

A. Evaluation Model Choices

We evaluate state-of-the-art proprietary models—GPT-4.1, GPT-4.1-mini [5], Gemini 2.0 Flash [3], and Claude 3.5 Sonnet [1]—as well as recent open-source models including InternVL-3 [6] and Qwen2.5-VL [2]. Our evaluation is limited to non-thinking models, excluding models such as o3 and Claude 3.7, which may benefit from additional reasoning steps.

B. Hyperparameters

We train both the Page Encoding and Modality Encoding variants for three epochs using distributed data-parallel training on $16 \times$ A100 40GB GPUs, which takes 27 and 17 hours, respectively. Table 1 summarizes the training hyperparameters. The two encodings use different maximum image resolutions (measured in total input pixels); in each case, we choose the largest resolution that can be processed on 40GB A100s without out-of-memory errors. Because Modality Encoding only feeds cropped visual elements, it can afford a higher effective resolution for each cropped region, whereas Page Encoding operates on full-page images, so even with a higher overall page image resolution, the effective resolution available for each visual element can be lower.

Configuration	<i>Page Encoding</i>	<i>Modality Encoding</i>
Epoch	3	3
Optimizer	AdamW	AdamW
Learning Rate	1e-05	1e-05
Learning Rate Scheduler	cosine	cosine
Warm-up Ratio	0.1	0.1
Global Batch Size	16	16
Grad Acc Steps	16	16
Numerical Precision	bfloat16	bfloat16
Image Resolution	2508800	1003520

Table 1. Hyperparameters for training Qwen2.5-VL on VinQA.

C. Details of VinQA dataset construction

C.1. Visual-element to textual form transformation

In the data preprocessing process, we generate class, caption, and description for the visual elements in the document page using GPT-4o. Each visual element is classified into categories such as chart, table, photo, diagram, icon, etc. In our dataset construction, we group photo and diagram under the type label “figure” and exclude elements classified as “icon” and “etc”. GPT-4o receives two images as input: one containing the full-page image marked with a red bounding box around the target visual element, and the other containing the cropped image of the target visual element. The corresponding prompt is presented in Figure 3.

C.2. Question Generation

To generate diverse and domain-balanced queries, we first sample page images uniformly across all domains in the corpus and assign each as a reference page. For every reference page, we utilize a ColQwen¹, multimodal retriever, to gather the ten most visually and semantically similar pages within the corpus. From these ten, four pages are randomly selected and combined with the reference page, yielding a five-page cluster. Each cluster thus encompasses a coherent but non-redundant context for question generation. From the 131,906 page train corpus, 30,000 reference pages are selected, and an equal number of clusters are consequently constructed; in parallel, 2,000 clusters are constructed from the 9,373 page test corpus.

We prompt Gemini 2.0 Flash Thinking with instructions focused on the following points for generating questions: 1) questions that target the core content of the given context; 2) questions whose answers can be derived from information distributed across multi pages; 3) when the context contains charts, tables, or figures, generate questions that integrate multiple modalities and contexts. We also include eight questions as few-shot examples in the prompt and direct the model to generate five questions for each context. Figure 1 shows an example of input context, and Figure 4 shows the prompt designed for

¹<https://huggingface.co/vidore/colqwen2-v1.0>

the generation of questions. Subsequently, we sample three of the five questions generated for each cluster and use Gemini to verify and filter them, removing any questions that are ambiguous, not self-contained, or that referenced unseen context (e.g., “based on the document” or “according to the table”), and retain only those that passed this filtering step. The prompt for filtering questions is present in Figure 5. Out of the 90,000 generated train questions, 66,988 remain after verification, and out of the 6,000 test questions, 4,632 are filtered.

[Page]:[1]
[figure_1]:Caption:None
Description:The image shows a tall, narrow tower with several levels. The structure is made of bricks, and there is a small spire at the top. The tower stands against a clear blue sky.
[chart_1]:Caption:Religion in Rome (2015)
Description:- A pie chart showing the distribution of religions in Rome as of 2015. - Red: Catholicism (82.0%) - Black: Other or non-religious (8.7%) - Blue: Eastern Orthodoxy (4%) - Pink: Protestant (0.8%) - Purple: Judaism (0.7%) - Green: Islam (3.8%)
[Context]:Religion in Rome
The Religio Romana (literally, the "Roman Religion") constituted the major religion of the city in antiquity. The first gods held sacred by the Romans were Jupiter, the highest, and Mars, the god of war, and father of Rome's twin founders, Romulus and Remus, according to tradition. ...

[Page]:[2]
[figure_2]:Caption:Forun Romanum
Description:This is a photo with the caption "Forun Romanum" indicating the location as Rome, Holy See and Italy. It is part of the UNESCO World Heritage Site, listed under various cultural criteria. The inscription year is 1980 (4th Session) with extensions noted in 1990 and 2015. The area is 1,430.8 ha (3,536 acres), and coordinates are 41°53'24.8"N 12°29'32.3"E.
[figure_3]:Caption:None
Description:The image shows a map of Rome with a red marker indicating a specific location within the city. There are various lines and markings typically representing roads and geographical features, along with a mini-map of Italy showing the location of Rome within the country.
[Context]:The image is a screenshot of a Wikipedia page titled "Culture of Rome." Here is the extracted text: ...

[Page]:[3]
[Context]:The Western religions are the religions that originated within Western culture, which are thus historically, culturally, and theologically distinct from Eastern, African and Iranian religions. The term Abrahamic religions (Christianity, Judaism and Islam) is often used instead of using the East and West terminology, as these originated in the Middle East. ...

[Page]:[4]
[figure_4]:Caption:Marcus Aurelius (head covered) sacrificing at the Temple of Jupiter
Description:The image shows a carved relief depicting a group of Roman figures in classical attire. The central figure, identified as Marcus Aurelius with his head covered, appears to be performing a sacrificial ritual at the Temple of Jupiter. The background includes architectural elements such as columns and a pediment structure typical of Roman temples. Several other figures surround Aurelius, engaged in the ceremonial act.
[table_1]:Caption:Religion in ancient Rome
Description:A header section with the title "Religion in ancient Rome" in bold white text on a dark red background. Below the title is a photo with the caption "Marcus Aurelius (head covered) sacrificing at the Temple of Jupiter" in blue italics and black text. Below the photo is a table divided into several categories with pink headers: 1. Practices and beliefs (bold): - libation - votum - temples - festivals - ludi - funerary practices - imperial cult - mystery religions 2. Priesthoods (bold): - Pontifices - Augures - Vestales - Flamines - Fetiales - Epulones - Fratres Arvales 3. Deities (bold): - Twelve major gods (bold) - Capitoline Triad - Aventine Triad - Underworld - indigitamenta (italic) - Agriculture - Birth Two subcategories under Deities: - Deified leaders (bold): - Julius Caesar - Augustus - Other deified persons (bold): - Antinous 4. Related topics (bold): - Glossary of ancient Roman religion (partially visible at the bottom)
[Context]:The text extracted from the image is as follows: ...

[Page]:[5]
[table_2]:Caption:Freedom of religion
Description:The table is titled "Freedom of religion" and contains clickable or expandable sections: "Concepts," "Status by country," and "Religious persecution" (with an option to hide). Below these sections, a list of related topics is provided, including Traditional African religions, Atheism, Bahá'í Faith, Buddhism, Christianity (Christophobia), post-Cold War era, Catholicism (Catholic Church), and Mormonism. The table has a light purple background with bold section headers and blue text for clickable links or items.
[Context]:Anti-Judaism
From Wikipedia, the free encyclopedia
Anti-Judaism describes a range of historic and current ideologies which are totally or partially based on opposition to Judaism, on the denial or the abrogation of the Mosaic covenant, and the replacement of Jewish ...

Figure 1. Input context example for Question and Answer generation.

C.3. Answer Generation

For 80% of the generated queries, we retrieve the top 5 pages using ColQwen to create answerable QA, while for the remaining 20%, we retrieve the pages ranked 15th to 20th to construct challenging unanswerable QA pairs. We specifically select these lower-ranked pages because they contain partially relevant contexts, making the resulting unanswerable QA more difficult and realistic. These 5 retrieved pages form the context.

We perform answer generation using Gemini 2.0 Flash Thinking and Claude 3.7. The model is provided with a context and a single question, along with the following instructions in the prompt: 1) generate an answer by utilizing as much relevant information as possible from the given context in relation to the question; 2) when citing content from a specific page, include the page index (e.g., [1], [2]) as unique identifier within the response sentence; 3) when referencing charts, tables, or figures include the unique identifiers (e.g., [chart_1], [table_2], [figure_3]) provided in the context within the response sentence; 4) structure the answer with an introduction, body, and conclusion, where the body is further divided into sections to provide a well-structured response format. For questions aimed at constructing unanswerable QA pairs, a different instruction is provided. While the rest of the process remains the same, the model is instructed to determine whether the question could be answered based solely on the given context, and to generate the reasoning behind this judgment to increase reliability. If the model determines the question to be answerable, it is instructed to generate an answer using the same instructions as for answerable QA pairs. We exclude the questions that are deemed answerable from this process. The prompts designed for the generation of answers are shown in Figure 6 and 7.

As a result, the train corpus contains 53,556 answerable and 10,554 unanswerable QA pairs, while the test corpus consists of 3,704 answerable QA and 751 unanswerable QA pairs. Figure 2 shows an example of the final answer generated from this process.

C.4. Data Verification

During the textual verification step, every Question–Context–Answer triple—where the context is in a textual form, as in the data generation process—is checked for 1) citation errors, 2) statements unsupported or contradicted by the context, 3) omissions of contextually relevant information, and 4) extraneous or query-irrelevant content; only data free of issues on all four criteria were preserved. For the training set, verification is performed solely by Gemini, whereas for the test set, verification is additionally conducted by Claude. Only data meeting all four criteria are retained. The prompt is shown in Figure 8.

During the visual verification step for the test set, whenever the supporting context included a chart, table, or figure, the corresponding image is supplied to the model as input. Visual verification is performed by both Gemini and Claude. The acceptance criteria are also adapted to the visual modality: the models must verify that 1) every image relevant to the question is properly used in the answer, and 2) each cited statement accurately reflects the information presented in its referenced image. Only QA pairs satisfying both visual-reasoning criteria are retained. The prompt is shown in Figure 10.

For unanswerable QA data in the test set, we perform a distinct textual verification with Claude 3.7: each Question–Context pair is inspected to determine whether the context provides enough precise information to answer the question definitively. Any pair that met this condition is deemed incorrectly labeled and discarded. The prompt is shown in Figure 9.

After multi-step machine filtering, the training set comprised 39,700 answerable and 10,554 unanswerable QA pairs, while the test set comprised 1,822 answerable and 723 unanswerable pairs. To investigate the appropriate quantity of unanswerable data, we experiment with training configurations containing either 10K or 3K unanswerable QA pairs. Because using 10K led to a drop in M-GroSE scores, we reduced the number of unanswerable QA pairs to 3K in the final training setup. Additionally, text-reference-only QA pairs are sampled to balance the dataset. As a result, the training set comprises 39,700 answerable and 3,000 unanswerable QA pairs, and the test set comprises 1,312 answerable and 400 unanswerable QA pairs.

For the test set, we apply an additional human verification stage on top of the textual and visual LLM-based checks described in the main paper. Four trained annotators manually inspect every QA pair following a structured guideline that checks (i) whether the bounding boxes of visual elements are correctly specified, (ii) whether visual citations are accurate and genuinely necessary to support the answer, and (iii) whether the supporting text is consistent with the cited visual elements and free of hallucinations. QA pairs that fail any of these checks are discarded. In total, 106 answerable and 1 unanswerable QA pairs are removed during this process.

D. Details of M-GroSE

GroUSE [4] was originally designed to evaluate text-grounded answers, focusing solely on textual contexts. For our setting, we first refine the released GroUSE evaluation pipeline and prompt, as we found that some components led to suboptimal

The religious landscape of Rome has undergone profound transformation from ancient times to the modern era, shifting from a polytheistic society to one predominantly Christian. This evolution reflects broader historical changes in the Mediterranean world and Europe, with significant implications for Rome's cultural and social development.

1. Religious Composition of Modern Rome (2015)

According to [chart_1], the religious demographics of Rome in 2015 showed:

- Catholicism dominates the religious landscape at 82.0%
- Other religions or non-religious individuals constitute 8.7%
- Eastern Orthodoxy represents 4.0% of the population
- Islam accounts for 3.8%
- Protestant denominations make up 0.8%
- Judaism represents 0.7% of Rome's population

This data reveals that modern Rome is overwhelmingly Christian, with the Catholic Church being the dominant religious institution. The presence of Eastern Orthodoxy and Protestantism demonstrates the diversification within Christianity itself.

...

2.2. Religious Practices and Institutions

Ancient Roman religion was highly ritualistic and integrated with civic life. [table_1] indicates that religious practices included libations, votum (vows), temple worship, festivals, and funerary practices. The Roman state maintained various priesthoods including the Pontifices, Augures, Vestales (Vestal Virgins), Flamines, and others to oversee religious affairs.

The religion was largely concerned with interpreting divine messages (auguries) through natural occurrences (omens), initially relying on Etruscan augurs for this purpose.[1]

2.3. Imperial Cult and Deification

During the Imperial period, the cult of the emperor became an important element of Roman religion. As seen in [figure_4], emperors like Marcus Aurelius performed sacrifices at temples, and some emperors were themselves deified after death. [table_1] specifically lists Julius Caesar and Augustus among deified leaders, showing how political power became intertwined with religious authority.

...

Figure 2. Answer example of VinQA. The blue part indicates the citations of either the page numbers or the visual element identifiers of the charts, tables, and within the context.

scoring behavior or were not directly suited to VinQA. Moreover, VinQA introduces a multimodal context where visual elements (e.g., tables, charts, figures) are interleaved within the retrieved pages, and the generated answers explicitly cite these visual elements. Therefore, we further adapt GroUSE’s evaluation methodology to this multimodal setting and refer to the resulting framework as Multimodal Grounded QA Scoring of Evaluators (M-GroSE).

Unlike the GroUSE pipeline, which uses an integrated scoring workflow, we directly evaluate all answerable QA instances along three criteria using an LLM-judge rubric based on GPT-5-mini: (1) Relevancy — whether the answer appropriately addresses the question; (2) Completeness — whether it incorporates all necessary information from the multimodal context; and (3) Faithfulness — whether it remains consistent with the cited sources without hallucination. We modify the evaluation prompt so that each criterion is assessed at the citation sentence level, enabling fine-grained judgments that consider every part of a long-form answer. The evaluation prompts for these three criteria are presented in Appendix E.4.

GroUSE also includes a fourth criterion, Usefulness, which measures the helpfulness of an answer when the context is insufficient (i.e., adversarial or unanswerable settings). In VinQA, unanswerable questions are explicitly defined, and models

are instructed not to provide additional speculative content. Therefore, Usefulness is not applicable. Instead, we separately evaluate the model’s behavior on unanswerable questions using the Unanswerability F1 metric, which measures whether the model correctly abstains when the provided context lacks sufficient evidence.

Finally, to extend the evaluation from text-only to multimodal grounding, we textualize both the textual and visual components of the context and adapt the LLM-judge prompts to evaluate not only textual reasoning but also the citations and explanations accompanying interleaved visual elements. This adaptation allows M-GroSE to jointly assess the groundedness of answers in multimodal document QA.

E. Prompt Template

E.1. Prompts for Data generation

Figures 3–7 show the prompts used during the VinQA dataset generation process.

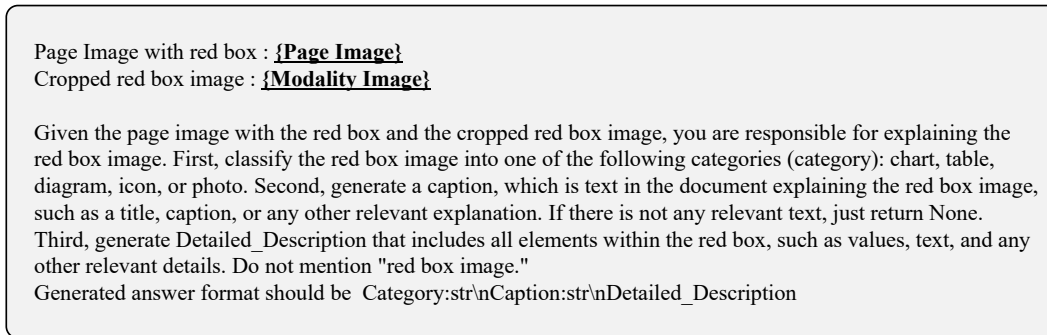


Figure 3. The prompt for generating visual-element’s class label, caption, and description.

E.2. Prompts for Data verification

Figures 8, 9, and 10 show the prompts used for data verification.

E.3. Prompt examples for encoding method

Figure 11 shows an example prompt input for the *Page Encoding* method, while Figure 12 corresponds to the *Modality Encoding* method.

E.4. Prompts for M-GroSE

Figures 13, 14, and 15 show the prompts used for M-GroSE.

E.5. Prompts for Visual G-Eval

Figures 16–21 show the prompts used for Visual G-Eval.

[Document]

{Input context}

[Instruction]

Please create 5 questions based on the document above, following these guidelines:

1. Reflecting Core Document Content

- You must examine the entire document and create questions related to its core content.
- Questions must be answerable by synthesizing information directly found throughout the provided document.
- Do not create questions that require interpretation or inference not explicitly answered within the document.

2. Multi-page Based Questions

- Prefer questions that can be answered by integrating information from multiple pages within the document rather than those limited to content on a single page.
- It is not necessary to utilize the entire content of the document, but questions reflecting information from multiple pages are preferable.
- Avoid generate questions that use 'and' to ask for two independent pieces of information within a single question as much as possible. (e.g., "What is ~, and how is ~?")

3. Questions Involving Various Modalities (Charts, Tables, Figures)

- If the document contains charts, tables, figures, you must create at least one question whose answer incorporates information from each modality type.
- Create questions that integrate multiple modalities and contexts within the document rather than focusing on only one modality.

4. Question Format

- Generate a total of 5 questions.
- Write each question in a numbered list format, with each question as a single sentence without newlines.
- Do not generate any explanations or statements outside of the question list.
- **The questions must not point to specific parts(e.g., [Document],[Page],[table_1],[chart_1],[figure_1],[Context]) of the provided document; do not include phrases such as "According to the document," "According to the table," "Based on the chart," "Based on figure," "Based on the document," "as mentioned in document," "shown in figure," "discussed in the document" in questions.**
- Please generate questions by referring to the examples below.
- The question should resemble a search query on GPT without any given context. GPT should be able to retrieve the relevant context using only this question.

Example 1: **{Example question}**

Example 2: **{Example question}**

Example 3: **{Example question}**

Example 4: **{Example question}**

Example 5: **{Example question}**

Example 6: **{Example question}**

Example 7: **{Example question}**

Example 8: **{Example question}**

Figure 4. The prompt for Question generation. The text marked with both bold and underline represents the parts provided as prompt inputs.

[Questions]

{Question}

Your role is to determine whether the given machine generated questions are suitable for a Retrieval-Augmented Generation (RAG) setting.

- The questions should not contain unclear and ambiguous information. For example, a question like "What datasets were utilized in the NLP experiments, and what are some examples of the custom prompts designed for tasks within these datasets?" is not clear, because the term "NLP experiment" is too generic and lacks specificity.
- Since this is a RAG environment, the questions must be self-contained. Question should be formulated as if it were a standalone search query submitted to GPT, without any accompanying context.
- Questions must not contain phrases that indicate something based on unknown context, such as "based on the document," "according to the table," or "provided chart."

For each question, generate a reasoning chain to assess whether it meets the above each criteria. Then, determine whether the question is appropriate using the following format:

[Final Response]

[Appropriate] or [Not appropriate]

Figure 5. The prompt for Question filtering.

[Document]
{Input context}

[Question]
{Question}

[Instruction]

Please structure and generate a detailed answer to the given question, referring to the provided document as much as possible.

1. Answer Guidelines

- Generate a detailed and information-rich response by including as much specific information from the provided document as possible.
- ****BUT, Do not include content that deviates from the intent of the question.****
- The response must be written in a professional tone and remain consistent throughout.
- In the entire response, "[Page]", "[Context]", "Caption" and "Description" must not be mentioned in any sentence.

2. Referencing Document Content

- For sentences that use context from the provided document, cite the page number (e.g., [1], [2]..) at the end of the sentence where the information is used:
 - * Apply this only to key sentences, and if the same citation is repeated, include it only in the last instance.
 - * Do not cite interpretations or insights that do not exist in the document.
 - * If you need to cite more than two pages in one sentence, you need to attach them consecutively as separate tags, as in the example below.
 - * Example citation format: This sentence is based on information from a specific page in the document.[1][2]
- For sentences that use modality information (e.g., chart, table, figure), directly mention the corresponding tag (e.g., [chart_1], [table_2], [figure_3]..) within the sentence:
 - * Modality tags should only be used as subjects or objects within a sentence and should not appear alone after a sentence.
 - * Avoid repeating citations for the same tag across consecutive sentences.
 - * Example citation format: According to [chart_1], the response is structured in this way.

3. Answer Format

- Structure your response into introduction, main body, and conclusion.
- Write in Markdown format to maximize readability.
- Do not generate an overly long response by elaborating excessively on the answer to the question.
- The introduction should be a short paragraph summarizing the key points of the response to the given question:
 - * Start directly with the introduction without creating a separate title for it.
- The main body should construct the overall content of the response:
 - * Divide the key points into sections for clarity and readability.
 - * Sections should be divided only to an extent that enhances readability.
 - * Avoid duplication of content across different sections.
 - * Use ## for main sections with numbered formatting (e.g., "## 1.")
 - * Use ### for subsections with decimal formatting (e.g., "### 1.1").
 - * Section titles must not include tags such as [chart_1] or [table_1] or [figure_1].
 - * If lists are required under subsections, use * for bullet points.
- The conclusion should summarize how the key points of the response align with the question in a short paragraph:
 - * Do not create a separate title for the conclusion.
 - * Do not mention any tags ([chart_1], [table_1], [figure_1], [Page], [Context], Caption, Description) from the provided document in the conclusion.

Figure 6. The prompt for Answer generation (Answerable QA).

[Document]
{Input context}

[Question]
{Question}

[Instruction]

****Step 1:** determine whether you can generate a detailed answer to the given question based on the given document. If even a part of the question cannot be answered completely, the response must be "[Unanswerable]".**

1. Carefully review the given document to check if it contains information relevant to the question.
2. When generating an answer to the question, determine whether a complete answer can be provided solely based on the content of the given document.
3. If question is unanswerable, first generate a brief paragraph explaining why it is unanswerable. (Starting with : "Unanswerable explanation: ")
4. After the explanation, write "Result: [Unanswerable]". Do not generate any other sentences afterward.
5. If the question is unanswerable, do not proceed to Step 2 and terminate at Step 1.
6. If the question is answerable, write "Result: [Answerable]", and proceed to Step 2.

****Step 2:** If question is not unanswerable, please structure and generate a detailed answer to the given question, referring to the provided document as much as possible. Generate answer following the guidelines below.**

1. Answer Guidelines
 - Generate a detailed and information-rich response by including as much specific information from the provided document as possible.
 - ****BUT**, Do not include content that deviates from the intent of the question.**
 - The response must be written in a professional tone and remain consistent throughout.
 - In the entire response, "[Page]" and "[Context]" must not be mentioned in any sentence.
2. Referencing Document Content
 - For sentences that use context from the provided document, cite the page number (e.g., [1], [2]..) at the end of the sentence where the information is used:
 - * Apply this only to key sentences, and if the same citation is repeated, include it only in the last instance.
 - * Do not cite interpretations or insights that do not exist in the document.
 - * If you need to cite more than two pages in one sentence, you need to attach them consecutively as separate tags, as in the example below.
 - * Example citation format: This sentence is based on information from a specific page in the document. [1][2]
 - For sentences that use modality information (e.g., chart, table, figure), directly mention the corresponding tag (e.g., [chart_1], [table_2], [figure_3]..) within the sentence:
 - * Modality tags should only be used as subjects or objects within a sentence and should not appear alone after a sentence.
 - * Avoid repeating citations for the same tag across consecutive sentences.
 - * Example citation format: According to [chart_1], the response is structured in this way.
3. Answer Format
 - Start the answer content with the phrase "[Generated Answer]:\n" and write the response below it.
 - Structure your response into introduction, main body, and conclusion.
 - Write in Markdown format to maximize readability.
 - Do not generate an overly long response by elaborating excessively on the answer to the question.
 - The introduction should be a short paragraph summarizing the key points of the response to the given question:
 - * Start directly with the introduction without creating a separate title for it.
 - The main body should construct the overall content of the response:
 - * Divide the key points into sections for clarity and readability.
 - * Sections should be divided only to an extent that enhances readability.
 - * Avoid duplication of content across different sections.
 - * Use ## for main sections with numbered formatting (e.g., "## 1.")
 - * Use ### for subsections with decimal formatting (e.g., "### 1.1").
 - * Section titles must not include tags such as [chart_1] or [table_1] or [figure_1].
 - * If lists are required under subsections, use * for bullet points.
 - The conclusion should summarize how the key points of the response align with the question in a short paragraph:
 - * Do not create a separate title for the conclusion.
 - * Do not mention any tags ([chart_1], [table_1], [figure_1], [Page], [Context]) from the provided document in the conclusion.

Figure 7. The prompt for Answer generation (Unanswerable QA).

Question: {Question}

Context: {Input context}

Answer: {Answer}

Given a Question, Context, and Answer, evaluate the following criteria:

1. Is there any citation ({Citation tag list}) incorrectly reference the corresponding information in the context? Verify each citation in the answer.
2. Does the answer include any statements that are not supported by the context or contradict it? Verify each statement in the answer.
3. Is there any information in the context that is relevant to the question but missing from the answer? Verify each page in the context.
4. Does the answer include content that is not directly related to the question or unnecessarily detailed? Verify each section in the answer.

First generate rationales for each criterion. And then respond to each number using the following format:

[Final Answer]

[1-Yes] or [1-No]

[2-Yes] or [2-No]

[3-Yes] or [3-No]

[4-Yes] or [4-No]

Figure 8. The prompt for Textual verification (Answerable QA). Citation tag list indicates all page number tags and the chart, table, and figure modality tags that appear in the answer.

Question: {Question}

Context: {Input context}

Answer: {Answer}

Given a Question, Context, and Answer, evaluate the following criteria:

1. Does the context contain enough precise information to answer the question definitively?
 - If YES, the pair is mislabeled as unanswerable.

First generate rationales for a criterion. And then respond to each number using the following format:

[Final Answer]

[1-Yes] or [1-No]

Figure 9. The prompt for Textual verification (Unanswerable QA).

Question: {Question}

Context: {Input context}

Answer: {Answer}

Given a Question, Context, and Answer, evaluate the following criteria:

1. Did the response fail to make appropriate use of any image in the context (\$image_list) that is relevant to the question and therefore should have been included in the answer? For each image in the context, verify whether it is relevant to the question and whether any relevant image was properly utilized in the answer.
2. Do any sentences in the answer that include a citation (\$citation_list) fail to match the facts shown in the image corresponding to each citation within the context? Verify each citation in the answer.

First generate rationales for each criterion. And then respond to each number using the following format:

[Final Answer]

[1-Yes] or [1-No]

[2-Yes] or [2-No]

Figure 10. The prompt for Visual verification. For visual verification, the input context differs from textual verification: any chart, figure, or table modality found in the context is replaced with its image, so the resulting context contains visual elements interleaved with the text.

```

[Page]:[1]
[figure_1]:<bbox>[602, 889, 940, 979]</bbox>
[chart_1]:<bbox>[731, 223, 939, 431]</bbox>
[Context]:<image>

[Page]:[2]
[figure_2]:<bbox>[673, 517, 934, 648]</bbox>
[figure_3]:<bbox>[677, 834, 930, 979]</bbox>
[Context]:<image>

[Page]:[3]
[Context]:<image>

[Page]:[4]
[figure_4]:<bbox>[762, 298, 914, 498]</bbox>
[table_1]:<bbox>[734, 228, 943, 977]</bbox>
[Context]:<image>

[Page]:[5]
[table_2]:<bbox>[672, 804, 939, 979]</bbox>
[Context]:<image>

```

Question:How does the religious composition of modern Rome in 2015 compare to the primary religions practiced in ancient Rome?

Find and use information related to the question in the given document to write an answer. If a page in the document contains a chart, table, or figure, the element's location(i.e., bounding box) on that page is provided. If you use information from a chart, table or figure in the given document, write an answer by directly mentioning the corresponding tag (e.g. [chart_1], [table_2], [figure_3]). Write an answer that is clear and systematic, and emphasizes key information.

Figure 11. The prompt example for *Page Encoding* method. The <image> part refers to the corresponding page image input that is converted into visual tokens.

[Page]:[1]
 [figure_1]:<image>
 [chart_1]:<image>
 [Context]:Religion in Rome
 The Religio Romana (literally, the "Roman Religion") constituted the major religion of the city in antiquity. The first gods held sacred by the Romans were Jupiter, the highest, and Mars, the god of war, and father of Rome's twin founders, Romulus and Remus, according to tradition. ...

[Page]:[2]
 [figure_2]:<image>
 [figure_3]:<image>
 [Context]:The image is a screenshot of a Wikipedia page titled "Culture of Rome." Here is the extracted text: ...

[Page]:[3]
 [Context]:The Western religions are the religions that originated within Western culture, which are thus historically, culturally, and theologically distinct from Eastern, African and Iranian religions. The term Abrahamic religions (Christianity, Judaism and Islam) is often used instead of using the East and West terminology, as these originated in the Middle East. ...

[Page]:[4]
 [figure_4]:<image>
 [table_1]:<image>
 [Context]:The text extracted from the image is as follows: ...

[Page]:[5]
 [table_2]:<image>
 [Context]:Anti-Judaism
 From Wikipedia, the free encyclopedia
 Anti-Judaism describes a range of historic and current ideologies which are totally or partially based on opposition to Judaism, on the denial or the abrogation of the Mosaic covenant, and the replacement of Jewish ...

Question:How does the religious composition of modern Rome in 2015 compare to the primary religions practiced in ancient Rome?
 Find and use information related to the question in the given document to write an answer. If you use information from a chart, table or figure in the given document, write an answer by directly mentioning the corresponding tag (e.g. [chart_1], [table_2], [figure_3]). Write an answer that is clear and systematic, and emphasizes key information.

Figure 12. The prompt example for *Modality Encoding* method. The <image> part refers to the input corresponding to each visual element identifier, which is converted into visual tokens.

[TASK]

Task: Answer Relevancy – Per-**cited-sentence** claim extraction & relevancy scoring

You will evaluate two answers (Answer 1 and Answer 2) ONLY for how well their **cited content** addresses the user's request. Truthfulness is NOT considered; evaluate adequacy/alignment to the request. Produce the requested JSON.

Rules (must follow):

- Relevancy measures alignment to the user request, not correctness.
- Evaluate ONLY sentences (or clauses) that include at least one explicit citation tag in the required style: [i], [chart_i], [table_i], [figure_i].
- Treat paraphrases/semantic equivalents as matches to the request intent.

[PROCEDURE]

1) Request intent mining.

Parse the user request and list the core intent facets as short items (e.g., entities, attributes, time ranges, operations, constraints).

2) Per-answer **cited-sentence** extraction (claims).

For each answer, extract atomic claims from sentences/clauses that contain a citation tag. Label them A1-1, A1-2, ... for Answer 1 and A2-1, A2-2, ... for Answer 2.

- Ignore uncited sentences for relevancy scoring.

3) Claim-to-intent mapping (run separately for Answer 1 and Answer 2).

For each cited claim, judge its relation to the request facets:

- Relevant: directly addresses one or more facets of the request (entity + attribute [+ time if present] match).
- Partially relevant: tangential/background context that is helpful but not directly answering a facet.
- Irrelevant: off-topic or addressing different entities/timeframes/constraints.

(No truthfulness judgment—only alignment to the request.)

4) Compute relevancy ratio and grade.

Let $R = \# \text{Relevant}$, $PR = \# \text{Partially relevant}$, $IR = \# \text{Irrelevant}$, $T = \text{total } \# \text{cited claims considered}$.

Score $S = (R + 0.5 * PR) / T$. (If $T = 0$, set $S = 0$.)

Map S to relevancy grade:

- 5 if $S \geq 0.8$
- 4 if $0.6 \leq S < 0.8$
- 3 if $0.4 \leq S < 0.6$
- 2 if $0.2 \leq S < 0.4$
- 1 if $0 \leq S < 0.2$

[EVALUATION INSTRUCTIONS]

Your output must be JSON. In the justification for each answer, first list the request facets, then enumerate the **cited** claims with their labels (Relevant/Partial/Irrelevant), and finally show the relevancy score calculation.

Output JSON format (must follow exactly):

```
{
  "answer_1": {
    "answer_relevancy_justification": "...",
    "answer_relevancy": Y
  },
  "answer_2": {
    "answer_relevancy_justification": "...",
    "answer_relevancy": Y
  }
}
```

Where $Y \in \{1,2,3,4,5\}$.

[SAMPLE]

User request: **{{ input }}**

[TO EVALUATE]

Answer 1: **{{ expected_output }}**

Answer 2: **{{ actual_output }}**

[TO EVALUATE]

Figure 13. The prompt used for evaluating relevancy in M-GroSE.

[TASK]
Task: Grounded Question Answering – Per-reference extraction & coverage scoring
You will evaluate two answers (Answer 1 and Answer 2) ONLY based on the provided references. Follow the steps strictly and produce the requested JSON.

Rules (must follow):

- Consider only information present in the references. Ignore any outside knowledge.
- Unrelated information in an answer does not affect completeness.
- Incorrect statements in the answer do not reduce completeness if they are irrelevant to the referenced facts.
- If the references contain no information that precisely answers the user request, set completeness to 'null'.
- Treat paraphrases/semantic equivalents as matches; require that the answer explicitly asserts the fact.

[PROCEDURE]

- 1) Per-reference relevance mining.
For each reference R_i , extract atomic facts relevant to the user request.
 - Represent each fact as a short clause (IDs: R_i-1 , R_i-2 , ...).
 - Include table/figure/chart items as atomic facts too.
 - Skip content in R_i that does not help answer the request.
- 2) Global fact list.
Concatenate all extracted facts into a single checklist F . Each fact keeps its origin R_i .
If F is empty \rightarrow completeness is 'null'.
- 3) Answer coverage check (run separately for Answer 1 and Answer 2).
For each fact in F , mark it as:
 - Covered: the answer states the fact (allowing paraphrase) AND uses the required citation style [i], [chart_i], [table_i], or [figure_i] for that fact's source.
 - Partial: the answer conveys only a subset of a composite fact.
 - Missing: otherwise.
If an answer contradicts a fact, mark it Missing.
- 4) Compute coverage ratio and grade.
Let $C = \# \text{Covered}$, $P = \# \text{Partial}$, $T = \text{total \#facts}$.
Score $S = (C + 0.5 * P) / T$.
Map S to completeness grade:
 - 5 if $S \geq 0.8$
 - 4 if $0.6 \leq S < 0.8$
 - 3 if $0.4 \leq S < 0.6$
 - 2 if $0.2 \leq S < 0.4$
 - 1 if $0 \leq S < 0.2$
If F is empty \rightarrow completeness = null.

[EVALUATION INSTRUCTIONS]
Your output must be JSON. In the justification, first summarize the mined facts F (grouped by reference), then report which facts are Covered/Partial/Missing for the answer and show the coverage calculation.

Output JSON format (must follow exactly):

```
{
  "answer_1": {
    "completeness_justification": "...",
    "completeness": X
  },
  "answer_2": {
    "completeness_justification": "...",
    "completeness": X
  }
}
```

Where $X \in \{1,2,3,4,5\}$ or null.

[SAMPLE]
List of references :
{%~ for ref_id, ref_text in contexts %}
Reference **{{ ref_id }}**: **{{ ref_text }}**
{%~ endfor %}

User request: **{{ input }}**

[TO EVALUATE]
Answer 1: **{{ expected_output }}**
Answer 2: **{{ actual_output }}**
[/TO EVALUATE]

Figure 14. The prompt used for evaluating completeness in M-GroSE.

[TASK]
Task: Faithfulness – Per-cited-sentence verification & graded scoring
Evaluate two answers (Answer 1 and Answer 2) ONLY for faithfulness to the provided references. Check whether statements THAT CITE a source are correctly attributed and not distorted. Produce the requested JSON.

Rules (must follow):

- Use ONLY the provided references; ignore any outside knowledge.
- Evaluate ONLY sentences that include an explicit citation tag in the required style: [i], [chart_i], [table_i], [figure_i].
- Paraphrases are acceptable if they preserve meaning and scope (no invented numbers, no altered relationships, no broadened/shifted scope).
- If a sentence contains multiple citations, ALL cited sources must be correct for the sentence to pass.

[PROCEDURE]

- 1) Extract cited sentences.
From the answer, list each sentence (or clause) that includes at least one citation tag. Ignore uncited sentences for faithfulness scoring.
- 2) Per-sentence verification.
For each cited sentence S:
- Criterion 1 (Attribution): Do the cited tags exist and correspond to the intended reference(s)? If multiple tags are present, each must be valid for the content asserted in S.
- Criterion 2 (Agreement): Does S faithfully convey what the cited reference(s) state (no invented numbers, no altered relationships, no broadened/shifted scope)? Paraphrase is fine; distortion is not.
- 3) Decide pass/fail for each sentence.
- A sentence PASSES if Criterion 1 = OK and Criterion 2 = OK for ALL of its citations.
- Otherwise the sentence FAILS.
- 4) Aggregate to a 1–5 grade.
Let P = #PASS sentences, F = #FAIL sentences, $T = P + F$ (number of cited sentences considered).
Define faithfulness score $S = P / T$. (If $T = 0$, set $S = 0$)
Map S to the faithfulness grade:
5 if $S \geq 0.8$
4 if $0.6 \leq S < 0.8$
3 if $0.4 \leq S < 0.6$
2 if $0.2 \leq S < 0.4$
1 if $0 \leq S < 0.2$

[EVALUATION INSTRUCTIONS]
Your output must be JSON. For each answer, list the cited sentences and explain results for both criteria. Then show P, F, T, S and the final grade.

Output JSON format (must follow exactly):

```
{
  "answer_1": {
    "content_analysis_sentence_by_sentence": [
      {
        "sentence": "...",
        "criterion_1": "...",
        "criterion_2": "...",
        "pass": true/false
      }
    ],
    "faithfulness_justification": "P=..., F=..., T=..., S=..., grade=..",
    "faithfulness": Y
  },
  "answer_2": {
    "content_analysis_sentence_by_sentence": [
      {
        "sentence": "...",
        "criterion_1": "...",
        "criterion_2": "...",
        "pass": true/false
      }
    ],
    "faithfulness_justification": "P=..., F=..., T=..., S=..., grade=..",
    "faithfulness": Y
  }
}
```

Where $Y \in \{1,2,3,4,5\}$.

[SAMPLE]
List of references :
{% for ref_id, ref_text in contexts %}
Reference {{ ref_id }}: {{ ref_text }}
{% endfor %}

[TO EVALUATE]
Answer 1: {{ expected_output }}
Answer 2: {{ actual_output }}
[TO EVALUATE]

Figure 15. The prompt used for evaluating answer faithfulness in M-GroSE.

[TASK]

Task: Image Effectiveness – Per-cited-image usage analysis & graded scoring

Evaluate two answers (Answer 1 and Answer 2) ONLY for how effectively each cited image and its supporting text are used to help answer the user's query.

The core focus of this evaluation is the visual evidence:

- how appropriate and informative the visual content of the cited image is, and
- how well the supporting text actually uses that visual content to support the answer.

Your goal is to evaluate, for each cited image as used in the answer, whether the image and its supporting text are relevant to what the question asks and whether they help explain the answer.

[INPUT FORMAT]

You are given:

- A Query (user question).
- Two Answers (Answer 1 and Answer 2).

Answer Input Format

Each answer consists of text interleaved with images: throughout the answer, an image is inserted, followed by supporting text that cites it using tags such as [chart_i], [table_i], and [figure_i].

For each tag:

- The tag refers to a specific visual element (e.g., a chart, table, or figure), provided together with the answer as an image.
- The supporting text immediately following the image explains, interprets, or uses that visual element to answer the query.

You may assume that the corresponding images are available to you (as a multimodal model) whenever these tags appear.

[GENERAL RULES] (must follow)

- Evaluate ONLY usage linked to explicit image citation tags: [chart_i], [table_i], [figure_i] that appear in the answer text.
- The primary evaluation target is the visual content itself:
- Judge how appropriate and informative the visual content of the cited image is for answering the query.
- The supporting text is evaluated only in how well it leverages the visual element:
- Focus on whether the supporting text uses what is visually shown (values, trends, entries, layout, etc.) to help the reader understand the answer.
- Do NOT judge overall factual correctness of the entire answer; focus on:
 - how useful the visual content is for the answer, and
 - how effectively the supporting text uses that visual content.
- If an answer contains no cited image tags, mark it as having no cited images and do not fabricate any scores.

[SCORING CRITERIA OF EFFECTIVENESS]

For each cited image usage, assign an effectiveness score from 1 to 5, based on two criteria:

Criterion 1 – Visual element helpfulness

- Does the visual content of the cited image contribute to understanding the answer or provide information that is part of the answer?
- Would the answer be weaker or less clear without this visual content?

Criterion 2 – Supporting text helpfulness

- Does the supporting text actively use the visual aspects of the image (e.g., specific numbers, trends, categories, regions, comparisons) to support the answer?
- Does it correctly interpret and connect the visual content to the query?

Then synthesize both criteria into a single 1–5 score per image usage:

- 1 – Harmful
 - The visual content does not support the answer and/or the supporting text misuses or misinterprets it.
 - Overall, the cited image and text make the answer more confusing or misleading.
- 2 – Irrelevant
 - The visual content is mostly unrelated to the answer, or the supporting text barely uses the visual information.
 - The image could be removed with almost no impact on understanding.
- 3 – Partially Effective
 - The visual content has some relation to the answer, and the supporting text uses it in a limited or vague way.
 - The image provides some help, but the contribution to understanding is weak or incomplete.
- 4 – Mostly Effective
 - The visual content is clearly relevant, and the supporting text makes good use of specific visual information.
 - Together, they noticeably help the reader understand or verify the answer.
- 5 – Highly Effective
 - The visual content is crucial for understanding the answer, and the supporting text strongly and accurately grounds the explanation in the visual details.
 - The answer would be significantly worse or unclear without this image and its explanation.

...

Figure 16. The Visual G-Eval prompt to assess Effectiveness. (Part 1)

...

[PROCEDURE]

1) Identify cited images per answer.
For each answer:
- Scan the text and find all image citation tags: [chart_i], [table_i], [figure_i].
- For each tag, define one image usage as:
- image_tag: the tag itself (e.g., "[chart_1]"), and
- local_answer_context: the nearby sentence(s) that describe, reference, or depend on that image.
- If there are no such tags, mark the answer as having no cited images.

2) Interpret each image usage.
For each image_tag in an answer:
- Inspect the visual content of the corresponding image:
- What does it show (structure, values, patterns, layout, etc.)?
- Read the supporting text around the tag:
- How does it reference and use that visual content to answer the query?

3) Evaluate each image usage using the two criteria.
For each image_tag:
- Evaluate Criterion 1 – Visual element helpfulness:
- Is the visual content itself useful and relevant for understanding or supporting the answer?
- Evaluate Criterion 2 – Supporting text helpfulness:
- Does the supporting text actually leverage the visual content (not just mention the tag) to help answer the query?

Then:
- Combine both criteria into a single 1–5 effectiveness score.
- Write a brief justification that explicitly mentions:
- how helpful the visual content is, and
- how well the supporting text uses that visual content.

[EVALUATION INSTRUCTIONS]

Your output must be JSON. For each answer:

- List every cited image usage with:
- "image_tag" – the tag string (e.g., "[chart_1]"),
- "reason" – a brief explanation referencing visual element helpfulness and supporting text helpfulness,
- "score" – an integer from 1 to 5.
- You do not need to compute any aggregated score per answer; scoring is done per cited image.
- If an answer has no cited image tags:
- Set "per_image_analysis": [],
- Set "no_cited_images": true.

Output JSON format (must follow exactly):

```
{
  "answer_1": {
    "per_image_analysis": [
      {
        "image_tag": "[chart_1]",
        "reason": "...",
        "score": X
      }
    ],
    "no_cited_images": false
  },
  "answer_2": {
    "per_image_analysis": [
      {
        "image_tag": "[table_3]",
        "reason": "...",
        "score": X
      }
    ],
    "no_cited_images": false
  }
}
```

Where:
- $X \in \{1, 2, 3, 4, 5\}$ is the per-image usage effectiveness score.

If an answer has no cited image tags, then:
"per_image_analysis": [],
"no_cited_images": true

[TO EVALUATE]
Query: **{{ query }}**

Answer 1: **{{ answer_1 }}**
Answer 2: **{{ answer_2 }}**
[TO EVALUATE]

Figure 17. The Visual G-Eval prompt to assess Effectiveness (Part 2).

[TASK]

Task: Contextual Placement – Per-cited-image position analysis & graded scoring

Evaluate two answers (Answer 1 and Answer 2) ONLY for how contextually appropriate the position of each cited image and its supporting text is within the answer.

You must check, for each cited visual element:

- whether the image and its supporting text appear at a natural, contextually appropriate place in the answer, and
- whether their placement helps the flow and coherence of the explanation, rather than disrupting or confusing it.

You are NOT evaluating factual correctness, faithfulness to the image, or overall helpfulness here – ONLY whether the image + supporting text are placed in the right context at the right moment in the answer.

[INPUT FORMAT]

You are given:

- A Query (user question).
- Two Answers (Answer 1 and Answer 2).

Answer Input Format

Each answer consists of text interleaved with images: throughout the answer, an image is inserted, followed by supporting text that cites it using tags such as [chart_i], [table_i], and [figure_i].

For each tag:

- The tag refers to a specific visual element (e.g., a chart, table, or figure), provided together with the answer as an image.
- The supporting text immediately following the image explains, interprets, or uses that visual element to answer the query.

You may assume that the corresponding images are available to you (as a multimodal model) whenever these tags appear.

[GENERAL RULES] (must follow)

- Evaluate ONLY usage linked to explicit image citation tags: [chart_i], [table_i], [figure_i] that appear in the answer text.
- The target of evaluation is CONTEXTUAL PLACEMENT, not correctness or usefulness:
 - Does the image appear at the point where the answer is actually talking about the information it contains?
 - Is the supporting text attached to the right part of the answer, or would it make more sense earlier, later, or elsewhere?
- Consider the surrounding context before and after each image:
 - Does the answer properly introduce why the image is relevant before showing it?
 - Does the answer refer back to or build on the image in a logical way after it appears?
- Ignore whether the content of the image is factually correct or helpful; focus on where and how it is inserted into the narrative.
- If an answer contains no cited image tags, mark it as having no cited images and do not fabricate any scores.

[SCORING CRITERIA OF CONTEXTUAL PLACEMENT]

For each cited image usage, assign a contextual placement score from 1 to 5, based on two criteria:

Criterion 1 – Local contextual fit

- Does the image appear exactly where the answer is discussing the concept, data, or detail that the image represents?
- Is there a clear, smooth transition from the preceding text into the image and its supporting text?

Criterion 2 – Flow and coherence

- Does the placement of the image and its supporting text maintain or improve the logical flow of the answer?
- Does it avoid disrupting the narrative (e.g., appearing too early, too late, or in a completely unrelated section)?

Use these criteria to assign one score per image usage:

- 1 – Very Poor Placement

- The image and its supporting text appear in a clearly wrong or confusing location.
- They interrupt the flow, are not connected to the surrounding discussion, or refer to content that is not yet introduced or already finished.

- 2 – Poor Placement

- The image is loosely related to the surrounding text but appears at an awkward time.
- The reader would likely be confused about why the image appears at that point or would expect it elsewhere in the answer.

- 3 – Mixed / Moderately Appropriate

- The placement is somewhat reasonable, but not ideal.
- The image and supporting text relate to the topic, but the transitions or ordering (introduction, explanation, follow-up) are weak or slightly misplaced.

- 4 – Mostly Appropriate

- The image appears in a generally good location relative to the explanation.
- The surrounding text largely prepares for and follows up on the image in a coherent way, with only minor issues.

- 5 – Highly Appropriate

- The image and its supporting text are placed at an excellent, contextually natural point in the answer.
- The answer smoothly introduces why the image is relevant, shows it at the right moment, and then immediately uses it in the surrounding text to continue the explanation.

...

Figure 18. The Visual G-Eval prompt to assess Position (Part 1).

...

[PROCEDURE]

1) Identify cited images per answer.
 For each answer:
 - Scan the text and find all image citation tags: [chart_i], [table_i], [figure_i].
 - For each tag, define one image usage as:
 - image_tag: the tag itself (e.g., "[chart_1]"), and
 - local_answer_context: the nearby sentence(s) before and after the image that introduce, cite, and follow up on that image.
 - If there are no such tags, mark the answer as having no cited images.

2) Analyze contextual placement.
 For each image_tag in an answer:
 - Examine the preceding text:
 - Does it introduce the concept or question that the image is supposed to support?
 - Examine the supporting text and the subsequent text:
 - Does it immediately use the image in a way that fits the ongoing explanation?
 - Does the answer move on naturally after the image, or does the image feel dropped in without connection?

3) Evaluate each image usage using the two criteria.
 For each image_tag:
 - Evaluate Criterion 1 – Local contextual fit:
 - Does the image appear exactly where it makes sense in the local narrative?
 - Evaluate Criterion 2 – Flow and coherence:
 - Does its placement preserve or enhance the logical flow of the answer?

Then:
 - Combine both criteria into a single 1–5 contextual placement score.
 - Write a brief justification that explicitly mentions:
 - how well the image fits the local context, and
 - how its placement affects the flow and coherence of the answer.

[EVALUATION INSTRUCTIONS]

Your output must be JSON. For each answer:

- List every cited image usage with:
 - "image_tag" – the tag string (e.g., "[chart_1]"),
 - "reason" – a brief explanation referencing local contextual fit and flow/coherence,
 - "score" – an integer from 1 to 5 (contextual placement score).
- You do not need to compute any aggregated score per answer; scoring is done per cited image.
- If an answer has no cited image tags:
 - Set "per_image_analysis": [],
 - Set "no_cited_images": true.

Output JSON format (must follow exactly):

```
{
  "answer_1": {
    "per_image_analysis": [
      {
        "image_tag": "[chart_1]",
        "reason": "...",
        "score": X
      }
    ],
    "no_cited_images": false
  },
  "answer_2": {
    "per_image_analysis": [
      {
        "image_tag": "[table_3]",
        "reason": "...",
        "score": X
      }
    ],
    "no_cited_images": false
  }
}
```

Where:
 - $X \in \{1, 2, 3, 4, 5\}$ is the per-image contextual placement score.

If an answer has no cited image tags, then:
 "per_image_analysis": [],
 "no_cited_images": true

[TO EVALUATE]
 Query: `{{ query }}`

Answer 1: `{{ answer_1 }}`
 Answer 2: `{{ answer_2 }}`
 [TO EVALUATE]

Figure 19. The Visual G-Eval prompt to assess Position (Part 2).

Task: Supporting-Text Faithfulness – Per-cited-image consistency check & graded scoring

You must check, for each cited visual element, whether the supporting text:

- You are NOT evaluating usefulness, completeness, or relevance to the question here – ONLY factual consistency between the supporting text and the visual element.

You are given:

- Answer Input Format**

For each tag:

- You may assume that the corresponding images are available to you (as a multimodal model) whenever these tags appear.

- Evaluate ONLY the supporting text linked to explicit image citation tags: [chart_i], [table_i], [figure_i] that appear in the answer text.

- [SCORING CRITERIA OF FAITHFULNESS]

Criterion 1 – Visual correctness

- Criterion 2 – Absence of hallucination or contradiction**

- Use these criteria to assign one score per image usage:

- ...

Figure 20. The Visual G-Eval prompt to assess Faithfulness (Part 1).

...

[PROCEDURE]

1) Identify cited images per answer.
For each answer:

- Scan the text and find all image citation tags: [chart_i], [table_i], [figure_i].
- For each tag, define one image usage as:
 - image_tag: the tag itself (e.g., "[chart_1]"), and
 - local_supporting_text: the nearby sentence(s) that are explicitly interpreting or referring to that image.
- If there are no such tags, mark the answer as having no cited images.

2) Inspect visual content and supporting text.
For each image_tag in an answer:

- Look at the visual content of the corresponding image:
- Identify what can be reliably read from it (axes, labels, legends, values, patterns, categories, spatial/layout structure).
- Carefully read the local_supporting_text:
- Extract all explicit claims about the visual element (numbers, comparisons, presence/absence of features, trends, etc.).

3) Evaluate each image usage using the two criteria.
For each image_tag:

- Evaluate Criterion 1 – Visual correctness:
 - Check if each claim in the supporting text is consistent with what can be seen in the image.
- Evaluate Criterion 2 – Absence of hallucination or contradiction:
 - Check whether the supporting text introduces details not seen in the image or contradicts the visible information.

Then:

- Combine both criteria into a single 1–5 faithfulness score.
- Write a brief justification that explicitly mentions:
 - which aspects of the supporting text are consistent or inconsistent with the visual content, and
 - whether there are any hallucinated or contradictory details.

[EVALUATION INSTRUCTIONS]

Your output must be JSON. For each answer:

- List every cited image usage with:
 - "image_tag" – the tag string (e.g., "[chart_1]"),
 - "reason" – a brief explanation referencing visual correctness and hallucination/contradiction,
 - "score" – an integer from 1 to 5 (faithfulness score).
- You do not need to compute any aggregated score per answer; scoring is done per cited image.
- If an answer has no cited image tags:
 - Set "per_image_analysis": [],
 - Set "no_cited_images": true.

Output JSON format (must follow exactly):

```
{
  "answer_1": {
    "per_image_analysis": [
      {
        "image_tag": "[chart_1]",
        "reason": "...",
        "score": X
      }
    ],
    "no_cited_images": false
  },
  "answer_2": {
    "per_image_analysis": [
      {
        "image_tag": "[table_3]",
        "reason": "...",
        "score": X
      }
    ],
    "no_cited_images": false
  }
}
```

Where:

- $X \in \{1, 2, 3, 4, 5\}$ is the per-image supporting-text faithfulness score.

If an answer has no cited image tags, then:

```
"per_image_analysis": [],
"no_cited_images": true
```

[TO EVALUATE]
Query: `{{ query }}`

Answer 1: `{{ answer_1 }}`
Answer 2: `{{ answer_2 }}`
[/TO EVALUATE]

Figure 21. The Visual G-Eval prompt to assess Faithfulness (Part 2).

Dimension	Spearman ρ	QW Cohen's κ
Effectiveness	0.7962	0.5333
Position	0.6775	0.6853
Faithfulness	0.7126	0.8676

Table 2. Correlation between Visual G-Eval and human judgments on 50 samples, and inter-annotator agreement (quadratic-weighted Cohen’s κ) between two human judges.

F. Correlation between Visual G-Eval and Human Judgments

M-GroSE extends human-aligned GroUSE [4] to multimodal LLM-based judging by textualizing visual elements. However, since Visual G-Eval lacks comparable validation, we conduct a targeted human study on 50 samples. Two human judges independently rate *effectiveness*, *position*, and *faithfulness*, and we measure rank correlation between Visual G-Eval and human scores. As shown in Tab. 2, Visual G-Eval correlates well with human judgments (Spearman $\rho = 0.68\text{--}0.80$), while human inter-annotator agreement is moderate to high (QW $\kappa = 0.53\text{--}0.87$).

G. Qualitative Analysis

Figures 22–25 provide the output examples of our model, including comparison with baselines and failure cases.

G.1. Evaluation

Figure 22 presents evaluation results from Qwen2.5-VL-7B and our model trained on VinQA. Compared to Qwen2.5-VL-7B, our model more effectively cites visual elements and generates descriptions in appropriate positions, resulting in more accurate and information-rich responses. Figure 23 shows successful results of our model. Specifically, our model improved readability by effectively citing user manual-related images and diagrams that are difficult to clearly comprehend through textual information alone.

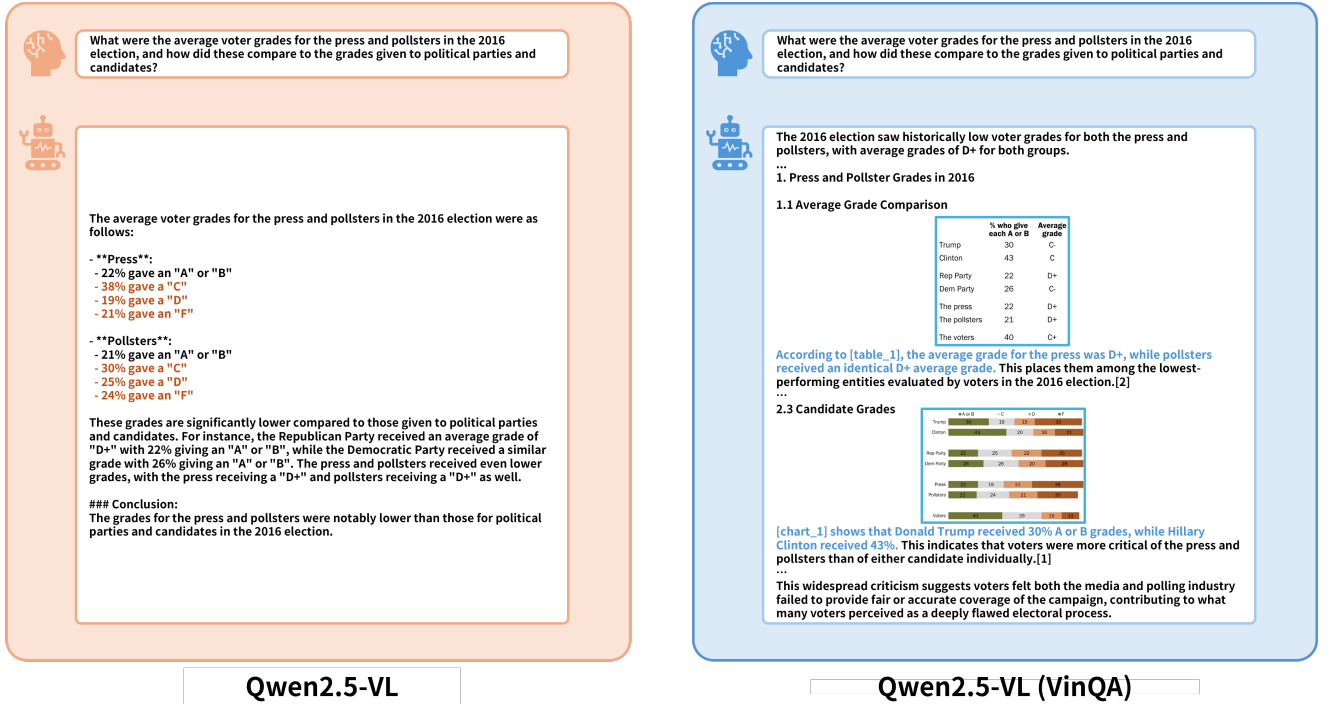


Figure 22. Comparison of predicted answers between Qwen2.5-VL-7B and our model trained on VinQA. The blue and red text respectively indicate the portions correctly and incorrectly predicted by the model.

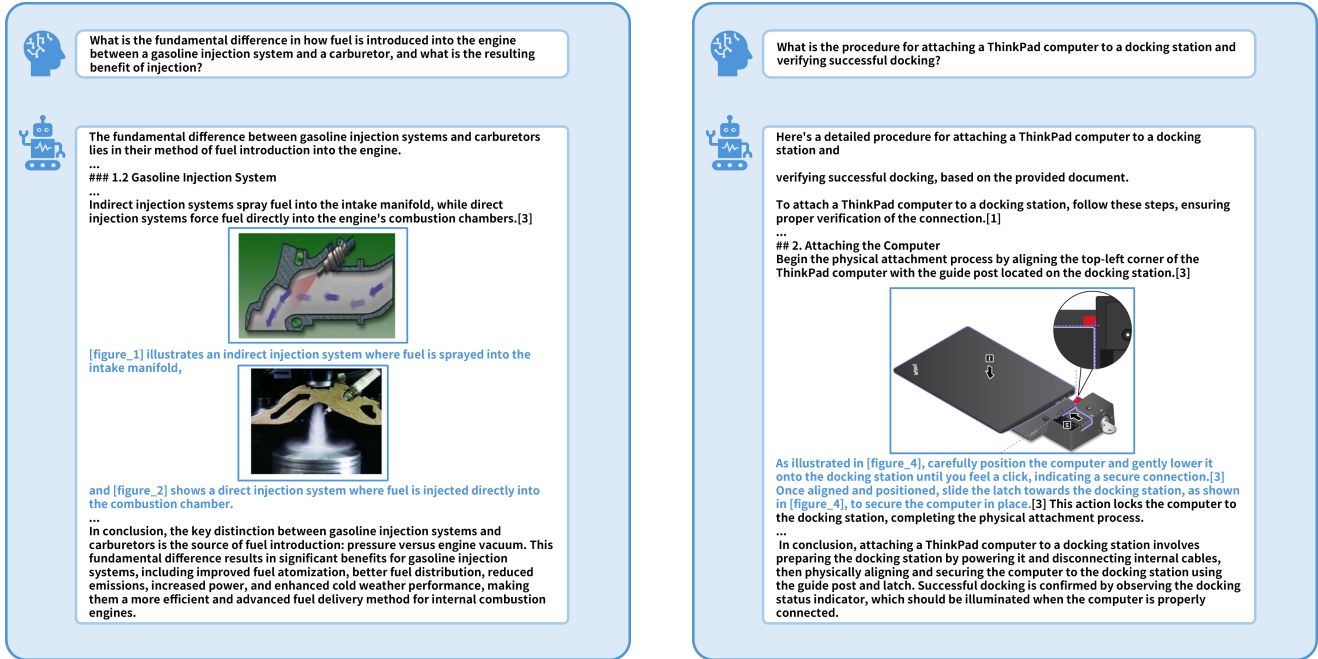


Figure 23. Inference samples of our model trained with VinQA.

G.2. Error Analysis

To identify the challenges, we analyze the inference results of our model. Representative failure cases are presented in Figure 24 and 25.

Faithfulness Our model generally shows strong performance in citing relevant visual elements and providing appropriate explanations. However, in some examples, hallucinations related to visual elements are observed. In Figure 24, our model successfully retrieved the visual element relevant to the given question but failed to recognize its detailed values. Such issues primarily occur when the input image exhibits a complex structure, such as charts containing a large amount of information in small text.

Citation While our model generally performs well, it occasionally cites incorrect images, leading to inconsistencies with the retrieved context. In Figure 25, the model correctly retrieved and utilized relevant textual content in the answer, but incorrectly cited an unrelated table visual element. This issue stems from the inherent complexity of integrating visual and textual information, particularly when dealing with ambiguous or highly detailed visuals.

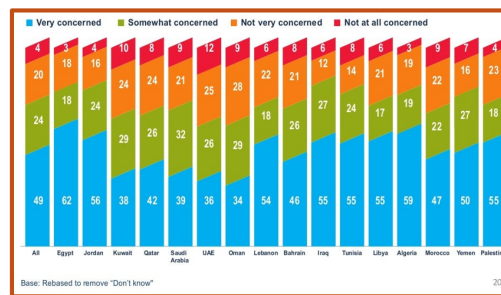


Which country had the highest percentage of respondents expressing very high concern about unemployment, and what was the corresponding perception of ease of starting a business in that same country compared to previous



Based on the Arab Youth Survey data, Algeria had the highest percentage of respondents expressing very high concern about unemployment at 59%. In contrast, 66% of Algerian respondents believed that people of their generation were more likely to start a business than previous generations.

1. Highest Concern About Unemployment



According to [chart_1], Algeria had the highest percentage of respondents expressing "Very concerned" about unemployment among all surveyed countries. Specifically, 59% of Algerian youth indicated they were very concerned about unemployment, making it significantly higher than other nations in the region.[1]

Algeria's high concern about unemployment (59%) was accompanied by a relatively positive perception of entrepreneurship (66%), suggesting that while Algerian youth were worried about finding jobs, they still maintained optimism about starting businesses. This combination of unemployment concerns and entrepreneurial optimism appears to be unique to Algeria among the surveyed countries, though the survey data does not provide information about specific reasons behind these contrasting perspectives.

Figure 24. Faithfulness error in the predicted answer from our model trained on VinQA. The text highlighted in red inaccurately describes the corresponding visual element.

H. M-GroSE performance by Context token length

In this work, we evaluate seven models on our VinQA test set using two encoding strategies—*Page Encoding* and *Modality Encoding*—across five context-token-length intervals (0–2.5k, 2.5–5k, 5–7.5k, 7.5–10k, 10k–). Table 3 presents the overall M-GroSE performance across all models by context token length.

I. Visual Source F1 performance by Modality type

We evaluate seven models on our VinQA test set using two encoding strategies—*Page Encoding* and *Modality Encoding*—across four modality types (Table, Chart, Figure, Mixed). Table 4 shows the Visual Source performance of all models by modality type.

Model	Context Token Length	Relevancy	Completeness	M-GroSE	Faithfulness	Unans. (F1)	Avg
Page Encoding							
GPT-4.1	0-2500	4.88	3.41	4.40	0.72	3.42	
	2501-5000	4.91	3.80	4.41	0.79	3.53	
	5001-7500	4.95	3.92	4.44	0.81	3.58	
	7501-10000	4.85	3.74	4.23	0.80	3.46	
	10001-	4.73	3.22	3.64	0.63	3.15	
GPT-4.1-mini	0-2500	4.69	3.38	4.21	0.57	3.32	
	2501-5000	4.66	3.62	4.07	0.49	3.34	
	5001-7500	4.61	3.83	4.14	0.45	3.4	
	7501-10000	4.45	3.61	3.91	0.50	3.25	
	10001-	4.35	3.20	3.54	0.31	3.02	
Gemini 2.0 Flash	0-2500	4.42	2.95	4.30	0.90	3.17	
	2501-5000	4.09	2.78	3.85	0.85	2.93	
	5001-7500	4.33	2.65	3.77	0.90	2.94	
	7501-10000	4.31	2.57	3.47	0.86	2.84	
	10001-	3.77	1.89	2.52	0.72	2.3	
Claude 3.5 Sonnet	0-2500	4.84	3.92	4.78	0.65	3.64	
	2501-5000	4.85	3.88	4.59	0.72	3.58	
	5001-7500	4.88	3.89	4.55	0.69	3.58	
	7501-10000	4.81	3.84	4.31	0.77	3.49	
	10001-	4.71	3.39	3.97	0.59	3.27	
InternVL3-8B	0-2500	4.05	1.96	3.01	0.79	2.51	
	2501-5000	3.35	1.74	2.35	0.82	2.11	
	5001-7500	3.32	1.71	2.14	0.81	2.04	
	7501-10000	2.98	1.58	2.14	0.78	1.93	
	10001-	2.63	1.25	1.61	0.53	1.62	
Qwen2.5-VL-7B	0-2500	3.42	1.76	2.89	0.74	2.27	
	2501-5000	2.79	1.61	2.41	0.67	1.95	
	5001-7500	3.18	1.68	2.31	0.70	2.04	
	7501-10000	3.03	1.62	2.25	0.76	1.97	
	10001-	2.68	1.34	1.74	0.70	1.69	
Qwen2.5-VL-7B (VinQA)	0-2500	4.59	3.69	4.29	0.87	3.4	
	2501-5000	4.71	3.95	4.35	0.91	3.5	
	5001-7500	4.71	3.75	4.24	0.96	3.43	
	7501-10000	4.44	3.55	3.87	0.91	3.22	
	10001-	4.40	3.03	3.47	0.81	2.98	
Modality Encoding							
GPT-4.1	0-2500	4.92	3.84	4.54	0.67	3.58	
	2501-5000	4.93	4.05	4.62	0.68	3.65	
	5001-7500	4.89	4.05	4.64	0.70	3.65	
	7501-10000	4.84	4.03	4.60	0.76	3.62	
	10001-	4.85	3.82	4.56	0.68	3.56	
GPT-4.1-mini	0-2500	4.74	3.72	4.34	0.75	3.45	
	2501-5000	4.63	3.98	4.33	0.77	3.49	
	5001-7500	4.69	4.00	4.22	0.77	3.48	
	7501-10000	4.55	4.08	4.06	0.77	3.42	
	10001-	4.51	3.78	4.11	0.72	3.35	
Gemini 2.0 Flash	0-2500	4.65	3.41	4.58	0.89	3.41	
	2501-5000	4.62	3.50	4.54	0.89	3.42	
	5001-7500	4.80	3.62	4.56	0.85	3.5	
	7501-10000	4.67	3.51	4.56	0.89	3.44	
	10001-	4.43	3.45	4.30	0.73	3.29	
Claude 3.5 Sonnet	0-2500	4.85	4.08	4.91	0.65	3.71	
	2501-5000	4.88	3.97	4.81	0.67	3.67	
	5001-7500	4.84	3.93	4.59	0.73	3.59	
	7501-10000	4.78	3.91	4.48	0.72	3.54	
	10001-	4.79	3.81	4.62	0.61	3.56	
InternVL3-8B	0-2500	3.94	2.06	3.07	0.70	2.52	
	2501-5000	3.61	2.08	2.74	0.80	2.36	
	5001-7500	3.57	1.84	2.46	0.85	2.22	
	7501-10000	3.32	1.82	2.41	0.79	2.14	
	10001-	2.86	1.57	2.17	0.54	1.90	
Qwen2.5-VL-7B	0-2500	3.22	1.79	2.66	0.64	2.17	
	2501-5000	2.72	1.66	2.32	0.75	1.93	
	5001-7500	3.01	1.85	2.53	0.79	2.10	
	7501-10000	3.03	1.87	2.60	0.81	2.12	
	10001-	2.95	1.96	2.72	0.65	2.16	
Qwen2.5-VL-7B (VinQA)	0-2500	4.65	3.67	4.34	0.90	3.42	
	2501-5000	4.70	3.86	4.29	0.93	3.46	
	5001-7500	4.65	3.74	4.24	0.95	3.41	
	7501-10000	4.46	3.30	3.84	0.89	3.15	
	10001-	4.40	3.14	3.57	0.80	3.03	

Table 3. Overall M-GroSE performance across context token length.

Model	Modal Type	Precision	Visual Source Recall	F1
<i>Page Encoding</i>				
GPT-4.1	Table	72.86	54.51	62.37
	Chart	70.12	58.81	63.97
	Figure	71.92	51.47	60.00
	Mixed	83.25	46.17	59.40
GPT-4.1-mini	Table	62.94	37.03	46.63
	Chart	68.21	54.84	60.80
	Figure	66.83	25.96	37.40
	Mixed	85.71	36.07	50.77
Gemini 2.0 Flash	Table	65.93	33.46	44.39
	Chart	68.18	40.94	51.16
	Figure	72.63	31.15	43.60
	Mixed	83.33	32.79	47.06
Claude 3.5 Sonnet	Table	72.87	53.01	61.37
	Chart	70.03	62.03	65.79
	Figure	70.63	62.98	66.59
	Mixed	83.69	53.28	65.11
InternVL3-8B	Table	50.72	6.58	11.65
	Chart	70.29	24.07	35.86
	Figure	71.25	12.87	21.80
	Mixed	84.62	15.03	25.52
Qwen2.5-VL-7B	Table	51.61	15.04	23.29
	Chart	70.83	33.75	45.71
	Figure	72.15	12.87	21.84
	Mixed	77.66	19.95	31.74
Qwen2.5-VL-7B (VinQA)	Table	77.29	39.66	52.42
	Chart	72.36	49.38	58.70
	Figure	74.60	41.76	53.55
	Mixed	85.80	41.26	55.72
<i>Modality Encoding</i>				
GPT-4.1	Table	84.30	65.60	73.78
	Chart	70.78	65.51	68.04
	Figure	72.83	72.01	72.42
	Mixed	83.65	63.93	73.49
GPT-4.1-mini	Table	77.26	56.20	65.07
	Chart	71.75	64.27	67.80
	Figure	65.95	34.54	45.33
	Mixed	85.71	45.90	59.79
Gemini 2.0 Flash	Table	84.49	57.33	68.31
	Chart	69.97	52.61	60.06
	Figure	75.10	42.21	54.05
	Mixed	88.44	41.80	56.77
Claude 3.5 Sonnet	Table	85.10	63.25	72.63
	Chart	72.83	66.50	69.52
	Figure	70.24	65.01	67.53
	Mixed	85.66	62.02	71.95
InternVL3-8B	Table	75.83	17.11	31.51
	Chart	76.05	31.51	44.56
	Figure	68.04	14.90	24.44
	Mixed	88.31	18.58	30.70
Qwen2.5-VL-7B	Table	84.80	27.26	41.25
	Chart	71.90	37.47	49.27
	Figure	63.30	15.58	25.00
	Mixed	83.61	27.87	41.80
Qwen2.5-VL-7B (VinQA)	Table	82.94	45.68	58.91
	Chart	71.94	55.33	62.55
	Figure	71.03	40.41	51.51
	Mixed	83.16	43.17	56.83

Table 4. Overall Visual Source performance across modality types.

References

- [1] Anthropic. Claude 3 model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf, 2024. Accessed: 2024-04-05. [1](#)
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. [1](#)
- [3] Google DeepMind. Gemini: Our december 2024 update. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>, 2024. Accessed: 2024-04-05. [1](#)
- [4] Sacha Muller, Antonio Loison, Bilel Omrani, and Gautier Viaud. GroUSE: A benchmark to evaluate evaluators in grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4510–4534, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. [3](#), [23](#)
- [5] OpenAI. Gpt-4.1 overview. <https://openai.com/index/gpt-4-1/>, 2024. Accessed: 2024-04-05. [1](#)
- [6] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. [1](#)