Evaluation Guide for AI-Generated Texts

Each text must be evaluated along four dimensions: cohesion, coherence, clarity, and informativeness.
The scale ranges from 1 to 5, where 1 represents the lowest level and 5 the highest.

1. Cohesion

Measures the linguistic and grammatical connectedness between phrases and sentences.

1: Disjointed text; isolated phrases, no connectors, hard to follow.

2: Minimal connections; some abrupt jumps between sentences.

3: Acceptable cohesion; sentences are linked but with occasional breaks.

4: Good fluency; proper use of connectors and clear idea progression.

5: Very high cohesion; the text flows naturally, smoothly, and with well-used connectors.

2. Coherence

Measures the logical and semantic consistency of the text in relation to the stated problem.

1: Incoherent; contradictions, nonsensical ideas, or unrelated to the question.

2: Partially incoherent; mix of valid ideas and illogical ones.

3: Moderate coherence; addresses the problem with some logic but includes flaws or digressions.

4: Good coherence; consistent ideas mostly aligned with the problem.

5: Very coherent; entirely logical, consistent, and directly relevant.

3. Clarity

Measures how easily the text can be understood by an average human reader.

1: Very confusing; tangled sentences, ambiguous or unintelligible language.

2: Unclear; partially understandable, requires rereading.

3: Intermediate clarity; understandable with effort, though some ambiguity remains.

4: Fairly clear; easy to understand, few difficulties.

5: Very clear; simple, direct, and immediately understandable language.

4. Informativeness

Measures the amount and usefulness of information provided in the response.

1: Very poor; provides almost no relevant information.

2: Limited; scarce or superficial information.

3: Moderate informativeness; includes relevant but incomplete data.

4: Good; sufficient information with useful details.

5: Very informative; complete, detailed, and enriching response.

Final Recommendations for Annotators

Evaluate each dimension independently.

3 = acceptable: basic idea linkage, roughly logical, understandable with some effort, and provides partial information.

Do not evaluate factual accuracy (e.g., whether a math result is correct), only textual quality.

Length does not imply quality: a long text is not necessarily better or worse.

Spelling and grammar primarily affect Clarity, not the other metrics.

Special cases:

Empty or irrelevant text → assign 1 in all dimensions.

Off-topic or hallucinatory text → set Coherence = 1, and adjust the other metrics as appropriate.