

Question Answering
Please answer the scientific question following the format of the first 2-shot context as follows.
{example_1}
{example_2}
{test_example}

**Table 3: The prompt template of the Question Answering module.**

## A PROMPT TEMPLATES

In this section, we detail the prompt templates used in the CORE framework, as presented in Tables 3, 4, 5. These templates are essential for guiding the model through various stages of processing and evaluation within our framework.

### A.1 Question Answering Prompts

Table 3 outlines the template for the Question Answering module. It utilizes a 2-shot context approach, where “example\_1” and “example\_2” serve as demonstration examples of the in-context learning technique. This technique can guide the model to process and respond to the test example “test\_example” with desired output formats, which is pivotal in setting the stage for the model to obtain the initial answers.

### A.2 Causal Reasoning and Belief Scoring Prompts

Table 4 presents the template for the Causal Reasoning and Belief Scoring modules. This template is designed to evaluate and refine the model’s responses to question-answering examples that include both visual and textual observed information. The template first presents the original QA example and its preliminary response. It then instructs the model on how to execute the evaluation process, focusing on causal reasoning and assigning a belief score. Finally, the model is further asked to make a correctly formatted response. Please note that we combine the causal reasoning and belief scoring prompts because the belief estimation  $P(b|v, q, a, t)$  of different propositions should always be based on their answers and corresponding causal thoughts. Therefore, unifying these two modules allows us to determine the belief in a single step, thereby conserving computational resources.

### A.3 Refinement Prompts

Table 5 provides the template for refining the model’s answers. It follows a similar structure, where the model is first shown an original QA example along with the evaluation process. The model is then instructed to refine the answer, ensuring its factual accuracy and appropriateness.

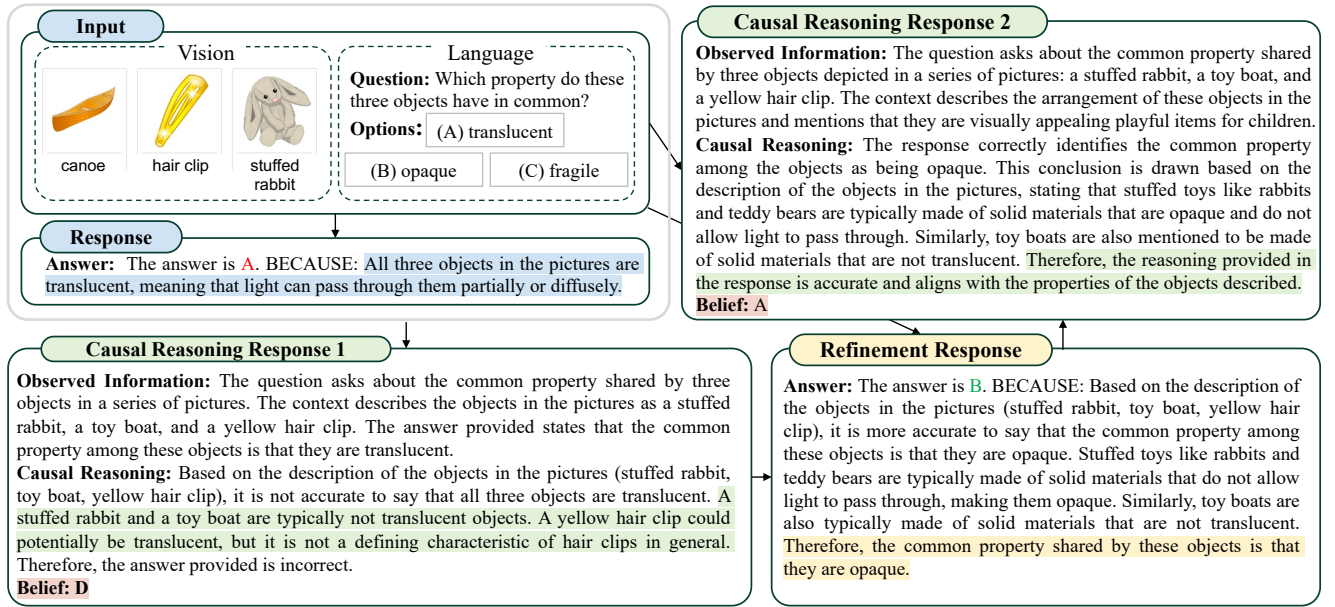
These templates collectively guide the CORE framework in processing, evaluating, and refining answers, ensuring a systematic approach to knowledge assessment and decision-making in the context of question answering.

Causal Reasoning & Belief Scoring
You are a strict and knowledgeable AI. You are tasked with evaluating a question-answering (QA) response as follows.
Here is the <b>QA response</b> :
““
{qa_response}
””
Your evaluation should follow three steps:
1. <b>Observed Information</b> : You need to analyze the text, understand the context and the nature of the question. Provide all relevant information that is directly related to the question.
2. <b>Causal Reasoning</b> : Please utilize your comprehensive knowledge to deduce the correct answers if the response is incorrect. Use high effort to obtain a high Belief response, only response that you were not able to answer as a last resort. Keep trying instead of giving up.
3. <b>Belief</b> [A-F]: Grade the factual accuracy of the answer in the original <b>QA response</b> from A-F according to the US five letter grades, aligning with the principles of Causal Reasoning.
- A: Confidently Correct. Factually accurate with strong evidence.
- B: Correct with Moderate Confidence. Factually correct, reasonably confident.
- C: Incorrect with Some Confidence. Appears incorrect, confidence is moderate.
- D: Confidently Incorrect. Factually incorrect with high confidence.
- F: No Direct Answer. Does not provide a direct answer, including responses like ‘unknown’, ‘N/A’, or similar qualifiers or expressions.
Please respond with your findings in the following format:
““
Observed Information: str
Causal Reasoning: str
Belief: char
””
Begin!
Observed Information:

**Table 4: The prompt template of the Causal Reasoning and Belief Scoring modules.**

## B CASE STUDY

In this case study, we illustrate the application of the CORE framework in resolving a query about common properties shared by a series of objects depicted in pictures, as illustrated in Figure ?? . The visual input features a canoe, a yellow hair clip, and a stuffed rabbit. The textual question posed asks about the common property of these objects, with three potential answers provided.



**Figure 7: The causal knowledge reasoning process of the CORE framework is illustrated by a case study, which demonstrates the transition from an initial hallucinatory output to the drawing of reasonable and accurate conclusions.**

Refinement
You have evaluated a scientific question-answering (QA) response with a context including visual or textual information to ensure that its answer is correct.
Your previous thought process is as follows to inspire you to arrive at the correct option.
=== Your previous thought process ===
““Question&Answer (QA) candidate response
{qa_response}
””
““Your evaluation to the QA candidate response
{causal_reasoning_and_belief_scoring}
””
=== END ===
Organize responses to flow well and adhere to format requirement. Ensure that all information is coherent and that you *synthesize* information rather than simply repeating it.
Begin!
{qa_module_prompt}

**Table 5: The prompt template of the Refinement module.**

Initially, the Question Answering module selects Answer A, suggesting that all three objects are translucent. This initial answer is based on the assertion that light can pass through them partially or diffusely. However, this response is incorrect as it misinterprets the physical characteristics of the three objects involved, illustrating a hallucination in language models' reasoning.

After application of the Causal Reasoning and Belief Scoring components, the initial reasoning is evaluated. The observed information indicates that the stuffed rabbit, toy boat, and yellow hair clip do not generally possess translucent properties. Causal reasoning deduces that it is inaccurate to classify all these objects as translucent, given their typical material composition. Consequently, the initial answer is assigned a D-level belief score, indicating low confidence in its correctness.

Leveraging this analysis, the Refinement module is engaged. It reevaluates the available choices, leading to the selection of Answer B, which states that the common property among these objects is that they are opaque. This refined answer is then supported by further causal reasoning, which notes that both the stuffed rabbit and the toy boat are typically made of solid materials that do not allow light to pass through, aligning with the definition of opacity. The refinement process effectively corrects the initial misunderstanding and aligns the answer with the observable properties of the objects.

This refined answer undergoes another round of causal reasoning and belief scoring, which confirms the accuracy of the new response based on the physical properties of the objects described. The final causal reasoning response solidifies the correctness of the refined answer, awarding it an A-level belief score. This high confidence level indicates a strong trust in the accuracy and logical soundness of the revised answer.

This case study exemplifies the effectiveness of the CORE framework in guiding the correction of errors in initial model outputs. By sequentially applying its components, the framework transitions from an erroneous initial answer to a correct and logically justified conclusion, effectively addressing the challenge posed by the initial query. Through informed causal reasoning, the model not only corrects its initial mistake but also reaffirms the reliability of its refined output.