Preprocessing of the 20th Conference on Computer
Science and Intelligence Systems (FedCSIS) pp. 669–677

reVISION: A Polish Benchmark for Evaluating Vision-Language Models on Multimodal National Exam Data

Michał Ciesiółka
0009-0009-8283-1118
Adam Mickiewicz University
Center for Artificial Intelligence AMU
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
Email: michal.ciesiolka@proton.me

Filip Graliński
0000-0001-8066-4533
Adam Mickiewicz University
Center for Artificial Intelligence AMU
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
Email: filipg@amu.edu.pl

***Abstract*—Vision-Language (VL) models have gained significant popularity in recent years for tasks involving data extraction and image recognition. In this paper, we introduce reVISION, a large-scale Polish benchmark for evaluating such models, comprising over 39k questions extracted from Polish national exams. We assess both the models' general knowledge and their ability to handle tasks that go beyond plain text, including tables and the recognition of specialized visual objects. Our study also examines the effects of image resizing and the inclusion of additional context via OCR-extracted text. To explore the correlation between human and model performance, we compare their results across multiple exam years.**

I. Introduction

FROM the very beginning, Large Language Models have been evaluated using exam-like tests based on multiple-choice questions, as exemplified by the MMLU benchmark created specifically for assessing capabilities of GPT-3[7]. MMLU was popular among LLM practitioners despite its inconsistencies and even outright errors [6]. A more principal way to build LLM benchmarks was proposed by [8], where the benchmark was created using Polish national exams.

Currently, many LLMs are actually not only about language, but they are multi-modal models, handling both language and vision. Again, a natural need to evaluate them in a systematic manner and an idea to use real-life exams remains valid, as actually a significant part of exam questions uses 2D layout and images (tables, graphs, diagrams, etc.). In this paper, we extend the ideas put forward by authors [8] to the bimodal (text and vision) domain, by preparing reVISION, a new comprehensive benchmark for measuring the quality of VL models.

Figure 1 shows an example of a difficult physics question extracted from the 2014 high school exam, which both Qwen2.5-VL-72B-Instruct and GPT-4.1 failed to answer correctly. In contrast, Figure 2 presents an easier question from the 2020 professional exam, which was correctly answered by both models.

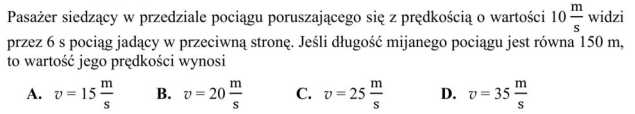

Fig. 1: Example of a difficult question from the 2013 high school physics exam. The question states: A passenger sitting in a train compartment moving at a speed of 10m/s sees, for 6 seconds, a train traveling in the opposite direction. If the length of the passing train is 150m, what is its speed?

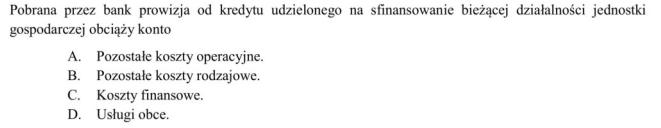

Fig. 2: Example of an easy question from the 2020 professional exam. The question states: The commission charged by the bank on a loan granted to finance the current operations of the business entity is recorded under the account: A. Other operating costs. B. Other cost types. C. Financial costs. D. External services.

Our contributions consist in stating and answering the following questions:

- How effectively do vision–language (VL) models answer Polish exam questions?
- How do VL models compare to human performance?
- How do visually grounded questions differ from text-based questions?
- Does resizing images to reduce computational cost affect model performance?
- Does providing additional context in the form of OCR-extracted text improve model accuracy?

II. Related work

Ever since the introduction of the first vision-language (VL) models [12], their performance has been continually challenged using a variety of benchmarks. Assessing their capabilities is

crucial for selecting the appropriate model for the task at hand. These evaluations range from testing object, attribute, and relation recognition [16] to performance on multi-disciplinary tasks [14]. Despite the diversity of tasks, all benchmarks share a common goal: to identify and compare models in order to determine which performs best in a given category.

LLMzSzŁ [8], a previous work that relied heavily on Polish exam questions, addressed the need for Polish-language benchmarks in the research landscape. It demonstrated the importance of academic datasets and benchmarks that capture the linguistic specificity of languages like Polish. By incorporating visual information, we aim to support the development and evaluation of multimodal models in the Polish-language context, extending the capabilities of previous text-only benchmarks.

Previous work such as EXAMS-V [4] emphasizes the need for benchmarks that assess curriculum-specific knowledge. The EXAMS-V dataset consists of thousands of multilingual exam questions. The authors highlight the dataset's difficulty, noting that even the most advanced models, such as GPT-4V and Gemini, struggle with it. While Polish professional exams are included in the benchmark, they account for just over 2.5k questions. In this paper, we focus on expanding the number and variety of questions while narrowing the language scope to a single language, while using clear criteria for inclusion and future extensions. Similarly, M3Exam [15] provides a multilevel benchmark composed of human exam questions. The study demonstrates that models face challenges when processing multilingual texts, particularly in low-resource languages.

III. Dataset

A. Data sources

Each year in Poland, three types of national exams are conducted to assess the knowledge of Polish students: the eighth-grade exam, the high school exam, and professional exams. The middle school exam, which was part of the Polish education system prior to recent educational reforms, was also considered when preparing the data sources. After each exam session, the exams are published by the Polish Central Examination Board (Pol. CKE, Centralna Komisja Egzaminacyjna). These archives provided us with access to both questions and answers from a reliable source, ensuring the high quality of our dataset.

B. Data selection

The primary constraint considered when selecting documents for dataset preparation was the number of closed-ended questions with a single correct answer. This approach eliminated the need for a larger model or human annotators as judges, enabling us to focus solely on extracting one type of question while ensuring objectivity by removing emotional bias and personal preferences. Since many types of exams were heavily focused on open-ended answers, we decided to exclude them from further processing. This strict elimination process resulted in the following categories: math, natural sciences, biology, physics, and the Polish language for pre-high school questions, and arts, mechanics (including mining and metallurgy), and agriculture (including forestry) for professional exams.

C. Dataset preparation

The preparation of the dataset began with downloading the exams that were made public over the years. We automated this stage using basic web scraping methods and Python scripts. All exams were published in PDF format, with their answers in separate files, making it crucial to track which question file corresponds to which answer file.

To extract all the questions from a single PDF file, we first had to determine the accurate X-axis positions of each question. We took a single page of the file in order to test the techniques for position extraction. Our initial approach was to prompt a multimodal model with an image and have it extract the bounding boxes of entire questions. However, this proved ineffective as multimodal models, although good for OCR text extraction, cannot accurately pinpoint and return text positions even on small images. Having this in mind, we decided to use an OCR engine, specifically PaddleOCR[1], that can extract text, precisely draw its bounding boxes, and return their positions. In this way, we identified the text indicating the beginning of a question. For most exams, the keyword was 'Task' (Pol. *Zadanie*). If the word appeared above the actual task, we used the lower X-axis of 'Zadanie' as the starting position of the task; otherwise, we used the upper X-axis. The upper X-axis of the next question was used as the ending position. The images were then cropped along those axes.

When initially testing on entire exam documents, we attempted to convert the entire PDF file into a single long image and then used an OCR engine to detect text. However, this approach proved ineffective, as the engine was unable to detect any text in the image. The issue arises primarily due to the extreme aspect ratio of the images. OCR engines are typically optimized for standard-sized documents and conventional aspect ratios. A more effective approach was based on the fact that each question had to be contained on a single page, making it possible to split the document into individual pages, convert them into images, and then detect tasks within them. After fine-tuning the detection and splitting process to match each exam type, we were able to extract almost 40,000 images, each containing exactly one question. This figure reflects the results after a cleaning process, which involved removing approximately 30% of the original questions, primarily due to duplication, with a smaller portion removed for issues such as poor cropping. The cleaning process was fully automated, involving the removal of duplicate images by comparing the OCR-extracted text from each image. Images containing no text or lacking answer options labeled A, B, C, and D were also removed. To ensure the accuracy of the cleaning process, random samples were extracted from the dataset for manual verification.

The second-to-last step in the preparation of the dataset was to match the questions with their corresponding answers. We

[1]https://github.com/PaddlePaddle/PaddleOCR

used a simple PDF text extraction tool and cleaning scripts to extract answers from Professional Exams files, as the format of the answer key has remained consistent over the years. For other exam types, the answers were manually extracted.

Once the dataset was nearly complete, we decided to add more metadata, specifically indicating whether a task requires additional data, such as an image or table, to answer the question, or if it is entirely text-based. To automate this task, we utilized the Qwen2-VL-7B-Instruct[2] model to analyze and classify the questions. We used the system and user prompt shown below.

```
System prompt:
    You will see a picture of a Polish
    task.If there is more than just a
    text in the image, like an image,
    table, etc, return true. Otherwise
    return false. Respond ONLY with
    'true' or 'false'.

User prompt:
    Does this task need an image
    for context?
```

Table I presents the hardest professional exams categories. A more detailed analysis of these categories is provided in Section V-C.

D. Documents cropping as a tool for building text-based datasets

Dividing a set of questions into individual items can be useful not only for creating image-based datasets but also for constructing textual datasets. When using an OCR engine to extract text from questions that do not require visual context for understanding, the post-processing effort can be significantly reduced compared to standard text extraction techniques, as the data is already segmented into more manageable chunks.

The growing popularity of using multimodal models as a means for data extraction, particularly with the use of Structured Outputs[3], opens up the possibility of generating structured datasets that require minimal or no post-processing. Feeding data to a model in smaller units, rather than entire document pages, can improve the accuracy of data extraction and analysis.

E. Dataset summary

To better understand the finalized dataset, a numerical representation of its potential size and actual results will be given now. Starting with the lowest extraction rate, the 8th-Grade Exams achieved a 27.5% extraction rate. This was followed by the High School Exams at 42.9%, the Middle School Exams at 61.7%, and the Professional Exams with the highest rate of 67.7%. Until 2012, Mathematics and Nature were combined in the middle school exams, resulting in a joint discipline. Table III presents the exact number of questions for each exam and discipline, while Table IV presents the distribution of answer options in the dataset.

F. Dataset availability

The dataset will be publicly available on the Hugging Face page.

IV. Evaluation

A. Evaluation harness

The tool we used to assist with model evaluation was lmms-eval [5] [10], a fork of LM Evaluation Harness[6] [5] framework. Each model was prompted with the following task (in English) in addition to being shown the image:

```
Answer the Polish exam question from
the image. Answer with the good answer
letter only. Possible answers are
A or B or C or D.
```

The evaluation was carried out using two different approaches. The first approach involved resizing the images so that their longest side was 512 pixels, while preserving the aspect ratio and readability. This resizing was performed not only to save memory and reduce computational overhead, but also to evaluate the consequences of image resizing, which are discussed later in the paper. In the second approach, models were evaluated using full-size images to assess their full capabilities. All analyses presented later in the paper are conducted using the outputs from full-sized images.

A common approach for selecting a model's answer involves extracting the log-likelihoods—or probabilities—of each possible option. However, this functionality was not implemented for all of the models evaluated. Therefore, we generated the answers in text format and processed them post hoc by stripping any punctuation or extraneous content beyond the answer letter.

Questions with answers other than A, B, C, or D were excluded from the evaluation. This prompting strategy also simulates a more realistic usage scenario, reflecting how such models are often used in everyday applications.

B. Evaluated models

For the general evaluation on resized images (the 1st approach), we selected 18 different open-weight models across 7 model families, with sizes ranging from 2 billion to 72 billion parameters. We also included one closed-weight GPT model in the evaluation.

GPT 4.1: gpt-4.1-2025-04-14 [7]

Qwen: Qwen2.5-VL (3B-Instruct, 7B-Instruct and 72B-Instruct) [2], Qwen2-VL (2B, 2B-Instruct, 7B and 7B-Instruct) [13]

InternVL: InternVL2_5-8B [3]

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

[3]https://platform.openai.com/docs/guides/structured-outputs?api-mode=responses

[4]Percentages are approximate and may not sum to exactly 100% due to rounding.

[5]https://github.com/EvolvingLMMs-Lab/lmms-eval

[6]https://github.com/EleutherAI/lm-evaluation-harness

[7]https://openai.com/index/gpt-4-1/

TABLE I: Top Hardest Professional Exam Categories based on top 10 models performance.

Code	Models	Visual tasks (%)	Domain	Category
R.13	8	24.0	Forest Technician	Protection and management of forest resources
MG.11	8	46.3	Underground Miner / Mining Technician	Exploitation of underground deposits
R.15	7	36.7	Inland Fish Farmer Technician	Organization of fish farming operations in aquaculture
MG.30	7	44.2	Optician Technician	Production and repair of visual aids
R.14	6	36.0	Forest Technician	Utilization of forest resources
M.21	6	55.3	Blacksmith	Production and repair of blacksmith products
R.02	5	55.3	Farmer / Agribusiness / Agricultural Technician	Agricultural production
RL.06	5	40.7	Horse Trainer	Horse riding and training

TABLE III: Distribution of Questions Across Exams and Disciplines

Exam	Discipline	Original Questions	Questions
8th-Grade Exam	Polish Language	101	9
8th-Grade Exam	Mathematics	99	46
Middle School Exam	Mathematics and Nature	346	152
Middle School Exam	Mathematics	184	96
Middle School Exam	Nature	192	79
High School Exam	Biology	169	21
High School Exam	Physics	399	154
High School Exam	Mathematics	492	280
Professional Exam	Arts	3480	2547
Professional Exam	Mechanical, Mining and Metallurgical	30240	21057
Professional Exam	Agriculture and Forestry	23200	14905

TABLE IV: Distribution of answer options in the dataset.[4]

Answer Option	Count	Percentage (%)
A	9,564	24.3
B	10,235	26.0
C	10,075	25.6
D	9,457	24.0
Other	15	0.04
Total	39,346	100.0

Llava-HF: 1.5 (7B and 13B), v1.6 (Mistral-7B and 34B), NeXT (Llama3-8B and 72B) [11]

Phi: Phi-4 [1]

Idefics2: Idefics2 (8B and 8B–chatty) [9]

InstructBLIP: Vicuna-7B

C. Cropped images results

The general accuracy of the evaluated models on resized images is summarized in Table V.

As shown in the table, the Qwen models dominate this part of the evaluation, despite being trained primarily on English and Chinese data. All other models tend to hover around the 25% random-guess baseline. The Qwen2-2B model began generating random tokens instead of answering the questions, which explains its low accuracy score.

The LLaVA models, even in their largest 72B variant, demonstrate that model size alone is insufficient; the quality and relevance of the training data are the primary determinants of performance. These models, trained on English data, achieve at most 28.06% accuracy.

TABLE V: Performance of models on images resized to 512px

Model	Size	Acc. (%)
Qwen/Qwen2.5-VL-72B-Instruct	72B	60.56
Qwen/Qwen2.5-VL-7B-Instruct	7B	44.82
Qwen/Qwen2.5-VL-3B-Instruct	3B	37.31
Qwen/Qwen2-VL-7B-Instruct	7B	34.16
Qwen/Qwen2-VL-7B	7B	29.04
OpenGVLab/InternVL2/5-8B	8B	28.69
llava-hf/llava-next-72b-hf	72B	28.06
microsoft/Phi-4-multimodal-instruct	14.7	27.65
Qwen/Qwen2-VL-2B-Instruct	2B	26.02
llava-hf/llava-v1.6-34b-hf	34B	25.91
HuggingFaceM4/idefics2-8b-chatty	8B	25.76
HuggingFaceM4/idefics2-8b	8B	25.09
random guessing	-	*25.00*
llava-hf/llava-1.5-7b-hf	7B	24.34
Salesforce/instructblip-vicuna-7b	7B	24.32
llava-hf/llava-1.5-13b-hf	13B	24.31
llava-hf/llava-v1.6-mistral-7b-hf	7B	24.2
llava-hf/llama3-llava-next-8b-hf	8B	24.16
Qwen/Qwen2-VL-2B	2B	20.18

TABLE VI: Performance of models on full-size images

Model	Size	Acc. (%)	No resize increase
GPT-4.1	-	69.77	-
Qwen/Qwen2.5-VL-72B-Instruct	72B	65.25	+7.44%
Qwen/Qwen2.5-VL-7B-Instruct	7B	48.95	+9.21%
Qwen/Qwen2.5-VL-3B-Instruct	3B	41.16	+10.32%
Qwen/Qwen2-VL-7B-Instruct	7B	46.69	+36.68%
Qwen/Qwen2-VL-7B	7B	40.58	**+39.74%**
OpenGVLab/InternVL2/5-8B	8B	28.56	-0.45%
llava-hf/llava-next-72b-hf	72B	28.24	+0.64%
microsoft/Phi-4-multimodal-instruct	14.7	28.77	+4.05%
Qwen/Qwen2-VL-2B-Instruct	2B	31.73	+21.94%
llava-hf/llava-v1.6-34b-hf	34B	26.01	+0.39%
HuggingFaceM4/idefics2-8b-chatty	8B	25.25	-1.98%
HuggingFaceM4/idefics2-8b	8B	25.18	+0.36%
llava-hf/llava-1.5-7b-hf	7B	24.32	-0.08%
Salesforce/instructblip-vicuna-7b	7B	24.32	0%
llava-hf/llava-1.5-13b-hf	13B	24.34	+0.12%
llava-hf/llava-v1.6-mistral-7b-hf	7B	24.32	+0.5%
llava-hf/llama3-llava-next-8b-hf	8B	24.19	+0.12%
Qwen/Qwen2-VL-2B	2B	20.05	-0.64%

D. Full image results

As seen in Table VI, model accuracy approaches practical usefulness only for larger models, such as GPT-4.1 and Qwen2.5-VL-72B-Instruct. Other models, even those that performed above the guess rate of 25%, still yield unsatisfactory results. These findings highlight the need for benchmarks like this one to rigorously evaluate model performance on real-world tasks. Moreover, they underscore the substantial room for improvement in smaller or mid-sized vision-language models.

E. To resize or not to resize

Depending on the model version, the performance gains from using high-resolution images vary. The Qwen2.5 models show modest improvements of just over 10%, while the Qwen2 models demonstrate increases approaching 40%. Notably, models that perform poorly on resized images tend to see limited improvement even when image quality is not altered.

These findings suggest that when image downscaling is necessary, newer model versions should be preferred to mitigate the negative effects of reduced image quality.

V. Results Analysis

A. Confusion analysis

To better understand the models' "thinking" processes—that is, the sequence of statistical and text-processing steps they use to generate answers—a deeper analysis of their answer choices is required. Each confusion matrix in Figure 3 illustrates the relationship between the true answers (y-axis) and the answers predicted by the models (x-axis). This enables a granular examination of how frequently the top three performing models confuse specific answer options.

One of the most notable insights from these matrices is the visual representation of the models' uncertain guesses. While

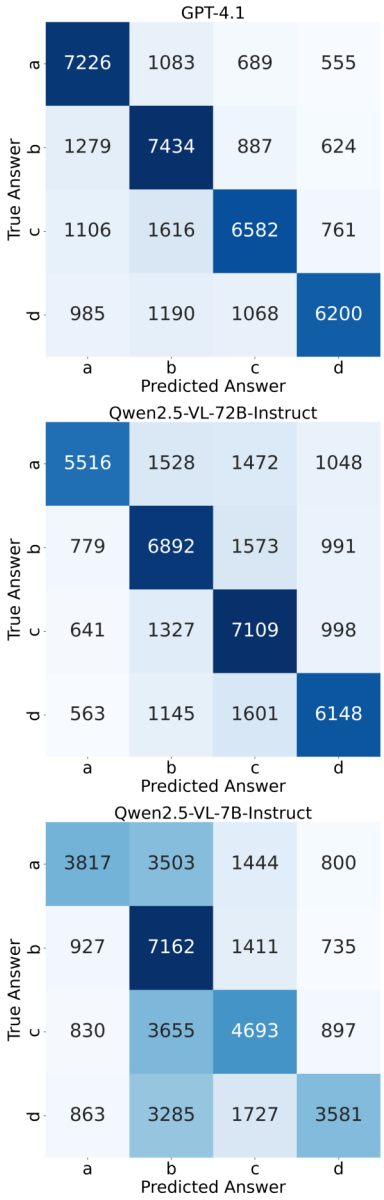

Fig. 3: Confusion matrices for different models.

the GPT model demonstrates the best overall performance across most answer classes, it is slightly outperformed by the 72B Qwen model on questions with "c" as the true answer. When uncertain, GPT exhibits a bias toward selecting earlier answer options, with the likelihood of selection decreasing progressively across the choices. Notably, it appears to almost avoid choosing the furthest option, "d", under uncertainty.

The Qwen models tend to favor the middle answer choices when in doubt. The larger 72B model most frequently guesses "b" or "c" under uncertainty, reflecting a moderate central bias. In contrast, the smaller 7B model exhibits a strong preference for answer "b," even to the extent that it accidentally outperforms the larger model in predicting this specific answer.

TABLE VII: Performance of vision-language models on visual- and text-based tasks.

Model	Visual Acc. (%)*	Text Acc. (%)*
GPT-4.1	65.47	73.07
Qwen/Qwen2.5-VL-72B-Instruct	59.09	69.98
Qwen/Qwen2.5-VL-7B-Instruct	44.72	52.19
Qwen/Qwen2-VL-7B-Instruct	42.74	49.71
Qwen/Qwen2.5-VL-3B-Instruct	38.35	43.32
Qwen/Qwen2-VL-7B	36.21	43.92
Qwen/Qwen2-VL-2B-Instruct	30.55	32.64
microsoft/Phi-4-multimodal-instruct	28.33	29.11
OpenGVLab/InternVL2_5-8B	28.37	28.71
llava-hf/llava-next-72b-hf	27.14	29.09
llava-hf/llava-v1.6-34b-hf	24.87	26.88
HuggingFaceM4/idefics2-8b-chatty	25.37	25.16
HuggingFaceM4/idefics2-8b	24.73	25.53
llava-hf/llava-1.5-13b-hf	23.67	24.86
llava-hf/llava-1.5-7b-hf	23.62	24.86
Salesforce/instructblip-vicuna-7b	23.53	24.92
llava-hf/llava-v1.6-mistral-7b-hf	23.72	24.71
llava-hf/llama3-llava-next-8b-hf	23.21	24.95
Qwen/Qwen2-VL-2B	23.42	17.47

*Task classification is based on automated labeling with an estimated 12.2% error rate. See Section V-B for details.

This pronounced bias suggests a reliance on heuristics or a limited ability to disambiguate between options in uncertain scenarios.

B. Visual vs. Text-Based Tasks

To confirm the accuracy of Qwen's classification of images that require additional context, such as a picture or table, a random sample of 1,000 images was selected for manual verification. The verification revealed that 122 of the 1,000 images were misclassified, with the vast majority being false positives (120 out of 122). This disproportion may indicate that the prompt did not sufficiently guide the model to identify tasks requiring additional context. All subsequent tables that report model results based on the image-context grouping should be interpreted in light of an estimated 12.2% labeling error, with a 95% confidence interval ranging from approximately 10.2% to 14.2%. While this level of error is acceptable for exploratory analysis and broad grouping, it is not sufficiently accurate for training downstream models. We therefore emphasize that findings based on this stratification are preliminary and should not be overinterpreted without more precise annotation.

Out of 39,331 questions considered when calculating the accuracy of the models, 17,058 were marked as visual-based and 22,273 were marked as text-based. Table VII shows results of each model both on visual- and text-based tasks.

With few exceptions, most of the tested models exhibit higher accuracy on text-based questions compared to those requiring visual context. This discrepancy stems from the fact that vision-language models are still primarily trained on textual data and are less focused on the deep analysis of figures, tables, or the recognition of specific and highly niche objects. Visual tasks often demand fine-grained perception, complex reasoning, and the integration of visual features with language understanding. Even the most advanced models evaluated remain far from achieving human-like reasoning and multidimensional understanding. This performance gap underscores the limitations of current multimodal training approaches and highlights the need for more diverse, richly annotated datasets that better capture the complexities of visual reasoning.

C. Professional Exams categories analysis

This analysis was conducted on the top 10 best-performing models, as it required a level of domain knowledge sufficient to demonstrate informed responses rather than mere guessing. There are 150 unique category codes, and for the easiest and hardest analyses, the codes that appeared most frequently among the top 20 easiest and hardest categories for the selected models were identified. The proportion of visually-based tasks was also considered when assessing question difficulty, as text-based tasks were generally easier for the models to answer as shown in Section V-B. This led to the exclusion of categories with more than 70% visual tasks among the hardest categories, and less than 30% visual tasks among the easiest categories. These thresholds were selected to create a clear contrast between predominantly visual and predominantly text-based categories, while also preserving a sufficient number of categories to support a robust and meaningful analysis. These thresholds were selected to create a clear contrast between predominantly visual and predominantly text-based categories, while also preserving a sufficient number of categories to support a robust and meaningful analysis.

The most challenging categories were closely tied to the limited availability of widely accessible resources for model training, particularly those that are highly niche and vocational. Eight categories fell into this group, the most difficult of which were *Protection and Management of Forest Resources* and *Exploitation of Underground Deposits*.

The knowledge required to become one of these specialists is heavily underrepresented in training corpora and typically demands extensive hands-on experience and practical training.

In contrast, most of the easiest categories belonged to the broader domain of mechanics. Notably, Shipbuilding works—a subcategory within this domain—appeared among the top 20 easiest categories in nine models. With approximately 1.64 billion cars in the world, mechanical data is widely available and has certainly been included in model training. The second easiest domain was *Environmental Protection*, a highly discussed topic, particularly in today's context of increasing global awareness and concern.

VI. Comparison against humans

Comparing the accuracy of Multimodal Language Models to exact human performance is not feasible, as human results are published only for entire exams rather than individual questions for all but the High School exams. All exams, except for the Professional exams, contain a significant number of open-ended questions, which were not included in our dataset.

Additionally, some closed-ended questions may have been excluded from the dataset due to the postprocessing described earlier. Data on human performance in professional exams were extracted from the WaszaEdukacja website[8], which is not affiliated with the Central Examination Board (CKE). Therefore, the exam content presented on the site may not fully correspond to the official questions used in the models' evaluation. The purpose of this section is not to compare human knowledge to model performance directly, but rather to evaluate whether the difficulty reflected in human performance for a given year aligns with the general performance of the models—even if the exact questions differ. This type of analysis applied to textual exams dataset has been proposed and shown to correlate well with human performance in [8]. For the visualized models performance, the top 3 performing models were chosen: GPT-4.1, Qwen2.5-VL-72B-Instruct and Qwen2.5-VL-7B-Instruct.

A. Eighth-grade exam

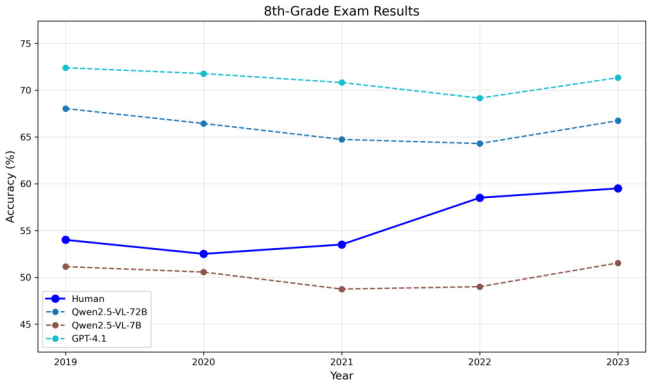

Fig. 4: 8-th grade selected models results compared to average human results.

In the combined results from all 8th-grade question categories—grouped due to the limited number of questions available for individual analysis—the GPT model emerges as the top-performing model. All models exhibit similar year-to-year fluctuations in accuracy, differing primarily in their overall performance levels. GPT achieves an average accuracy of 69.87%, while the Qwen models follow with 64.48% and 48.99%, respectively. A notable divergence between human and model performance occurred between 2020 and 2022, where the average human accuracy improves compared to the previous years, all the evaluated models experience a decline in performance.

[8]https://waszaedukacja.pl/egzaminy/zawodowy

B. Middle school exam

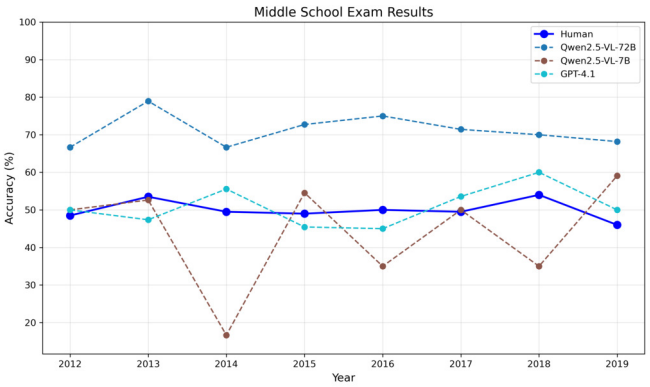

Fig. 5: Middle school selected models results compared to average human results.

Qwen2.5-VL-72B performs exceptionally well on the Middle School exams, achieving an impressive average accuracy of 77.14%, clearly leading among the evaluated models. In contrast, GPT-4.1 achieves a more modest average accuracy of 58.02%, aligning more closely with the smaller Qwen2.5-VL-7B model, which scores 51.98%. Interestingly, the performance of the 72B Qwen model mirrors the trend of human results across the years, with the exception of 2018, where a noticeable deviation occurs. GPT-4.1 maintains accuracy levels generally comparable to human performance, while the Qwen 7B model exhibits high variability—ranging from below 20% to as high as 60%—indicating inconsistency in its performance across exam years.

C. High school exam

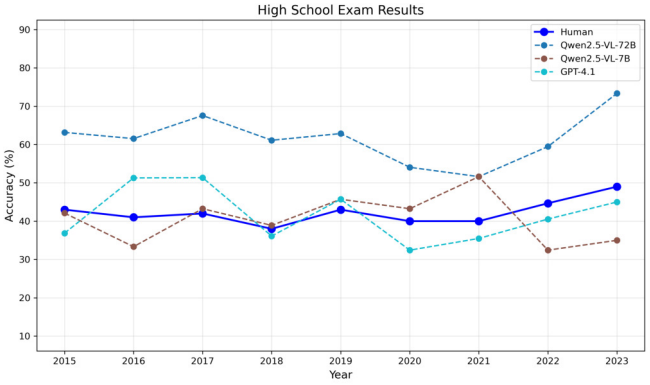

Fig. 6: High school subjects selected models results compared to average human results.

For the high school exams, the Qwen2.5-VL-72B model's performance mirrors the overall trend of human performance over the years, with rises and declines occurring in parallel, though at a different absolute level of accuracy. In contrast, the other two models hover around the human performance baseline, with GPT-4.1 achieving an average accuracy of 53% and Qwen2.5-VL-7B scoring 47.34%.

D. Professional exam

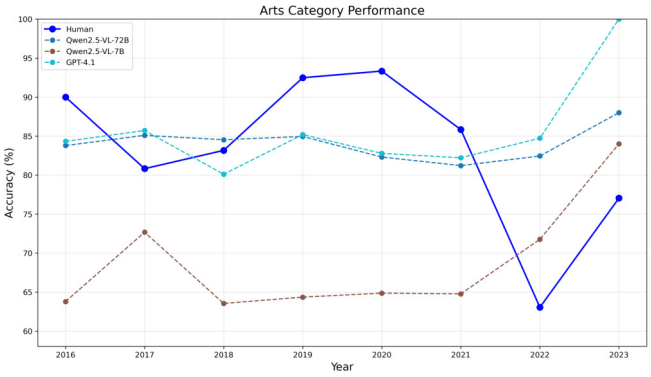

Fig. 7: Arts selected models results compared to average human results.

1) Arts: In contrast to the models' performance, fluctuations in human accuracy are particularly noticeable in the arts professional exams. The two best-performing models consistently maintain accuracy above 80%, while human performance occasionally drops below 65%, highlighting a significant performance gap in this category.

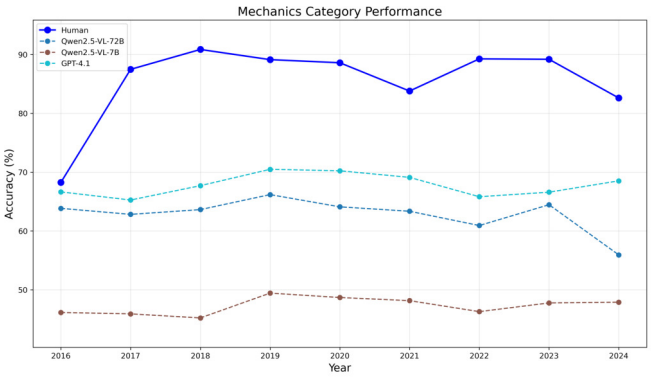

Fig. 8: Mechanics (including mining and metallurgy) selected models results compared to average human results.

2) Mechanical, Mining and Metallurgical: In the mechanical category, human performance significantly surpasses that of the models, indicating that in highly specialized domains, models still underperform and have considerable room for improvement. All models exhibit similar trends of rising and falling performance over time, though at different levels of average accuracy. This consistent fluctuation pattern across models suggests a correlation between model size (i.e., number of parameters) and overall performance, with larger models achieving higher accuracy while following the same performance trajectory.

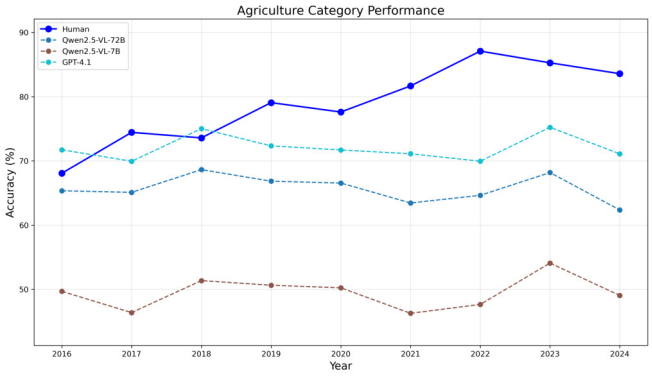

Fig. 9: Agriculture (including forestry) selected models results compared to average human results.

3) Agriculture and Forestry: The accuracy trends for models and humans on agricultural exam questions appear to move in opposite directions: while human performance steadily improves over the years, model performance consistently declines. This contrast highlights the limitations of using vision-language models as a benchmark for assessing exam difficulty. It suggests that what may seem harder for a model may, in fact, be easier for a human—and vice versa—underscoring the fundamental differences in reasoning strategies and domain familiarity. This emphasizes the need for caution when relying on model-based metrics in educational evaluation or exam design.

VII. Additional OCR context

In this section we will try to answer the question if prompting the model with additional context from the image—in the form of Optical Character Recognition (OCR)-extracted text—can improve its accuracy and overall performance. Including OCR-extracted text from the image in the prompt itself can enhance model performance by providing additional context that may not be captured through visual input alone, bridging the gap between modalities. To encourage the model to utilize this additional information when answering the question, we appended the following text to the prompt:

```
Use the OCR text as well as the image
to answer the question. OCR text:
```

The following models were selected to evaluate the benefits of additional context: **Qwen2.5-VL-7B-Instruct**, **Qwen2-VL-7B**, **Qwen2-VL-7B-Instruct**, **Llama3-llava-next-8b-hf**, **Phi-4**.

For the OCR engine, EasyOCR [9], was used, as it handles Polish characters more effectively than the previously used PaddleOCR. The text indicating the beginning of each question did not contain any Polish letters, so the earlier use of PaddleOCR did not result in processing errors. The extracted text was provided in the prompt verbatim, without any post-processing. The OCR was performed on the original, pre-resize image.

[9]https://github.com/JaidedAI/EasyOCR

A. Results

Across the tested models, only the Phi-4 model exhibited notable improvements when prompted with additional OCR context. This outcome may be attributed to its modular architecture, which employs separate adapters (LoRAs) for vision, speech, and text. Phi-4 processes each modality independently and integrates them later, enabling more effective fusion of external textual input with visual features.

Interestingly, the Qwen2-VL-7B-Instruct model showed a decrease in performance with the additional OCR context. This suggests that the injected text may have conflicted with the model's internal visual interpretation, potentially leading to confusion or interference during reasoning. Unlike Phi-4, Qwen models use a single vision encoder that processes visual inputs and aligns them with textual representations via rotary embeddings before feeding the combined data into a unified Transformer stream.

The overall results suggest that, while promising in theory, incorporating OCR context when processing images of questions does not consistently improve model performance and may, in some cases, degrade it. Nevertheless, the observed gains in Phi-4 underscore that, when properly integrated, OCR context can enhance performance. Therefore, it is important to continue exploring modality-specific architectures and training strategies.

TABLE VIII: General performance of selected models with additional OCR context

Model	Acc.	OCR Acc.	Increase
Qwen2.5-VL-7B-Instruct	48.95	50.24	+2.64%
Qwen2-VL-7B	40.58	40.84	+0.64%
Qwen2-VL-7B-Instruct	46.69	46.16	-1.14%
Llama3-llava-next-8b-hf	24.19	25.07	+3.64%
Phi-4	28.77	31.46	**+9.35%**

VIII. Limitations

While this benchmark is a necessary step toward assessing a model's knowledge, it evaluates only the model's final answers to closed-ended questions and does not account for the reasoning process behind those answers. Although the provided responses may be sufficient for tasks such as assisting a student in finding a correct answer, they fall short in educational contexts where the underlying reasoning is also evaluated. In such scenarios, a model might provide the correct answer but fail to justify it coherently, potentially misleading users about its understanding. Additionally, the dataset used in this benchmark represents idealized conditions in which all questions are fully visible and neatly typewritten. As a result, the benchmark assesses visual recognition only within these controlled parameters and does not reflect the broader variability encountered in real-world visual inputs, such as handwritten notes, partially obscured text, or classroom environments.

One of the previously mentioned issues is the error rate in Qwen's visual-context task recognition. Although the classification was sufficient for simple comparisons and provided some insight into the increased difficulty compared to text-only tasks, the error rate may render the data unsuitable for other applications.

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang. Phi-4 technical report, 2024.

[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[4] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, I. Koychev, and P. Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024.

[5] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, 07 2024.

[6] A. P. Gema, J. O. J. Leang, G. Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, et al. Are we done with MMLU? *arXiv preprint arXiv:2406.04127*, 2024.

[7] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[8] K. Jassem, M. Ciesiółka, F. Graliński, P. Jabłoński, J. Pokrywka, M. Kubis, M. Jabłońska, and R. Staruch. LLMzSzŁ: a comprehensive LLM benchmark for Polish, 2025.

[9] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models?, 2024.

[10] B. Li, P. Zhang, K. Zhang, F. Pu, X. Du, Y. Dong, H. Liu, Y. Zhang, G. Zhang, C. Li, and Z. Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024.

[11] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2015.

[13] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[14] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, 2024.

[15] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, and L. Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023.

[16] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023.