Can Multimodal LLMs Generate Pedagogical Questions?

Thomas Gerald[1], **Sahar Ghannay**[1], **Julie Lascar**[1], **Paul Lerner**[2], **Anne Vilnat**[1]
Université Paris-Saclay, CNRS, LISN[1] - ISIR CNRS[2]
{firstname.lastname}@lisn.upsaclay.fr[1] - lerner@isir.upmc.fr[2]

Abstract

Educational materials frequently combine text, diagrams, tables, and charts to convey complex concepts. Understanding such materials often requires reasoning across modalities rather than relying solely on textual descriptions, both to ask interesting questions or to answer them. In educational contexts, the main challenge lies in assessing the relevance and quality of the questions themselves. This raises a key issue: what defines a good question in a specialized learning environment? By comparison, evaluating answers is a more conventional task, although it requires examining criteria consistent with the targeted educational level. To the best of our knowledge, the use of LLMs for assessing the pedagogical relevance of questions remains unexplored. This gap highlights the need to define pedagogical relevance more clearly and to investigate the consistency of LLM judgments, as well as their alignment with human evaluations. We investigate whether multimodal LLMs may generate pedagogical questions, and whether they can judge a question as pedagogical or not, in English and French. Contrary to most of QA Multimodal corpora, we focus on questions that could be asked by a teacher in his/her class, and that need to deal with different parts of the document to be answered. Results show that while LLMs as a judge is an efficient framework, many problems could arise, and that aligning prediction with human annotators is a difficult task for complex criteria.

Keywords: visual question generation, education, LLM-as-a-judge

1. Introduction

Educational materials frequently combine text, diagrams, tables, and charts to convey complex concepts, particularly in domains such as biology, physics, history, and geography. Understanding such materials often requires reasoning across modalities rather than relying solely on textual descriptions. Question answering (QA) systems capable of operating in these Multimodal settings have potential applications in intelligent tutoring, automated assessment, and adaptive learning platforms. However, despite rapid progress in natural language processing, automated generation and evaluation of Multimodal educational QA remain relatively underexplored, especially in a multilingual setting.

In educational contexts, the main challenge lies in assessing the relevance and quality of the questions themselves. This raises a key issue: *what defines a good question in a specialized learning environment?* By comparison, evaluating answers is a more conventional task, although it requires examining a terminology and criteria consistent with the targeted educational level. To the best of our knowledge, the use of LLMs for assessing the pedagogical relevance of questions remains unexplored. This gap highlights the need to define pedagogical relevance more clearly and to investigate the consistency of LLM judgments, as well as their alignment with human evaluations.

Large language models (LLMs) such as GPT (Radford et al., 2018), LLaMA (Touvron et al., 2023), and Bloom (BigScience Workshop and others, 2023) have achieved notable results in text-based QA. LLMs have rapidly expanded beyond the text modality and most modern LLMs feature a Multimodal version (Llama Team, 2024; Malartic et al., 2024; Team et al., 2024; Wang et al., 2024; OpenAI, 2024). In their simplest form, Multimodal LLMs keep the text parameters frozen and learn a linear projection from the image to the text embedding space (Tsimpoukelli et al., 2021). The LLM then inputs visual tokens just as textual prompts. The visual projection is learned using the same language modeling loss as for pretraining, although conditioned on the input image (i.e., effectively learning to caption the image). Multimodal LLMs such as Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023), and Pixtral (Agrawal et al., 2024) have demonstrated promising Multimodal reasoning capabilities, but their open-domain training data may limit robustness in specialized educational settings. Domain adaptation techniques—such as LoRA (Hu et al., 2021), prefix tuning (Li and Liang, 2021), and prompt engineering (Wei et al., 2023)—offer potential solutions, yet their effectiveness for Multimodal educational QA remains insufficiently examined.

Traditional visual question answering (VQA) datasets, including VQA (Antol et al., 2015) and CLEVR (Johnson et al., 2016), predominantly target photographic imagery and shallow factual queries, limiting their applicability for domains requiring deeper cross-modal reasoning. Several

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5506–5515
11–16 May 2026.

datasets attempt to bridge this gap. ScienceQA (Lu et al., 2022) integrates diagrams with textbook-style questions, while ChartQA (Masry et al., 2022) focuses on chart comprehension with a mix of human-authored and synthetic QA pairs. DiagramQG (Zhang et al., 2025) targets diagram-based question generation but omits the surrounding textual context, which is often crucial for semantic interpretation. In this paper, we investigate whether multimodal LLMs may generate pedagogical questions, and whether they can judge a question as pedagogical or not.

Contrary to most QA Multimodal corpora, we focus on questions that could be asked by a teacher in their class (i.e., that have an interest in the educational context) and that require dealing with different parts of the document to be answered. While examining different documents used for teaching, it becomes obvious that most of them are composed of texts *and* images, and specifically not photos, but, for example, diagrams (as in Figure 1) or maps (as in Figure 2). Thus, most questions in education focus on reasoning among these different types of media. To evaluate whether LLMs can generate interesting questions from multimodal documents in the education domain, we need to build a corpus of such documents and define new criteria to assess this ability. We begin by proposing possible annotation pipelines—either generating the question and answer in two steps or in a single step. We propose evaluating LLMs' ability to generate an answer to a question based on specific parts of the context, thereby determining which parts are necessary to answer the question. In a second stage, we propose evaluating different criteria across many configurations for automatic judgment using LLMs as judges. We compare these results with human judgments to assess the relevance of LLMs for automatically evaluating pedagogical criteria in the question-answering setting.

The paper is organised as follows. In Section 2, we present existing QA datasets and methods to leverage LLMs in this context. In Section 3, we describe the protocol to collect and extract data from selected English and French resources. We then introduce our generation pipelines in Section 4, along with the evaluation protocol and criteria that align with educational objectives. Finally, the experimental results are discussed in 5, before concluding and discussing future works.

2. Related Works

2.1. QA dataset in education domain

Many Multimodal datasets have been proposed with the advent of Multimodal LLMs. However, most rely on straightforward questions and answers on image and text. Kembhavi et al. (2016) pioneered the extension of Visual Question Answering from natural images depicting cats and dogs (Antol et al., 2015) to complex diagrams. They then extended their work in Kembhavi et al. (2017) by also adding text related to the diagram, much like in our own work. However, their focus is more on benchmarking models and advancing machine learning rather than generating questions to help students. Similar datasets include Sampat et al. (2020); Mathew et al. (2022); Masry et al. (2022). Other datasets, such as DocVQA (Mathew et al., 2021), rely on scanned images of documents containing handwritten, typewritten or printed textual elements, hence containing textual information within the image.

Closer to our topic, the ScienceQA dataset (Lu et al., 2022) gathers questions and answers on images and text with reasoning objectives. In addition, evaluation is eased by the fact that a MCQ format is considered for answers. However, while ScienceQA is based on educational content, graphics, and text, explanations alongside the MCQ answer are relatively short (an average 47 words). However, this dataset, along with DiagramQG (Zhang et al., 2025), could be helpful in our automatic annotation step, as corpora for models fine-tuning. Additionally, we would apply the schema of annotations to provide insights into the task's difficulties and (material) needs. Moreover, we will consider the course page as a whole and not only a small context that encompasses the answer.

2.2. Automatic Annotation

Recent works have explored the paradigm of "LLM-as-a-Judge", in which large language models evaluate generated outputs based on contextual alignment, factuality, and fluency (Zheng et al., 2023; Gu et al., 2025). Comparative judgment methods (Liusie et al., 2024), where two candidate answers are ranked relative to each other, have been shown to improve evaluation consistency, yet concerns persist regarding prompt sensitivity, positional bias (Shi et al., 2025), and reliability in specialized domains. Still, LLM-as-a-Judge proved to be an effective way to evaluate text generation, for benchmarking (Zheng et al., 2023) but also to *annotate* new data, as we do in this article. For example, Röttger et al. (2025) leveraged LLMs to filter queries about political issues, annotate their framing, as well as the stance of (generated) answers to these queries. In our setting of Multimodal educational QA, LLM-as-a-Judge issues are compounded by the need to assess whether visual and textual modalities are essential to answer the question, a dimension often overlooked in existing automatic evaluation pipelines. Another line of our work coming from the Computer Vision community, Visual Question Generation (see Zhang et al.

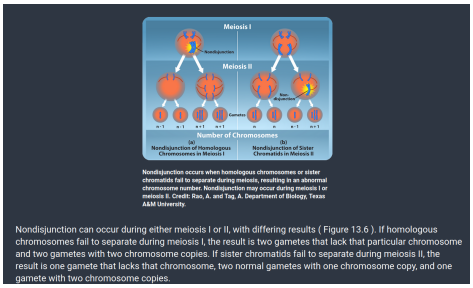

Figure 1: Section of OpenStax concerning meiosis

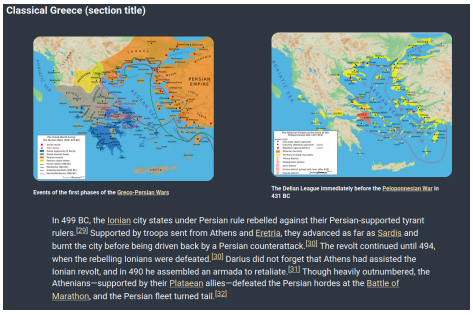

Figure 2: Section of Wikipedia on classical Greece

(2025) and references therein), proposed several methods to generate visual questions, sometimes training a VQA module jointly with the question generation module (Li et al., 2018). Closely related to our work, Zhang et al. (2025) introduce DiagramQG, a large-scale dataset with 8,372 diagrams and 19,475 questions spanning 169 educational concepts, to enable the generation of concept-focused questions from diagrams. They further propose a framework that leverages hierarchical knowledge integration to generate concept-focused questions. However, their method requires training a model specifically for question generation. In contrast, we propose to prompt a Multimodal LLM, as it proved to be effective in other settings (Zou et al., 2024).

3. Resources

We first look for existing resources to construct the QA dataset. We focused on three subject areas: history, geography, and biology, because their educational materials frequently include images that convey meaningful information. For example, history and geography often feature maps and timelines, while biology includes diagrams such as cell structures or biological processes. We also consider several domains to avoid being too specific. For the same reasons, we take into account two languages: English and French. To build questions and answers, we leveraged two resources that we will furnish as context for the different generation pipelines (see section 4):

- **OpenStax**[1] : a resource of high school educational content produced by the RICE University. It consists of HTML pages of course and exercises (see Figure 1).
- **Wikipedia** [2] : Wikipedia is a well-known encyclopedic resource containing articles on a wide range of topics. For our dataset, we selected Wikipedia articles that align with OpenStax content, as well as additional topics relevant to the French high school (see Figure 2). Articles were collected in both English and French.

The preprocessing steps vary depending on the data source. For **Wikipedia**, we split the selected documents following the sections. We only kept sections where there is an image with a tag related to diagrams, maps, or timelines. The final subset of Wikipedia amounts to 100 relevant sections from 20 articles to experiment with the automatic annotation process. We collected information from the **OpenStax** platform, providing educational material (in English) in HTML. In total, we collected 522 text-image pairs related to Biology from OpenStax (in English) and 100 related to Biology and History from Wikipedia (in French).

4. QA generation and Evaluation

4.1. QA generation

To create the dataset, we generate question–answer pairs from a given section, consisting of a (figure, body-text) pair. In this work, we propose to generate question–answer pairs using two main strategies. First, 2-steps generation, where (i) a first prompt, including the context, asks to generate a question and (ii) a second prompt, including the context and the question, asks to generate the answer. Second, 1-step generation, where a single prompt including the context asks to generate the question and answer given the context. The model is fed with the body-text in HTML. This allows to capture the structure of the document, especially titles (<hi> tags), the different paragraphs (<p> tags), and the highlighted content (bold, italic, ...). The images are provided before the body-text. We provide the prompt used for generation in supplementary material.

4.2. Assessing the Multimodality and Document Relevance

As the instructions in the prompt do not guarantee that the model leverages both the figure and

[1] https://openstax.org/books/biology-2e/pages/18-review-questions

[2] https://fr.wikipedia.org and https://en.wikipedia.org

#	Prompt
F	Considering the document: [doc] Answer the question: [quest]. You should write the output in json format with the field "answer" containing the answer to the question
B	Considering the image(s) and the caption(s) associated: [captions-list] Answer the question: [quest]. You should write the output in json format with the field "answer" containing the answer to the question
N	Answer the question: [quest] You should write the output in json format with the field "answer" containing the answer to the question

Table 1: Prompts used to assess the multimodality level of the generated questions under each ablation.

body-text to generate the QA pair, we designed an experiment in which LLMs were prompted to answer the questions under different content ablations of the (figure, body-text, question) triplets. We ablate three different inputs: (i) the figure F (leaving only body-text and question); (ii) the body-text B (leaving only figure and question); (iii) both (denoted N, leaving only the question). The logical formula F (resp. B, N) indicates that the answer is correct under ablation F. On the contrary, $\neg F$ (resp $\neg B, \neg N$) indicates an incorrect answer under the ablation. Prompts for each ablation are given in Table 1.

Because of the limits of traditional QA metrics (see, e.g. Bulian et al., 2022, and references therein), we follow an LLM-as-a-Judge approach. The judge inputs the question, the 'ground-truth' answer[3], and the answer generated with ablated context. We considered for this experiments the model $Qwen2.5-VL$ prompted with the instruction provided in the supplementary material.

We interpret the results depending on the correctness of the different ablation configuration. We report our interpretation in Table 2. For instance $F \wedge B \wedge \neg N$ means that the figure ablation led to a correct answer, the body-text ablation as well, but the ablation of both led to an incorrect answer. In other words, either the figure *or* the body-text is enough to correctly answer the question.

Logical formula	Interpretation
$F \wedge B \wedge N$	No context is needed to answer the question
$F \wedge B \wedge \neg N$	The figure *or* the text is required to answer the question
$F \wedge \neg B \wedge \neg N$	Only the body-text is required to answer the question
$\neg F \wedge B \wedge \neg N$	Only the figure is required to answer the question
$\neg F \wedge \neg B \wedge \neg N$	The image *and* the text are required to answer the question
Other cases	Hallucination

Table 2: The different interpretations of combinations for answer generation settings F, B and N as defined in Table 1

[3]Ground-truth was generated automatically

4.3. Assessing the Pedagogical Relevance

Education questions (and consequently answers) are generally constructed to evaluate or improve students' analytical competencies. Generic criteria, such as correctness or fluency, which are typically designed for QA, are insufficient to assess the relevance of both questions and answers.

One key objective of this study is to state the relevance of LLMs in judging the appropriateness of the content to education domains. And especially to answer the question: *can Multimodal LLMs faithfully judge pedagogical criteria?*

Based on the French national curricula[4] in History-Geography and Life Sciences, we established additional pedagogical criteria that are essential to consider in this context. The competencies targeted in our dataset include the ability to extract information from documents, analyze and interpret these documents, establish connections between multiple sources, engage in scientific reasoning, represent knowledge through structured forms such as diagrams or tables, and apply acquired knowledge to specific examples or problem situations. Table 3 lists the criteria that should be considered for assessing the relevance and quality of questions and answers in this context. These criteria concern the question or the answer. Some of them are binary (such as correctness, which is true or false), and others are evaluated on a scale of 3 or 4 classes (for example, the question may be more or less appropriate for the students).

We evaluate these criteria according to the document provided, comparing two methods: (a) using a pool of 4 human annotators and (b) LLMs-as-Judges. Human annotators annotate each criterion summarized in Table 3 given the contextual document, question, and ground-truth answer, using a simple spreadsheet interface and guidelines precising how to evaluate the different criteria[5].

For each QA generation pipeline (2-step and 1-step), 40 annotations were required to judge and shared across annotators.

4.4. Experimental protocol

Model We experimented with two models:

- Qwen2.5-7B VL instruct[6] using the default parameters for generation. We used this model in our three experiments (generating question-

[4]Cycle 4 and Core Competencies: Accompanying Document for Assessment (Cycles 2-4)

[5]https://gitlab.lisn.upsaclay.fr/multimodal-qa/guidelines-2025-qa

[6]https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

Criterion	Objective	#Classes
correctness (COR)	the answer is correct	2
relevance (REL)	the question is aligned with the document content	4
clarity (CLA)	the question is clearly worded and easy to understand	3
appropriateness (APP)	the question is suitable for the target (students aged 12–18)	3
visual-interpretation-1 (VIS-1)	the question induces understanding diagrams, charts, or tables	2
visual-interpretation-2 (VIS-2)	the question induces understanding chronological references	2
visual-interpretation-3 (VIS-3)	the question induces understanding spatial references	2
language (LAN)	the answer use disciplinary language and representations	2
reasoning-1 (REA-1)	the question induces applying reasoning and methods	2
reasoning-2 (REA-2)	the question is cognitively complex	3
creating (CRE)	the question induces creating and modeling information	2

Table 3: Summary of the criteria evaluated on the corpus, with the number of possible classes

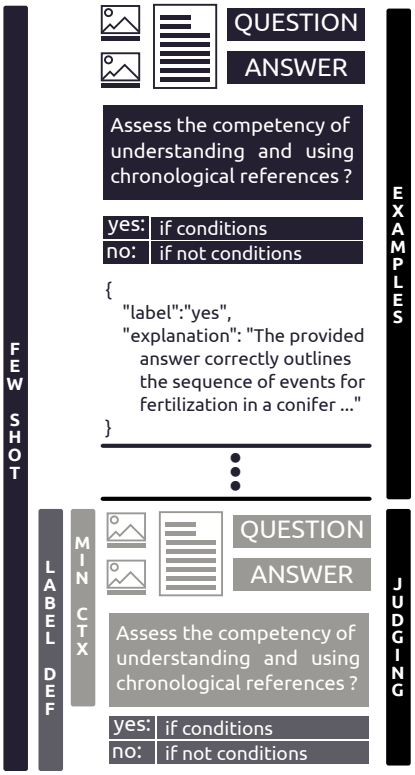

Figure 3: LLM-as-a-judge pipelines: few-shot, label-definition and min-context

answer pairs, answering under content ablation, and judging).

- Mistral-Small-3.2-24B instruct[7] with recommended parameters (temperature of 0.6). The model is only used in the LLM-as-a-judge experiment.

Even while using sampling, both models will have little randomness in their predictions due to the low temperature.

[7] https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

Prompting LLM-as-a-judge We defined different configurations when prompting the model to judge the different aspects of the generated question and/or answer. More precisely, we provide three configurations that depend on the knowledge of the task we provide to the model:

- **MC** (minimum-context): We only provide as instruction to the model the definition of the task, and the possible labels without explicit specifications on each class.
- **LD** (label-definition): We increment the knowledge of the task provided to the model within the prompt by adding the specification of the labels, i.e. in which case we select one label preferentially to another.
- **FS** (few-shot): We increment the information provided within the previous configuration (LD) by adding different examples and their expected prediction. One example is given for each possible label.

We summarize the different configuration in Figure 3, depicting an example for the criterion visual-interpretation-2[8].

Dataset We considered a subset of documents for the evaluations to validate or not the proposal.

Gen. pipe.	Examples	Wikipedia	OpenStax
1-step	151	23	128
2-steps	154	33	121

Table 4: Number of question-answer pairs generated by each generation pipeline (gen. pipe.) that were judged by LLMs

5. Experimental results

In this section, we present the two evaluations described in Sections 4.2 and 4.3. We then provide

[8] The complete prompt is not depicted in the figure

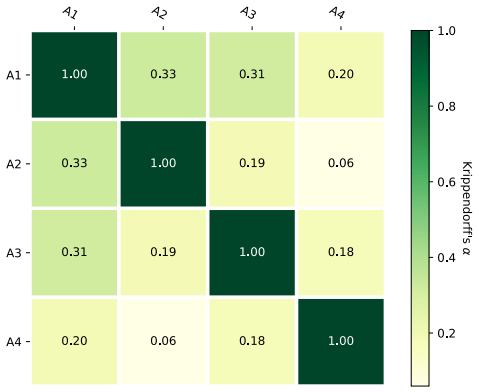

Figure 4: Human annotator agreement

insights on automatic annotation (section 5.3)[9].

5.1. Multimodal Questions

We first discuss the results of content ablation (described in Section 4.2), using Qwen2.5 VL. We report the proportion of examples that fall in each of the categories in Table 5 with the two pipelines and the two resources (Wikipedia and OpenStax). First, we find that the 1-step pipeline generates 20% of questions that fall into $(\neg F \wedge \neg B \wedge \neg N)$, i.e., both figure and body text are necessary to correctly answer the question, against 12.5% for the 2-step pipeline. Thus, suggesting that producing questions and answers in a single step would help the model produce genuinely multimodal questions (requiring information from the figure and body text). More globally, the 1-step pipeline seems to mainly rely on the figure with 56.6% of generated questions falling into one of $(\neg F \wedge \neg B \wedge \neg N)$, $(F \wedge B \wedge \neg N)$, or $(\neg F \wedge B \wedge \neg N)$, against 32% for the 2-steps pipeline. Additionally, the 1-step pipeline generates fewer questions that can be answered without any context $(F \wedge B \wedge N)$, only 31% vs 45.3% for the 2-step pipeline. Thus, the 1-step pipeline seems to be most effective for generating multimodal or visual questions.

5.2. Evaluation of pedagogical relevance

In this section, we report and discuss results for the human and model evaluation of the criteria presented in Table 3. To ensure the relevance of model judgment and configuration, we will discuss inter-annotator agreement. Annotation was carried out by four annotators (among the authors of the paper). Before analyzing agreement per criterion, we discuss the global annotator agreements.

[9] https://gitlab.lisn.upsaclay.fr/multimodal-qa/can_multimodal_llms_generate_pedagogical_relevant_question

Overall Agreement In Figure 4, we report Krippendorff's alpha inter-annotator agreement averaged over all criteria. At first glance, we can observe a disparity among the annotators, especially between annotator A1 and the others. The A3 annotator gets a slight agreement with annotators A2 and A4, while A2 and A4 get very poor agreement between them (0.06). This phenomenon could highlight a difference in interpretation between annotators or underline the difficulty or subjectivity of the criteria. Notice that $A4$ only annotated 20 examples; this disagreement could have been alleviated with a greater number of annotations.

Agreement per criterion To verify that we can rely on criteria judgments, we evaluate (i) the number of total complete agreement (i.e. percent where all annotators annotated the criterion with the same label); (ii) Cohen's κ (averaged across annotators); and (iii) the ordinal Krippendorff's α that is meaningful for scale liberationist (in our case the multi-label criteria can be interpreted as a scale). We report these different metrics in Table 6 for each criterion. A first observation is that some criteria show moderate agreement, especially the COR (whether the answer is correct given the provided context) and the VIS-1 binary criteria (which assess whether the question requires interpretation/understanding of diagrams). It is worth noting that many documents contain diagrams, and consequently, many examples are annotated positively for the criteria, leading to sufficiently diverse annotations to compute the agreement. Especially, it could explain the higher disagreement for VIS-2 or VIS-3, where fewer documents contain chronological or map-based figures, so a different annotation has a greater impact on agreement. For criterion REA-1 (does the question induce reasoning), annotators' agreement is slight ($.25$ cohen's κ). Although the agreement is rather low, it is far from random. For multi-label criteria, especially the relevance of the question to the context, the ordinal metric (α) yields much higher agreement than the nominal one (κ). The same behaviour is observed for the reasoning level (REA-2), but in this case with a smaller difference between the two metrics (κ and α). However, for APP (appropriateness for the audience) and CREA (the question needs to summarise/model new content or information), no agreement is reached. This highlights a gap in our annotation guide or an insufficient level of training for our annotators (e.g. knowledge of the course program is necessary for judging appropriateness), so these criteria will not be considered in the rest of the paper.

Model-Annotator Agreement To state whether or not model judgment is relevant for our task of judging the QA dataset, we first report the human-

Pipeline	Source	figures & body-text $\neg F \wedge \neg B \wedge \neg N$	image or text $F \wedge B \wedge \neg N$	body-text $F \wedge \neg B \wedge \neg N$	figures $\neg F \wedge B \wedge \neg N$	no context $F \wedge B \wedge N$	other
1-step	OS	17.1%	15.4%	9.8%	10.6%	34.1%	13.0%
	WP	36.4%	9.1%	13.6%	22.7%	13.6%	4.5%
	Total	20.0%	14.5%	10.3%	12.4%	31.0%	11.7%
2-steps	OS	13.0%	13.0%	11.0%	6.0%	46.0%	11.0%
	WP	10.7%	21.4%	10.7%	3.6%	42.9%	10.7%
	Total	12.5%	14.8%	10.9%	5.5%	45.3%	10.9%

Table 5: Evaluating content for answering the question according to the protocol described in section 4.2

	% equals	κ	α
APP	46.34	0.06	0.07
CLA	58.54	0.07	0.21
COR	70.73	0.42	0.57
CRE	80.49	0.06	0.06
LAN	97.56	–	0.00
REA-1	43.90	0.25	0.28
REA-2	43.90	0.15	0.19
REL	29.27	0.21	0.44
VIS-1	60.98	0.50	0.53
VIS-2	82.93	0.28	0.31
VIS-3	75.61	0.17	0.13

Table 6: Annotator agreement averaged by users considering κ (Cohen) and α (Krippendorff).

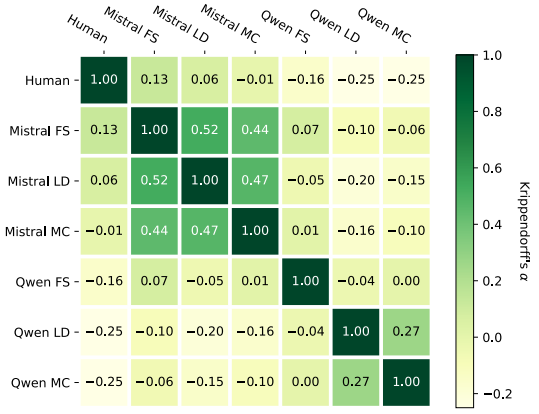

Figure 5: Human-model annotator agreement

model alignment using Krippendorff's α averaged by criteria. In Figure 5, we report these results, measuring agreement between human annotators and the different models across the few-shot (FS), label-definition (LD), and min-context (MC) configurations (see section 4.4). For human annotation, we associate each example and criterion with the label chosen by the majority of annotators. Contrary to the human criteria evaluation, the α metric is especially low, such that in most cases the agreement cannot be differentiated from random ($\alpha \leq 0$). Especially the Qwen model response, regardless of configuration, shows no agreement with human annotations. However, the Mistral model reached $\alpha = .13$ with annotators in the few-shot setting, meaning a slight agreement. Interestingly, the Mistral model shows greater agreement with human annotators as the instruction increases (FS $>$ LD $>$ MC), suggesting that feeding the model more task information is important for the task. In general, Mistral better aligns with humans than Qwen; it is worth noting that the Qwen (7B) model used in the experiment is much smaller than Mistral (24B). Further experiments should be conducted to confirm whether larger models align more closely with human decisions or perform better at following few-shot instructions. In the next experiment, we propose comparing alignment by criterion between humans and the Mistral model in the few-shot setting. The results are reported in the Table 7, where we proceed as in the human evaluation. First, we can see that the model does not align as well as humans (comparing table 6 and 7). Nevertheless, both humans and models align moderately on the $REA-1$ and $VIS-1$ criteria, indicating that, at least for $VIS-1$, which has greater human-human agreement, the model prediction is quite reliable. Both $VIS-2$ and $VIS-3$ are poorly aligned, and therefore, model results should not be used for dataset evaluation. Interestingly, while nominal alignment is very low for relevance (REL), ordinal alignment leads to a drastic increase in alignment. However, the other criteria in the current state cannot be considered reliable.

Criteria	% equals	κ	α
APP	53.66	0.09	-0.04
CLA	80.49	0.15	0.09
COR	63.41	0.01	-0.02
CRE	87.80	-0.04	-0.05
LAN	97.56	0.00	0.00
REA-1	75.61	0.52	0.51
REA-2	41.46	0.04	0.03
REL	51.22	0.06	0.23
VIS-1	73.17	0.46	0.44
VIS-2	73.17	0.17	0.11
VIS-3	78.05	0.14	0.07

Table 7: Mistral FS - Human agreement

Resource	Generation Pipeline	COR	CRE	LAN	REA-1	VIS-1	VIS-2	VIS-3
openstax	2-steps	0.95	0.26	1.00	0.94	0.87	0.23	0.09
	1-step	0.92	0.03	0.98	0.79	0.73	0.22	0.08
wikipedia	2-steps	0.94	0.45	1.00	0.88	0.82	0.42	0.48
	1-step	0.83	0.04	0.96	0.87	0.78	0.57	0.61
Total		0.91	0.20	0.98	0.87	0.80	0.36	0.32

Table 8: Proportion of generated question-answer pairs that have fulfilled the pedagogical criterion according to Mistral-as-a-judge (FS setting), for each criterion

Model judgment While the model does not fully align with the criteria, some of them can be used to verify or judge question-answer pairs for educational resources; in our case, the $REA-1$ and $VIS-1$ criteria. In this last experiment, we report the results in Table 8 as the percentage of examples labelled as yes for the binary criteria. In particular, the objective is to evaluate the proportion of satisfied criteria for each pipeline and resource. We first can observe that the generated QA often rely (80%) on interpreting diagrams. As most of the page contains diagrams (biology class), this is an expected result when the model (question-answer generation model) follows the instruction rules. Furthermore, we can show that most questions (87%) require reasoning (REA-1) regardless of the resource or the pipeline. While it is difficult to state the adequacy of other criteria (due to the low model human agreement), the proportion of VIS-2 and VIS-3 labels is coherent with the resource, where on Wikipedia we extracted history articles, which are more likely to contain chronological diagrams (such as a timeline) and geographical figures (such as a map). This last observation could indicate that the model partially labelled these criteria correctly; however, to verify this hypothesis, we should conduct a larger human evaluation campaign.

5.3. Qualitative Analysis

From qualitative evaluation, regarding the different criteria concerning the use of graphics or images, we observed that generated questions are mostly based on textual paragraphs, and the answers primarily utilize the text. Even if the prompts indicate to use the figures, it is generally not the case. Moreover, when it is the case, the question is not really correct, and the answer is false, as we can see in the Figure 6. The generated question was: "*Using the diagram and text provided, explain how the Senate exerted its influence over magistrates and assemblies in the Roman Republic*"[10]. The diagram illustrates that the Senate is elected by different assemblies, indicating how these assemblies influence the Senate, not the reverse case. The generated answer explains that the Senate elects the

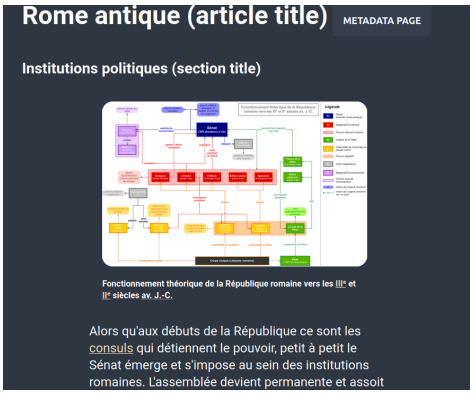

Figure 6: Rome antique (wikipedia)

Consul, which is not mentioned in the text nor in the diagram, illustrating the fact that other knowledge has been used by the model to answer. We see here the illustration of two current limitations: (a) the models make very little use of diagrams, even if they are prompted to do so, and (b) when they try to use them, the result is not good. This qualitative analysis needs to be deepen, but it clearly shows the use of outside knowledge to answer.

6. Conclusion

In this work, we evaluate LLMs to assess the quality of generated question-answer pairs in the education domain. The different evaluations show that, at least for the model used in this work, LLMs as judges' predictions remain difficult to be confident about. In our first experiment, we show that, for one of the designed pipelines, the model tends to use context (and figures) to answer questions; a large number of questions remain answerable after content ablation. An interpretation is that models often rely on internal knowledge (acquired during training) rather than the provided context. It seems tempting to use LLMs as automatic judges to evaluate dataset quality to build new corpora. However, the alignment gap persists between humans and models, even in our best setting using few-shot examples and a larger model. We consequently show that automatically judging pedagogical criteria in an LLM-as-a-judge setting remains challenging. In the future, we will explore methods to strengthen align-

[10] original question was in french

ment between human judges and between humans and models. For instance, to align with human models, PEFT and low-resource fine-tuning approaches could be leveraged. Additionally, we plan to evaluate the usability of the generated dataset on a larger scale, either by organizing a larger evaluation campaign or by conducting user studies involving both teachers and students.

7. Acknowledgment

We would like to thank the anonymous reviewers for their constructive feedback. This work was funded by the ANR-25-CE23-2916 EQUATION project. It also benefited from access to IDRIS's computing resources through resource allocation 103226/AD011014532R2 granted by GENCI.

This work was also co-funded by the European Union's Horizon Europe Research and Innovation programme through the project UTTER – Unified Transcription and Translation for Extended Reality under Grant Agreement No. 101070631.

8. Bibliographical References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, et al. 2024. Pixtral 12b.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

BigScience Workshop and others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models.

Meta Llama Team. 2024. The Llama 3 Herd of Models.

Quentin Malartic, Nilabhra Roy Chowdhury, Ruxandra Cojocaru, Mugariya Farooq, Giulia Campesan, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Maksim Velikanov, Basma El Amel Boussaha, Mohammed Al-Yafeai, Hamza Alobeidli, Leen Al Qadi, Mohamed El Amine Seddik, Kirill Fedyanin, Reda Alami, and Hakim Hacid. 2024. Falcon2-11B Technical Report.

OpenAI. 2024. Gpt-4o system card.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. 2024. VGBench: Evaluating large language models on vector graphics understanding and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3647–3659, Miami, Florida, USA. Association for Computational Linguistics.

9. Language Resource References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile. IEEE.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pages 235–251, Cham. Springer International Publishing.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. Issuebench: Millions of realistic prompts for measuring issue bias in llm writing assistance.

Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-linguistic question answering (VLQA) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4606–4616, Online. Association for Computational Linguistics.

Xinyu Zhang, Lingling Zhang, Yanrui Wu, Muye Huang, Wenjun Wu, Bo Li, Shaowei Wang, Basura Fernando, and Jun Liu. 2025. DiagramQG: Concept-Focused Diagram Question Generation via Hierarchical Knowledge Integration.