

Instructions for the Evaluation Process

Artworks Labels to Evaluate:

- Rubens, Susanna and the Elders
- Rubens, Heracles in the Garden of the Hesperides
- Rubens, Deianira Tempted by the Fury
- Anton Van Dyck, The Children of King Charles I of England
- Anton Van Dyck, Equestrian Portrait of Prince Thomas of Savoy-Carignano

Prompting Methods employed:

- **0-shot prompting:** The LLM receives only the general instructions
- **1-shot prompting:** The LLM receives the instructions plus an example
- **1-shot prompting with annotated example:** The LLM receives instructions, an example with specific annotations inside squared brackets.

How to evaluate (0-2 Likert scale evaluation)

Optimization of contents grid

1. Completeness of Information from a Cultural Accessibility Perspective

This criterion evaluates the completeness of the information necessary for understanding the depicted scene from a cultural accessibility standpoint.

Important: This evaluation considers the text as a whole. Any gaps in the logical flow of information across sections are assessed separately under the “Logical Progression of Information” criterion.

- **2 – Fully/Mostly Achieved**
All essential information regarding the artwork and its story is fully covered, despite some eventual minor gaps that do not compromise overall understanding.
- **1 – Partially Achieved**
Some major gaps are present (i.e., omission of key information about the main subjects of the scene), and/or irrelevant elements are included, partially compromising the understanding of the artwork.

- **0 – Not Achieved/Mostly Ignored**

The essential information to fully comprehend the episode narrated within the artwork is lacking.

Definitions:

- **Major gaps:** Missing or incomplete information about the main subjects of the scene.
- **Minor gaps:** Missing or incomplete information concerning secondary elements that contribute to a full understanding of the depicted scene.
- **Not relevant elements:** Information or elements that do not contribute meaningfully to the understanding of the artwork's scene or its historical context.

2. Logical Progression of Information

This criterion evaluates the sequential organization of the text based on adherence to the structure requested in the prompt instructions, including the correct inclusion and ordering of all required subjects.

The evaluation follows the 3-point scale:

- **2 – Fully Absent/Mostly Absent**

Full adherence to the prompt instructions and/or example. All required sections are present, correctly ordered, and generally fully developed with no relevant omissions.

- **1 – Present**

All required sections are present; however, instructions are not completely followed due to one or more of the following reasons:

- The order of the sections might not fully match the prompt instructions or the example provided.
- Non-required sections might be included.
- Some elements are included in the wrong section.
- Internal gaps may exist within sections (e.g., missing key topics or incomplete treatment of subjects).

- **0 – Consistently present**

The overall structure does not follow the provided instructions for one or more of the following reasons:

- At least one required section is missing and/or significant internal gaps are present within sections.
- Section titles and structure are not properly followed.

- Significant internal gaps affect the logical progression of information.

3. Absence of hallucinations and biases

This criteria evaluate the presence of incorrect, or imprecise informations and stereotypical information not relevant in this context

- **2 – Fully followed/mostly followed**

The interpretation is generally aligned with the correct understanding of the artwork with no major hallucination or biases. However, there can be few minor imprecisions that do not compromise the overall interpretation of the artwork but may introduce slightly misleading indications.

- **1 – Partially followed**

Presence of some hallucinations or biases that provide misleading information and partially distort the interpretation of the artwork.

- **0 - Not followed/most ignored**

Presence of multiple major hallucinations or biases that severely distort the correct interpretation of the painting (e.g., mention of non-existent subjects or incorrect identification of key elements, systematic misidentification of subjects, attribution of non-existent elements)

Definitions:

- **Major hallucinations or biases:** Errors that fundamentally alter the correct interpretation of the artwork, such as the mention of non-existent subjects or misidentification of key elements.
- **Minor hallucinations or biases:** Errors that provide misleading but non-critical information, not sufficient to fully compromise the understanding of the painting.

Example of Bias:

Misinterpreting the Greek mythological figure Ladon (the dragon/serpent) through a Christian lens as a symbol of evil, when this symbolism is not accurate for the mythological context.

***Notes:** This is the section where qualitative comments on the outputs produced should be inserted.

Readability and formatting grid

1. **Overall length** (250-280 words)

- **2 - Fully/mostly followed**

The text fully respects the specified word range (250–280 words) or it deviates slightly from the required range, exceeding or falling short by no more than 10 words.

- **1- Partially followed**

The text deviates by more than 10 but no more than 50 words from the specified range.

- **0 – Not followed/mostly ignored**

The text deviates by more than 50 words from the specified range.

2. Text Structure and formatting (4-5 lines, line lengths of less 45 characters, subtitles, bullet points, bolded keywords)

- **2 Fully/mostly followed**

At least four out of five formatting requirements are respected.

- **1 – Partially followed**

Three out of five formatting requirements are respected.

- **0 – Not followed/mostly ignored**

Only one or two formatting requirements are respected or none of the formatting requirements are respected.

3. Sentence simplicity and correctness (active verbs, subject at sentence start, no grammatical errors)

- **2 - Fully/Mostly followed**

Sentence simplicity is generally respected: active voice is mainly used, subjects are usually placed at the beginning of sentences, and there are no grammatical errors.

- **1 – Partially followed**

In some parts of the text, the subject is not placed at the beginning of the sentence, or the passive voice is used, and/or grammatical errors are present.

- **0 – Not followed/Mostly ignored**

In several parts of the text, the subject is not placed at the beginning of the sentence, the passive voice is used, and/or grammatical errors are found.

The text extensively uses passive voice, lacks clear subject placement at the beginning of sentences throughout, and presents frequent grammatical errors.

4. Simple Syntax (avoid subordinates, spoken-like language, one concept per sentence, related info proximity, direct questions)

2 - Fully followed

The syntax is simple and clear: almost all the listed requirements are fully respected, making the text easy to read.

1 – Partially followed

The text poses some challenges to reading comprehension, although certain requirements have been respected.

0 - Not followed/Mostly ignored

The text is complex and difficult to read; most requirements are not respected.

5. Technical Terms (presence and explanation).

2 - Fully Absent/mainly absent

The text ensures full accessibility: it avoids technical terms by using synonyms or paraphrasing or if one is present is clearly explained within the context.

1 – Present

The text presents some limited terms and their usage may still affect overall accessibility.

0 - Consistently present

The text includes several technical terms without providing explanations, limiting comprehension for non-expert to understand .