

## A. LLM as a judge prompting

### A.1. Model prompt

Model are prompted with figures and body-text, all the prompt start by the same prefix

#### DOC-PROMPT

Considering the image(s) and the caption(s) associated:

"[CAPTIONS FIGURE 1]"

"[CAPTIONS FIGURE 2]"

:

"[CAPTIONS FIGURE N]"

The document:

"[BODY-TEXT]"

Each line are separated by two line breaks, and figures (image tokens) are provided before the DOC-PROMPT. Thus, it means that in our case, the placement of the image within the document does not correspond to the original document. The exact prompt global prompt is thus :

#### PREFIX-PROMPT

[UTTERANCE START TOKEN] system

[MODEL-PROMPT] [UTTERANCE END TOKEN]

[UTTERANCE START TOKEN] user

[IMAGE START TOKEN] [IMAGE TOKEN 1]  
...[IMAGE TOKEN L] [IMAGE END TOKEN]

:

[IMAGE START TOKEN] [IMAGE TOKEN 1]  
...[IMAGE TOKEN L] [IMAGE END TOKEN]

[DOC-PROMPT]

[CRITERION-PROMPT]

[UTTERANCE END TOKEN]

[UTTERANCE START TOKEN] assistant

Notice that [SYSTEM-PROMPT] depends on the model, we follow original system prompt proposed for each model.

### A.2. Criterion Prompt

For each model the task description prompt is provided, the output format expected is a json containing the predicted label followed by an explanation of the choice.

#### Correctness

**Description :** The model is asked to decide if the answer to the question is correct according to the content provided

#### CORRECTNESS-PROMPT

The question: "[QUESTION]" decide whether the answer "[ANSWER]" is correct, answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the label chosen.

#### Relevance

**Description :** The relevance judges if the question is aligned with document content or the objective of the page course

#### RELEVANCE-PROMPT

evaluate if the question "[QUESTION]" is well aligned with the content of the document. Choose between the following labels "fully-aligned", "aligned", "poorly-aligned" and "not-aligned". You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

#### Clarity

**Description :** Estimate if the question is clearly worded and easy to understand without ambiguities.

#### CLARITY-PROMPT

evaluate if the question "[question]" is clearly worded and easy to understand. Choose between the following labels "clear", "understandable", and "unclear". You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Appropriateness

**Description :** Estimate if the question is appropriate for the targeted audience, in our case students aged from 12-18

### APPROPRIATENESS-PROMPT

evaluate if the question "[QUESTION]" is suitable for the target learning (students aged 12–18). Choose between the following label "suitable", "acceptable", and "inappropriate". You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Visualisation 1 (Diagrams)

Estimate if the question and the answer could assess the competency of analysing and understanding diagrams, charts, or tables.

**Description :**

### VISUALISATION-1-PROMPT

evaluate the pedagogical content by determining whether the question "[QUESTION]" and the answer "[ANSWER]" assess the competency of analyzing and understanding diagrams, charts, or tables (extract relevant information, compare sources). Answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Visualisation 2 (Chronological)

**Description :** Estimate if the question and the answer could assess the competency of understanding and using Chronological references (situate events in times, interpret/use timelines, recognizing continuities and changes).

### VISUALISATION-2-PROMPT

evaluate the pedagogical content by determining whether the question "[QUESTION]" and the answer "[ANSWER]" assess the competency "understanding and using Chronological references (situate events in time, use timelines, recognize continuities and changes)". Answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Visualisation 3 (Maps)

**Description :** Estimate if the question/answer can assess the competency of understanding spatial references (interpret maps and legends, locate places, interpret diagrams flows)

### VISUALISATION-3-PROMPT

evaluate the pedagogical content by determining whether the question "[QUESTION]" and the answer "[ANSWER]" assess the competency "understanding and using Spatial References (read and interpret maps and legends, locate places, understand scales and distances)". Answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Language

**Description :** Assess the following competency: Use disciplinary language and representations (e.g., appropriate vocabulary, symbols, and conventions specific to the subject).

### LANGUAGE-PROMPT

evaluate the pedagogical content by determining whether the question "[QUESTION]" and the answer "[ANSWER]" assess the following competency: Use disciplinary language and representations (e.g., appropriate vocabulary, symbols, and conventions specific to the topic). Answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Reasoning-1

**Description :** The criteria is judged positive if it uses multi-hop steps to answer the question, for instance combining information from multiple sections or figures

### REASONING-1-PROMPT

evaluate the pedagogical content by determining whether the question [QUESTION] and the answer [ANSWER] assess the following competency: Apply reasoning and methods (e.g., observing, hypothesizing, interpreting data, drawing conclusions). Answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating.

## Reasoning-2

**Description :** Assessing the cognitive complexity of the question

### REASONING-2-PROMPT

evaluate the cognitive complexity of the task required to answer the question "[QUESTION]". Consider the following aspects: 1. Type of reasoning involved. 2. Whether it combines information from multiple sources. 3. Number of steps or operations required to reach the answer. 4. Level of abstraction. You should write the output in json format with the field "label" containing one of the following label: "no" meaning no reasoning involved, "moderate" meaning that it involve moderate reasoning (1 step reasoning for instance) "heavy" meaning the question involve difficult reasoning capacities. The JSON response must also contains a field "explanation" containing the justification of the rating.

meaning, factual content, and intent. If both answers express the same meaning, respond with the label "yes". If the meanings differ in any way (including being unrelated), respond with the label "no". You should write the output in json format with the field "label" containing the label "yes" or "no"

## Creating

**Description :** Assess the competency : Create, represent, and model information (e.g., designing diagrams, graphs, timelines, or conceptual maps based on data).

### CREATING-PROMPT

evaluate the pedagogical content by determining whether the question "[QUESTION]" and the answer [ANSWER] assess the competency : Create, represent, and model information (e.g., designing diagrams, graphs, timelines, or conceptual maps based on data). Answer by labels "yes" if correct and by "no" if not. You should write the output in json format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the rating

## B. Guidelines

The guidelines for evaluation of the QAs pairs can be found at

## C. QA generation prompts

### C.1. Evaluating Similarity Prompt

To evaluate if to answer are similar, we prompted the Qwen model with the following instruction:

Considering the question: [question]  
determine whether the predicted answer:  
[ablated\_answer] conveys the same meaning as the gold (reference) answer: [answer]  
Do not consider grammar, wording, or phrasing differences. Only compare