Task: Analyze a revised email for agreement with the rewrite instruction.

You will be given this XML input:
<natural_prompt>The original prompt given to the LLM to generate the initial email.</natural_prompt>
<raw_generated_email>The initial email generated by the LLM based on the natural prompt.</raw_generated_email>
<rewrite_instruction>The instruction provided for rewriting the initial email.</rewrite_instruction>
<rewritten_email>The revised email that has been altered according to the rewrite instruction.</rewritten_email>

Instructions:
1. Carefully compare the 'Raw Generated Email' and the 'Rewritten Email' based on the 'Rewrite Instruction'. Note whether the structure, tone, and content changes were necessary to integrate the instruction's requirements while maintaining coherence and consistency.
2. Check Agreement: Meticulously examine the 'Rewritten Email'. Your focus is to determine if ALL requirements mentioned in the 'Rewrite Instruction' are correctly and effectively incorporated.
3. Calculate Percentage: Calculate the following:
Total Number of Requirements: The number of distinct requirements mentioned in the 'Rewrite Instruction'.
Correct Requirements: The number of requirements accurately integrated into the 'Rewritten Email'.
Percentage: (Correct Requirements / Total Number of Requirements) * 100
4. Output Percentage.

You will output in XML form:
<output_explanation>Explain your reasoning for each requirement, indicating whether it's satisfied, and why.</output_explanation>
<output_percentage>Provide the final accuracy percentage. Do not include anything besides a percentage.</output_percentage>

Begin!

<natural_prompt>[NATURAL PROMPT]</natural_prompt>
<raw_generated_email>[EMAIL]</raw_generated_email>
<rewrite_instruction>[REWRITE INSTRUCTION]</rewrite_instruction>
<rewritten_email>REWRITTEN EMAIL</rewritten_email>

Your XML output with explanation and percentage:
<output_explanation>

Table 22: Prompt template used to generate conversational rewrite agreement score. The [NATURAL PROMPT], [EMAIL], [REWRITE INSTRUCTION], and REWRITTEN EMAIL placeholders will be replaced by a given natural prompt, its raw generated email, rewrite instruction, and rewritten email.

agreement evaluation, which focuses on compliance with rewrite instructions, coherence evaluation directly measures whether a revised response contains contradictions, logical inconsistencies, or abrupt structural breaks. The AutoRater employs a structured prompt (Table 23) that systematically evaluates coherence by requiring explicit yes/no judgments alongside explanatory reasoning. This approach enhances interpretability and ensures reliable scoring. (i) Binary Coherence Decision: Each rewritten response is judged as either internally consistent ("YES") or inconsistent ("NO"), ensuring clarity in evaluation. (ii) Justification for Each Judgment: The prompt mandates a reason for the decision, encouraging the model to articulate why a response is or is not coherent. (iii) Few-Shot Learning with Examples: The prompt provides multiple illustrative cases, demonstrating correct coherence judgments across different scenarios, including numerical inconsistencies, logical contradictions, and unsupported claims.

AutoSxS evaluation. AutoSxS is designed to directly compare rewritten responses by evaluating how well they satisfy a given rewrite instruction. Unlike individual scoring metrics (e.g., agreement, coherence), AutoSxS provides pairwise judgments, making it a more fine-grained and human-aligned evaluation method. As shown in Table 24, the AutoSxS framework follows a structured decision-making process: (i) The model is presented with two rewritten responses (A, B) for the same raw email and rewrite instruction. (ii) It must analyze both responses and select the one that better satisfies the rewrite instruction (or declare them as equal). (iii) It provides explicit reasoning for its choice and assigns numerical scores (0–1) to each response, ensuring that the relative ranking is interpretable. To enhance reliability, the prompt includes: (i) Clear task instructions emphasizing rewrite alignment. (ii) Demonstrations with diverse examples covering different types of modifications (removal, addition, tone change, etc.). (iii) Structured output, ensuring consistency in comparisons.

Table 25 presents two examples illustrating AutoSxS decisions. In Example 1, Response A is rated 0.91, significantly higher than Response B (0.58), as it strictly adheres to the instruction of removing unnecessary details. In Example 2, Response B is preferred (0.83 vs. 0.76) for expanding on the religious holiday's significance in a more formal and comprehensive way. These results demonstrate AutoSxS's ability to: (i) Identify minor instruction violations (e.g., residual details in Response B of Example 1). (ii) Capture nuanced preferences in tone and detail level (e.g., preference for a more formal explanation in Example 2). (iii) Provide interpretable ranking beyond binary correctness. By leveraging structured evaluation, explicit reasoning, and numerical scoring, AutoSxS enhances automated rewrite assessment, ensuring more human-aligned and context-aware comparisons across rewriting tasks.

Your task is to evaluate whether or not a piece of text is internally consistent.

Please provide your answer in the following format:
<text>[text to evaluate for internal consistency]</text>
<answer>[YES or NO]</answer>
<reason>[reason for giving the answer above]</reason>

Example #1:
<text>Paul has 5 daughters named Ava, Brittney, and Claire.</text>
<answer>NO</answer>
<reason>The response states that Paul has 5 daughters but only names 3 of them.</reason>

Example #2:
<text>IPhones are better than Samsung. The reason is because Samsung has a shorter battery life.</text>
<answer>YES</answer>
<reason>Everything is internally consistent.</reason>

Example #3:
<text>While Bob Ross and Annettte Kowalski's exact age difference is not publicly known, they likely had several years age difference.

Here's what we found online:
* Bob Ross was born on October 29, 1942 and died at age 52 on July 4, 1995.
* Annette Kowalski was born on January 26, 1936.
</text>
<answer>NO</answer>
<reason>The beginning says that the age difference is not publicly known, but later the text says that both birth dates were found online.</reason>

Question:
<text>`[REWRITTEN RESPONSE]`</text>
<answer>

Table 23: Prompt template used to generate rewrite coherence score. The `[REWRITTEN RESPONSE]` placeholder will be replaced by a rewritten response.