Human Evaluation Protocol

The following text is ready to be delivered to human evaluators, already containing operational instructions and rubrics (DoR and ORI) for scoring on a 0–10 scale (integers). The item below is listed as an example and must be evaluated solely based on the content presented, without external consultation.

Evaluation Objective

The evaluator must assign two scores, DoR (Depth of Reasoning) and ORI (Originality), for the model's response to the item below. The evaluation focuses on the quality of reasoning and the originality of the textual content, not on the accuracy of the answer key.

General Rules

- Evaluate only the model text under "MODEL OUTPUT".
- Do not use external sources, nor research the topic. General knowledge may be used only to judge basic coherence.
- Do not adjust the score based on personal agreement or disagreement with the topic.

Scoring Rubrics

1) DoR: Depth of Reasoning (0–10)

Measures depth, logical chaining, coverage of alternatives, and justification based on the prompt. Select the score range that best fits the description:

- (9,10]: Multi-step, stratified reasoning with strong inferential structure.
- (7,9]: Solid, structured reasoning with some complexity.
- (5,7]: Basic logical sequence with minimal elaboration.
- (3,5]: Superficial reasoning with gaps or simplifications.
- (1,3]: Fragmented logic or single-step heuristic response.
- [0,1]: No trace of reasoning or incoherent justification.

---

2) ORI: Originality (0–10)

Measures textual and intellectual originality, avoiding mere paraphrase of the prompt, and observing if the model adds content beyond the obvious. Select the score range that best fits the description:

- (9,10]: Highly original and creative formulation, abstract reasoning, no lexical overlap with the prompt.
- (7,9]: Strong reformulation and innovative phrasing with clear semantic transformation.
- (5,7]: Some novel elements yet retains the input's structure or vocabulary.
- (3,5]: Predominantly generic or templated response; moderate echo of the prompt.
- (1,3]: Prompt echo dominates; low semantic variation.
- [0,1]: Literal or near-literal reproduction of the prompt or reference.