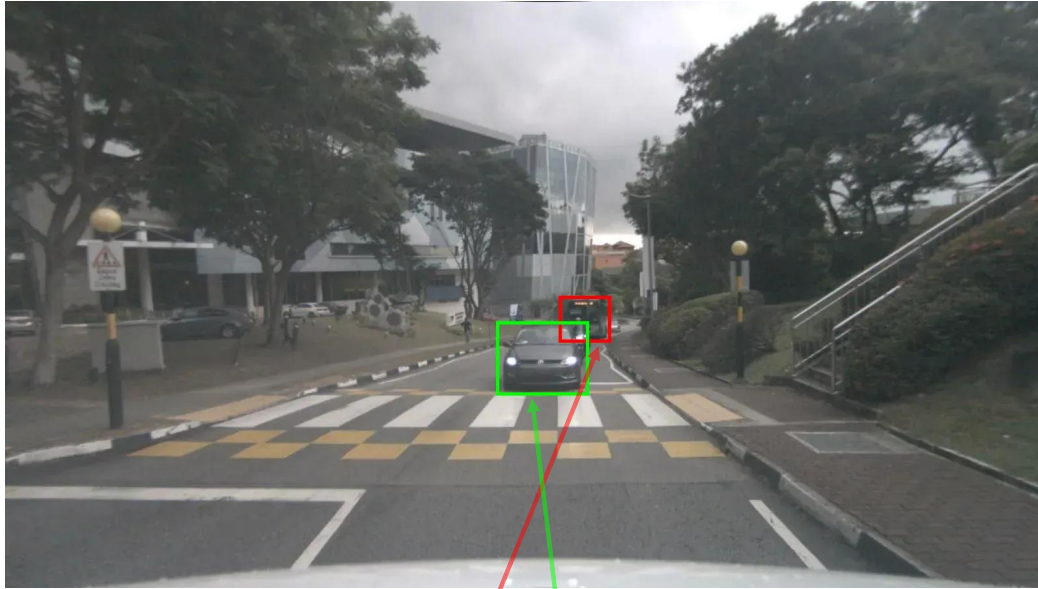


Figure 5: Example of Left/Right Determination task.



#### Question:

##### Task Description:

The primary goal of this task is to determine the relative front-back positioning of the two objects from the camera's perspective, **where the object farther from the camera is considered to be more forward.**

Question: **Is the blue rigid bus behind the gray car?**

Options: - Yes - No - Almost the same in terms of front-back position

First, detect and list the target objects mentioned in the question from the image. Output their bounding boxes under the `<location>` tag. Next, reason carefully and step-by-step under the `<think>` tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the `<answer>` tag. Use the following format: `<location>[[Detected target object]: [xmin, ymin, xmax, ymax], ...]</location>`  
`<think>[Step-by-step reasoning with attention to detail and potential error checks]</think>`  
`<answer>[Final answer]</answer>`

#### Answer:

`<location>`blue rigid bus: [850, 452, 930, 523], gray car: [754, 491, 896, 604]`</location>`  
`<think>`

Step 1: Identify the Key Objects and Their Descriptions

- The key objects are the blue rigid bus and the gray car.
- **The blue rigid bus is visible in the distance, further down the road.**
- **The gray car is closer to the camera, positioned on the road in front of the bus.**

Step 2: Understand the Camera Perspective

- The camera is positioned at a lower angle, possibly from a vehicle, and is facing forward along the road.
- **The objects are positioned along the road, with the bus being farther away and the car being closer to the camera.**

Step 3: Interpret Relative Positioning Terms

- **According to the task, "forward" means farther from the camera, and "backward" means closer to the camera.**
- **The blue rigid bus is farther from the camera compared to the gray car.**

Step 4: Analyze Object Positions Relative to the Camera

- **The gray car is closer to the camera than the blue rigid bus.**
- **Therefore, the blue rigid bus is more forward relative to the camera.**

Step 5: Consider Possible Scenarios and Visualize

- Visualizing the scene confirms that the gray car is in front of the blue rigid bus from the camera's perspective.

Step 6: Evaluate Given Options and Select the Best Answer

- Given the options, "No" is the correct answer because the blue rigid bus is not behind the gray car; it is more forward from the camera's perspective.

Step 7: Verify Accuracy and Provide Final Answer

- The reasoning aligns with the task requirements and the provided information.

`</think>`

`<answer>`No`</answer>` ☒

Figure 6: Example of Front/Behind Determination task.

5 **2 Template for Generating VQA Tasks**

**Template for Yaw Angle Determination Task**

**Task Description:**

The primary goal of this task is to identify the direction that the specified object is facing in the given image. The camera in the image is facing {}, and you need to analyze the object's orientation based on this reference.

**Question:**

Which direction is {} facing in the image?

Options: - {} - {} - {} - {}

6

**Template for Pixel Location Estimation Task**

**Task Description:**

The primary goal of this task is to accurately identify and provide the coordinates of a specified object within a given image. Your task is to analyze the image, locate the object, and return its position in the form of coordinates [x, y].

**Question:**

Where is {} located in the image?

7

**Template for Depth Range Determination Task**

**Task Description:**

The primary goal of this task is to estimate the vertical distance of the specified object in the image from the camera, which is positioned at the origin. You need to analyze the image and choose the correct range of distance from the camera based on the visual cues provided.

**Question:**

How far is the vertical distance of {} in the picture from the camera?

Options: - {} - {} - {}

8

### Template for Distance Estimation Task

#### Task Description:

The primary goal of this task is to determine which of the two objects is closer to the camera that captured the image below. You need to assess the relative distance between the two objects based on the camera's perspective.

#### Question:

Which object, {} or {}, is {} to the camera?

Options: - {} - {} - Almost the same

9

### Template for Left/Right Determination Task

#### Task Description:

The primary goal of this task is to determine the relative left-right positioning of the two objects from the camera's perspective.

#### Question:

Which is further {}, {} or {}?

Options: - {} - {} - Almost the same

10

### Template for Front/Behind Determination Task

#### Task Description:

The primary goal of this task is to determine the relative front-back positioning of the two objects from the camera's perspective, where the object farther from the camera is considered to be more forward.

#### Question:

Is {} {} {}?

Options: - Yes - No - Almost the same in terms of front-back position

11



12 **3 Structured Response Format for the VQA Task**

**Structured Response Format with Location Tag**

First, detect and list the target objects mentioned in the question from the image. Output their bounding boxes under the <location> tag. Next, reason carefully and step-by-step under the <think> tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the <answer> tag.

Use the following format:

```
<location>[[Detected target object]: [xmin, ymin, xmax, ymax],
...]</location>
<think>[Step-by-step reasoning with attention to detail and po-
tential error checks]</think>
<answer>[Final answer]</answer>
```

13

**Structured Response Format without Location Tag**

Reason carefully and step-by-step under the <think> tag to ensure logical accuracy and robustness, including any relevant error checks. Finally, provide a concise and definitive response in the <answer> tag.

Use the following format:

```
<think>[Step-by-step reasoning with attention to detail and po-
tential error checks]</think>
<answer>[Final answer]</answer>
```

14

15 **4 Prompts for High-Quality Chain-of-Thought Generation**

**Prompt for Generating Chain-of-Thought**

Analyze the following task step by step to derive the best possible answer.

Task: {task}

Answer: {answer}

Please provide a detailed reasoning process, verify its accuracy, and then give your final answer clearly.

16

### Prompt for Summarizing Rules from Examples

You are given the following reasoning examples. Analyze these examples to identify the underlying, generalizable problem-solving principles.

Examples: {examples}

Present your findings as bullet points in this format:

- Step 1: [core principle] - Step 2: [core principle] ...

Ensure these rules can be applied broadly to similar questions.

17

### Prompt for Generating Answers Using Extracted Rules

Use the following principles to answer the question:

{rules}

Question: {question} Answer: {answer}

Provide a concise solution with key reasoning steps in the following format: <think>[Your step-by-step reasoning]</think>  
<answer>[Final answer]</answer>

18

### Prompt for Verifying and Refining Reasoning and Answers

{response}

Evaluate the structured response above for logical consistency and completeness. Specifically:

1. Does the reasoning in <think> logically support the conclusion in <answer>?
2. Are there any internal contradictions, logical errors, or missing steps in the reasoning?
3. Is the reasoning chain complete and valid?

Provide your evaluation in the following format:

<reason>[A concise justification of your assessment or a brief note confirming the reasoning's validity]</reason> <validation>Valid / Invalid</validation>

Then, regardless of validity, output the full response in the following format:

- Keep <answer> unchanged.

- Modify <think> only if necessary to ensure logical soundness.

<think>[final version of reasoning steps]</think>

<answer>[original final answer]</answer>

19