EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models

Rocktim Jyoti Das[1] **Simeon Emilov Hristov**[2] **Haonan Li**[1]
Dimitar Iliyanov Dimitrov[2] **Ivan Koychev**[2] **Preslav Nakov**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2]FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
`{Rocktim.JyotiDas, haonan.li, Preslav.Nakov}@mbzuai.ac.ae`,
`s.e.hristov99@gmail.com`. {ilijanovd, koychev}@fmi.uni-sofia.bg

Abstract

We introduce EXAMS-V, a new challenging multi-discipline multimodal multilingual exam benchmark for evaluating vision language models. It consists of 20,932 multiple-choice questions across 20 school disciplines covering natural science, social science, and other miscellaneous studies, e.g., religion, fine arts, business, etc. EXAMS-V includes a variety of multimodal features such as text, images, tables, figures, diagrams, maps, scientific symbols, and equations. The questions come in 11 languages from 7 language families. Unlike existing benchmarks, EXAMS-V is uniquely curated by gathering school exam questions from various countries, with a variety of education systems. This distinctive approach calls for intricate reasoning across diverse languages and relies on region-specific knowledge. Solving the problems in the dataset requires advanced perception and joint reasoning over the text and the visual content of the image. Our evaluation results demonstrate that this is a challenging dataset, which is difficult even for advanced vision–text models such as GPT-4V and Gemini; this underscores the inherent complexity of the dataset and its significance as a future benchmark.[1]

1 Introduction

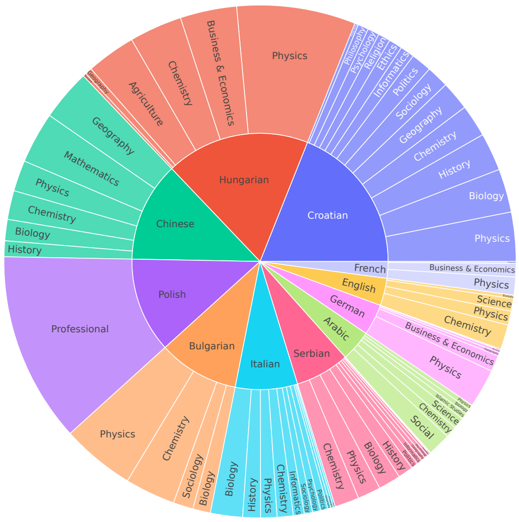


Figure 1: Data distribution for our EXAMS-V dataset: languages and subjects.

Large Language Models (LLMs) have recently demonstrated impressive skills in understanding and generating natural languages (Brown et al., 2020; Zhang et al., 2022; Scao et al., 2022; Zeng et al., 2023; Touvron et al., 2023b). This progress has paved the way for significant advancements in LLM-based vision models (Zhu et al., 2023; Liu et al., 2023a). Notable developments like GPT-4V (OpenAI, 2023) and Gemini (Anil et al., 2023) represent a new era in image understanding, exhibiting remarkable proficiency in interpreting and analyzing visual data alongside textual information. However, as Vision Language Models (VLMs) grow more sophisticated, existing benchmarks are becoming outdated, and unable to accurately assess these models' performance.

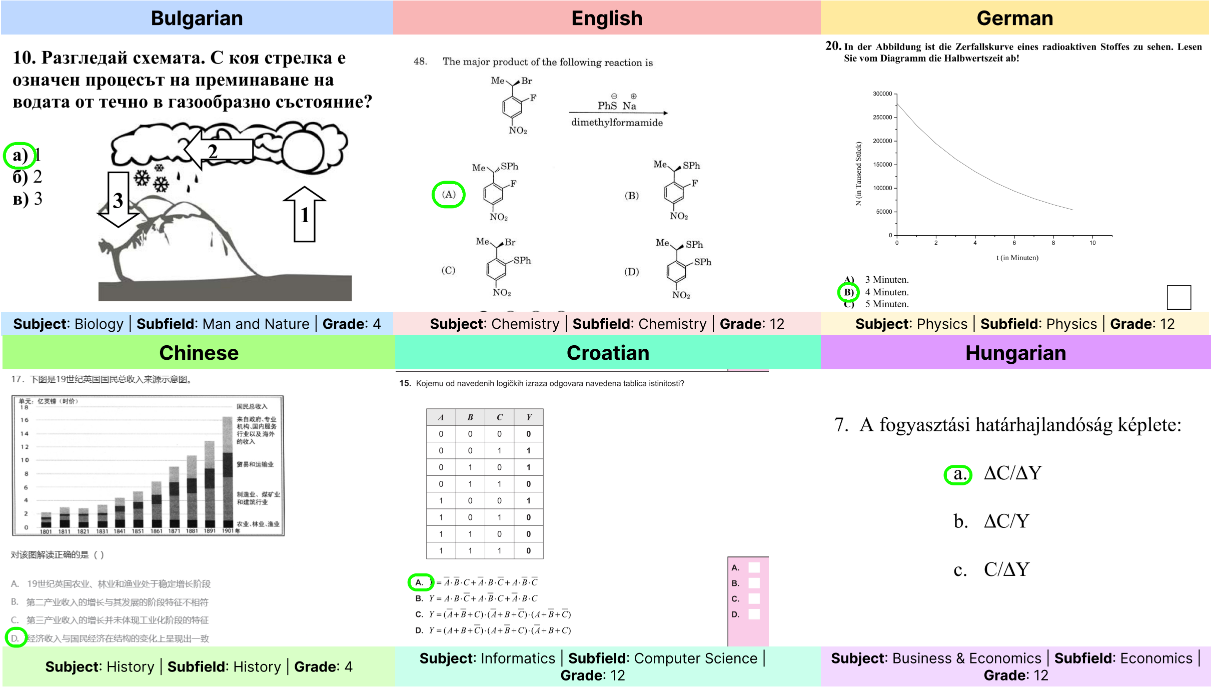
For LLM evaluation, standardized testing akin to school examinations has proven to be an effective measure of a model's capabilities. A typical benchmark MMLU (Hendrycks et al., 2021), which contains 57 subjects across science, engineering, and humanities, has become a de facto benchmark for LLM evaluation. Several other school exam datasets have also set the standard in evaluating LLMs in different languages (Hardalov et al., 2020; Li et al., 2023b; Koto et al., 2023).

In terms of VLM, a comparable benchmarking framework is conspicuously absent. Existing benchmarks are (1) primarily monolingual, focused on English; (2) mostly not from school exams, leading to differences in methods of examining humans; (3) tend to keep images and text separate,

---

[1] Datasets are available at `https://github.com/mbzuai-nlp/EXAMS-V`

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7768–7791
August 11-16, 2024 ©2024 Association for Computational Linguistics


Bulgarian
10. Разгледай схемата. С коя стрелка е означен процесът на преминаване на водата от течно в газообразно състояние?
а) 1
б) 2
в) 3
2
3
1
Subject: Biology | Subfield: Man and Nature | Grade: 4
English
48. The major product of the following reaction is
PhS Na
dimethylformamide
(A)
(B)
(C)
(D)
Subject: Chemistry | Subfield: Chemistry | Grade: 12
German
20. In der Abbildung ist die Zerfallskurve eines radioaktiven Stoffes zu sehen. Lesen Sie vom Diagramm die Halbwertszeit ab!
N (in Tausend Stück)
t (in Minuten)
A) 3 Minuten.
B) 4 Minuten.
C) 5 Minuten.
Subject: Physics | Subfield: Physics | Grade: 12
Chinese
17．下图是19世纪英国国民总收入来源示意图。
单元：亿英镑（时价）
国民总收入
来自政府、专业机构、国内服务行业以及海外的收入
贸易和运输业
制造业、煤矿业和建筑行业
农业、林业、渔业
对该图解读正确的是（ ）
A. 19世纪英国农业、林业和渔业处于稳定增长阶段
B. 第二产业收入的增长与其发展的阶段特征不相符
C. 第三产业收入的增长并未体现工业化阶段的特征
D. 经济收入与国民经济在结构的变化上呈现出一致
Subject: History | Subfield: History | Grade: 4
Croatian
15. Kojemu od navedenih logičkih izraza odgovara navedena tablica istinitosti?
A. $Y = \overline{A}\cdot\overline{B}\cdot C + \overline{A}\cdot B\cdot\overline{C} + A\cdot\overline{B}\cdot\overline{C}$
B. $Y = A\cdot B\cdot\overline{C} + A\cdot\overline{B}\cdot C + \overline{A}\cdot B\cdot C$
C. $Y = (\overline{A}+\overline{B}+C)\cdot(\overline{A}+B+\overline{C})\cdot(A+\overline{B}+\overline{C})$
D. $Y = (A+B+\overline{C})\cdot(A+\overline{B}+C)\cdot(\overline{A}+B+C)$
Subject: Informatics | Subfield: Computer Science | Grade: 12
Hungarian
7. A fogyasztási határhajlandóság képlete:
a. ΔC/ΔY
b. ΔC/Y
c. C/ΔY
Subject: Business & Economics | Subfield: Economics | Grade: 12

A	B	C	Y
0	0	0	**0**
0	0	1	**1**
0	1	0	**1**
0	1	1	**0**
1	0	0	**1**
1	0	1	**0**
1	1	0	**0**
1	1	1	**0**

Figure 2: Sampled EXAMS-V examples from different languages. The questions require the ability to understand multiple languages in addition to expert perception and reasoning capabilities.

which fails to challenge models with more complex tasks involving integrated visual elements like tables, symbols, and scientific notations.

We introduce EXAMS-V, which addresses all these issues. First, this dataset represents a significant leap forward, treating visual and text content as a cohesive unit. This forces models to engage in more sophisticated processing, including distinguishing, preprocessing, and logical reasoning over combined textual and visual information. Additionally, EXAMS-V has a multilingual reach, covering 7 language families, further enhancing its complexity and applicability.

The key contributions of our paper include:

- We introduce a novel dimension to benchmarking vision language models, requiring them to reason over a unified snapshot that includes text, images, tables, graphs, and more. For this, we propose a new multimodal multilingual dataset, EXAMS-V, comprising 20,932 questions, spanning 11 languages and 20 subjects.
- We evaluate the performance of state-of-the-art large language models and vision language models on our proposed dataset.

Through EXAMS-V, we aim to set a new standard in evaluating VLMs, providing a more realistic and challenging benchmark that mirrors the complexity and diversity of real-world information processing.

2 Related Work

LLM witnessed remarkable advancements in recent years, enabling them to generate human-like text, answer complex questions, and perform a wide range of NLP tasks (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a; Chiang et al., 2023; Liu et al., 2023c). Simultaneously to the rapid development of English-centroid LLMs, researchers have also focused on extending monolingual language models to multilingual (Scao et al., 2022; Zeng et al., 2023; Li et al., 2023a; Sengupta et al., 2023) and multimodal (Alayrac et al., 2022; Chen et al., 2022; Liu et al., 2023a; Li et al., 2023c; Bai et al., 2023). Models, such as GPT-4 (OpenAI, 2023), Gemini (Anil et al., 2023) have demonstrated exceptional performance on various benchmarks and have been widely adopted in academia and industry. However, the evaluation of these models is a critical aspect that requires careful consideration to ensure reliable and comprehensive assessments.

Several benchmarks have been proposed to assess the multimodal capabilities of LLMs (Antol et al., 2015; Hudson and Manning, 2019b; Gurari

Dataset	Size	Source	Answer
MMBench (Liu et al., 2023b)	2974	Repurposed from 12 existing datasets	MC
MM-Vet (Yu et al., 2023)	200	Internet images and annotated questions	Open
ScienceQA (Lu et al., 2022)	21,198	Textbooks	MC
MMMU (Yue et al., 2023)	11,550	Textbooks, Internet, Annotated	Open/MC
MathVista (Lu et al., 2023)	6,141	Repurposed from 28 existing dataset	Open/MC
M3Exam (Zhang et al., 2023)	12,317	Exam Papers	MC
EXAMS-V	20,932	Exam Papers	MC

Table 1: Comparison of EXAMS-V with existing benchmarks. Here, "repurposed" means the benchmark is a compilation of prior datasets, MC refers to multi-choice type questions, and "open" refers to open-ended generation questions.

	M3Exam	EXAMS-V
Interleaved	No	Yes
Languages	9	11
Min Sub. in a lang	1	3
Max Sub. in a lang	12	13
Avg. Sub. per lang	5	7.1
Samples	12,317	20,946
Multimodal samples	2,816	5,086

Table 2: Comparison of M3Exams with EXAMS-V. Here, interleaved means that multimodal elements, like tables, figures, etc., are interleaved with the textual information in the image. The average subject per language for EXAMS-V is reported by excluding Polish because Polish is a collection of 55 different professional exams that cannot be directly mapped to conventional subjects.

et al., 2018; Singh et al., 2019; Lu et al., 2022; Yue et al., 2023; Lu et al., 2023). Most early-stage benchmarks consist of photos as images, and the questions ask about the objects, attributes, or relationships between objects in the image.

Recently, inspired by the use of school exams as benchmarks for LLMs, researchers have begun to collect curriculum-based questions with images for VLM benchmarking. ScienceQA (Lu et al., 2022) is one of the most popular datasets in this area. It contains 21,208 multimodal multiple-choice questions with rich domain diversity across 26 topics, collected from elementary and high school science curricula. To answer these science questions, a model needs to understand multimodal content and extract external knowledge to arrive at the correct answer. MMMU (Yue et al., 2023) is another benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. It includes 11,550 questions from college exams, quizzes, and textbooks, covering six core disciplines: art, business, science, health, humanities, and technology. Similarly, MathVista (Lu et al., 2023) is a benchmark with 6,141 samples for evaluating the mathematical reasoning capabilities in a visual context. However, like all previous benchmarks, these two exam benchmarks are in English.

M3Exam (Zhang et al., 2023) is the first multilingual multimodal exam benchmark that covers 9 languages. It includes 12,317 questions, with 2,816 questions requiring information from an image to arrive at the answer. One main difference between M3Exam and our dataset is that, like all other VLM benchmarks, M3Exam separates text and images for a single question, while we embed the question in the images.

Unlike the above benchmarks, our dataset boasts a broader linguistic scope, placing a particular emphasis on low-resource languages like Croatian, Hungarian, Spanish, and French. Notably, our examination benchmark surpasses others by featuring a greater number of questions, encompassing a diverse range of types and topics. This variety includes questions with accompanying images, tables, and graphs, as well as mathematical and chemistry equations. For a detailed quantitative analysis, please refer to Table 1.

3 EXAMS-V Dataset

EXAMS-V is a multimodal extension of the EXAMS dataset (Hardalov et al., 2020), which is collected from official state examinations crafted by the ministries of education across different countries. These assessments, taken by high school graduates, cover diverse subjects, including core disciplines like Biology, Chemistry, Geography, History, and Physics, as well as specialized areas such as Economics and Informatics. The original EXAMS dataset was intended for multilingual question answering and thereby ignored the questions requiring visual information. We included additional data for English and Chinese in our EXAMS-V dataset. The subject coverage and statistics are detailed in Table 8.

3.1 Data Collection and Analysis

Collection and Preparation of Dataset. The dataset collection process consisted of three main steps. Initially, we retraced the original PDFs used for creating the original EXAMS dataset. For English and Chinese, we gathered high school and entrance exam questions (specifically, Gaokao and JEE Advanced Questions) from China and India, respectively.

Then, we converted each PDF document to a series of cropped images, with each image having a single question and possible answers with accompanying tables, images, graphs, etc. This required the conversion of each page in the PDF document to an image and then the use of an open-sourced labeling pipeline to place bounding boxes around each question and its answers for each page.[2]

The third step involved the creation of metadata for each cropped question. This metadata includes a unique ID, file path to the question snapshot, subject, grade, language, and the correct answer for the question. Each set of metadata is stored as a JSON file corresponding to a specific subject in a particular language.

Annotation Guidelines. All the bounding box annotations are done manually by the authors with the following agreed-up guidelines: Only multiple-choice questions with 3 to 5 options and exactly **one** correct answer are considered, as they allow for a standard automatic evaluation of the correctness of model outputs.

Along with placing the bounding boxes, we marked whether the context within the bounding box is pure text or has visual context like tables, graphs, figures, or symbols. As the result of annotation, each question sample is an image that contains the question text and candidate options, along with other vision information such as figures, tables, graphs, etc. It also includes meta-information, as mentioned beforehand. This rigorous process allowed us to maintain the high quality of the dataset.

Data Quality Assessment. After the completion of our annotation process, we conducted a data quality assessment on seven languages based on the availability of an annotator with language expertise. In this evaluation, we randomly selected 50 questions from each language and requested annotators to assess each image sample based on four binary criteria:

- Image Clarity: Clarity of visual elements such as images, diagrams, or tables.
- Question Clarity: Clarity of textual information in the question.
- Single Correct: The image contains a single Multiple Choice Question (MCQ) with precisely one correct option.
- Others: Identification of other issues. The other issues encompass factors that render a question invalid, such as the presence of the answer within the question snapshot.

A question is deemed completely valid only if it meets all four criteria.

Upon thorough review, all annotators unanimously deemed the samples to exhibit exceptionally high quality across all annotated languages. Specifically, all the samples in Bulgarian, Croatian, Serbian, Italian, and Arabic met the four quality assessment criteria. However, in the case of Chinese, one sample exhibited unclear image information, and another displayed a question that lacked clarity. Similarly, an English sample was deemed invalid due to the presence of the answer within the image sample. This proves the high quality of our dataset. The annotation guideline used by the annotators is provided as Figure 8 in the Appendix section C.

3.2 Data Statistics

The EXAMS-V dataset contains 20,932 samples in total, spanning 20 subjects from grade 4-12. It encompasses a total of 11 languages from 7 language families, while it contains parallel data in more than three languages.[3] The statistics are presented in Table 3, and Table 8 per language per subject details.

---

[2] `https://github.com/Cartucho/OpenLabeling`

[3] Parallel data means that questions are semantically the same but in different languages.

Language	ISO	Family	Grade	# Subjects	# Questions	# visual Q.	# text Q.
English	en	Germanic	11, 12	4	724	181	543
Chinese	zh	Sino-Tibetan	8-12	6	2,635	1,991	644
French	fr	Romance	12	3	439	50	389
German	de	Germanic	12	5	819	144	675
Italian	it	Romance	12	11	1,645	292	1,353
Arabic	ar	Semitic	4-12	6	823	117	706
Polish	pl	Slavic	12	1	2,511	422	2,089
Hungarian	hu	Finno-Ugric	12	6	3,801	495	3,306
Bulgarian	bg	Slavic	4, 12	4	2,132	435	1,697
Croatian	hr	Slavic	12	13	3,969	700	3,269
Serbian	sr	Slavic	12	11	1,434	259	1,175

Table 3: Statistics of EXAMS-V dataset. The languages are ordered from high-resource to low-resource languages. Here, # visual Q. refers to questions with multimodal context and # text Q. refers to text only questions.

Language Diversity. Table 3 provides an overview of the various languages featured in the dataset, along with the number of questions and subjects available for each language. The dataset includes high-resource languages like English and Chinese and low-resource languages such as Bulgarian, Croatian, and Serbian. It offers a diverse linguistic landscape, spanning Germanic, Slavic, and Sino-Tibetan language families. We also include Arabic, which has a script directionality from right to left. Additionally, Slavic and Romance language families exhibit multiple language representations, enabling the evaluation and understanding of closely related languages. These characteristics make EXAMS-V a great fit for multimodal multilingual assessment of any LLMs and VLMs.

Parallel Questions. Examinations in Croatia and the United Arab Emirates are administered in multiple languages, facilitating the development of parallel question sets for two language groups. Specifically, for Croatian examinations, we have parallel questions available in both Serbian and Italian. Additionally, Arabic questions are paired with English counterparts for four subjects: Science, Physics, Chemistry, and Biology. This process resulted in the creation of 1,207 Serbian questions and 1,147 Italian questions in parallel with Croatian. Furthermore, for Arabic, we have developed 262 parallel questions in English.

Subject Diversity. Each education system has its own specifics, leading to some differences in curricula, topics, and even the naming of the subjects. As a result, we initially collected 83 different subjects from different countries. Since different naming conventions for subjects in different countries, the values of the subjects were very sparse and non-uniformly populated. We performed subject aggregation to club similar subjects into one single subject and finally got 20 aggregated subjects. They were further grouped into three major categories, based on the main branches of science: Natural Sciences – the study of natural phenomena; Social Sciences – the study of human behaviour and societies; others – Applied Studies, Arts, Religion, etc. The distribution of the major categories is Natural Sciences (53.02%), Social Sciences (27.15%), and Others (19.82%).

Question Complexity. The dataset is compiled from high school examinations administered in various countries, primarily featuring questions from grades 4 to 12. Questions in natural sciences such as Physics, Chemistry, Biology, and Mathematics, demand foundational knowledge of these subjects and intricate reasoning skills. Questions related to Geography and History necessitate specific knowledge about particular regions or countries. Additionally, the Polish section comprises a compilation of 55 diverse professional exam questions across various fields, spanning from accounting to the motor vehicle service process. Answering these questions requires precise understanding of these professions.

3.3 Comparison with Existing Datasets

EXAMS-V differs from other datasets by mainly introducing a new way of benchmarking VLMs – passing an entire question snapshot that contains both the visual and the text components instead of passing the parsed and processed text with the image. This leaves the model to the work of text extraction and representation. Moreover, the dataset has questions of varying complexity and diversity with most of the questions coming from high school matriculation exams. Most previous benchmarks normally require commonsense knowledge or simple physical or temporal reasoning. In contrast, the EXAMS-V benchmark requires deliberate reasoning with high school-level subject and region-specific knowledge. Lastly, EXAMS-V aims to cover high school-level knowledge with different forms of visual features, including diagrams, tables, charts, chemical structures, paintings, geometric shapes, etc. This means that a well-performing model on EXAMS-V could be considered to surpass a human adult on general-purpose tasks. We have included a detailed comparison of EXAMS-V dataset with other benchmarks in Table 1.

4 Experimental Setup

As we see in Table 8, the original data appears sparse and imbalanced. To ensure a more balanced benchmark, we split EXAMS-V into training and test sets, with careful consideration for language and subject representation in the test set. We sampled 20 to 100 questions for each subject-language pair based on availability. For languages with parallel data like Croatian, Serbian, and Italian, we performed parallel splits to maintain question consistency across training and test sets. Finally, we got 16,724 training and 4,208 test instances.

We evaluate state-of-the-art LLMs and VLMs on EXAMS-V benchmark. Our evaluation is conducted under a zero-shot setting without model finetuning or in-context learning, using either APIs or NVIDIA A100 GPUs.[4]

4.1 Models

VLMs. We consider various large vision language models. We evaluated two open-source models, which have shown remarkable performance on multiple multimodal tasks: (i) LLaVA-1.5 (Liu et al., 2023a) which integrates visual embeddings with Vicuna's linguistic space. (ii) Qwen-VL-Chat (Bai et al., 2023), a multilingual multimodal chat model trained on Chinese and English data, which possesses excellent grounding, text-reading, and text-oriented question-answering performance. We also evaluated two proprietary multimodal models: GPT-4V and Gemini-Pro-Vision (denoted as Gemini-V) (Anil et al., 2023). GPT-4V is the best-performing multimodal model by OpenAI, and Gemini-V is the mid-range model among the Gemini family of multimodal models.

Augmented LLMs. To evaluate text-only LLMs, we augment language models with two image-to-text tools, namely Optical Character Recognition (OCR) and Image Captioning (IC). We employ Google Tesseract for OCR and GPT-4V for image captioning. We treat LLM augmented with OCR and IC as a vision language system. This setup was applied to GPT-3.5-Turbo, GPT-4, and Gemini Pro.

4.2 Evaluation Setup

Given the multiple-choice nature of the questions, accuracy served as our primary metric. Models were instructed to format answers as JSON objects {"answer": "choice"}, allowing for straightforward prediction extraction from the outputs. Based on our observations, all models under consideration can adhere to the instructions and produce the answer in a JSON format.

5 Main Results

We present the results across languages in Table 4. To gain a clearer understanding of model performance, we establish a random baseline by assigning an option randomly from the available choices for each question. The random baseline for all languages ranges between 19-26%.

VLM Results. Among the various VLMs, GPT-4V stands out with the highest performance, achieving an overall average score of 42.78%. This score, being only 20 percentage points above the random baseline, indicates significant potential for improvement in VLM capabilities. Gemini-V, following GPT-4V in our evaluation, achieves an overall average of 31.13%.

In comparison to commercial VLMs, open-source VLMs such as LLaVA-1.5-13B and Qwen-VL-7B fall short in terms of language support and model performance. According to our findings, open-source VLMs are limited in language support (2 for Qwen and 1 for LLaVA) and their performance in these languages is close to the random baseline. On the other hand, commercial models exhibit broader language support, as evidenced by

---

[4] Our experiments were conducted in Dec-2023 and attached to the latest version of the commercial models.

Model	**bg**	**zh**	**hr**	**fr**	**de**	**hu**	**en**	**sr**	**it**	**ar**	**pl**	**Avg**
Random	25.23	24.13	25.58	24.14	22.56	19.55	24.40	24.50	24.76	19.83	23.00	23.62
Vision Language Models (VLMs)												
LLaVA-1.5-13B	–	–	–	–	–	–	26.00	–	–	–	–	–
Qwen-VL-7B	–	15.72	–	–	–	–	23.60	–	–	–	–	–
GPT-4V	**36.00**	22.20	55.47	**60.34**	**51.24**	**44.77**	29.27	39.84	62.07	24.29	30.00	42.78
Gemini-V	30.46	**24.56**	29.39	47.70	47.80	27.05	29.20	28.29	43.03	19.38	28.00	31.13
Augmented Large Language Models (LLMs): OCR + Captioning												
GPT-3.5 Turbo	27.08	22.20	52.08	39.08	34.81	37.73	30.00	48.61	55.48	26.36	33.00	39.47
GPT-4	30.46	23.57	**66.58**	36.71	23.76	34.09	**32.40**	**73.51**	**75.95**	**26.47**	30.00	**47.11**
Gemini Pro	32.00	23.97	58.90	38.51	28.09	43.41	31.20	59.96	64.38	23.25	**42.00**	43.99

Table 4: Overall results for different models on EXAMS-V test set. Besides reporting performance for VLMs, we additionally add text-only LLM baselines. The best-performing model in each category is in **bold**, and the second-best is underlined.

Subject	**GPT-4V**			**Gemini-V**			**GPT-4 (w/ OCR, captions)**		
	hr	**sr**	**it**	**hr**	**sr**	**it**	**hr**	**sr**	**it**
Biology	**72.04**	37.64	66.90	32.26	31.18	**43.41**	75.27	73.11	**77.55**
Chemistry	48.00	28.00	**53.33**	25.33	26.67	**34.67**	**72.00**	68.00	72.00
History	59.26	45.68	**61.73**	29.63	23.46	**37.03**	**85.19**	77.78	76.54
Informatics	38.39	33.33	**40.74**	42.59	33.34	**46.29**	34.00	57.41	**66.67**
Politics	**82.22**	46.67	73.33	46.67	31.11	**64.44**	**97.78**	91.11	86.67
Psychology	85.19	55.56	**88.89**	33.33	29.63	**59.26**	92.59	**100.00**	92.59
Sociology	**63.33**	53.33	60.00	33.33	30.00	**56.67**	**80.00**	73.33	70.00
Average	**62.56**	40.44	62.11	33.33	28.76	**45.25**	**80.53**	75.29	76.60

Table 5: Fine-grained subject-wise comparison on the parallel Croatian–Serbian–Italian examples. For a particular VLM or augmented LLM, the best-performing language for each subject among the three languages is in **bold**.

their performance surpassing random outcomes in almost all languages.

LLMs Augmented with OCR and Captioning. Large language models enhanced with OCR and image captioning show superior average performance compared to standalone vision-language models. GPT-4, when augmented with OCR and captioning, demonstrates the highest overall performance among both VLMs and LLMs. This can be attributed to the precise OCR capabilities of Google Tesseract, the detailed captions produced by GPT-4V, and GPT-4's robust textual reasoning abilities. Furthermore, unlike prompting GPT-4V to directly generate the answer, augmented GPT-4 decouples the difficulties in visual information extraction and text reasoning.

5.1 Analysis from a Language Perspective

Comparing model performance in different languages, we find that all models show random-level results for Chinese (zh), which might be due to the inherent challenges associated with the Chinese subset. The Chinese subset, derived from the Gaokao exam, contains the highest proportion of vision features such as figures, tables, or graphs. This makes it difficult not only for single VLM but also for OCR and image captioning techniques to capture the visual information in text form fully.

Following Chinese, Arabic (ar) and English (en) emerge as the next most challenging languages. For Arabic, the low performance is associated with the image sample itself. Figure 9 shows an image sample of Arabic where we can see that, unlike other subjects, the answer choices do not have any letter associated with it. Thus, when the evaluated VLMs and LLMs are instructed to return the answer in the form of A, B, C or D, they find it very difficult to pinpoint the correct option.

For English, the top-performing models only scored about 8 % above the random baseline. The difficulty in English might be attributed to its sourcing from the Joint Entrance Exam (JEE) conducted every year for admission to Engineering Institute in India. To solve these questions, the model needs to be able to demonstrate complex multi-step reasoning along with a very good understanding of fundamental science- Physics, Chemistry, and mathematics.

Overall, the best performing VLM, i.e. GPT-4V, outperforms in four languages, whereas LLMs with OCR and captioning capabilities excel in several languages. These four languages include Bulgarian (bg), French (fr), German (de) and Hungarian (hu). If we refer to the test set distribution reported in Table 9 of the Appendix, we observe that most of the samples in these languages have very few multimodal questions. Additionally, they have very few graphical and tabular questions on which GPT-4V tends to show poor performance according to Table 6. Other languages like Croatian (hr), Serbian (sr), Italian (it), and Polish (pl) have a fair distribution of multimodal and textual questions.

5.2 Parallel Data Evaluation

Since Croatian, Serbian, and Italian data come from the same examination, we conducted a parallel sample experiment for these languages. The GPT-4V, Gemini-V, and augmented GPT4 results are reported in Table 5. The results show dependence on the language for all the models under consideration. Augmented GPT4 has the least performance variance. This can be attributed to the accurate OCR capabilities of Google-Tesseract.

For GPT-4V, there is a significant performance gap between Croatian and Serbian with Croatian outperforming Serbian by 20.12 %. Although both languages are very similar and often mutually intelligible, their scripts differ significantly. Serbian is Cyrillic, whereas Croatian is Latin. Latin script is more widely used, and the majority of the most spoken languages in the world have Latin script. This can be attributed to the strong performance of GPT-4V in languages with Latin script, like Croatian and Italian.

Even for Gemini-Vision-Pro , there is a gap in performance between Croatian and Serbian. The accuracy for Croatian is better than Serbian by 4.57%. Gemini-V exhibits a notable performance disparity between Italian and Croatian, as well as Serbian. This discrepancy is likely because Italian, as a high-resource language, enjoys greater representation within the Gemini family of models.

Feature	Samples	GPT-4V	Gemini-V
Symbol	36	52.78	25.00
Figure	50	60.00	22.00
Graph	50	42.00	26.00
Table	40	27.50	37.50
Text	50	62.00	48.00

Table 6: Model performance on different vision features.

Model	Accuracy on English
Random	24.40
LLaVA-1.5-7B	23.60
LLaVA-1.5-13B	26.00
Moondream-1.6B	24.40

Table 7: The performance of models of different scales on the English subset of EXAMS-V.

5.3 Vision Feature Evaluation

We compared the performance of GPT-4V and Gemini-V for four different vision features: scientific symbols, figures, graphs, and tabular data. We compared it against image samples with only textual information. For the evaluation, we curated samples of different vision features from the Croatian subset. The results are reported in Table 6.

GPT-4V shows fairly good performance for questions that involve scientific symbols and figures. However, it demonstrates poor performance for questions with graphs and tabular. Surprisingly, Gemini-V can show better performance for tabular data when compared to GPT-4V. Nevertheless, its performance across the remaining three vision features — scientific symbols, figures, and graphs, was subpar.

5.4 Effect of Model Scale

To understand the scale effect, we compared the performance of LLaVA-1.5-7B and LLaVA-1.5-13B on the English subset of the EXAMS-V dataset. We also evaluated the model on Moondream (Vikhyat, 2024), a 1.6B parameter model demonstrating strong performance in visual reasoning benchmarks such as GQA (Hudson and Manning, 2019a). The results are reported in the Table 7. The table shows that the performance of all three models is near random. However, the larger LLaVA-1.5-13b model performs better than LLaVA-1.5-7b by 2.4%.

6 Conclusion and Future Work

The development of EXAMS-V as a benchmark for assessing the multilingual and multimodal capabilities of VLMs marks a significant milestone in the journey towards multilingual models. Furthermore, the EXAMS-V introduces a new dimension to visual question answering where the textual information is a part of the image. Thus, EXAMS-V not only tests the multimodal reasoning capability of current VLMs but also their ability to do OCR in a multilingual context. This requires a strong perception capability to draw boundaries between textual questions and multimodal contexts like tables, figures, graphs, etc. Furthermore, the questions with in-depth knowledge of multiple disciplines or subjects and questions from physics, chemistry, and mathematics require intricate reasoning. These features collectively contribute to the considerable complexity of the EXAMS-V. We believe the evaluation of VLMs on this dataset can directly contribute to our understanding of the progress toward the expert vision language model with multilingual capability.

In future work, we plan to extend the dataset with more image samples, subjects, languages, and modalities.

Limitations

Despite its comprehensive nature, EXAMS-V, like any benchmark, is not without limitations. For ease of evaluation and analysis, we only considered and collected multiple-choice questions. We limited our multimodal analysis to four broad categories, which are scientific symbols, figures, graphs, and tabular data. But this can be further extended to finer-grained analysis. For example, scientific symbols can be further broken down into mathematical notions and chemical symbols, while figures can be broken down into maps, figures, paintings, diagrams, etc. However, this requires the collection of more data, which is difficult, particularly for low-resource languages like Croatian, Serbian, and Arabic under consideration. Furthermore, since we are collecting exam questions from different regions of the world, the difficulty of the questions varies depending on the region they originate from. This hurts the comparability of the dataset across languages. Although we tried to include parallel questions for direct comparability, but it was feasible only for three European languages: Croatian, Serbian, and Italian.

Ethical Consideration

- **Copyright and Licensing:** All data in EXAMS-V are collected from public sources.
- **Ethics and Data Privacy:** All testing instances in EXAMS-V are carefully scrutinized to exclude any examples with ethical concerns. Since all the data are collected from exam papers there is no privacy issue.

Acknowledgments

This research is partially funded by the EU NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

References

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

Gemini Team Google Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models.

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.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities. *ArXiv preprint*, abs/2308.12966.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18009–18019. IEEE.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. IEEE Computer Society.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Drew A. Hudson and Christopher D. Manning. 2019a. Gqa: a new dataset for compositional question answering over real-world images. *ArXiv preprint*, abs/1902.09506.

Drew A. Hudson and Christopher D. Manning. 2019b. 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.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-X: A multilingual replicable instruction-following model with low-rank adaptation. *ArXiv preprint*, abs/2305.15011.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. CMMLU: Measuring massive multitask language understanding in chinese.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. MMBench: Is your multi-modal model an all-around player?

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023c. LLM360: Towards fully transparent open-source llms.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. MathVista: Evaluating math reasoning in visual contexts with GPT-4V, Bard, and other large multimodal models. *ArXiv preprint*, abs/2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2023. GPT-4 technical report.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *ArXiv preprint*, abs/2308.16149.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Vikhyat. 2024. Moondream: A tiny vision language model. Accessed: 2024-06-03.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv preprint*, abs/2311.16502.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3EXAM: A multilingual, multimodal, multilevel benchmark for examining large language models.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592.

A Dataset statistics

The statistics of the EXAMS-V dataset for all languages and subjects are presented in Table 8. Table 9 shows the distribution of multimodal data in the test dataset used for the evaluation of VLMs and LLMs. One point to note is that there are instances where a single image might have multiple modalities, e.g. figure and table. We count them in each of the categories in the table.

Subjects	**bg**	**zh**	**hr**	**sr**	**it**	**fr**	**de**	**hu**	**en**	**ar**	**pl**
Physics	970	408	649	305	215	235	510	1,570	185	67	
Chemistry	665	381	427	322	212		14	697	347	150	
Biology	233	281	574	294	424				47	67	
Geography		678	383	54	40	24	46	92			
Sociology	264		295	30	109					306	
Business				6		180	216	747			
History		209	500	200	235						
Philosophy			140	12	34						
Psychology			154	47	105						
Politics			270	90	100						
Informatics			188	74	146						
Mathematics		678							145		
Ethics			180		25						
Tourism							33	43			
Science									20	150	
Professional											2,511
Islamic Studies										83	
Religion			161								
Fine Arts			48								
Agriculture								652			
Overall	2,132	2,635	3,969	1,434	1,645	439	819	3,801	744	823	2,511

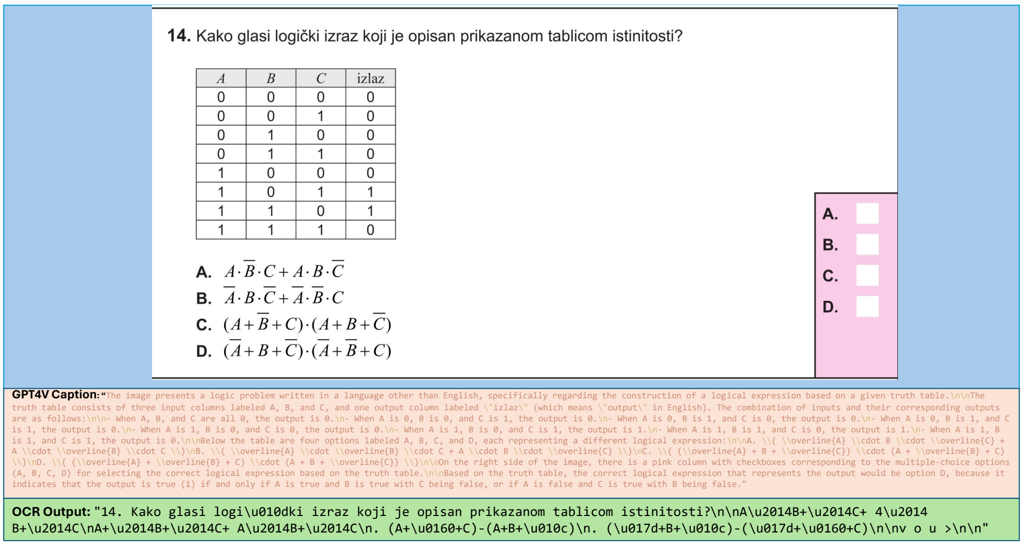
Table 8: Detailed statistics of EXAMS-V.

Languages	Table	Figure	Graph	Symbol	Text
English	0	38	28	43	159
Chinese	82	246	56	56	126
French	0	37	6	9	124
German	0	35	10	8	179
Italian	11	151	13	39	382
Arabic	11	81	12	29	211
Polish	14	36	0	0	50
Hungarian	0	36	7	80	270
Bulgarian	6	70	10	51	200
Croatian	19	330	24	62	780
Serbian	11	151	13	39	357

Table 9: Vision feature distribution of EXAMS-V test set

B Example of OCR and GPT-4V Caption Output

This section shows examples of OCR and GPT-4V for different vision features.



14. Kako glasi logički izraz koji je opisan prikazanom tablicom istinitosti?

A	*B*	*C*	izlaz
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

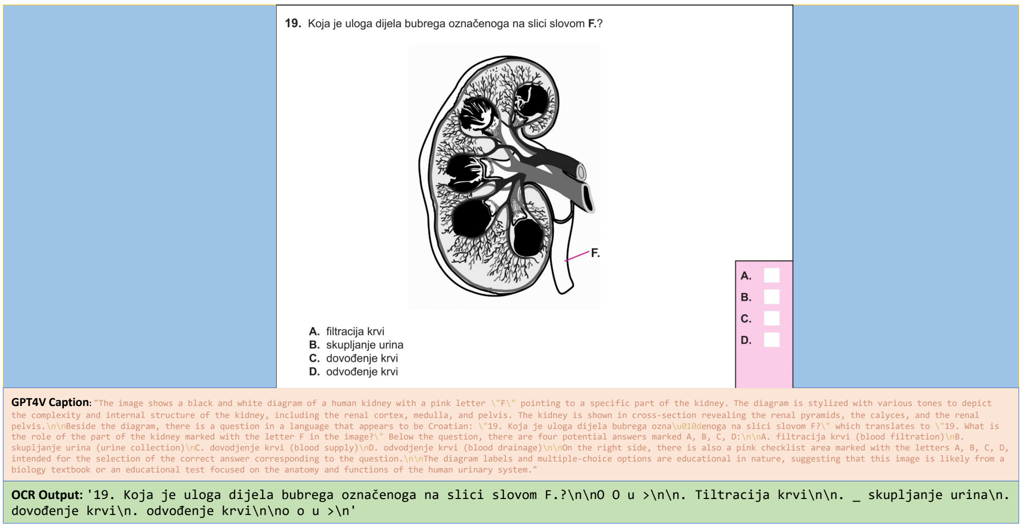
A. $A\cdot\overline{B}\cdot C + A\cdot B\cdot\overline{C}$
B. $\overline{A}\cdot B\cdot\overline{C} + \overline{A}\cdot\overline{B}\cdot C$
C. $(A+\overline{B}+C)\cdot(A+B+\overline{C})$
D. $(\overline{A}+B+\overline{C})\cdot(\overline{A}+\overline{B}+C)$

A. ☐
B. ☐
C. ☐
D. ☐

GPT4V Caption:
```
“The image presents a logic problem written in a language other than English, specifically regarding the construction of a logical expression based on a given truth table.\n\nThe
truth table consists of three input columns labeled A, B, and C, and one output column labeled \"izlaz\" (which means \"output\" in English). The combination of inputs and their corresponding outputs
are as follows:\n\n- When A, B, and C are all 0, the output is 0.\n- When A is 0, B is 0, and C is 1, the output is 0.\n- When A is 0, B is 1, and C is 0, the output is 0.\n- When A is 0, B is 1, and C
is 1, the output is 0.\n- When A is 1, B is 0, and C is 0, the output is 0.\n- When A is 1, B is 0, and C is 1, the output is 1.\n- When A is 1, B is 1, and C is 0, the output is 1.\n- When A is 1, B
is 1, and C is 1, the output is 0.\n\nBelow the table are four options labeled A, B, C, and D, each representing a different logical expression:\n\nA. \\( \\overline{A} \\cdot B \\cdot \\overline{C} +
A \\cdot \\overline{B} \\cdot C \\)\nB. \\( \\overline{A} \\cdot \\overline{B} \\cdot C + A \\cdot B \\cdot \\overline{C} \\)\nC. \\( (\\overline{A} + B + \\overline{C}) \\cdot (A + \\overline{B} + C)
\\)\nD. \\( (\\overline{A} + \\overline{B} + C) \\cdot (A + B + \\overline{C}) \\)\n\nOn the right side of the image, there is a pink column with checkboxes corresponding to the multiple-choice options
(A, B, C, D) for selecting the correct logical expression based on the truth table.\n\nBased on the truth table, the correct logical expression that represents the output would be option D, because it
indicates that the output is true (1) if and only if A is true and B is true with C being false, or if A is false and C is true with B being false."
```

OCR Output:
```
"14. Kako glasi logi\u010dki izraz koji je opisan prikazanom tablicom istinitosti?\n\nA\u2014B+\u2014C+ 4\u2014
B+\u2014C\nA+\u2014B+\u2014C+ A\u2014B+\u2014C\n. (A+\u0160+C)-(A+B+\u010c)\n. (\u017d+B+\u010c)-(\u017d+\u0160+C)\n\nv o u >\n\n"
```

Figure 3: Example of OCR and GPT-4V caption output when provided image with tabular data



19. Koja je uloga dijela bubrega označenoga na slici slovom **F.**?

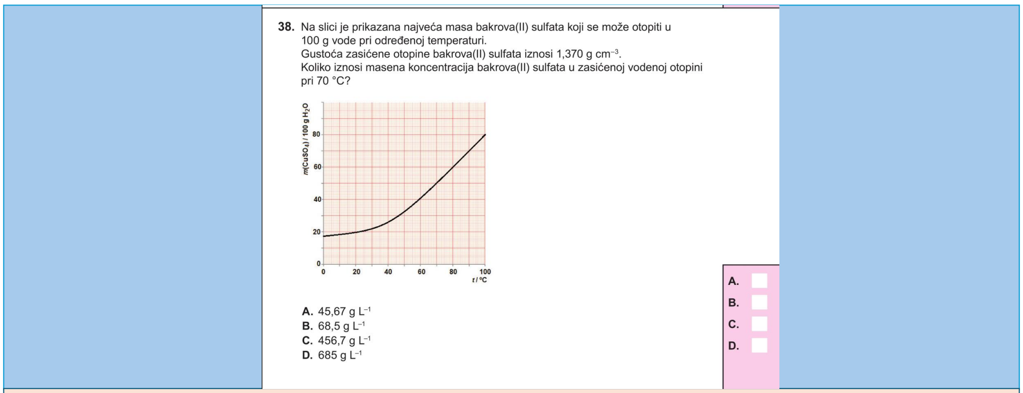
A. filtracija krvi
B. skupljanje urina
C. dovođenje krvi
D. odvođenje krvi

A. ☐
B. ☐
C. ☐
D. ☐

GPT4V Caption:
```
"The image shows a black and white diagram of a human kidney with a pink letter \"F\" pointing to a specific part of the kidney. The diagram is stylized with various tones to depict
the complexity and internal structure of the kidney, including the renal cortex, medulla, and pelvis. The kidney is shown in cross-section revealing the renal pyramids, the calyces, and the renal
pelvis.\n\nBeside the diagram, there is a question in a language that appears to be Croatian: \"19. Koja je uloga dijela bubrega ozna\u010denoga na slici slovom F?\" which translates to \"19. What is
the role of the part of the kidney marked with the letter F in the image?\" Below the question, there are four potential answers marked A, B, C, D:\n\nA. filtracija krvi (blood filtration)\nB.
skupljanje urina (urine collection)\nC. dovodjenje krvi (blood supply)\nD. odvodjenje krvi (blood drainage)\n\nOn the right side, there is also a pink checklist area marked with the letters A, B, C, D,
intended for the selection of the correct answer corresponding to the question.\n\nThe diagram labels and multiple-choice options are educational in nature, suggesting that this image is likely from a
biology textbook or an educational test focused on the anatomy and functions of the human urinary system."
```

OCR Output:
```
'19. Koja je uloga dijela bubrega označenoga na slici slovom F.?\n\nO O u >\n\n. Tiltracija krvi\n\n. _ skupljanje urina\n.
dovođenje krvi\n. odvođenje krvi\n\no o u >\n'
```

Figure 4: Example of OCR and GPT-4V caption output when provided image with figure



38. Na slici je prikazana najveća masa bakrova(II) sulfata koji se može otopiti u 100 g vode pri određenoj temperaturi.
Gustoća zasićene otopine bakrova(II) sulfata iznosi 1,370 g cm^{-3}.
Koliko iznosi masena koncentracija bakrova(II) sulfata u zasićenoj vodenoj otopini pri 70 °C?

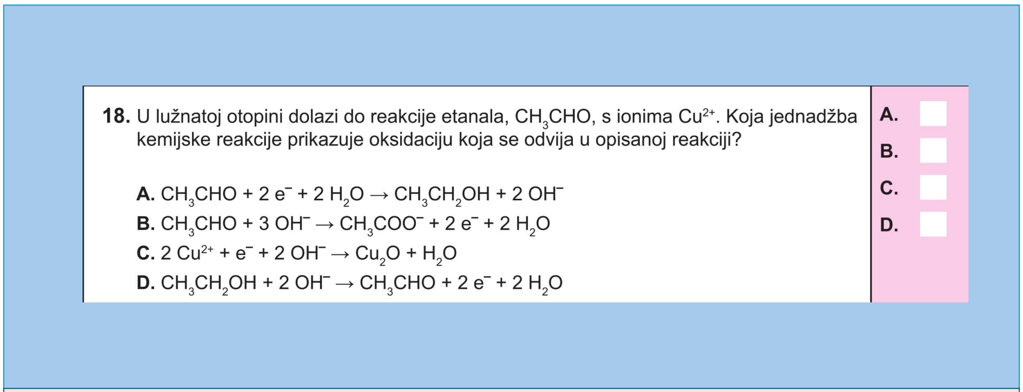
A. 45,67 g L^{-1}
B. 68,5 g L^{-1}
C. 456,7 g L^{-1}
D. 685 g L^{-1}

A. ☐
B. ☐
C. ☐
D. ☐

GPT4V Caption: 'This image contains a graph and a set of multiple-choice questions written in a language that appears to be Croatian. The graph is plotting \'mass of CuSO4\' on the vertical y-axis, measured in grams per 100 grams of H2O, against \'t / °C\' on the horizontal x-axis, which seems to represent temperature in degrees Celsius. It shows a curve that starts lower on the y-axis at 0°C and progressively increases as the temperature rises, indicating that the solubility of copper(II) sulfate increases with temperature.\n\nThe text above the graph reads as a question about solubility: "The picture shows the maximum mass of copper(II) sulfate that can be dissolved in 100 g of water at a certain temperature. The density of the saturated solution of copper(II) sulfate is 1.370 g cm^-3. How much is the mass concentration of copper(II) sulfate in the saturated aqueous solution at 70 °C?"\n\nBelow the graph, there are four multiple-choice answers labeled A, B, C, and D, offering different values for the mass concentration in g L^-1:\n\nA. 45,67 g L^-1\nB. 68,5 g L^-1\nC. 456,7 g L^-1\nD. 685 g L^-1\n\nAs an AI, I cannot solve the question but the graph and options presented suggest it is a chemistry problem related to solubility and concentration calculation. To find the correct answer, one would use the graph to determine the solubility of copper(II) sulfate at 70°C, apply density information of the solution, and then calculate the mass concentration in g L^-1.'

OCR Output: “38. Na slici je prikazana najveća masa bakrova(ll) sulfata koji se može otopiti u\n100 g vode pri određenoj temperaturi.\nGustoća zasićene otopine bakrova(ll) sulfata iznosi 1,370 g cm—.\nKoliko iznosi masena koncentracija bakrova(ll) sulfata u zasićenoj vodenoj otopini\npri 70 “ C?\n\nm(CuSs04)/ 100 g H—O\n\nA. 45,67 g —“\nB. 68,5 g —\nC. 456,7 g —\nD. 685 g —\n\nA.\nB.\nC.\nD.\n”

Figure 5: Example of OCR and GPT-4V caption output when provided image with graphical data



18. U lužnatoj otopini dolazi do reakcije etanala, CH_3CHO, s ionima Cu^{2+}. Koja jednadžba kemijske reakcije prikazuje oksidaciju koja se odvija u opisanoj reakciji?

A. $CH_3CHO + 2\,e^- + 2\,H_2O \rightarrow CH_3CH_2OH + 2\,OH^-$
B. $CH_3CHO + 3\,OH^- \rightarrow CH_3COO^- + 2\,e^- + 2\,H_2O$
C. $2\,Cu^{2+} + e^- + 2\,OH^- \rightarrow Cu_2O + H_2O$
D. $CH_3CH_2OH + 2\,OH^- \rightarrow CH_3CHO + 2\,e^- + 2\,H_2O$

A. ☐
B. ☐
C. ☐
D. ☐

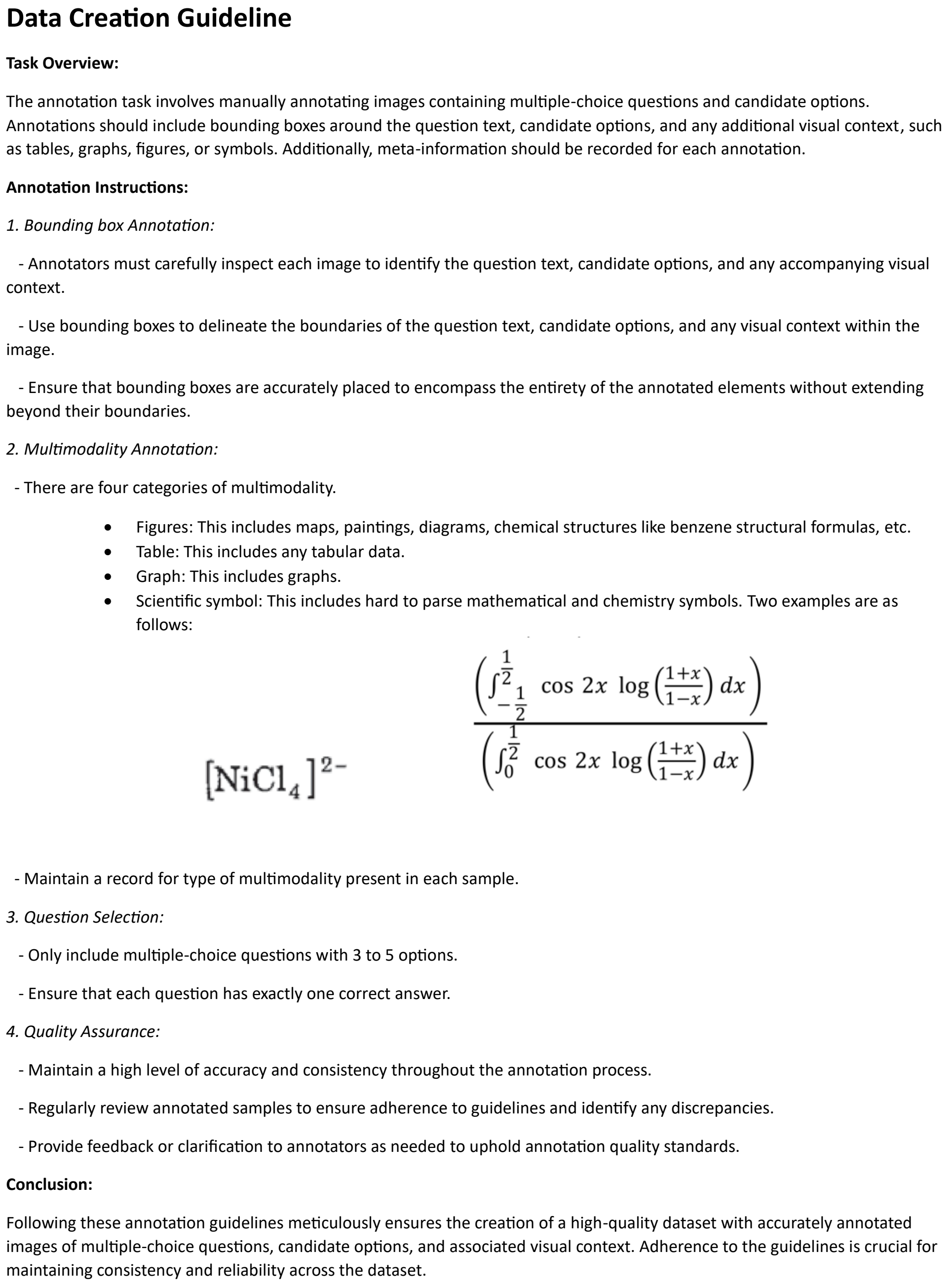
GPT4V Caption: 'This image displays a numbered question (number 18) from what appears to be a chemistry exam or homework, written in Croatian. The question is related to the oxidation reaction of ethanal (CH_3CHO) in an alkaline solution with copper ions (Cu^2+). The question asks which chemical reaction equation demonstrates the oxidation taking place in the described reaction. Below the question, four potential reaction equations labeled A, B, C, and D are provided, presumably as multiple-choice answers. To the right of the equations, there is a vertical pink column with checkboxes corresponding to each of the answer choices, but no marks have been made. There are no diagrams in the image; it contains only text and multiple-choice answer options.'

OCR Output: "'18. U lužnatoj otopini dolazi do reakcije etanala, CH,CHO, s ionima Cu**. Koja jednadžba\nkemijske reakcije prikazuje oksidaciju koja se odvija u opisanoj reakciji?\n\nA. CH,CHO + 2 e + 2 H O — CH,CHOH + 2 OH\nB. CH,CHO + 3 OH — CH,COO +2 e + 2 H O\nC. 2 Cu* + e + 2 OH — Cu, O + H O\n\nD. CH,CH OH + 2 OH — CH,CHO +2 e +2 H .O\n\no o u >\n'"

Figure 6: Example of OCR and GPT-4V caption output when provided image with chemistry symbols data

C Data Quality assessment Guideline

Figure 7 shows a snapshot the annotation guideline shared with the annotators who created the data. The data creation annotators are two authors of the paper, one of which is from India and the other from Bulgaria.



Data Creation Guideline

Task Overview:

The annotation task involves manually annotating images containing multiple-choice questions and candidate options. Annotations should include bounding boxes around the question text, candidate options, and any additional visual context, such as tables, graphs, figures, or symbols. Additionally, meta-information should be recorded for each annotation.

Annotation Instructions:

1. Bounding box Annotation:

- Annotators must carefully inspect each image to identify the question text, candidate options, and any accompanying visual context.

- Use bounding boxes to delineate the boundaries of the question text, candidate options, and any visual context within the image.

- Ensure that bounding boxes are accurately placed to encompass the entirety of the annotated elements without extending beyond their boundaries.

2. Multimodality Annotation:

- There are four categories of multimodality.
    - Figures: This includes maps, paintings, diagrams, chemical structures like benzene structural formulas, etc.
    - Table: This includes any tabular data.
    - Graph: This includes graphs.
    - Scientific symbol: This includes hard to parse mathematical and chemistry symbols. Two examples are as follows:

$[NiCl_4]^{2-}$

$$\frac{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \cos 2x \ \log\left(\frac{1+x}{1-x}\right) dx\right)}{\left(\int_{0}^{\frac{1}{2}} \cos 2x \ \log\left(\frac{1+x}{1-x}\right) dx\right)}$$

- Maintain a record for type of multimodality present in each sample.

3. Question Selection:

- Only include multiple-choice questions with 3 to 5 options.

- Ensure that each question has exactly one correct answer.

4. Quality Assurance:

- Maintain a high level of accuracy and consistency throughout the annotation process.

- Regularly review annotated samples to ensure adherence to guidelines and identify any discrepancies.

- Provide feedback or clarification to annotators as needed to uphold annotation quality standards.

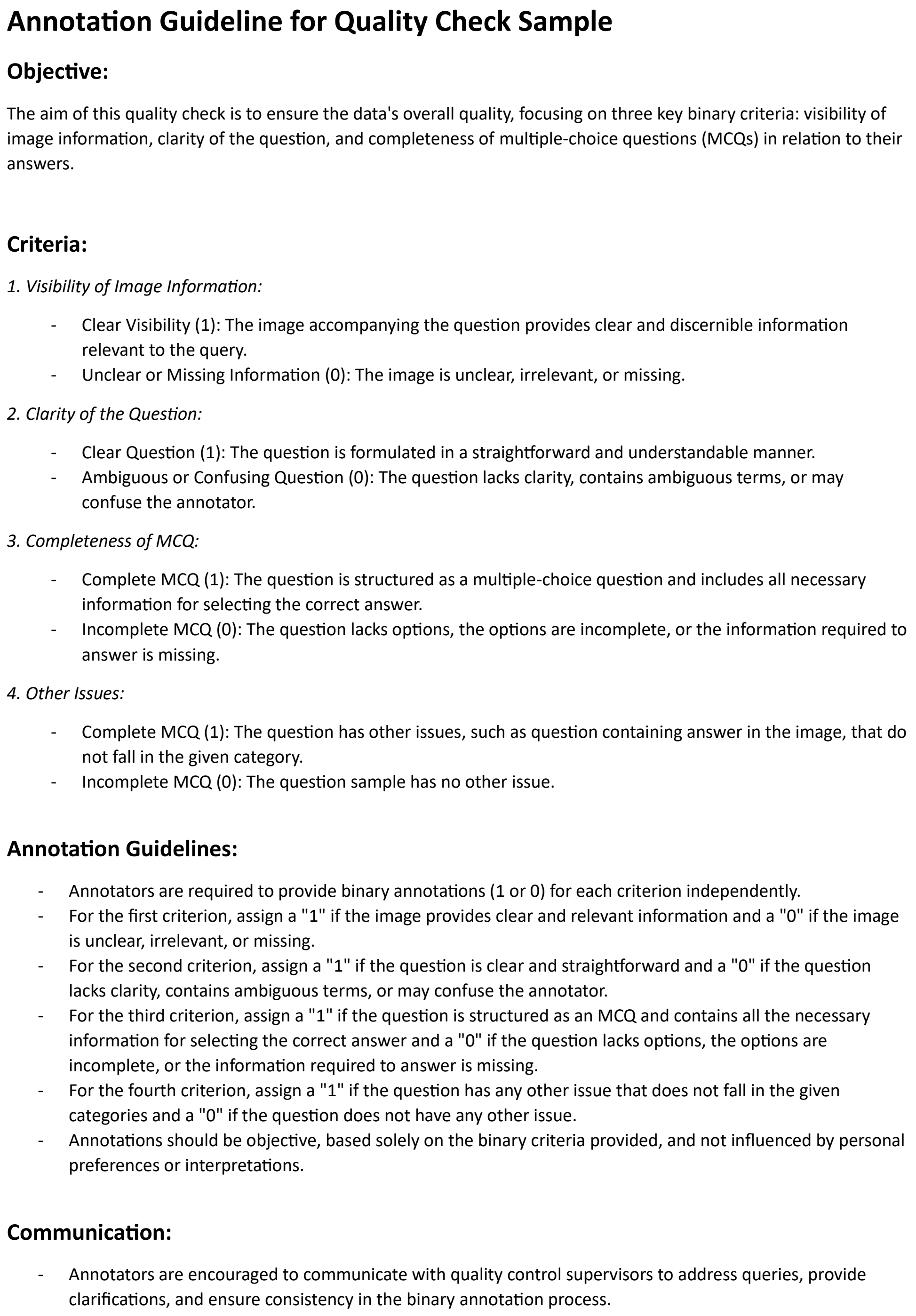
Conclusion:

Following these annotation guidelines meticulously ensures the creation of a high-quality dataset with accurately annotated images of multiple-choice questions, candidate options, and associated visual context. Adherence to the guidelines is crucial for maintaining consistency and reliability across the dataset.

Figure 7: The annotation guideline provided to the annotators while creating the dataset.

Figure 8 shows a snapshot of the annotation guideline shared with annotators for quality assessment.

The data quality assessment annotators are authors and colleagues with bachelor's degrees and native speakers of the corresponding language.



Annotation Guideline for Quality Check Sample

Objective:

The aim of this quality check is to ensure the data's overall quality, focusing on three key binary criteria: visibility of image information, clarity of the question, and completeness of multiple-choice questions (MCQs) in relation to their answers.

Criteria:

1. Visibility of Image Information:

- Clear Visibility (1): The image accompanying the question provides clear and discernible information relevant to the query.
- Unclear or Missing Information (0): The image is unclear, irrelevant, or missing.

2. Clarity of the Question:

- Clear Question (1): The question is formulated in a straightforward and understandable manner.
- Ambiguous or Confusing Question (0): The question lacks clarity, contains ambiguous terms, or may confuse the annotator.

3. Completeness of MCQ:

- Complete MCQ (1): The question is structured as a multiple-choice question and includes all necessary information for selecting the correct answer.
- Incomplete MCQ (0): The question lacks options, the options are incomplete, or the information required to answer is missing.

4. Other Issues:

- Complete MCQ (1): The question has other issues, such as question containing answer in the image, that do not fall in the given category.
- Incomplete MCQ (0): The question sample has no other issue.

Annotation Guidelines:

- Annotators are required to provide binary annotations (1 or 0) for each criterion independently.
- For the first criterion, assign a "1" if the image provides clear and relevant information and a "0" if the image is unclear, irrelevant, or missing.
- For the second criterion, assign a "1" if the question is clear and straightforward and a "0" if the question lacks clarity, contains ambiguous terms, or may confuse the annotator.
- For the third criterion, assign a "1" if the question is structured as an MCQ and contains all the necessary information for selecting the correct answer and a "0" if the question lacks options, the options are incomplete, or the information required to answer is missing.
- For the fourth criterion, assign a "1" if the question has any other issue that does not fall in the given categories and a "0" if the question does not have any other issue.
- Annotations should be objective, based solely on the binary criteria provided, and not influenced by personal preferences or interpretations.

Communication:

- Annotators are encouraged to communicate with quality control supervisors to address queries, provide clarifications, and ensure consistency in the binary annotation process.

Figure 8: The annotation guideline provided to the annotators to assess the quality of the samples.

D Fine-Grained Evaluation

Table 10 to Table 20 present a detailed, subject-wise performance analysis of various VLMs and augmented LLMs across all the languages under consideration. For a particular subject, the best-performing model is in **bold**.

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Mathematics	12.00	15.0	18.00	**20.00**	14.00
Chemistry	31.00	28.0	30.00	31.00	**42.00**
Physics	**31.00**	30.0	30.00	24.00	25.00

Table 10: Fine-Grained Subject Wise Evaluation of English

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Biology	22.00	17.00	17.00	17.00	**24.00**
Chemistry	30.00	**32.00**	24.00	20.00	27.00
Geography	14.00	**26.00**	24.00	24.00	18.00
History	19.00	23.00	22.00	23.00	**27.00**
Physics	22.86	24.29	27.14	**28.57**	14.29
Science	21.35	25.84	**31.46**	23.59	25.84

Table 11: Fine-Grained Subject Wise Evaluation of Chinese

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Business & Economics	**62.00**	48.00	38.00	44.00	34.00
Geography	**79.17**	54.17	58.33	45.83	62.00
Physics	**55.00**	26.00	34.00	35.00	32.00

Table 12: Fine-Grained Subject Wise Evaluation of French

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Business & Economics	16.67	22.92	16.67	**35.42**	20.83
Geography	**78.26**	56.52	41.30	30.43	21.74
Physics	**52.00**	52.00	14.00	33.00	25.00
Tourism	**63.64**	60.61	30.30	45.45	27.27

Table 13: Fine-Grained Subject Wise Evaluation of German

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Biology	66.90	43.41	64.73	59.39	**77.55**
Chemistry	53.33	34.67	54.67	52.00	**72.00**
Ethics	84.00	36.00	96.00	80.00	**100.00**
Geography	60.00	44.00	60.00	36.00	**72.00**
History	61.73	37.04	56.79	50.62	**76.54**
Informatics	40.74	46.29	53.70	46.30	**66.67**
Philosophy	76.47	44.12	76.47	73.53	**85.29**
Physics	51.39	31.94	51.39	34.72	**62.50**
Politics	73.33	64.44	82.22	68.69	**86.67**
Psychology	88.89	59.26	85.19	77.78	**92.59**
Sociology	60.00	56.67	76.67	66.67	**70.00**

Table 14: Fine-Grained Subject Wise Evaluation of Italian

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Biology	**29.82**	21.05	10.52	29.82	28.07
Chemistry	25.67	21.62	**28.38**	16.22	20.27
Islamic Studies	12.00	14.00	**32.00**	28.00	24.00
Physics	16.42	16.42	25.37	22.39	**31.34**
Science	**34.25**	26.03	27.40	26.03	31.51
Social	24.24	15.15	15.15	**37.88**	23.45

Table 15: Fine-Grained Subject Wise Evaluation of Arabic

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Professional	30.00	28.00	**42.00**	33.00	30.00

Table 16: Fine-Grained Subject Wise Evaluation of Polish

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Business & Economics	37.14	37.14	**52.86**	41.43	45.71
Geography	44.00	26.00	**54.00**	46.00	42.00
Physics	**52.00**	28.00	42.00	45.00	34.00
Tourism	**60.47**	25.58	55.81	41.86	32.56
Landscaping	40.74	22.22	**44.44**	33.33	29.63
Chemistry	**34.00**	23.00	30.00	25.00	28.00
Agriculture	**52.00**	24.00	38.00	34.00	26.00

Table 17: Fine-Grained Subject Wise Evaluation of Hungarian

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Biology	**42.67**	24.00	42.67	30.67	29.33
Chemistry	35.00	27.00	**43.00**	23.00	27.00
Physics	28.00	32.00	14.00	25.00	**33.00**
Sociology	**44.00**	44.00	30.00	34.00	34.00

Table 18: Fine-Grained Subject Wise Evaluation of Bulgarian

Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Biology	72.04	32.26	45.00	62.36	**75.27**
Chemistry	48.00	25.33	50.00	48.00	**72.00**
Ethics	76.00	24.00	50.00	84.00	**100**
Fine Arts	41.30	41.30	32.61	36.96	**47.83**
Geography	**46.93**	20.04	31.00	34.02	30.99
History	59.26	29.63	50.00	51.85	**85.19**
Informatics	38.89	42.59	49.40	50.00	**62.96**
Philosophy	70.59	00.00	66.00	67.65	**88.24**
Physics	45.84	27.78	37.00	40.28	**61.11**
Politics	82.22	46.67	32.00	82.22	**97.78**
Psychology	85.19	33.33	52.00	81.48	**92.59**
Religion	26.00	28.00	**56.00**	28.00	30.00
Sociology	63.33	33.33	60.00	70.00	**80.00**

Table 19: Fine-Grained Subject Wise Evaluation of Croatian

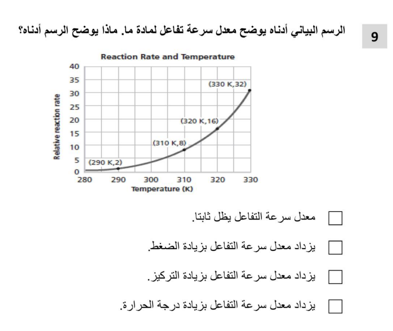
Subject	GPT-4V	Gemini-Pro-Vision	Gemini-Pro	GPT3.5	GPT4
Biology	37.64	31.18	69.89	51.61	**73.12**
Chemistry	28.00	26.67	52.00	38.67	**68.00**
Geography	36.00	32.00	72.00	40.00	**80.00**
History	45.68	23.46	55.56	46.91	**77.78**
Informatics	33.33	33.34	50.00	48.15	**57.41**
Physics	38.89	23.62	38.39	38.89	**63.89**
Politics	46.67	31.11	77.78	60.00	**91.11**
Psychology	55.56	29.63	85.19	62.96	**100.00**
Sociology	53.33	30.00	70.00	70.00	**73.33**

Table 20: Fine-Grained Subject Wise Evaluation of Serbian

E Sample Example from Different languages

Figure 10 to 19 shows sample examples from different languages in EXAMS-V.

Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.



9 الرسم البياني أدناه يوضح معدل سرعة تفاعل لمادة ما. ماذا يوضح الرسم أدناه؟

Reaction Rate and Temperature
Relative reaction rate
40
35
30
25
20
15
10
5
0
280
290
300
310
320
330
Temperature (K)
(290 K,2)
(310 K,8)
(320 K,16)
(330 K,32)

☐ معدل سرعة التفاعل يظل ثابتا.

☐ يزداد معدل سرعة التفاعل بزيادة الضغط.

☐ يزداد معدل سرعة التفاعل بزيادة التركيز.

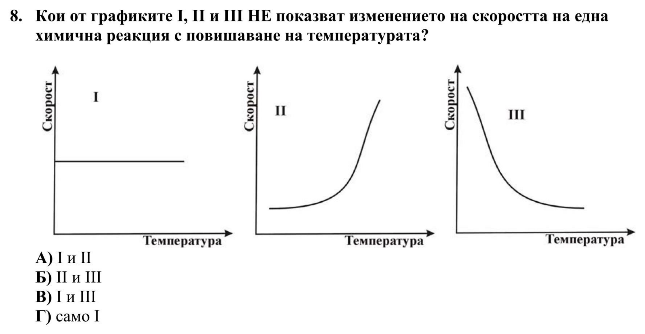
☐ يزداد معدل سرعة التفاعل بزيادة درجة الحرارة.

GPT4V Answer: `"{'answer': ‘B'}"`

Correct Answer: `"{'answer': 'C'}"`

Figure 9: Example from the Arabic test set with a GPT-4V output.

Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.



8. Кои от графиките I, II и III НЕ показват изменението на скоростта на една химична реакция с повишаване на температурата?

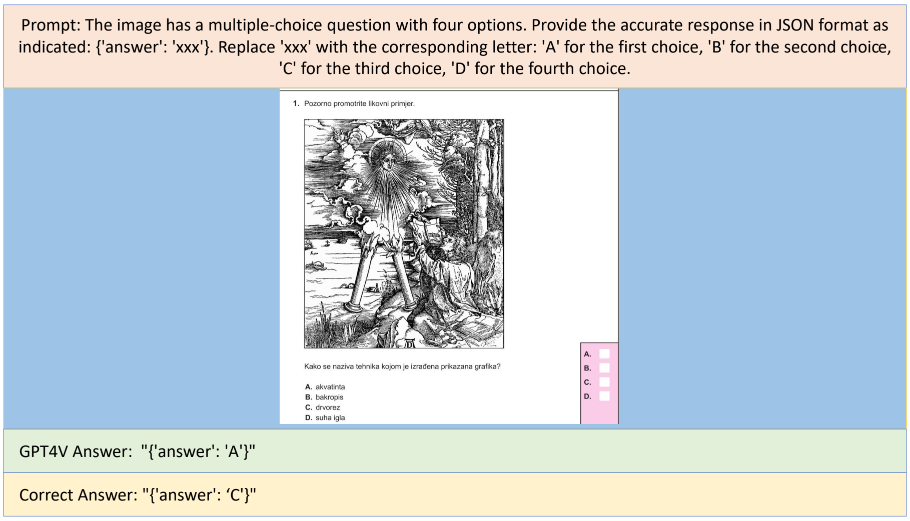
Скорост
I
Температура
Скорост
II
Температура
Скорост
III
Температура

А) I и II
Б) II и III
В) I и III
Г) само I

GPT4V Answer: `"{'answer': 'D'}"`

Correct Answer: `"{'answer': 'B'}"`

Figure 10: A sample from Bulgarian test set with GPT-4V output



Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.

1. Pozorno promotrite likovni primjer.

Kako se naziva tehnika kojom je izrađena prikazana grafika?

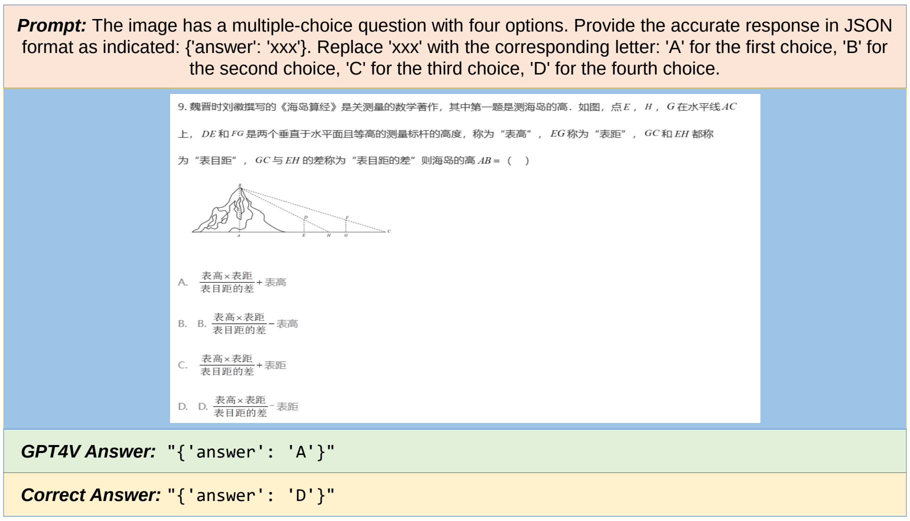
- **A.** akvatinta
- **B.** bakropis
- **C.** drvorez
- **D.** suha igla

A. ☐
B. ☐
C. ☐
D. ☐

GPT4V Answer: "{'answer': 'A'}"

Correct Answer: "{'answer': ‘C'}"

Figure 11: A sample from Croatian test set with GPT-4V output



Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.

9. 魏晋时刘徽撰写的《海岛算经》是关测量的数学著作，其中第一题是测海岛的高．如图，点E，H，G在水平线AC上，DE和FG是两个垂直于水平面且等高的测量标杆的高度，称为“表高”，EG称为“表距”，GC和EH都称为“表目距”，GC与EH的差称为“表目距的差”则海岛的高$AB=$（　）

A. $\frac{\text{表高}\times\text{表距}}{\text{表目距的差}}+\text{表高}$

B. B. $\frac{\text{表高}\times\text{表距}}{\text{表目距的差}}-\text{表高}$

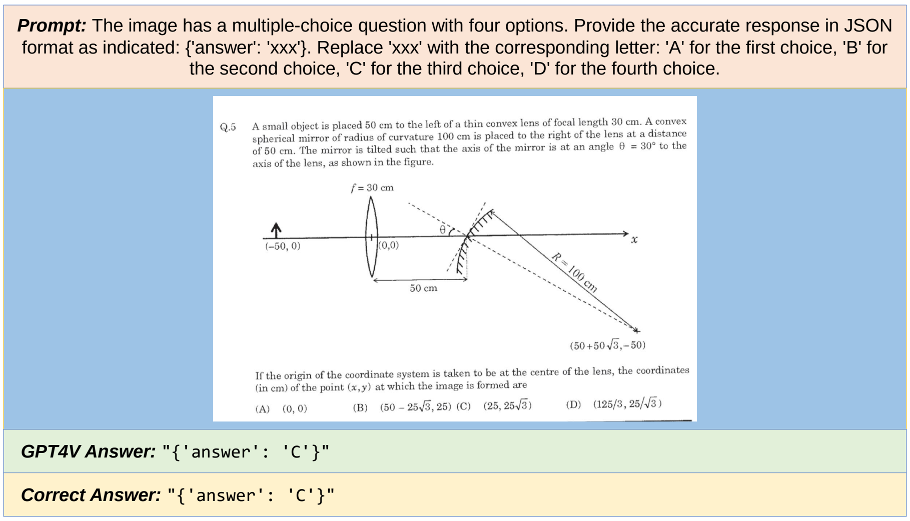
C. $\frac{\text{表高}\times\text{表距}}{\text{表目距的差}}+\text{表距}$

D. D. $\frac{\text{表高}\times\text{表距}}{\text{表目距的差}}-\text{表距}$

GPT4V Answer: `"{'answer': 'A'}"`

Correct Answer: `"{'answer': 'D'}"`

Figure 12: A sample from Chinese test set with GPT-4V output



Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.

Q.5 A small object is placed 50 cm to the left of a thin convex lens of focal length 30 cm. A convex spherical mirror of radius of curvature 100 cm is placed to the right of the lens at a distance of 50 cm. The mirror is tilted such that the axis of the mirror is at an angle $\theta = 30°$ to the axis of the lens, as shown in the figure.

If the origin of the coordinate system is taken to be at the centre of the lens, the coordinates (in cm) of the point (x, y) at which the image is formed are

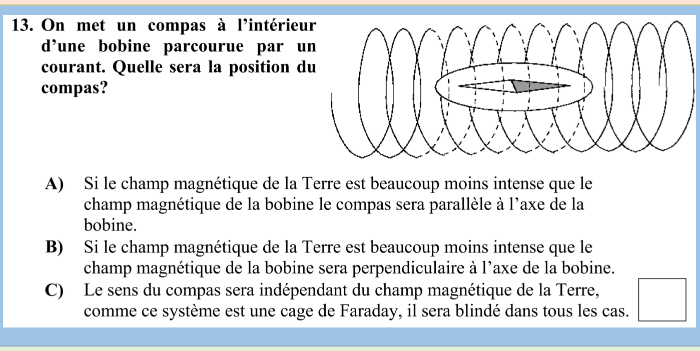
(A) (0, 0) (B) $(50-25\sqrt{3}, 25)$ (C) $(25, 25\sqrt{3})$ (D) $(125/3, 25/\sqrt{3})$

GPT4V Answer: `"{'answer': 'C'}"`

Correct Answer: `"{'answer': 'C'}"`

Figure 13: A sample from English test set with GPT-4V output

Prompt: The image has a multiple-choice question with three options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice.



13. On met un compas à l'intérieur d'une bobine parcourue par un courant. Quelle sera la position du compas?

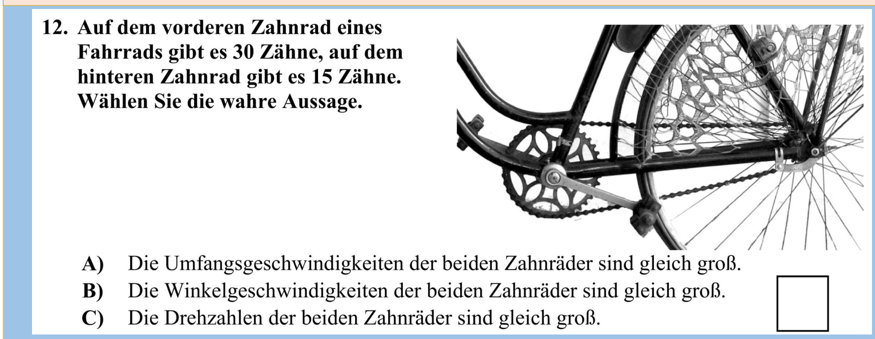
A) Si le champ magnétique de la Terre est beaucoup moins intense que le champ magnétique de la bobine le compas sera parallèle à l'axe de la bobine.
B) Si le champ magnétique de la Terre est beaucoup moins intense que le champ magnétique de la bobine sera perpendiculaire à l'axe de la bobine.
C) Le sens du compas sera indépendant du champ magnétique de la Terre, comme ce système est une cage de Faraday, il sera blindé dans tous les cas. ☐

GPT4V Answer: `"{'answer': 'B'}"`

Correct Answer: `"{'answer': 'A'}"`

Figure 14: A sample from French test set with GPT-4V output

Prompt: The image has a multiple-choice question with three options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice.



12. Auf dem vorderen Zahnrad eines Fahrrads gibt es 30 Zähne, auf dem hinteren Zahnrad gibt es 15 Zähne. Wählen Sie die wahre Aussage.

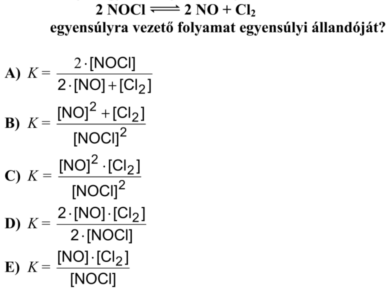
A) Die Umfangsgeschwindigkeiten der beiden Zahnräder sind gleich groß.
B) Die Winkelgeschwindigkeiten der beiden Zahnräder sind gleich groß.
C) Die Drehzahlen der beiden Zahnräder sind gleich groß. ☐

GPT4V Answer: `"{'answer': 'C'}"`

Correct Answer: `"{'answer': 'A'}"`

Figure 15: A sample from German test set with GPT-4V output

Prompt: The image has a multiple-choice question with five options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice, 'E' for the fifth choice.



5. Melyik egyenlet fejezi ki helyesen a
$2\,NOCl \rightleftharpoons 2\,NO + Cl_2$
egyensúlyra vezető folyamat egyensúlyi állandóját?

A) $K = \frac{2 \cdot [NOCl]}{2 \cdot [NO] + [Cl_2]}$

B) $K = \frac{[NO]^2 + [Cl_2]}{[NOCl]^2}$

C) $K = \frac{[NO]^2 \cdot [Cl_2]}{[NOCl]^2}$

D) $K = \frac{2 \cdot [NO] \cdot [Cl_2]}{2 \cdot [NOCl]}$

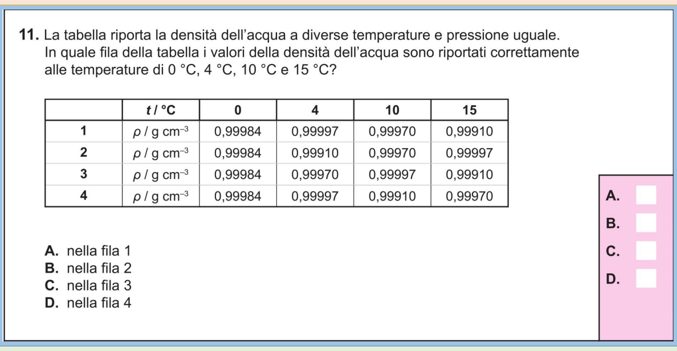
E) $K = \frac{[NO] \cdot [Cl_2]}{[NOCl]}$

GPT4V Answer: `"{'answer': 'C'}"`

Correct Answer: `"{'answer': 'C'}"`

Figure 16: A sample from Hungarian test set with GPT-4V output

Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.



11. La tabella riporta la densità dell'acqua a diverse temperature e pressione uguale. In quale fila della tabella i valori della densità dell'acqua sono riportati correttamente alle temperature di 0 °C, 4 °C, 10 °C e 15 °C?

	t / °C	0	4	10	15
1	ρ / g cm^{-3}	0,99984	0,99997	0,99970	0,99910
2	ρ / g cm^{-3}	0,99984	0,99910	0,99970	0,99997
3	ρ / g cm^{-3}	0,99984	0,99970	0,99997	0,99910
4	ρ / g cm^{-3}	0,99984	0,99997	0,99910	0,99970

A. nella fila 1
B. nella fila 2
C. nella fila 3
D. nella fila 4

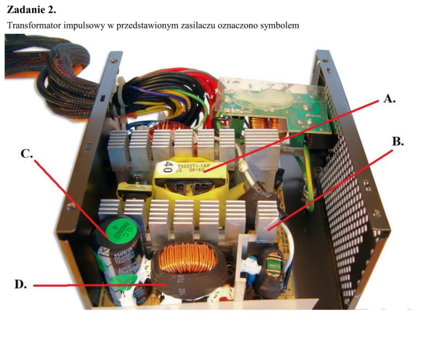
A. ☐
B. ☐
C. ☐
D. ☐

GPT4V Answer: `"{'answer': 'C'}"`

Correct Answer: `"{'answer': ‘A'}"`

Figure 17: A sample from Italian test set with GPT-4V output

Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.



Zadanie 2.

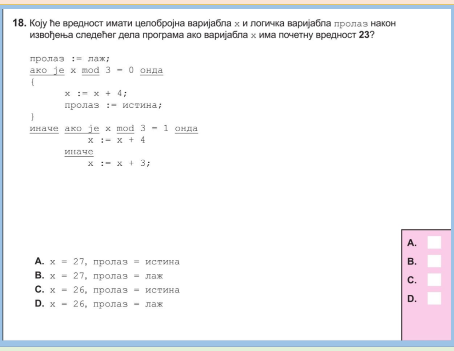
Transformator impulsowy w przedstawionym zasilaczu oznaczono symbolem

GPT4V Answer: `"{'answer': 'B'}"`

Correct Answer: `"{'answer': 'B'}"`

Figure 18: A sample from Polish test set with GPT-4V output

Prompt: The image has a multiple-choice question with four options. Provide the accurate response in JSON format as indicated: {'answer': 'xxx'}. Replace 'xxx' with the corresponding letter: 'A' for the first choice, 'B' for the second choice, 'C' for the third choice, 'D' for the fourth choice.



18. Коју ће вредност имати целобројна варијабла `x` и логичка варијабла `пролаз` након извођења следећег дела програма ако варијабла `x` има почетну вредност **23**?

```
пролаз := лаж;
ако је x mod 3 = 0 онда
{
      x := x + 4;
      пролаз := истина;
}
иначе ако је x mod 3 = 1 онда
          x := x + 4
      иначе
          x := x + 3;
```

A. x = 27, пролаз = истина
B. x = 27, пролаз = лаж
C. x = 26, пролаз = истина
D. x = 26, пролаз = лаж

A. ☐
B. ☐
C. ☐
D. ☐

GPT4V Answer: `"{'answer': 'B'}"`

Correct Answer: `"{'answer': 'D'}"`

Figure 19: A sample from Serbian test set with GPT-4V output