

MaDOVE: a Dataset for Multilingual Document VQA Evaluation

Elio Musacchio^{1,2}, Lucia Siciliani², Pierpaolo Basile², Giovanni Semeraro²,
Mark Federico Zampedroni³, Asia Beatrice Uboldi³

¹University of Pisa, ²University of Bari Aldo Moro, ³Fastweb SpA

elio.musacchio@phd.unipi.it

Abstract

Visual Question Answering (VQA) is one of the most popular tasks in the fields of Computer Vision and Natural Language Processing. The task is to provide an answer to a question related to the contents of a given image. VQA has been expanded to several domains. One of these is document VQA, where the image is a page of a document and the question relates to the text in the image. In this case, the system that is being evaluated has to understand text written within the image, unlike common VQA datasets where images are natural scenes. However, document VQA datasets are exclusively in English; hence, the system’s performance is evaluated only for English. In light of this, we introduce MaDOVE (Multilingual DOcument VQA Evaluation), a novel dataset designed to evaluate document VQA in multiple languages. Moreover, to establish a strong benchmark, we conduct an extensive set of experiments by evaluating a diverse range of models on this task.

1 Introduction

Visual Question Answering (VQA) is one of the most explored tasks in the history of Computer Vision and Natural Language Processing (Kim et al., 2025). The task tests the ability of a system to: 1) understand the visual contents of an image; 2) reason jointly about the contents of an image and a given question; 3) provide an answer that is correct w.r.t. the contents of the image and the question. Solving VQA is a non-trivial task, as it requires multimodal understanding and reasoning.

In light of the success of Large Language Models (LLMs), a new type of architecture has emerged that extends LLMs with multimodal capabilities. These models are called Large Vision-Language Models (LVLMs), and they can handle data from both the visual and textual domains. Evaluating LVLMs has become extremely important and several benchmarks have been designed for this pur-

pose (Xu et al., 2024; Cao et al., 2024). VQA is one of the most important tasks in this context. However, while there are works focusing on natural scenes and text-centric images, none focus on multilingual document VQA. This omission creates a relevant gap, as it limits our understanding of these models’ capabilities, since they are tested on a narrow range of visual domains.

In this work, we focus on text-centric document images. While there are some efforts in the state of the art for document VQA, they are exclusively in English. In light of this, the contributions of this work are the following:

- We introduce MaDOVE, a large-scale multilingual document VQA dataset designed to bridge the gap in evaluating multimodal reasoning capabilities over documents. To address the need for a large amount of data during training, we generate answers synthetically, while ensuring high-quality evaluation through two different test sets: one human-annotated and another LLM-annotated for scalable comparison¹.
- We study the effect of fine-tuning on a synthetically generated dataset of multilingual document VQA pairs and how this affects model performance on both text-centric and natural scene VQA datasets.
- We benchmark a diverse set of models on MaDOVE, offering a systematic study on multilingual document VQA and highlighting an important issue for current practices in VQA evaluation.

¹<https://huggingface.co/collections/m-elio/madove>

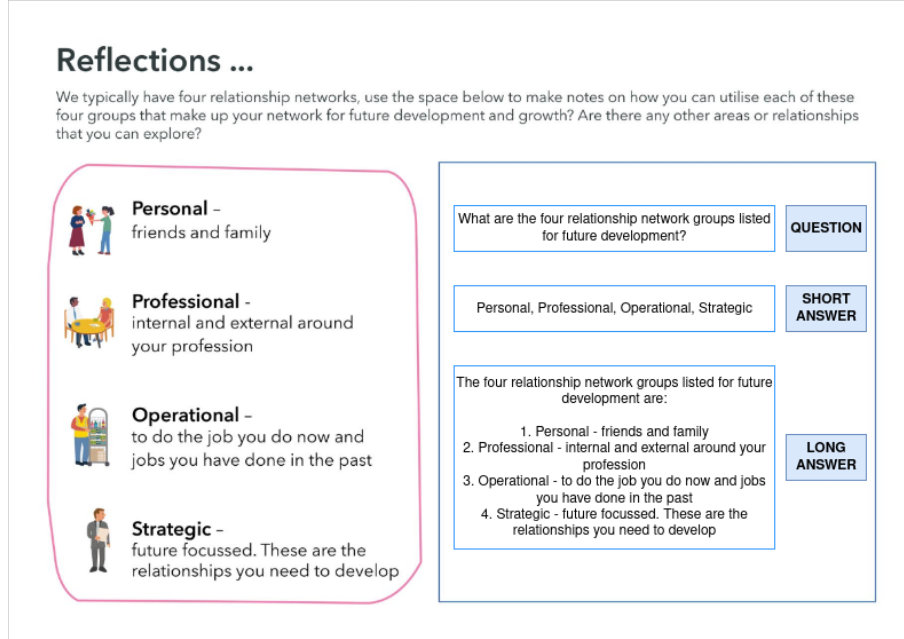


Figure 1: Example of an instance from our dataset. Each instance for each language consists of an image containing text in that language, a question related to the text in the image, and two answers: a long one (meaningfully detailed) and a short one (concise).

2 Related Works

2.1 Visual Question Answering

The original task of free-form and open-ended VQA was proposed by Antol et al. (2015), who introduced a dataset comprising over 250K images, 760K questions and 10M answers. Following this work, several datasets have been proposed to further evaluate model performance on VQA. These datasets include GQA (Hudson and Manning, 2019), which focuses on scene understanding of everyday images, TEXTVQA (Singh et al., 2019), which focuses on text embedded within natural scene images. The task was then specialized in specific domains or sources, for example, Shabtay et al. (2025) proposed LIVEXIV, a VQA dataset over scientific ArXiv papers, while Lau et al. (2018) proposed VQARAD, a VQA dataset of question-answer pairs on radiology images. Another text-centric benchmark is ST-VQA (Biten et al., 2019), which focuses on semantic information conveyed by text within a natural scene image.

The focus of this work is on text-centric document VQA. For this type of task, there are some existing works in the literature. DocVQA (Mathew et al., 2021) is a VQA dataset where images are pages of documents downloaded from UCSF Industry Documents Library, with questions and answers manually written by remote workers. This

is the first dataset to introduce document VQA with a focus on understanding the document structure. INFOVQA (Mathew et al., 2022) is similar to DocVQA but focuses on infographics, which are documents designed to effectively communicate information using a combination of textual, graphical and visual elements. Other relevant resources include CHARTQA (Masry et al., 2022) and MP-DOCVQA (Tito et al., 2022). CHARTQA focuses on visual and logical reasoning on charts, while MP-DOCVQA focuses on multi-image document VQA, rather than single image. For multilingual document VQA, relevant works include HW-MLVQA (Pal et al., 2025) and MTVQA (Tang et al., 2025). HW-MLVQA is an extensive collection of handwritten documents with expert human validation of question-answer pairs. MTVQA provides high-quality human expert annotations across nine diverse languages. Images are obtained from both publicly available datasets and Internet crawls. Annotations for questions and answers are obtained by human experts. Remarkably, this dataset also contains document images. However, MTVQA does not focus exclusively on document images, which limits the number of instances for this specific task in the train and test splits of the dataset, while HW-MLVQA focuses on handwritten documents and only covers the English and Hindi languages.

Despite the extensive effort and research in VQA, no benchmark specifically targets document images containing multilingual text. This highlights the relevance and diversity of MaDOVE w.r.t. other VQA datasets, allowing for the broadening of the scope of VQA evaluation.

2.2 Large Vision-Language Models

LVLMS are an extension of pre-trained LLMs to expand their inputs so that they can understand images together with text. The most relevant approach is that of LLAVA (Liu et al., 2023), where image embeddings extracted from a vision encoder are projected into the latent space of the LLM. This methodology was then further enhanced in LLAVA 1.5 HD (also known as LLAVA NEXT) (Liu et al., 2024). While initial works, including LLAVA, only focused on the English language, modern LVLMS support multiple languages. These models include GEMMA (Kamath et al., 2025), LLAMA 4 (Meta AI, 2025) and QWEN 2.5 VL (Bai et al., 2025), the most recent models developed by Google, META and Qwen, respectively. Finally, several studies have examined the application of LVLMS to VQA. Yang et al. (2025) propose a three-step process to integrate commonsense knowledge in LVLMS, while Shiri et al. (2024) study LVLMS for their spatial understanding and reasoning capabilities.

No work has investigated the effect of fine-tuning a LVLMS on a multilingual document VQA mixture or has established benchmarks in this setting. Hence, we will evaluate the performance of both pre-trained and fine-tuned models on MaDOVE and provide the first benchmark study for multilingual document VQA.

3 Dataset Creation

3.1 Data Collection

We collect the *vdr-multilingual* (Multilingual Visual Document Retrieval)² dataset train and test splits. This dataset is designed for visual document retrieval, where the objective is to find the image that answers the query. It is important to note that here the task is missing the last step of VQA, i.e., exploiting the information provided by the image to generate a factual answer to the question. In *vdr*, images were collected from online resources, while queries were generated synthetically, with additional manual verification for the

test set. The train set comprises 280,769 instances (across all languages), while the test set consists of 300 instances (one for each language). The dataset does not provide queries for all images, since additional images serve as hard negatives. In the aforementioned statistics, we considered only instances paired with a query. This represents a solid foundation for building our document VQA dataset. In addition, images were collected with a focus on multilinguality. The dataset supports five languages, with the text in each image written in a specific language. The supported languages are: English, French, German, Italian and Spanish. Hence, we will focus on these languages for creating our dataset. However, while the dataset provides queries and images, textual answers to the given queries are not provided. In light of this, we need to augment the dataset with textual answers to the queries. Given the dataset’s scale, manual annotation is infeasible; hence, we focus on synthetic generation using LVLMS.

3.2 Data Annotation

We perform synthetic generation over the queries for each language. Specifically, we leverage QWEN2.5-VL-72B-INSTRUCT. We chose this model since it represents a good trade-off between performance and efficiency. Furthermore, according to the technical report, the Qwen2.5 VL family of models shows powerful document parsing capabilities. We perform inference by providing the model with the image and its associated query from the original dataset. The model is asked to generate a response to the given question-image pair. Since document VQA datasets, that are DOCVQA and INFOVQA, focus on short answers, we append the string “*Answer the question using a single word or phrase.*” to the end of the prompt, forcing the model to generate short and concise answers. We will refer to this type of answer as “short”. We also repeat this process without this special string appended to the end of the prompt. We will refer to this type of answer as “long”. We perform annotation and validation for both “short” and “long” responses, but we will focus on the “short” responses during evaluation to follow the same style as DOCVQA and INFOVQA. We also add a condition to the prompt, where we ask the model to generate a specific string if the question cannot be answered, allowing us to easily filter these cases after completing the inference process. We use greedy decoding as the inference strategy

²<https://huggingface.co/blog/vdr-2b-multilingual>

to maximize efficiency and reproducibility. After filtering cases where the model was unable to provide the corresponding answer, we report statistics for the train set in Table 1. We show an annotated instance from our dataset for English in Figure 1. Finally, to further augment the test set, we also leverage the model to generate questions for images that are not associated with a query. We then perform the same inference process as before to generate the answers. We report complete technical details for synthetic generation in Appendix A. All the prompts used for each step described in this section are reported in Appendix F. Error analysis reporting common error patterns in the train set is shown in Appendix C. To further verify the effectiveness of this generation process, we validate the generated instances for the test set.

3.3 Data Validation

We perform validation by leveraging both: 1) a human annotator; 2) an LLM-as-a-judge methodology (Zheng et al., 2023) using the GPT-4o (Hurst et al., 2024) model. In both cases, we use a survey to evaluate three aspects of our generated VQA instances: 1) Grammatical correctness of questions and answers (both short and long); 2) Coherence of the answers (both short and long) w.r.t. the question; 3) Coherence of the question and answers (both short and long) w.r.t. the contents of the image. This allows us to cover all possible error cases that could be caused by synthetic generation. The survey consists of eight statements for each instance to evaluate. We report complete details for them in Appendix D. Each statement has to be paired with a judgment on a Likert scale from 1 to 5, where 1 is "Absolutely Disagree", 2 is "Disagree", 3 is "Uncertain", 4 is "Agree", and 5 is "Absolutely Agree". Additionally, we ask both the human annotator and the GPT-4o model to provide a brief explanation for their validation when a negative rating is assigned to any statement, as well as possible solutions to address the issue. For human annotation, we recruit a human annotator who is proficient in English, Italian and Spanish. For the LLM-as-a-judge methodology, we leverage GPT-4o for all languages in the dataset: English, French, German, Italian and Spanish. The human annotator only covers instances in the test set for which a query was already provided in the original *vdr* dataset, whereas for the LLM-as-a-judge methodology, we augment the overall cardinality by also considering instances for which we generated the

question. To strike a balance between efficiency and statistical significance, we limit each language subset to approximately one thousand instances. The prompt used for the LLM-as-a-judge methodology is provided in Appendix F. After applying both methodologies, we (the authors) manually validate the instances. We review all cases in which the validator (either the human annotator or GPT-4o) assigned a negative or neutral rating (1 to 3) to any statement. If the issue can be resolved easily, we manually revise the instance based on the validator’s explanation. When the issue cannot be easily fixed (e.g., the question is completely incoherent w.r.t. the answer), we remove it from the test set directly. We perform this step for all annotated test sets for each validator; therefore, we obtain eight test sets for each language (three annotated by humans and five annotated by the GPT-4o model). We report statistics for the final test sets in Table 2.

Lang	Answer Type	
	Short	Long
DE	44,580	47,990
EN	42,158	46,303
ES	47,733	45,261
FR	42,049	45,558
IT	43,109	45,316
TOTAL	219,629	230,428

Table 1: Train set statistics

Lang	Validator	
	Human	GPT-4o
DE	X	983
EN	230	990
ES	221	991
FR	X	988
IT	213	1,000
TOTAL	664	4,952

Table 2: Test set statistics

3.4 Annotation Comparison

We study the agreement between the human annotator and GPT-4o ratings. Specifically, we aim to understand whether annotations provided by GPT-4o are aligned with those provided by the human annotator. To do so, we study the ratings provided by the human and GPT-4o on the same instances (the subset of instances annotated by the human annotator in English, Italian and Spanish). In Figure 2, we present a radar chart showing the average ratings for the three languages across the eight statements. AVG refers to the agreement obtained by

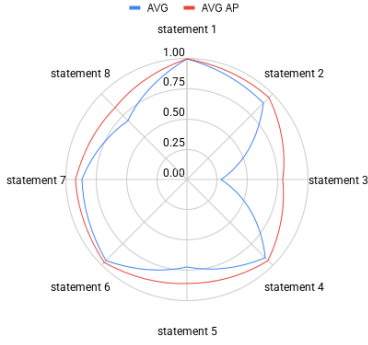


Figure 2: Radar chart for the average agreement between human and GPT-4o for the Likert scale rating obtained on each statement for the three languages.

hard matching the given ratings (that is, for each instance, the result is 1 if the given rating is exactly the same between the two annotators). AVG AP refers to the agreement obtained by matching the ratings on a scale from 1 to 3 rather than 1 to 5. Specifically, "Disagree" and "Agree" ratings (2 and 4, respectively) are merged into the "Absolutely Disagree" and "Absolutely Agree" ratings (1 and 5, respectively). We can appreciate that, except for Statements 3 and 8, which focus on grammatical correctness of the short answer and coherence of the short answer with respect to the image, there is a good degree of alignment between human and model for AVG. When considering AVG AP, the alignment significantly increases. Remarkably, the alignment increases for both Statements 3 and 8. However, we find that the degree of alignment is very low (for some statements, it is close to 0%) when considering only instances where the human annotator provided a negative rating for any statement (either 1 or 2). The same pattern also holds for negative ratings given by GPT-4o. We find that this is due to the following reasons: 1) The number of instances with negative ratings is low (average of around 10% for the three languages for both the human and the GPT-4o model) w.r.t. the number of instances with positive ratings. After manually verifying these instances, we find that most of them are limit cases, where it is reasonable to expect different answers depending on the annotator’s background and sensibility. For example, the human annotator provided “Absolutely Disagree” ratings for instances where the answer was a LaTeX formula, since they did not know LaTeX; 2) The GPT-4o model provided higher ratings w.r.t. the human annotator. For example, in the English subset, the GPT-4o model has never provided an

“Absolutely Disagree” rating, whereas the human annotator did so for several instances. We also find that the model treated the “Uncertain” rating as a negative rather than a neutral rating in limit cases.

We emphasize that there are cases where the human annotator gave ratings to instances better than the GPT-4o model and vice versa, highlighting that both methodologies should be considered when performing this type of annotation. Still, the LLM-as-a-judge approach is a suitable methodology in our use case, especially considering that the outputs are further refined and filtered after the data validation step.

4 Experiments

We evaluate LVLMs on both the human-annotated and model-annotated test sets. We consider two strategies: 1) *zero-shot inference*, where we leverage pre-trained LVLMs to perform inference directly; 2) *fine-tuning*, where we fine-tune a LVLM on the train split of our dataset. We will also evaluate the fine-tuned models on state-of-the-art multimodal tasks to understand the performance drop on tasks unrelated to document VQA. Finally, we will also evaluate both zero-shot and fine-tuned models on other English language document VQA benchmarks, specifically DOCVQA and INFOVQA, to understand whether fine-tuning on our training mixture can also enhance performance on other related benchmarks. We will evaluate outputs generated on our benchmark using the same metric as DOCVQA and INFOVQA, that is *Average Normalised Levenshtein Similarity* (ANLS), originally proposed for evaluation on ST-VQA, which takes into account minor answer mismatches.

For fine-tuning, we consider two different types of models: QWEN 2.5 VL 3B, 7B and 32B, which are smaller versions of the original model used for synthetic generation (72B); LLAVA NEXT using a checkpoint trained with LLaMA 3 as the LLM backbone. We consider smaller versions of QWEN 2.5 VL 72B to assess whether distillation from a large model to smaller models can yield significant performance gains, while we examine a LLAVA NEXT model to determine whether a different architecture can still achieve remarkable performance improvements after fine-tuning. We train for one epoch over the entire training mixture, combining both short and long answers. For short answers, we append a string to the question to request a concise response. During fine-tuning, we use LORA (Hu

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.8668	.7379	.5194	.3087	.6082	.9103	.7923	.5280	.2971	.6319
ES	.8389	.6334	.3109	.1227	.4765	.8583	.6494	.2623	.1201	.4725
IT	.8000	.6359	.3691	.2080	.5033	.8538	.7298	.2830	.0604	.4817
	Qwen2.5-VL-32B-Instruct					gemma-3-4b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.8981	.7525	.4808	.1887	.5800	.7246	.3941	.0355	.0000	.2885
ES	.8732	.7718	.2530	.0955	.4984	.5929	.5319	.0606	.0000	.2963
IT	.8152	.7533	.3564	.1251	.5125	.6164	.5317	.0793	.0000	.3068
	gemma-3-12b-it					gemma-3-27b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.7896	.5404	.0825	.0179	.3576	.7005	.4859	.1182	.0212	.3314
ES	.7555	.6889	.1781	.0000	.4056	.6018	.6115	.1279	.0329	.3435
IT	.5827	.7877	.3456	.0250	.4353	.6733	.5876	.2310	.0458	.3844
	Llama-4-Scout-17B-16E-Instruct					Llama-4-Maverick-17B-128E-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.8289	.6245	.3474	.1200	.4802	.8425	.7204	.4228	.1425	.5321
ES	.8635	.6683	.2435	.2220	.4993	.8278	.7180	.2563	.1128	.4787
IT	.7878	.7074	.3628	.2005	.5146	.8741	.7593	.4382	.1803	.5630
	llama3-llava-next-8b-hf									
	Q1	Q2	Q3	Q4	AVG					
EN	.5075	.3232	.0598	.0000	.2226					
ES	.3958	.3122	.0528	.0147	.1939					
IT	.4572	.3181	.0495	.0000	.2062					

Table 3: Complete results for the pre-trained models on the human-annotated test set.

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9343	.7907	.6556	.4797	.7151	.9294	.7883	.6652	.5671	.7375
ES	.9142	.8048	.6215	.5084	.7122	.8742	.8398	.7031	.6395	.7642
IT	.8201	.7045	.4573	.4907	.6182	.8731	.7732	.5406	.5979	.6962
	Qwen2.5-VL-32B-Instruct					llama3-llava-next-8b-hf				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9602	.7852	.7382	.5985	.7705	.7155	.4532	.4832	.2431	.4737
ES	.9214	.8055	.6998	.6318	.7646	.6059	.5136	.3135	.2033	.4091
IT	.8433	.7769	.5660	.5882	.6936	.6267	.5416	.3619	.2961	.4566

Table 4: Complete results for the fine-tuned models on the human-annotated test set.

et al., 2022) to improve efficiency while retaining the model’s original knowledge. Complete fine-tuning details, including hyperparameters, can be found in Appendix B. Additionally, we benchmark multimodal and multilingual models, specifically GEMMA 3 4B, 12B and 27B, as well as LLAMA 4 17B with 16 experts (Scout) and 17B with 128 experts (Maverick). During inference, the prompt consists of: the image tokens, the question and an additional string to force the model to generate a short response. The images are provided in their original resolution as in the *vdr* dataset. We also apply the chat template expected by each model. At inference time, we use greedy decoding to maximize efficiency and reproducibility. Additional details regarding the prompt used during inference can be found in Appendix F.

4.1 Results

After a preliminary evaluation of the results, we found that Qwen2.5 VL 3B and 7B outperformed all other models on most test sets. This was surprising, since the benchmark also considers models that are bigger in terms of total parameter count, specifically Qwen2.5 VL 32B and Llama Maverick. Intuitively, one would expect performance to improve with model scale. After validating the outputs, we find that this is caused by the mismatch in answer length and style across models. Specifically, the bigger the difference in length between the label and the generated answer, the more likely it was for ANLS to miss possibly correct answers. This is due to the response style that each model adopted for each instance. We provide a complete example in Appendix E. Hence, we decided to study how label length affects the evaluation process. We split each test set into quartiles according to the length of the ground truth answer. Therefore, Q1 will con-

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.7353	.6257	.3581	.1731	.4730	.8419	.5765	.1751	.0947	.4221
EN	.8777	.6076	.3853	.1655	.5090	.8935	.6480	.4336	.2184	.5484
ES	.8476	.6283	.2823	.1544	.4781	.8615	.6136	.2251	.0990	.4498
FR	.8393	.6365	.3578	.2106	.5111	.9027	.6447	.2981	.1228	.4921
IT	.7781	.5851	.3160	.1596	.4597	.8269	.5315	.1704	.0555	.3961
	Qwen2.5-VL-32B-Instruct					gemma-3-4b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.7656	.6515	.3397	.2026	.4899	.6848	.4736	.1533	.0345	.3366
EN	.8804	.5647	.2313	.1000	.4441	.6640	.0608	.0000	.0000	.1812
ES	.8734	.6209	.2399	.1137	.4620	.6618	.3363	.0148	.0037	.2542
FR	.8249	.6126	.4098	.2401	.5219	.6509	.4173	.1247	.0205	.3034
IT	.8067	.5460	.2299	.1068	.4223	.6519	.1915	.0061	.0052	.2137
	gemma-3-12b-it					gemma-3-27b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.4261	.4999	.3485	.1035	.3445	.7316	.5208	.1821	.0486	.3708
EN	.7536	.1878	.0065	.0000	.2370	.6730	.2290	.0502	.0150	.2418
ES	.7702	.5175	.0908	.0120	.3476	.7151	.4149	.0864	.0146	.3077
FR	.7138	.6267	.3249	.0864	.4380	.7343	.4626	.1978	.0433	.3595
IT	.6914	.4800	.0862	.0217	.3198	.7039	.3776	.0733	.0117	.2916
	Llama-4-Scout-17B-16E-Instruct					Llama-4-Maverick-17B-128E-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.6965	.5940	.3421	.2213	.4635	.7972	.6463	.2838	.1417	.4673
EN	.8304	.4475	.1905	.0832	.3879	.8783	.5270	.2012	.0788	.4213
ES	.8235	.5900	.2823	.1837	.4699	.8529	.6596	.2549	.1372	.4762
FR	.7243	.6050	.3105	.2395	.4698	.7697	.5935	.2731	.1011	.4344
IT	.8080	.5577	.2639	.2009	.4576	.8247	.5945	.3126	.1346	.4666
	llama3-llava-next-8b-hf									
	Q1	Q2	Q3	Q4	AVG					
DE	.4272	.2666	.0456	.0046	.1860					
EN	.5345	.2158	.0133	.0206	.1960					
ES	.4702	.3373	.0759	.0200	.2258					
FR	.5230	.2792	.0830	.0094	.2236					
IT	.4259	.2171	.0339	.0064	.1708					

Table 5: Complete results for the pre-trained models on the GPT-4o annotated test set.

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.8299	.7177	.5585	.4162	.6306	.8414	.7682	.5711	.4625	.6608
EN	.9231	.7059	.5977	.3663	.6483	.9147	.7150	.6528	.4075	.6725
ES	.8633	.8123	.6421	.5357	.7134	.8772	.8361	.6807	.5670	.7403
FR	.8751	.7707	.6447	.5172	.7020	.8808	.7974	.6456	.5405	.7161
IT	.8356	.7148	.5981	.4156	.6410	.8620	.7367	.6391	.4656	.6758
	Qwen2.5-VL-32B-Instruct					llama3-llava-next-8b-hf				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.8508	.7972	.6026	.4966	.6868	.6158	.5389	.3064	.1546	.4039
EN	.9341	.7609	.7156	.4440	.7137	.7140	.5825	.4249	.2340	.4889
ES	.8919	.8298	.7194	.5809	.7555	.6149	.5746	.4685	.2933	.4878
FR	.9203	.8202	.6810	.5649	.7466	.6429	.5424	.4044	.2717	.4653
IT	.8589	.7421	.6710	.5054	.6943	.6403	.4431	.3501	.2188	.4131

Table 6: Complete results for the fine-tuned models on the GPT-4o annotated test set.

	Model	Document VQA		Scene VQA		General
		DOC VQA	INFO VQA	GQA	TEXT VQA	MME
PT	Qwen2.5-VL-3B-Instruct	.9220	.7185	.6013	.7787	622.5
	Qwen2.5-VL-7B-Instruct	.9358	.8293	.6042	.8295	645.71
	Qwen2.5-VL-32B-Instruct	.9275	.8103	.5924	.7715	713.93
FT	Qwen2.5-VL-3B-Instruct	.9251	.7481	.5625	.7466	611.79
	Qwen2.5-VL-7B-Instruct	.9477	.8179	.5693	.7663	622.86
	Qwen2.5-VL-32B-Instruct	.9350	.8419	.5535	.7629	715

Table 7: Complete results for the SOTA benchmarks.

tain the results for the shortest labels, while Q4 will contain the results for the longest labels. We report complete results for the human-annotated test set in Table 3 and Table 4 for the pre-trained and fine-tuned models, respectively. We report complete results for the GPT-4o annotated test set in Table 5 and Table 6 for the pre-trained and fine-tuned models, respectively. Overall, for pre-trained models, we find that performance is remarkable for the QWEN 2.5 VL models, both in zero-shot inference and with fine-tuning. LLAMA 4 models also perform well, while GEMMA models have the lowest performance, excluding LLAVA which performs worst out of all the models. We also highlight that, due to the previously described issue, models tend to perform better in Q1 and Q2 and worse in Q3 and Q4. Furthermore, given that the labels with the longest length are in Q4, it is effectively the most difficult quartile to match the ground truth answer and the LLM-generated output. Still, it is remarkable that Llama-4-Scout achieves competitive performance on both Q3 and Q4. Despite this, we find that Llama-4-Scout has low performance on Q3 and Q4 for the English language only. Following this, we decided to plot the average performance of pre-trained models over the quartiles for each language. We report complete plots, performance per language and discuss the results in Appendix E. Fine-tuned Qwen models perform best across all test sets. Fine-tuned LLaVA NeXT also improves markedly over its pre-trained version; hence, fine-tuning improves performance regardless of the underlying architecture. To further improve the soundness of our findings, we perform evaluation with BertScore F1 (Zhang et al., 2020) as well, using the evaluate library implementation³. We report complete results for BertScore F1 in Appendix E. Overall, the metric follows patterns similar to ANLS, highlighting that current metrics have difficulty evaluating open-ended VQA.

4.2 Results for State of the Art Datasets

We evaluate the fine-tuned models on popular multimodal state-of-the-art benchmarks. We consider the following categories: *Document VQA*, *Scene VQA* and *General*. *Document VQA* refers to tasks where the images are document pages, *Scene VQA* to tasks where the images are natural scenes or everyday images, *General* to comprehensive multimodal benchmarks. For *Document VQA* we use

DOCVQA and INFOVQA, for *Scene VQA* we use GQA and TEXTVQA, for *General* we use MME (Fu et al., 2023). MME is a benchmark that measures both perceptual and cognitive abilities across 14 subtasks. We perform evaluation on these benchmarks by using the *lmms-eval*⁴ library. We report the main metric used by each task, for MME we report the perception score. Overall, performance either increases or remains stable for *Document VQA* benchmarks, while it becomes worse for *Scene VQA* and *General* ones. The decrease in performance is expected, since the models have been extensively fine-tuned for the document VQA task. However, it is remarkable that the fine-tuned models still achieve strong performance on both the *Scene VQA* and *General* benchmarks. Specifically for *General* benchmarks, the decrease in performance is lower than that for *Scene VQA*, demonstrating that the generalization ability of the fine-tuned LVLM remains remarkable. Furthermore, there is a similar pattern to the one previously explained for our benchmark for state-of-the-art datasets. For the pre-trained Qwen2.5 VL models, performance is inconsistent with model scale across all VQA benchmarks. This further highlights the importance of analyzing and studying the outputs in open-ended VQA tasks.

5 Conclusions

In this work, we proposed **MaDOVE**, the first multilingual document VQA dataset. We consider the English, German, French, Italian and Spanish languages, for which data collection has been performed starting from the *vdr* dataset. After collecting the data and generating synthetic answers, we perform extensive experiments on state-of-the-art LVLMs, considering both *zero-shot inference* for thirteen models from four different families and *fine-tuning*. Results show that, while LVLMs are capable of handling multilingual document VQA, fine-tuning can lead to superior performance. However, we also highlight that extracting the generated answers from zero-shot pre-trained models and comparing them with the ground-truth labels proved difficult for both our dataset and state-of-the-art datasets. In the future, we plan to further extend the dataset to additional languages and study new techniques to improve LVLM’s effectiveness for document VQA. We also plan to design and

³<https://github.com/huggingface/evaluate>

⁴<https://github.com/EvolvingLMs-Lab/lmms-eval>

implement a metric that accounts for style and language variations in VQA.

Limitations

We recognize the following limitations in our work:

- **Human annotation:** we considered only three of the original dataset’s languages, and a single human annotator performed the annotations. Due to the high annotation cost required in this task, since annotating even a single instance can require a lot of time, depending on the amount of text within the image, we were unable to further extend the human annotation process. Hence, we note that the absence of multiple annotators limits the assessment of annotation consistency. Furthermore, LLM validation may affect perceived dataset quality, and agreement metrics should be interpreted with caution due to potential asymmetry in rating bias.
- **Languages covered in the dataset:** while this benchmark does not focus on the English language, it provides instances for only four other languages. However, since we relied on the *vdr* dataset, we would have had to extend it to cover more languages. We plan to pursue this as future work, building on the successful experiments we conducted in this study.
- **Limited cardinality of the test set:** to strike a balance between efficiency and effectiveness, we considered only about 1,000 instances per language in the GPT-4o-annotated test set. We recognize that extending the dataset using the LLM-as-a-judge methodology could further improve the soundness of the results.

Ethics statement

In this work, we have performed manual annotation with a human annotator. The annotator is a woman aged 25-65. The annotator is a native speaker of Italian. The annotator was instructed to prioritize their native language and then proceed only to languages in which they were fluent (English and Spanish). The participant is an employee of Fastweb SpA and performed this task as part of their standard professional duties. Informed consent was obtained regarding the purpose of the study and the utilization of the resulting data. Compensation was provided via the participant’s standard employment contract, which adheres to local

labor laws and exceeds the national minimum wage. To ensure high-quality data and prevent fatigue, the annotation workload was limited to six hours per day over a one-week period. The *vdr* dataset is on Huggingface under the Apache License 2.0, a permissive license whose main conditions require preservation of copyright and license notices. We will release MaDOVE under the same license.

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 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. [Scene text visual question answering](#). In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4290–4300. IEEE.
- Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. 2024. [VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12161–12176, Bangkok, Thailand. Association for Computational Linguistics.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). *CoRR*, abs/2306.13394.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: A new dataset for real-world visual reasoning and compositional question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20,*

- 2019, pages 6700–6709. Computer Vision Foundation / IEEE.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. [Gpt-4o system card](#). *CoRR*, abs/2410.21276.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 191 others. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. [Visual question answering: A survey of methods, datasets, evaluation, and challenges](#). *ACM Comput. Surv.*, 57(10):249:1–249:35.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. 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, Dublin, Ireland, May 22-27, 2022*, pages 2263–2279. Association for Computational Linguistics.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. [Infographicvqa](#). In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. [Docvqa: A dataset for VQA on document images](#). In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.
- Meta AI. 2025. [The llama 4 herd: The beginning of a new era of natively multimodal ai innovation](#).
- Aniket Pal, Ajoy Mondal, Minesh Mathew, and C. V. Jawahar. 2025. [HW-MLVQA: elucidating multilingual handwritten document understanding with a comprehensive VQA benchmark](#). *CoRR*, abs/2507.15655.
- Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, Muhammad Jehanzeb Mirza, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. 2025. [Livexiv - A multi-modal live benchmark based on arxiv papers content](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. [An empirical analysis on spatial reasoning capabilities of large multimodal models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA. Association for Computational Linguistics.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.
- Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025. [MTVQA: benchmarking multilingual text-centric visual question answering](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 7748–7763. Association for Computational Linguistics.
- Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. [Hierarchical multimodal transformers for multi-page docvqa](#). *CoRR*, abs/2212.05935.
- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. [Lvlm-hub: A comprehensive evaluation benchmark for large vision-language models](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Shuo Yang, Caren Han, Siwen Luo, and Eduard H. Hovy. 2025. [MAGIC-VQA: multimodal and grounded inference with commonsense knowledge for visual question answering](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16967–16986. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Synthetic Generation Details

For synthetic generation, we use the Qwen2.5 VL 72B model by loading it on two H100 80 GB VRAM GPUs. We use the vLLM framework⁵ to optimize local inference on this model. Generating the long answers for each language in the train set takes at most three days, so a total of fifteen days is required. Generating the short answers for each language in the train set takes at most two days for each language, so a total of ten days is required. Generating the long answers for each language in the test set takes at most two hours for each language, so a total of ten hours is required. Generating the short answers for each language in the test set takes at most one hour for each language, so a total of five hours is required.

B Training Details

We perform fine-tuning with LoRA. We train for one epoch over the entire train split of the dataset, considering both short and long answers. For short answers, we append the strings to constrain short answer generation that are reported in Appendix F. We use the following hyperparameters: 64 effective batch size, 2e-5 learning rate, 0.0 warmup ratio, adamw optimizer. For LoRA we use the following hyperparameters: 8 rank, 16 alpha, 0.05 dropout, targeting all linear modules. All images are resized to 672x672 for efficiency. Fine-tuning is performed on a single H100 80 GB VRAM, except for the 32B model where two H100 were used. For all models, fine-tuning took at most 24 hours in total.

C Error Analysis in Train Set Generation

We show common error patterns that were manually found from the train set of the generated dataset. The two patterns that occurred the most are shown in Figure 3. In the first pattern, the model has provided text by “copy-pasting” the textual contents present in the image. However, while the textual contents are correct, the answer is not complete

w.r.t. the contents of the question. In the example provided, the model did not provide the four fundamental aspects of Computational Thinking, which are provided in a later section of the text. In the second pattern, the model fails to generate correct answers when the question is very specific and the visual contents of the images are complex. In the example provided, the model provided the answer “2”, which refers to the Standard Group (“Standard-Gruppe” in the image) rather than the Isokinetic Group (“Isokinetik-Gruppe” in the image). We want to underline that these patterns are consistent across all languages included in the dataset. These errors imply that: 1) the Qwen2.5-VL-72B model fails to provide complete and detailed answers when the textual contents of the images are long; 2) the Qwen2.5-VL-72B model is capable of understanding simple layouts and images, but may fail for complex images and specific questions.

D Annotation Details

Both the human annotator and the LLM were provided with the following eight statements to evaluate on a Likert scale in the 1-5 range:

1. The question is grammatically correct.
2. Answer A is grammatically correct.
3. Answer B is grammatically correct.
4. Answer A is coherent with respect to the question.
5. Answer B is coherent with respect to the question.
6. The question is coherent with respect to the image.
7. Answer A is coherent with respect to the image.
8. Answer B is coherent with respect to the image.

In Table 15, we report the guidelines provided to the human annotator.

Screenshots of the annotation platform used by the human annotator are provided in Figure 7 and in Figure 8.

⁵<https://github.com/vllm-project/vllm>

Computational Thinking: di cosa si tratta e perchè è importante?

Il termine Computational Thinking (pensiero computazionale) è stato coniato da Jeannette Wing nel 2006 alla Carnegie Mellon University, per descrivere una tipologia di approccio alla risoluzione dei problemi.

In breve «il pensiero computazionale è un approccio alla soluzione dei problemi».

Non si tratta di un'abilità specifica ma una gamma di concetti, applicazioni, strumenti e strategie di pensiero che vengono utilizzati per risolvere i problemi. Inoltre, è possibile praticare il pensiero computazionale senza utilizzare il computer.

Jeannette Wing definisce quattro aspetti principali del pensiero computazionale:

- 1) Decomposizione: scomporre un problema in parti più piccole;
- 2) Riconoscimento del modello: trovare somiglianze e differenze tra le diverse parti per essere in grado di fare previsioni;

3) Astrazione: capacità di trovare i principi generali dietro le parti scomposte in precedenza e gli schemi nei problemi;

4) Progettazione dell'algoritmo: sviluppare le istruzioni passo dopo passo per risolvere i diversi problemi.

Siccome la tecnologia continua a progredire e trasformare la nostra società è imperativo per gli studenti e i giovani imparare a pensare in modo critico ed essere in grado di controllare e sviluppare la propria esperienza digitale. I giovani devono aspirare a diventare produttori di contenuti e tecnologie digitali piuttosto che semplici consumatori.

Inoltre, insegnare ai giovani il pensiero computazionale permette loro di capire come funzionano le tecnologie digitali ed è importante per garantire un utilizzo attivo e corretto di questi strumenti.

Il contatto precoce con le sfide del pensiero computazionale può ispirare gli studenti a scegliere diversi campi di studio legati alle discipline STEAM.

Quale definizione Jeannette Wing ha dato del pensiero computazionale nel 2006?

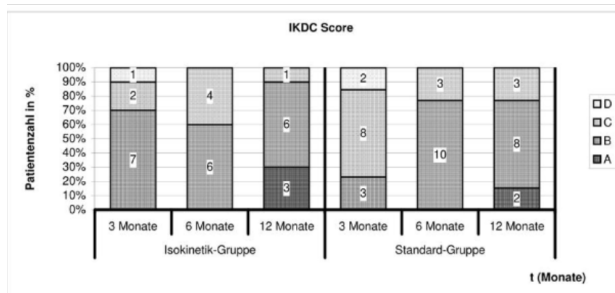
QUESTION

(How did Jeannette Wing define computational thinking in 2006?)

Jeannette Wing ha coniato il termine Computational Thinking (pensiero computazionale) nel 2006 alla Carnegie Mellon University per descrivere una tipologia di approccio alla risoluzione dei problemi. In breve, il pensiero computazionale è un approccio alla soluzione dei problemi.

ANSWER

(Jeannette Wing coined the term "computational thinking" in 2006 at Carnegie Mellon University to describe a particular approach to problem-solving. In short, computational thinking is an approach to problem-solving.)



Wie viele Patienten befanden sich nach 12 Monaten in der Isokinetik-Gruppe in der IKDC Kategorie A?

QUESTION

(How many patients in the isokinetic group were in IKDC Category A after 12 months?)

2

ANSWER

Figure 3: Two error patterns found in manual validation of the generated data. Translation to English of the questions and answers is provided in brackets.

E Additional Results and Examples

We report complete results for **BertScore** in Table 8 and Table 9 for the pre-trained and fine-tuned models on the human annotated test set, while in Table 10 and Table 11 for the pre-trained and fine-tuned models on the GPT-4o annotated test set.

We report per-language results in Figure 4 for the human-annotated test set and in Figure 5 for the GPT-4o annotated test set. These radar chart show that there are models which consistently achieve better performance across all languages. For example, both Llama-4-Scout and Llama-4-Maverick show optimal performance for non-English languages, but tend to perform worse on English (except for the human-annotated test set for Llama-4-Scout). Similar trends can also be seen in other models, for example Qwen2.5 VL 7B overperforms in English but has lower performance for the other languages. Mostly, these charts highlight that model performance across languages and label length is not consistent, and that these factors need to be studied when evaluating multilingual and multimodal models. Furthermore, in terms of average performance across quartiles, the model with the largest scale doesn't always performs best,

further highlighting the issue of evaluating pre-trained models due to the style and length of their responses.

We did not consider LLM-as-a-judge as a possible alternative metric since the labels are mostly short answers, while the LLM-as-a-judge methodology was introduced to address long and conversational formats.

We report an example of linguistic style variation across model scale in Figure 6.

F Prompts

We report all prompts used in this work in Table 12, Table 13, Table 14.

The following strings are used in the prompts provided in Table 12 and Table 13.

Negative message strings for answer generation ({NEGATIVE_MESSAGE_LANG_ANSWER}):

- DE: "Es gibt keine Informationen zu dieser Frage im Bild"
- EN: "There is no information regarding this question in the image"

- ES: “No hay información sobre esta pregunta en la imagen”
- IT: “L’immagine non contiene informazioni su questa domanda”
- FR: “Il n’y a pas d’information concernant cette question dans l’image”

Negative message strings for question generation ({NEGATIVE_MESSAGE_LANG_QUESTION}):

- DE: “Auf dem Bild ist kein Text zu sehen”
- EN: “There is no text in the image”
- ES: “No hay texto en la imagen”
- FR: “Il n’y a pas de texte dans l’image”
- IT: “L’immagine non contiene testo”

Finally, the prompt used at training and inference time consists of the question for the instance, followed by a string to force short answer generation (this string is not appended when the answer is a long form one). The short answer generation strings are the following:

- DE: “\nBeantworten Sie die Frage mit einem einzigen Wort oder Satz.”
- EN: “\nAnswer the question using a single word or phrase.”
- ES: “\nResponda a la pregunta con una sola palabra o frase.”
- IT: “\nRispondi alla domanda utilizzando una sola parola o frase.”
- FR: “\nRépondez à la question en utilisant un seul mot ou une seule phrase.”

G Licenses

The dataset will be licensed under Apache License 2.0, that is the original license of the *vdr* dataset.

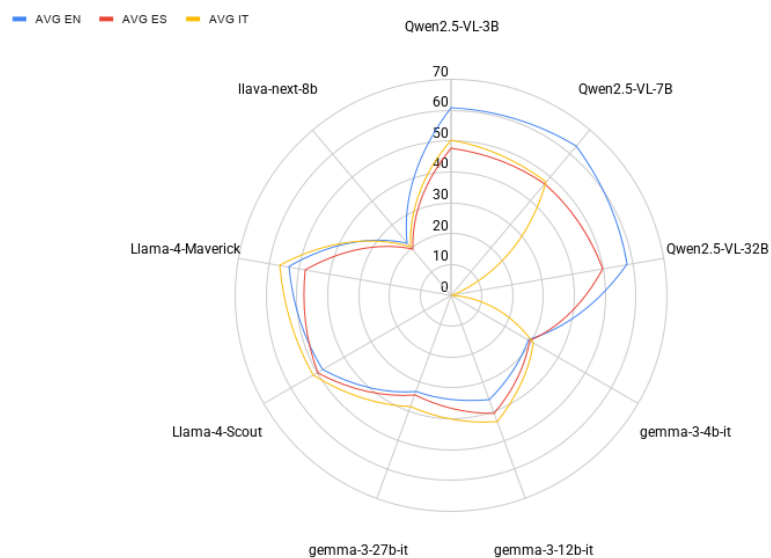


Figure 4: Radar chart for average performance over the quartiles for each pre-trained model on the human-annotated test set.

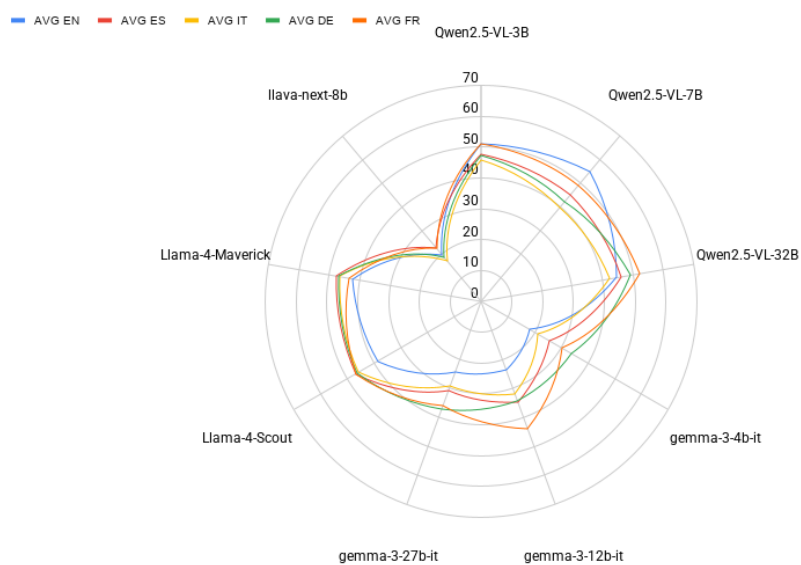


Figure 5: Radar chart for average performance over the quartiles for each pre-trained model on the GPT-4o annotated test set.

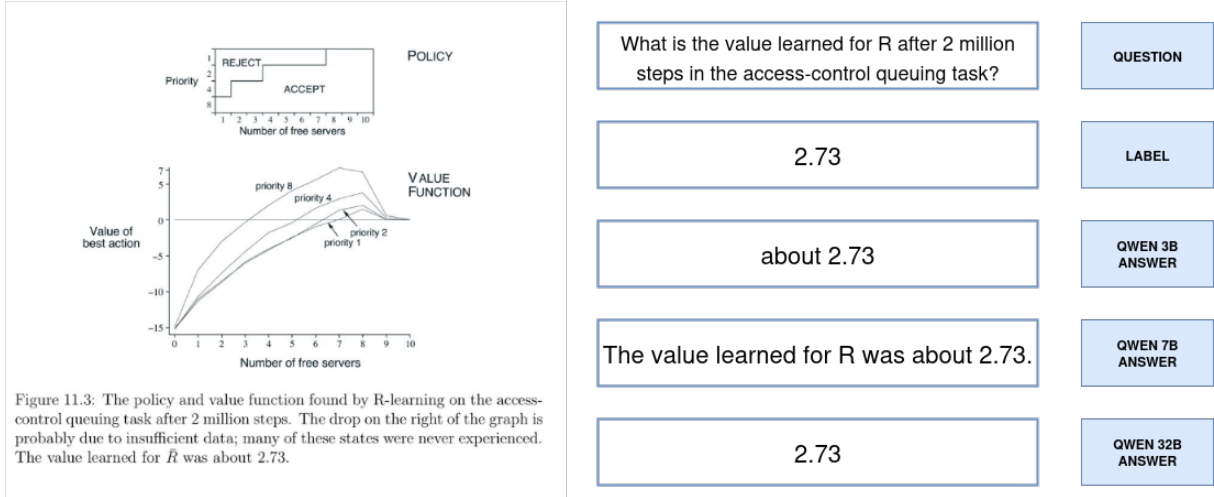


Figure 6: Example of different styles of answer obtained for three models. All answers are correct, but all metrics would provide a higher score to the response that perfectly matches the ground truth.

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9881	.9650	.9303	.9095	.9482	.9908	.9638	.9387	.9123	.9514
ES	.9603	.8991	.8055	.7453	.8525	.9695	.9069	.7838	.7309	.8478
IT	.9652	.8982	.7954	.7591	.8545	.9640	.9129	.7711	.6832	.8328
	Qwen2.5-VL-32B-Instruct					gemma-3-4b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9907	.9647	.9291	.8813	.9414	.9778	.9122	.8473	.8283	.8914
ES	.9661	.9292	.7765	.7111	.8457	.9186	.8781	.6967	.6250	.7796
IT	.9553	.9252	.8043	.7067	.8479	.9234	.8443	.6875	.6208	.7690
	gemma-3-12b-it					gemma-3-27b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9806	.9353	.8745	.8428	.9083	.9742	.9249	.8722	.8418	.9033
ES	.9512	.9105	.7595	.6454	.8166	.9251	.8847	.7332	.6498	.7982
IT	.8991	.9221	.7848	.6783	.8211	.9293	.8774	.7620	.6617	.8076
	Llama-4-Scout-17B-16E-Instruct					Llama-4-Maverick-17B-128E-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9859	.9495	.9081	.8742	.9294	.9865	.9560	.9226	.8808	.9365
ES	.9648	.9004	.7729	.7689	.8517	.9500	.9100	.7836	.7294	.8432
IT	.9344	.8969	.7900	.7362	.8394	.9705	.9191	.8311	.7488	.8674
	llama3-llava-next-8b-hf									
	Q1	Q2	Q3	Q4	AVG					
EN	.9324	.8980	.8518	.8305	.8781					
ES	.8081	.7721	.6774	.6349	.7231					
IT	.8313	.7640	.6573	.6151	.7169					

Table 8: Complete **BertScore** results for the pre-trained models on the human annotated test set.

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9927	.9727	.9583	.9408	.9661	.9921	.9734	.9614	.9433	.9676
ES	.9793	.9387	.8983	.8707	.9218	.9809	.9551	.9039	.8987	.9347
IT	.9515	.8895	.8612	.8706	.8932	.9643	.9142	.8780	.8959	.9131
	Qwen2.5-VL-32B-Instruct					llama3-llava-next-8b-hf				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
EN	.9961	.9725	.9669	.9497	.9713	.9733	.9308	.9278	.9159	.9369
ES	.9859	.9468	.9091	.8966	.9346	.9346	.8677	.8053	.8103	.8545
IT	.9601	.9274	.8861	.8988	.9181	.9076	.8511	.8194	.8075	.8464

Table 9: Complete **BertScore** results for the fine-tuned models on the human annotated test set.

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.9176	.8860	.8362	.7774	.8543	.9544	.8728	.7362	.7173	.8202
EN	.9845	.9462	.9249	.9007	.9391	.9825	.9557	.9346	.9123	.9463
ES	.9451	.8745	.7871	.7490	.8389	.9523	.8694	.7657	.7117	.8248
FR	.9520	.8868	.8193	.7925	.8626	.9663	.8863	.8018	.7569	.8528
IT	.9354	.8605	.8003	.7419	.8345	.9413	.8395	.7293	.6660	.7940
	Qwen2.5-VL-32B-Instruct					gemma-3-4b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.9247	.8877	.8227	.7780	.8533	.9119	.8331	.7488	.7038	.7994
EN	.9826	.9388	.9007	.8712	.9233	.9600	.8654	.8360	.8218	.8708
ES	.9540	.8696	.7764	.7265	.8316	.9011	.7910	.6779	.6330	.7508
FR	.9356	.8774	.8293	.7986	.8602	.8954	.8079	.7380	.6981	.7849
IT	.9344	.8450	.7544	.7080	.8105	.8934	.7461	.6506	.6132	.7258
	gemma-3-12b-it					gemma-3-27b-it				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.8167	.8483	.8310	.7670	.8157	.9202	.8469	.7596	.7190	.8114
EN	.9675	.8927	.8499	.8263	.8841	.9594	.8955	.8597	.8453	.8900
ES	.9275	.8405	.7308	.6682	.7918	.9205	.8156	.7066	.6491	.7730
FR	.9137	.8743	.8185	.7637	.8426	.9190	.8356	.7735	.7333	.8153
IT	.9093	.8245	.7186	.6637	.7790	.9157	.8038	.7004	.6419	.7654
	Llama-4-Scout-17B-16E-Instruct					Llama-4-Maverick-17B-128E-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.9054	.8778	.8212	.7871	.8479	.9343	.8862	.8008	.7556	.8442
EN	.9730	.9221	.8923	.8676	.9138	.9796	.9361	.8951	.8674	.9195
ES	.9369	.8537	.7879	.7588	.8343	.9447	.8780	.7922	.7326	.8369
FR	.9172	.8671	.8040	.7846	.8432	.9254	.8754	.8023	.7539	.8392
IT	.9258	.8409	.7694	.7425	.8197	.9460	.8633	.7939	.7331	.8341
	llama3-llava-next-8b-hf									
	Q1	Q2	Q3	Q4	AVG					
DE	.8132	.7353	.6521	.6384	.7098					
EN	.9271	.8815	.8444	.8349	.8720					
ES	.8135	.7467	.6937	.6485	.7256					
FR	.8299	.7551	.6842	.6656	.7337					
IT	.8095	.7245	.6534	.6197	.7018					

Table 10: Complete **BertScore** results for the pre-trained models on the GPT-4o annotated test set.

	Qwen2.5-VL-3B-Instruct					Qwen2.5-VL-7B-Instruct				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.9507	.9183	.8859	.8480	.9007	.9511	.9347	.8970	.8580	.9102
EN	.9869	.9637	.9563	.9319	.9597	.9869	.9662	.9609	.9373	.9628
ES	.9506	.9305	.9065	.8806	.9170	.9592	.9408	.9207	.8895	.9275
FR	.9605	.9332	.9062	.8779	.9194	.9643	.9420	.9074	.8823	.9240
IT	.9549	.9052	.8897	.8494	.8998	.9602	.9109	.9045	.8678	.9108
	Qwen2.5-VL-32B-Instruct					llama3-llava-next-8b-hf				
	Q1	Q2	Q3	Q4	AVG	Q1	Q2	Q3	Q4	AVG
DE	.9543	.9388	.9017	.8709	.9164	.8938	.8509	.8066	.7716	.8307
EN	.9895	.9711	.9674	.9422	.9676	.9640	.9432	.9343	.9146	.9390
ES	.9641	.9430	.9319	.8912	.9325	.8941	.8594	.8587	.8178	.8575
FR	.9744	.9531	.9185	.8866	.9331	.9028	.8560	.8415	.8070	.8518
IT	.9605	.9212	.9093	.8766	.9169	.8961	.8253	.8150	.7870	.8308

Table 11: Complete **BertScore** results for the fine-tuned models on the GPT-4o annotated test set.



Figure 7: Screenshot of the annotation platform. The user is shown the image, the question and the answers.

1. The question is grammatically correct.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

2. Answer A is grammatically correct.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

3. Answer B is grammatically correct.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

4. Answer A is coherent with respect to the question.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

5. Answer B is coherent with respect to the question.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

6. The question is coherent with respect to the image.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

7. Answer A is coherent with respect to the image.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

8. Answer B is coherent with respect to the image.

☐ Absolutely Disagree ☐ Disagree ☐ Uncertain ☐ Agree ☐ Absolutely Agree

Notes

Figure 8: Screenshot of the annotation platform. The user is shown the eight statements to evaluate and a notes section to provide an explanation for their ratings.

```

<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|>placeholder<|vision_end|>Provide an answer in {LANG} to the given question
using the textual content in the image. If the answer is not in the image, just respond with “{NEGA-
TIVE_MESSAGE_LANG_ANSWER}”. {QUESTION}
Answer the question using a single word or phrase.<|im_end|>
<|im_start|>assistant

```

Table 12: Prompt used for the generation of short and long answers. {LANG} is one of “English”, “German”, “French”, “Italian”, “Spanish”. {NEGATIVE_MESSAGE_LANG_ANSWER} is a negative message generated when the model cannot answer the question. {QUESTION} is a placeholder for the query. Note that the string “Answer the question using a single word or phrase.” is not present in the prompt when performing long answer generation.

```

<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|>placeholder<|vision_end|>Write a question in {LANG} that re-
lates to the text in the image. If the image contains no text, respond with “{NEGA-
TIVE_MESSAGE_LANG_QUESTION}”.
<|im_start|>assistant

```

Table 13: Prompt used for generation of questions for images in the *vdr* dataset without a query. {LANG} is one of “English”, “German”, “French”, “Italian”, “Spanish”. {NEGATIVE_MESSAGE_LANG_QUESTION} is a negative message generated when the image doesn’t contain text.

```

1. The question is grammatically correct.
2. Answer A is grammatically correct.
3. Answer B is grammatically correct.
4. Answer A is coherent with respect to the question.
5. Answer B is coherent with respect to the question.
6. The question is coherent with respect to the image.
7. Answer A is coherent with respect to the image.
8. Answer B is coherent with respect to the image.
Given these 8 statements, provide a rating in a range for 1 to 5 (where 1 is “Absolutely Disagree”, 2 is
“Disagree”, 3 is “Uncertain”, 4 is “Agree” and 5 is “Absolutely Agree”) for each of them. Answer A
will be longer, while answer B more concise. The question and answers are the following:

Question: {QUESTION}
Answer A: {ANSWER_A}
Answer B: {ANSWER_B}

If you give a negative rating for any of the statements, provide a short explanation in English of a
possible solution to fix the question or the answers. If the question or answers cannot be fixed, answer
with “The data is impossible to fix.”.

```

Table 14: Prompt used for the LLM-as-a-judge methodology. Note that, while the special tokens for the images are not present in this prompt, the GPT-4o model is still provided with the image at inference time. {QUESTION} is a placeholder for the question, {ANSWER_A} is a placeholder for the long answer and {ANSWER_B} is a placeholder for the short answer.

In this task, you will be given:

- One image: the given image will contain text in the annotation language
- One question: the given question will reference the text that is present in the image
- Two answers: two possible answers for the given question.

Answer A will contain a longer and more descriptive text, while answer B will be direct and concise. Given these resources, you will have to evaluate eight statements (with ratings ranging from “Absolutely Disagree” to “Absolutely Agree”).

You have to evaluate the quality of questions and answers both for grammar and logical coherence between each other.

Finally, you have to evaluate the logical coherence of the questions and the answers with respect to the provided image.

For “Absolutely Disagree”, “Disagree” and “Uncertain” ratings, please add a short note in English regarding your annotation, describing how the instance can be fixed (e.g. “Answer A is correct, but Answer B contains a different date w.r.t. Answer A”).

Finally, consider these additional guidelines:

- A “Disagree” rating should be given when the information is incomplete and unsatisfactory. For example, if Answer A doesn’t mention all relevant aspects in the image that are needed to answer the question.
- An “Agree” rating should be given when the information is incomplete and satisfactory. For example, if Answer A mentions everything that is needed to answer the question, but there are some additional aspects that could be covered to enhance the quality of the answer.

Table 15: Complete guidelines given to the annotator for manual annotation of the dataset instances.