Here’s how the **prompt was changed** for calculating **prompt sensitivity**, along with the reasoning behind each modification:

### **1. Rephrasing for Clarity**

* **Original**: "A multi-choice clarification question has high coverage if its potential answers cover every potential aspect of the query."
* **Revised**: "The goal is to determine how comprehensively the clarification question and its options address every possible aspect of the query."

**Reason**: This rephrasing introduces the task in more explicit terms, aiming to remove ambiguity. "Comprehensively address" is used instead of "cover" to emphasize depth.

### **2. Simplified Rating Description**

* **Original**: Descriptions of ratings were verbose and repeated "high coverage" multiple times.
* **Revised**: Descriptions were streamlined for brevity:
  + "1: The clarification question and options do not adequately cover the aspects of the query."
  + "5: The clarification question and options fully cover all potential aspects of the query."

**Reason**: Reducing verbosity ensures the focus remains on the task and avoids redundancy, which may influence the model's performance by overemphasizing certain terms.

### **3. Examples Modified for Context**

* **Original Examples**: Included detailed phrases like "Coverage label (Ground Truth Label):".
* **Revised Examples**:
  + Simplified to "Coverage label (Ground Truth): [Value]".
  + Added labels as part of a bulleted list for readability.

**Reason**: Simplified examples ensure the model doesn’t overly rely on formatting while keeping the task context intact.

### **4. Clear Task Instruction**

* **Original**: "You are tasked to rate the extent to which the Clarification Question and its Options cover every potential aspect of the query."
* **Revised**: "Given the following query and clarification question with its options, rate the extent to which the clarification question and its options cover all aspects of the query."

**Reason**: The revised instruction explicitly mentions "Given the following query," ensuring the model understands the sequence of input it will process.

### **5. Adjusted Confidence Explanation**

* **Original**: "Rate your confidence in your response on a scale of 0% to 100%, where 0% means no confidence and 100% means absolute confidence."
* **Revised**: Retained but moved under "Instructions" for clear separation of the task and response format.

**Reason**: Structuring confidence as part of instructions clarifies its role as a secondary output metric.

### **6. Structuring Variations**

* **Additional Variation**: The revised prompt allows variations in:
  + **Formal vs. Conversational tone**: For testing sensitivity to tone.
  + **Prompt Length**: Examples can be shortened or extended.
  + **Rating Descriptions**: Use simpler numeric ranges (e.g., "Low/Medium/High") for testing.
  + **Examples Inclusion**: Run experiments with and without examples.

### **Impact of Changes**

* The changes ensure the prompt:
  + Tests model sensitivity to phrasing and structure.
  + Prevents overfitting to specific words like "high coverage."
  + Allows measurable comparisons across variants (e.g., accuracy, confidence, F1-scores).