Open-VEQA: A Multilingual Multimodal OpenQA Benchmark for Educational Reasoning

Anonymous ACL submission

Abstract

Educational examination questions are a demanding evaluation setting for vision-language models because they require reading dense text, interpreting document-like layouts, understanding diagrams, graphs, and tables, and producing precise answers. We introduce Open-VEQA, a multilingual multimodal OpenQA benchmark constructed from real examination images. The dataset spans 2,406 questions across six languages, grades 4-12, 18 subjects, two difficulty levels, and both text-only and text-with-visual inputs. It includes a stratified test split for controlled evaluation across educational and visual factors, as well as 828 paired MCQ-to-OpenQA examples that isolate how model performance changes when answer choices are removed. Zero-shot experiments demonstrate that open-ended answer generation remains challenging, with large differences across difficulty, language, subject, question type, and visual element category, while supervised fine-tuning shows that task-specific adaptation can substantially improve open-weight models. Furthermore, our paired analysis exposes a large gap between multiple-choice selection and direct generation. We also assess the limitations of current open-ended evaluation metrics. Open-VEQA provides a focused benchmark for evaluating multilingual, visually grounded educational reasoning in realistic OpenQA settings. We release data and code at http://anonymous.for.review.

1 Introduction

Vision-language models have made strong progress on multimodal understanding tasks, including visual question answering, document understanding, chart reasoning, and problem solving (Singh et al., 2019; Masry et al., 2022; Lu et al., 2024a). Educational examination questions provide a realistic test case for these models because they incorporate basic visual elements characteristic commonly found in real-world documents and require reasoning beyond basic visual recognition. To answer them correctly, models may need to read dense text, understand structured layouts, interpret diagrams, graphs, and tables, reason over subject-specific knowledge, and produce precise answers.

Despite recent progress, multimodal examination evaluation remains limited as most of them rely on multiple-choice questions (MCQ), where models select an answer from a small set of candidates (Dimitrov et al., 2025; Yue et al., 2024a; Li et al., 2023). While this enables simple and reliable evaluation, it can overestimate model capabilities by allowing models to rely on visible answer choices rather than generate answers independently. In contrast, OpenQA requires models to produce answers directly without candidate options, placing greater emphasis on reasoning grounded in the problem formulation itself.

To further investigate this difference, we repurpose 828 MCQ questions into open-ended questions (described in Section 3.3) and evaluate model performance without answer choices. The results show a consistent performance decrease once the answer options are removed (Table 4). To address this gap, we introduce Open-VEQA, a multilingual multimodal benchmark for open-ended question answering (OpenQA) in educational settings. The dataset is constructed from real examination materials and preserves each question in its original visual form via screenshots or cropped images of the source documents, as shown in Figure 1. Each input is an image that may contain only text or combine text with visual elements such as diagrams, graphs, tables, and other illustrative figures.

A central goal of Open-VEQA is to support detailed evaluation across educational and visual di-

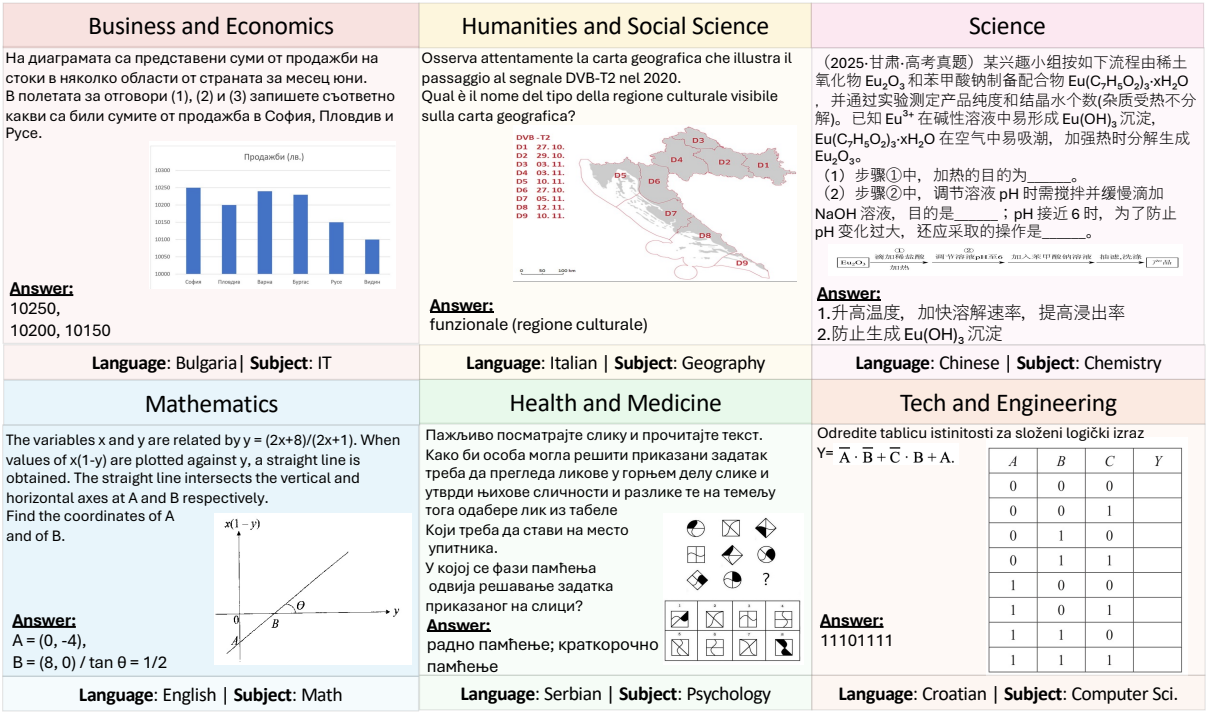


Figure 1: Real-world multilingual questions across diverse academic subjects.

mensions beyond MCQ. The benchmark covers 6 languages, grades from 4 to 12, 18 subjects, and two difficulty levels: easy and hard. To support reliable comparison across models, we construct stratified train and test splits that preserve the distribution of key attributes, including difficulty, language, and question type. This enables both overall evaluation and detailed subgroup analysis across languages, grades, subjects, difficulty levels, question types, and visual element categories.

We evaluate proprietary and open-weight vision-language models on OPEN-VEQA under a zero-shot setting. In addition to evaluating performance on the original OpenQA benchmark, we use the repurposed MCQ-to-OpenQA subset to analyze how answer choices affect model behavior and quantify the gap between answer selection and answer generation.

Our contributions are as follows:

- We develop an automatic data processing pipeline which enables easier scaling of data collection. Furthermore, we conduct manual verification of the automatic processing and analyze the error rate for each step. We release the pipeline code and data verification application code[1].
- We introduce OPEN-VEQA, a multilingual multimodal dataset consisting of 2,406 questions, covering 6 languages, 18 subjects, two difficulty levels, text-only and text-with-visual questions, and four visual element categories. We also release approximately *828* repurposed MCQ questions into OpenQA format, enabling a controlled analysis of how answer choices affect vision-language model performance.
- We evaluate open-weight and proprietary models under zero-shot and supervised fine-tuning settings, and complement the results with qualitative error analysis and an assessment of automatic metrics for semantic correctness.

2 Related Work

Visual and document-based OpenQA: Visual QA evaluates whether models can answer questions about images (Antol et al., 2015). Later benchmarks added readable text, document layouts, and structured visual content: TextVQA and OCR-VQA focus on text in images (Singh et al., 2019; Mishra et al., 2019); InfographicVQA studies document and infographic images (Mathew et al., 2022); and ChartQA and FlowVQA target charts and flowcharts (Masry et al., 2022; Singh et al., 2024). Free-form answer evaluation remains difficult: SimpleQA Verified studies

[1] anonymous.for.review

short-form open-ended QA for parametric factuality (Haas et al., 2025), QAEval combines answer extraction, rule-based matching, and a mixture-of-evaluators model for OpenQA scoring (Yue et al., 2025), and AutoConverter converts open-ended VQA questions into multiple-choice form with challenging distractors for reproducible VLM evaluation (Zhang et al., 2025). OPEN-VEQA addresses a complementary gap by evaluating direct answer generation from multilingual educational exam images, where models must handle document-like layouts, embedded text, and visual elements without candidate answers.

Educational and exam-based multimodal benchmarks. Educational and examination benchmarks evaluate subject knowledge, multilingual understanding, and visual reasoning. M3Exam introduced a multilingual, multimodal, and multilevel benchmark from real human exams (Zhang et al., 2023); EXAMS-V collected multilingual school-exam question images across 20 disciplines and 11 languages as unified snapshots with text, answer options, diagrams, tables, and other visual elements (Das et al., 2024; Dimitrov et al., 2025); and MULTI evaluates authentic Chinese examination questions across single-answer choice, multiple-answer choice, fill-in-the-blank, and open-ended writing formats (Zhu et al., 2025). Other benchmarks expand coverage through massively multilingual in-language exam evaluation, as in Kaleidoscope (Salazar et al., 2026), or target specific educational domains, such as Chinese K–12 multimodal mathematics in CMMaTH (Li et al., 2025) and Chinese multimodal scientific reasoning in MMSciBench (Ye et al., 2025). OPEN-VEQA extends this line of work with multilingual educational exam images in a direct OpenQA setting, stratified and difficulty-aware splits across key educational and visual factors, and a paired MCQ-to-OpenQA analysis that isolates how performance changes when answer choices are removed.

Multidisciplinary multimodal reasoning benchmarks: Multidisciplinary benchmarks evaluate broader reasoning across disciplines, skills, and visual contexts. MathVista targets visual mathematical reasoning over charts, plots, tables, geometry diagrams, scientific figures, and puzzle-style images (Lu et al., 2024b), while MMMU evaluates expert-level college-domain reasoning across six disciplines, 30 subjects, and diverse image types, including diagrams, tables, charts, maps, chemical structures, medical images, and sheet music (Yue et al., 2024b). More specialized benchmarks include MathVerse, which tests diagram use in visual math problems by varying textual and visual information (Zhang et al., 2024), and OlympiadBench, which covers bilingual Olympiad-level mathematics and physics problems with open-ended answers and step-by-step solution annotations (He et al., 2024). General-purpose suites such as MMBench and SEED-Bench assess broad multimodal capabilities through structured ability taxonomies and multiple-choice evaluation (Liu et al., 2024; Li et al., 2024), while MMIE evaluates interleaved multimodal comprehension and generation across multiple fields (Xia et al., 2025). OPEN-VEQA centers multilingual school-exam images and direct OpenQA, with breakdowns across language, subject, grade, difficulty, question type, and visual element category.

3 Dataset Construction

3.1 Data Collection and Processing Pipeline

We collected publicly available exam PDFs across six languages, mostly from high-school materials, with additional primary and middle-school Olympiad materials for Bulgarian due to limited availability of open-ended high-school questions. The Croatian, Italian, and Serbian subsets are parallel, containing the same Croatian national exam questions translated into the respective languages. Additional source details are provided in Appendix A.

We developed an automated processing pipeline, shown in Figure 2, which takes a PDF document and extracts: *(i)* a crop of each identified question, *(ii)* the corresponding answer, and *(iii)* question-specific metadata. The components run independently, enabling parallel execution and improving collection efficiency. Although our work focuses on open-ended questions, the pipeline supports other formats, including MCQ-style questions.

To reduce faulty outputs, we added a verification module with three steps: *(i)* filename verification, *(ii)* crop verification, and *(iii)* answer & metadata verification. We manually verified the outputs to assess the quality of the extraction modules. After verification, all extracted information is consolidated into a unified entry for each question.

Question Cropping: This module extracts individual question crops from each input PDF. We

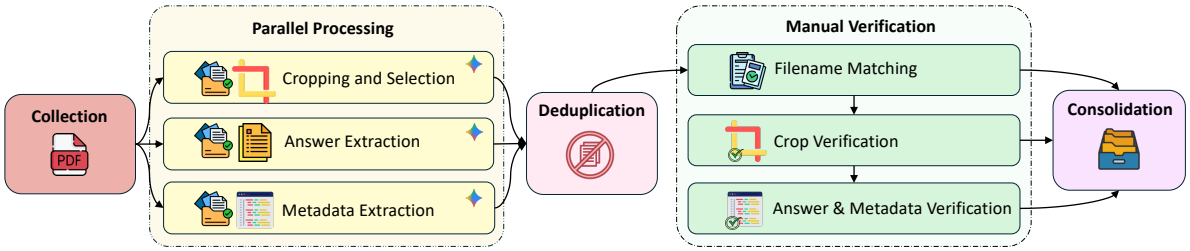


Figure 2: Overview of the OPEN-VEQA data collection and processing pipeline. The pipeline begins with examination material collection, followed by deduplication, parallel processing, question cropping and selection, answer and metadata extraction, manual verification, and final consolidation into the benchmark.

prompt a LVLM to identify bounding boxes for questions satisfying the selection criteria, convert the PDF into page images, and store each crop as a separate image. After testing several VLM systems, we used Gemini-3.1 Pro for this step. Additional details are provided in Appendix A.2.

Answer Extraction: This module extracts the answer corresponding to each cropped question. When answers were provided in separate files, we merged the exam and answer PDFs to improve consistency. To standardize outputs, answers were extracted in LaTeX format and restricted to final answers, excluding explanations and intermediate steps. We used Gemini 3.1-Flash for this step to reduce inference cost and processing time. Additional details are provided in Appendix A.3.

Metadata Extraction: This module identifies whether a question is text-only or contains visual elements, and assigns visual-element categories when applicable. We categorize visual elements as diagrams, graphs, tables, or other figures. We used Gemini 3.1-Flash for this step. Additional details are provided in Appendix A.4.

Deduplication: We removed exact, cropped, and near-duplicate question images using a three-stage deduplication pipeline. The pipeline retained the largest file in each duplicate group and removed 44 duplicates in total. Additional details are provided in Appendix A.5.

Verification: Three annotators, employed as part-time researchers and compensated in accordance with local labor regulations, manually verified filenames, crop boundaries, extracted answers, and metadata. This process enabled correction of crop boundaries, rejection of invalid questions, and validation of extracted information. Additional details on the verification interface and procedure are provided in Appendix A.6.

Data Consolidation: Finally, we consolidated all verified information into a unified dataset file containing the question crop, answer, metadata, and source-level information such as language, subject, and grade. We also filtered out rejected questions and entries labeled as explanations, drawings, or proofs.

3.2 Pipeline Quality Analysis

Table 1 summarizes the quality analysis of the processing pipeline. The analysis is based on manual verification of 509 processed PDFs, resulting in 2,406 open-ended questions. The PDFs averaged 18.68 pages and 6.3 extracted open-ended questions per document. Excluding initial experimentation, processing all 509 PDFs cost 44$.

Overall, the pipeline produced high-quality outputs. After performing human correction, only 4.23% of processed questions were rejected, and metadata errors were rare at 0.6%. Most corrections involved crop adjustments and answer formatting, especially LaTeX formatting in extracted answers. These results show that VLM-based processing can scale dataset construction, but is not yet reliable enough to remove verification entirely. Additional analysis is provided in Appendix A.7.

3.3 Dataset Statistics

As shown in Table 2, the finalized dataset consists of 2,406 open-ended questions in 6 languages across 18 subjects. Of these, 1,551 questions contain no visual elements, while 855 include visual elements in the question crop. The detailed distribution of visual element instances is provided in Appendix A.9; diagrams are the most common visual element.

Language	# PDFs	Rejected (%)	Crop Mod. (%)	Answer Mod. (%)	Metadata Err. (%)
Bulgarian	289	13.9	3.5	15.3	1.0
Chinese	60	0.3	26.8	100.0	0.0
Croatian	39	3.4	9.1	26.9	2.6
English	76	2.7	10.3	22.1	0.0
Italian	24	2.7	16.7	0.0	0.0
Serbian	21	2.4	0.6	0.0	0.0
Total / Avg	**509**	**4.23**	**11.16**	**27.38**	**0.6**

Table 1: Processing statistics by language; **Mod.**=Modification; **Err**=Error.

Question Difficulty: Many open-ended exam questions, especially from recent exams, contain multiple subquestions. In total, 1,421 questions, or 59.1% of the dataset, contain at least two subquestions, increasing the complexity of the benchmark and making it suitable for evaluating multi-step educational reasoning.

Subject Variety: The dataset covers a diverse range of subjects, with the largest portions from Mathematics, Physics, Chemistry, Geography, and Biology, as shown in Appendix A.10. Additional subjects include Informatics, History, Psychology, Sociology, Philosophy, Fine Arts, Music, and several social science and humanities domains. The distribution is uneven due to limited availability of standardized open-ended exam materials, but the dataset still captures a broad range of question styles across disciplines.

Repurposed MCQ: We also repurposed MCQ-style questions from EXAMS-V (Das et al., 2024) into open-ended questions by removing answer choices from the image and replacing the correct answer key with the corresponding answer text. This enables controlled analysis across question formats while keeping the underlying question content unchanged. The repurposed subset contains 828 questions, of which 528 contain visual elements and 300 do not. Detailed statistics are provided in Appendix A.8.

Language	# Subj.	w/ Visual	w/o Visual	Total
Bulgarian	11	221	410	631
Chinese	4	144	99	243
Croatian	13	147	293	440
English	5	164	374	538
Italian	10	94	194	288
Serbian	11	85	181	266
Total	**18**	**855**	**1,551**	**2,406**

Table 2: Number of questions with and without visual elements per language in the OPEN-VEQA dataset.

4 Experiments

We evaluate OPEN-VEQA in a zero-shot setting. Our experiments address two questions: (i) how well current vision-language models answer open-ended educational questions from images, and (ii) whether visible answer options make the task easier than direct answer generation, using a separate MCQ subset repurposed into OpenQA.

4.1 Experimental Setup

OpenQA task. For the OPEN-VEQA evaluation, each model receives an image of an exam question. The image may contain only text, or it may contain text together with visual elements such as diagrams, graphs, tables, or other figures. Models are instructed to inspect the image and return concise final answers in the same language as the question.

We split the 2,406 open questions into train and test sets using a 50/50 split. The split is stratified by language, difficulty, and question type. We use the test split for zero-shot evaluation. The train split is released to support future supervised experiments.

Informed OpenQA task: To evaluate the effect of answer choices, we conducted experiments using both the original and repurposed MCQs from Exams-V. This setup enables a comparison between answer selection (i.e., original MCQs) and answer generation (i.e., repurposed MCQs).

Supervised fine-tuning: We additionally fine-tune Qwen3.5-9B and Gemma-3-12B-IT on the OpenQA training split to measure the effect of task-specific adaptation. We use the same multimodal input-output format as evaluation, with the table image and question as input and the gold answer as the target response. The training set contains 1,203 examples, and the test set is kept disjoint from training. We use QLoRA for efficient adaptation and evaluate the fine-tuned models with the same decoding and scoring protocol as the zero-shot models. Implementation details are provided in Appendix B.1.

Models: We evaluate both open-weight and closed models. The open-weight models are InternVL3.5-14B (Wang et al., 2025), InternVL3.5-8B-Instruct (Wang et al., 2025), Qwen3.5-27B (Qwen Team, 2026), Qwen3.5-9B (Qwen Team, 2026), Gemma-3-12B-IT (Team et al., 2025), Gemma-3-27B-IT (Team et al., 2025), and Llama-3.2-11B-Vision-Instruct (Grattafiori et al., 2024). The closed models are Claude-Sonnet-4.6, GPT-5.4-Mini, and Gemini-3.1-Flash-Lite. All models are evaluated with greedy decoding where supported. We additionally use Claude-Haiku-4.5, GPT-5-Mini (Singh et al., 2026), and Gemini-3-Flash-Preview for difficulty tagging.

Evaluation measures: For OpenQA, we evaluate generated answers against the gold answers using COMET (Rei et al., 2020), METEOR (Banerjee and Lavie, 2005), BLEU-avg (Papineni et al., 2002), and ROUGE-L (Lin, 2004). To better understand how well COMET scores aligned with our expectations, we conducted a manual analysis and observed several inconsistencies in the resulting scores (discussed further in Section 5.4). To address these inconsistencies, we additionally report a combined Mean Score M, defined in Equation 1, where y denotes the gold answer and $\hat{y}$ the model prediction.

$$M(y, \hat{y}) = \tfrac{1}{2}\text{COMET}(y, \hat{y}) + \tfrac{1}{2}\text{METEOR}(y, \hat{y}) \quad (1)$$

We also report OpenQA Accuracy as the proportion of predictions with $M(y, \hat{y}) \geq 0.5$.

5 Results

5.1 Main OpenQA Results

Overall performance: Closed models achieve the strongest OpenQA performance, but the task remains difficult for all systems. Gemini-3.1-Flash-Lite performs best overall with 0.51, followed by Claude-Sonnet-4.6 with 0.48. Among open-weight models, Qwen3.5-27B performs best with 0.46, close to GPT-5.4-Mini at 0.45. The remaining open-weight models show a clear gap, with scores ranging from 0.25 to 0.43. These results show that open-ended educational QA from images remains challenging, even for the strongest evaluated models (Table 3).

Effect of visual element: The presence of visual element adds a clear source of difficulty.

Model	COMET	METEOR	Mean
Open Models			
InternVL3.5-14B	0.62	0.19	0.40
InternVL3.5-8B-IT	0.54	0.13	0.34
Qwen3.5-27B	**0.66**	**0.26**	**0.46**
Qwen3.5-9B	0.64	0.22	0.43
Gemma-3-12B-IT	0.52	0.10	0.31
Gemma-3-27B-IT	0.51	0.12	0.32
Llama-3.2-11B-Vision-IT	0.43	0.08	0.25
Closed Models			
Claude-Sonnet-4.6	0.66	0.31	0.48
GPT-5.4-Mini	0.66	0.25	0.45
Gemini-3.1-Flash-Lite	**0.69**	**0.34**	**0.51**
SFT			
Qwen3.5-9B	0.70	0.38	0.54
Gemma-3-12B-IT	0.75	0.39	**0.57**

Table 3: Zero-shot and SFT OpenQA results on the test set for open-weight and closed models. Mean is computed as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

Most models perform better on Text-only questions than on Text+Visual questions. For example, Gemini-3.1-Flash-Lite improves from 0.47 to 0.54 Mean Score, Claude-Sonnet-4.6 from 0.46 to 0.50, and Qwen3.5-27B from 0.44 to 0.48 (Table 14). The visual-element breakdown suggests that performance differs by visual type. Within this subset, the strongest models tend to score higher on table questions and lower on diagram or graph questions. For example, Gemini-3.1-Flash-Lite reaches 0.53 Accuracy on tables, compared with 0.36 on diagrams and 0.30 on graphs; Qwen3.5-27B shows a similar pattern, with 0.47 on tables, 0.28 on diagrams, and 0.31 on graphs (Table 12).

Difficulty and language: Question difficulty strongly affects performance. All models perform better on easy questions than on hard questions. Gemini-3.1-Flash-Lite drops from 0.69 Mean Score on easy questions to 0.41 on hard questions, followed by Claude-Sonnet-4.6 , and Qwen3.5-27B drops from 0.62 to 0.37 (Table 15). Performance also varies by language, although the strongest models are relatively stable. Gemini-3.1-Flash-Lite obtains the best closed-model Mean Scores across most languages, ranging from 0.49 on Bulgarian to 0.55 on Croatian, while Qwen3.5-27B is the strongest open-weight model across most languages, with Mean Scores between 0.45 and 0.48 (Table 16). A subject-level inspection shows variation across disciplines. On this subset, the strongest models score higher on subjects such as mathematics, informatics, history, and politics/economics. For example, Gemini-3.1-Flash-Lite obtains 0.52 on Mathematics, 0.63 on Informatics,

0.65 on History and Civilization, and 0.75 on Politics and Economics. In contrast, chemistry and physics are lower, with 0.37 on Chemistry and 0.41 on Physics.

Effect of supervised fine-tuning: We additionally evaluate supervised fine-tuned variants of Qwen3.5-9B and Gemma-3-12B-IT to measure whether task-specific adaptation improves OpenQA performance. SFT is compared against the corresponding zero-shot base model under the same evaluation protocol. Qwen3.5-9B improves from 0.64 to 0.70 COMET, 0.22 to 0.38 METEOR, and 0.43 to 0.54 Mean, while Gemma-3-12B-IT improves more substantially from 0.52 to 0.75 COMET, 0.10 to 0.39 METEOR, and 0.31 to 0.57 Mean. These results show that SFT provides clear gains for both models, with especially large improvements for Gemma-3-12B-IT, suggesting that task-specific adaptation improves both semantic answer quality and lexical overlap.

5.2 Informed OpenQA Task Results

Answer options substantially improve performance: Across all four models, performance is much higher when answer options are visible in the MCQ image than when the same questions are repurposed as OpenQA. Gemini-3.1-Flash-Lite obtains the best MCQ Mean of 0.80, followed closely by Qwen3.5-27B at 0.79. In the repurposed OpenQA setting, the same models drop to 0.47 and 0.45, respectively (Table 4). This large gap shows that visible answer choices make the task substantially easier than direct answer generation.

Language and question type: The MCQ advantage is consistent across languages and question types. Across languages, Gemini-3.1-Flash-Lite remains strong in MCQ, reaching up to 0.88 on Croatian, while Qwen3.5-27B reaches up to 0.86 on Italian; in the repurposed OpenQA setting, their best language scores drop to 0.54 and 0.52, respectively (Tables 17 and 19). A similar pattern holds for question type: Gemini-3.1-Flash-Lite scores 0.84 on Text-only MCQ but 0.49 on Text-only OpenQA, while Qwen3.5-27B scores 0.80 on Text+Visual MCQ but 0.44 on Text+Visual OpenQA (Tables 18 and 20). Notably, Qwen3.5-27B slightly outperforms Gemini-3.1-Flash-Lite on Text+Visual MCQ questions, with 0.80 versus 0.78, although Gemini remains stronger overall on MCQ and repurposed OpenQA. These breakdowns show that the gap is driven mainly by the answer format rather than by a single language or question type.

5.3 Error analysis

We conducted a thorough error analysis and observed examples reveal a recurring failure mode: models often produce plausible answers without executing the reasoning required by the specific instance as shown in Figure 3a. In the chart example, both models identify the salient country, China, but fail when the question requires domain knowledge or regional interpretation.

The same pattern appears in the calculus example in Figure 3a: models return simple scalar answers while failing to track the shaded regions, integration bounds, and curve-line relationships. These are not merely formatting errors. They show that the evaluated VLMs can often recognize the genre of an educational task, but remain brittle when connecting visual extraction, symbolic reasoning, and numerical verification.

A more concerning pattern is that the models often fail in structurally similar ways, as shown in Figure 3b. When different systems produce nearby wrong answers or follow the same flawed solution template, their errors should not be treated as independent noise. Instead, they suggest convergence on shared shortcuts: extracting salient numbers, applying a familiar formula, and producing an exam-like answer without verifying the specific visual instance.

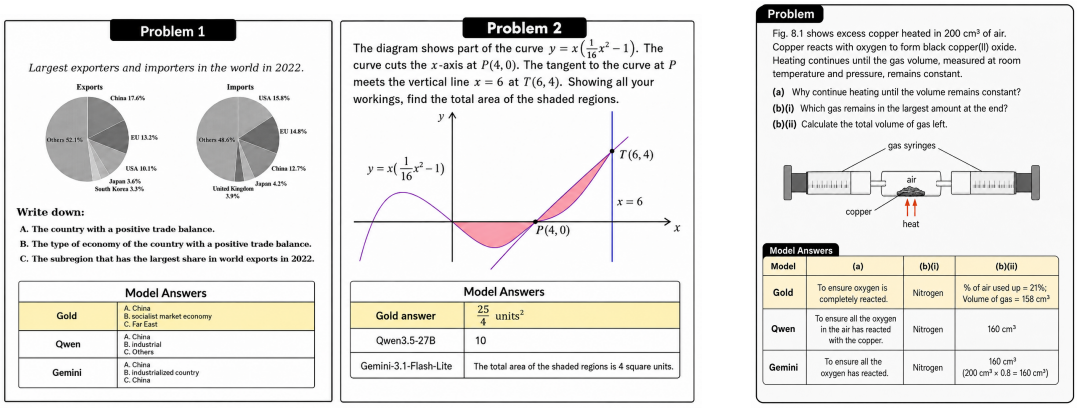
This has direct implications for OpenQA evaluation. Agreement among models may create the appearance of reliability, but in these cases it can reflect correlated failure driven by the same brittle heuristics.

5.4 Analyzing Metric Inconsistencies

Our error analysis further revealed that COMET does not consistently reflect semantic correctness in OpenQA. In particular, the metric can penalize answers that are correct but differ only in surface form. For example, COMET assigns a score of 0.37 when the gold answer is $2400\ m^3$; $1800\ m^3$, and the prediction is *2400; 1800*, even though the required numerical values are preserved. At the same time, it can assign relatively higher scores to predictions that are semantically incorrect. In another example with gold answer *132* m^2, the VLM produces *120* m^2 instead of a meaningful answer, yet the metric assigns this prediction a score of 0.78. Together, these examples suggest

	MCQ			OpenQA		
Model	**COMET**	**METEOR**	**Mean**	**COMET**	**METEOR**	**Mean**
Open Models						
Qwen3.5-27B	**0.92**	**0.67**	**0.79**	**0.68**	**0.22**	**0.45**
Gemma-3-27B-IT	0.79	0.38	0.59	0.61	0.12	0.37
Closed Models						
Gemini-3.1-Flash-Lite	**0.92**	**0.68**	**0.80**	**0.69**	**0.25**	**0.47**
GPT-5.4-Mini	0.84	0.49	0.66	0.66	0.19	0.42

Table 4: Comparison of MCQ and repurposed OpenQA results for open-weight and closed models. MCQ uses the original question images with visible answer options, while OpenQA removes answer options and requires direct answer generation. Mean is computed as $0.5 \times \text{COMET} + 0.5 \times \text{METEOR}$.



(a) Models produce plausible but incorrect answers when visual interpretation and task-specific reasoning are required.

(b) Different models converge on the same semantically wrong answer.

Figure 3: Examples from the qualitative error analysis. Models often recognize the broad structure of the educational task, but fail to connect visual extraction, symbolic reasoning, and numerical verification.

that COMET is sensitive to superficial formatting differences while failing to reliably detect omissions that affect answer interpretation. Therefore, it should not be used on its own to evaluate semantic correctness in open-ended QA.

Similarly, METEOR, as a lexical-overlap metric, is also likely to fail in cases where the prediction is an alternative correct answer, a valid paraphrase, or a partially correct response that does not share sufficient surface overlap with the reference.

6 Conclusion and Future Work

We introduce Open-VEQA, a multilingual multimodal OpenQA benchmark built from real examination materials, with 2,406 questions across six languages, 18 subjects, grades 4–12, two difficulty levels, and both text-only and visually grounded inputs. We release stratified train/test splits, a scalable data processing pipeline, and 828 paired MCQ-to-OpenQA examples for analyzing answer-choice effects. Our experiments show that open-ended multimodal QA remains challenging for current VLMs, with performance differences across difficulty levels and between text-only and text-with-visual questions. SFT improves open-weight models, suggesting that the released training split can support useful task-specific adaptation.

The paired MCQ-to-OpenQA analysis further shows that visible answer choices can change measured model behavior. Error analysis suggests that models often recognize the structure of educational tasks, but remain brittle when combining visual understanding, symbolic reasoning, and numerical verification. Automatic evaluation also remains a key challenge: even semantic metrics can assign low scores to correct or equivalent answers when formatting, numeric precision, or answer variants differ. Future work should develop QA-specific evaluation methods that combine semantic matching with syntactic or numerical verification, and extend OpenQA benchmarks toward longer responses such as explanations and proofs.

Limitations

Data processing pipeline. Our pipeline analysis focuses on the manually verified construction process used to create OPEN-VEQA. This allows us to estimate extraction quality and identify recurring errors, while leaving fully automated verification for future work. The implementation prioritizes correctness and auditability over runtime optimization, but its modular design supports further parallelization. The pipeline currently uses proprietary VLMs for stages where reliable bounding box detection was critical. To control cost, we use cheaper models for simpler extraction tasks and report the total construction cost.

Limited benchmark size. While the benchmark contains 2,406 real examination questions, scaling the dataset uniformly across all languages remains challenging due to the limited availability of publicly accessible examination materials, particularly those containing open-ended questions. In many educational systems, standardized assessments are dominated by MCQ-style formats, resulting in comparatively fewer high-quality open-ended examples suitable for collection. One possible direction for future expansion is increasing the number of covered languages and educational systems, which could improve both dataset scale and diversity.

Model selection. Although we aimed to conduct a comprehensive experimental evaluation, we did not test every possible model. In particular, we focused primarily on widely available open-weight and proprietary models.

Ethics and Broader Impact

Copyright and Licensing: All data included in OPEN-VEQA was collected from publicly accessible educational sources, primarily standardized examination materials released by educational institutions or publicly available preparation resources. The dataset is intended strictly for non-commercial research purposes in multimodal reasoning and educational AI. We preserve only the information necessary for benchmarking and research reproducibility and release metadata describing the original collection sources whenever possible.

Ethics and Data Privacy: The dataset does not contain personal or sensitive information, as all collected materials originate from examination documents rather than user-generated data. No student records, identities, or personally identifiable information are included in the benchmark. Additionally, the questions represent educational content already intended for broad public distribution and academic use.

Data Annotation. We employed three annotators as part-time researchers, compensated in accordance with local labor regulations, to conduct manual verification of the data processing pipeline quality. The annotators were informed about the intended research use of the collected data and the goals of the benchmark. Since the dataset was constructed exclusively from publicly available examination materials and does not contain personal or sensitive information, no additional consent procedures were required.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Dimitar Iliyanov Dimitrov, Mingshan Hee, Zhuohan Xie, Rocktim Jyoti Das, Momina Ahsan, Sarfraz Ahmad, Nikolay Paev, Ivan Kolev Koychev, and Preslav Ivanov Nakov. 2025. Overview of imageclef 2025–multimodal reasoning.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. 2025. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *CoRR*, abs/2509.07968.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850. Association for Computational Linguistics.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Zhongzhi Li, Ming-Liang Zhang, Pei-Jie Wang, Jian Xu, Rui-Song Zhang, Yin Fei, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Jiaxin Zhang, and Cheng-Lin Liu. 2025. CMMaTH: A Chinese multi-modal math skill evaluation benchmark for foundation models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2690–2726, Abu Dhabi, UAE. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. 2024. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, Lecture Notes in Computer Science, pages 216–233. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Findings of ACL, pages 2263–2279. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ronald Rivest. 1992. The md5 message-digest algorithm. Technical report.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, and 26 others. 2026. Kaleidoscope: In-language exams for massively multilingual vision evaluation. In *The Fourteenth International Conference on Learning Representations*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry,

Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. 2024. Flowvqa: Mapping multimodal logic in visual question answering with flowcharts. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 1330–1350. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. 2025. MMIE: massive multimodal interleaved comprehension benchmark for large vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xinwu Ye, Chengfan Li, Siming Chen, Wei Wei, and Robert Tang. 2025. MMSciBench: Benchmarking language models on Chinese multimodal scientific problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14621–14663, Vienna, Austria. Association for Computational Linguistics.

Tan Yue, Rui Mao, Xuzhao Shi, Shuo Zhan, Zuhao Yang, and Dongyan Zhao. 2025. QAEval: Mixture of evaluators for question-answering task evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14717–14730, Vienna, Austria. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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, and 3 others. 2024b. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, page 169–186, Berlin, Heidelberg. Springer-Verlag.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. 2025. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 29580–29590. Computer Vision Foundation / IEEE.

Zichen Zhu, Yang Xu, Lu Chen, Jingkai Yang, Yichuan Ma, Yiming Sun, Hailin Wen, Jiaqi Liu, Jinyu Cai, Yingzi Ma, Situo Zhang, Zihan Zhao, Liangtai Sun, and Kai Yu. 2025. MULTI: Multimodal understanding leaderboard with text and im-

ages. *SCIENCE CHINA Information Sciences*, 68(10):200107.1–200107.26.

A Additional Dataset Construction Details

A.1 Data Sources

The dataset was collected from publicly available exam materials ranging from the primary to high-school levels, with the majority of questions targeting the high-school level. Due to a limited availability of open-ended questions in Bulgarian high-school exams, we supplemented this subset with primary and middle-school Olympiad materials. For Chinese, we collected materials from the Gaokao examinations. English data was sourced from Singaporean primary and secondary school test papers. Finally, the Croatian, Italian, and Serbian data were all collected from Croatian national high-school exams. The Italian and Serbian subsets are parallel to the Croatian data, consisting of the exact same exam questions translated into their respective languages.

A.2 Question Cropping

This module is responsible for extracting crops of individual exam questions from a given input PDF file. First, by prompting a LVLM model to identify the bounding box of each question satisfying the predefined selection criteria. The input PDF is then converted into a sequence of page images, after which each identified question is cropped and stored as a separate image file using the OpenCV Python library.

We evaluated several recent VLM systems for the task and concluded that Gemini-3.1 Pro was the most accurate model at the time. The full prompt is shown in Figure 4.

A.3 Answer Extraction

The answer extraction module is responsible for locating and extracting the answer corresponding to each question. In cases where the answers were not included in the original PDF document and instead provided in separate file(s), we merged the exam PDF with the corresponding answer PDF(s), as extraction proved more consistent when performed on a unified document.

To ensure consistency, we introduced several extraction rules, with the most important being: all answers must be returned in LaTeX format, only the final correct answers should be extracted, excluding explanations or intermediate steps. The full prompt with extraction rules is provided in Figure 5.

For this task, we used Gemini 3.1-Flash, as it extracted answers effectively while significantly reducing both inference cost and processing time.

A.4 Metadata Extraction

The metadata extraction module is responsible for identifying whether the question contains visual elements and, if so, which categories of visual elements are present in the cropped question. We define four visual-element categories:

- **Diagram**: schematic or illustrative drawings representing physical systems, setups, objects, or spatial relationships.
- **Graph**: visuals representing quantitative relationships.
- **Table**: structured data presented in rows and columns.
- **Other/Figure**: visual elements that do not fit the above categories.

For this module, we again used Gemini 3.1-Flash, as the task was comparatively simpler and no errors were observed during testing. The full prompts and extraction rules are provided in Figure 6.

A.5 Deduplication

We applied a three-stage image deduplication pipeline to remove redundant examination question images. First, exact duplicates were identified using MD5 hashes (Rivest, 1992), which detect byte-identical files. Second, cropped duplicates were detected using OpenCV template matching with the structural similarity index measure (SSIM) (Wang et al., 2004), allowing the system to find images that appear as sub-regions of larger images. Third, near-duplicate images were detected using perceptual hashing (pHash) (Zauner, 2010).

For each duplicate group, the largest file was retained as the representative image, under the assumption that it preserves the highest visual quality. This approach removes exact copies, cropped variants, and near-duplicate scans while keeping one high-quality version of each unique question. In total, this pipeline flagged and removed 44 duplicates, where the questions are either exact copies or cropped copies.

A.6 Verification and Consolidation Details

Question IDs were constructed in the format <question_number>_<paper_number>. For example, question_31_paper_2.png.

Verification Interface. To improve verification efficiency, we developed a lightweight Python application that served as a dedicated *verification interface*. The application was implemented in Python using the Streamlit[2] framework to provide an interactive interface.

Filename Verification. In the first step, we verify whether filenames produced by the three parallel extraction modules match and modify them if necessary. Filenames generated by the question cropping module were treated as the reference, as Gemini 3.1-Pro proved highly consistent in this task. Additional information and an example of the interface is shown in Figure 7.

Crop Verification. After filename verification, we manually inspected each detected bounding box to ensure that the question and all associated visual elements were correctly enclosed. Instead of discarding incorrect crops, we provided functionality for modifying bounding boxes in order to preserve as much data as possible. We also included a rejection option. For rejections, we provided a free-text field for specifying the rejection reason, primarily used in cases where the identified question was not genuinely open-ended and included answer choices within the question text. An example of the interface is shown in Figure 8.

Answer and Metadata Verification. After confirming the question boundaries, we verified: *(i)* whether the extracted metadata was correct and *(ii)* the quality of the extracted answers. The interface provided an editable text field preloaded with the extracted answer together with a LaTeX preview component to help validate formatting correctness as shown in Figure 9

The task focused exclusively on extracting final answers rather than full explanations or intermediate reasoning steps. Consequently, when an answer could not be extracted within one or two sentences, we labeled it as EXPLANATION and handled it separately during data consolidation. The same procedure was applied to proofs and drawings.

[2] https://streamlit.io/

Final Consolidation. During consolidation, we filtered out rejected questions as well as entries labeled as explanations, drawings, or proofs in order to maintain a consistent answer format across languages and subjects. In total, this filtering removed 246 drawing-based answers, 231 explanation-based answers, and 206 proof-based answers from the final dataset.

A.7 Pipeline Quality Analysis Details

Rejection Rates. We observe comparatively high rejection rates for Bulgarian. Upon further inspection, approximately 50% of the rejected questions originated from Olympiad materials that were not truly open-ended, while the remaining cases primarily involved overly complex answers. Overall, the relatively low rejection rate of 4.23% indicates that only a small portion of the processed questions had to be discarded.

Crop Modifications. Crop modifications were necessary across all languages. Most corrections were caused by missing question numbers in the crop or overly large bounding boxes, particularly when questions appeared at the end of a page and additional whitespace was included. Chinese required substantially more corrections, with 26.8% of questions needing crop adjustments. We attribute that to the structure of many Chinese exam formats, which often contain minimal spacing between consecutive questions, causing bounding boxes to occasionally include parts of neighboring questions.

Answer Modifications. We observe a relatively high answer modification rate of 27.38%, primarily caused by recurring LaTeX formatting issues produced by Gemini 3.1-Flash. The most common problems included incorrectly formatted \frac expressions, typically due to missing backslash symbols or unbalanced dollar delimiters. In some cases, answers containing dollar symbols caused the model to incorrectly interpret subsequent text as LaTeX content. We also observed occasional extraction errors involving superscripts and subscripts rendered in very small fonts. For Chinese, every extracted answer required manual modification, primarily due to incorrect LaTeX formatting, with a large number of \frac expressions containing compilation errors.

Metadata Errors. Metadata extraction errors were extremely rare, indicating that Gemini 3.1-

Flash reliably detected the presence and category of visual elements within question crops. We could not identify a consistent cause for the observed errors, although we speculate that they were likely caused by occasional model hallucinations.

A.8 Repurposed MCQ-to-OpenQA Statistics

In addition to the collected open-ended questions, we repurposed MCQ-style questions from EXAMS-V (Das et al., 2024) into open-ended questions by removing the answer choices from the question image and replacing the correct answer key with the corresponding answer text. This repurposing allows us to analyze differences in model behavior across question formats while keeping the underlying question content unchanged. Table 5 6 summarizes the language-wise statistics of the repurposed subset, including the number of samples with and without visual content.

Language	w/ Visual	w/o Visual	Total
Bulgarian	84	50	134
Chinese	139	50	189
Croatian	96	50	146
English	66	50	116
Italian	87	50	137
Serbian	56	50	106
Total	**528**	**300**	**828**

Table 5: Statistics of the MCQ subset repurposed into OpenQA.

Language	Diagram	Graph	Table	Other
Bulgarian	35	20	21	8
Chinese	51	30	30	28
Croatian	35	20	20	21
English	51	9	2	4
Italian	35	20	13	19
Serbian	34	8	6	8
Total	**241**	**107**	**92**	**88**

Table 6: Visual element counts for the OpenQA versus MCQ task per language.

A.9 Visual Element Distribution

Table 7 reports the distribution of visual element instances in the Open-VEQA dataset. Diagrams are the most frequent visual element, followed by graphs, other figures, and tables. The total number of visual element instances is larger than the number of questions with visual content because a single question may contain more than one visual element.

A.10 Subject distribution

Figure 10 summarizes the distribution of questions across academic subjects. Mathematics forms the largest share of the benchmark, with 979 questions, followed by Physics with 368 questions, Chemistry with 237 questions, Geography with 205 questions, and Biology with 192 questions. Together, these five subjects account for the majority of the dataset, reflecting the benchmark's emphasis on quantitative, scientific, and visual reasoning tasks. The remaining subjects contribute smaller but diverse portions, covering Informatics, Politics and Economics, History, Combined Science, Psychology, Latin, Sociology, Entrepreneurship, Philosophy, Fine Arts, Music, Religious Education, and Ethics.

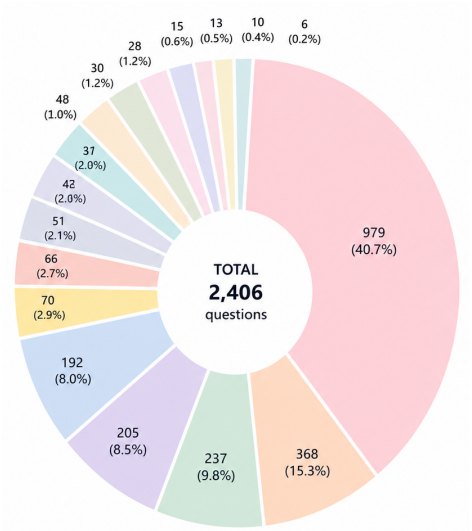


Figure 10: Subject distribution of the Open-VEQA benchmark. The chart shows the number and relative share of questions for each academic subject, with Mathematics forming the largest category and the remaining subjects providing broad coverage across science, humanities, social science, arts, and ethics.

B Additional Experiment Details

This section includes additional experiment details of our work.

B.1 Supervised Fine-Tuning Details

We fine-tune Qwen3.5-9B and Gemma-3-12B-IT on the OpenQA training split using QLoRA. Each training example is formatted as a chat-style multimodal instance containing the table image and question in the user message and the gold answer in the assistant message. The training set contains 1,203 examples, while the validation and test sets are kept disjoint from training.

Visual Element	Bulgarian	Chinese	Croatian	English	Italian	Serbian	Total
Diagram	118	124	54	108	42	35	481
Graph	28	42	30	36	22	18	176
Tables	36	14	29	24	18	17	138
Other/Figure	45	8	47	8	22	26	156
Total	**227**	**188**	**160**	**176**	**104**	**96**	**951**

Table 7: Distribution of visual element instances per language in the OPEN-VEQA dataset. A question may contain more than one visual element.

Hyperparameter	Value
Training examples	1,203
Fine-tuning method	QLoRA
Quantization	4-bit NF4
Double quantization	Yes
Compute dtype	bfloat16
Base model weights	Frozen
Trainable parameters	LoRA adapters
Loss masking	Assistant answer tokens only
Trainer	HuggingFace Trainer
Gradient checkpointing	Yes
Decoding at evaluation	Greedy decoding
Metrics	COMET, METEOR, Mean

Table 8: Supervised fine-tuning setup. Mean is computed as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

We load each base model with 4-bit NF4 quantization, double quantization, and bfloat16 computation. We train LoRA adapters while keeping the base model weights frozen. During training, the loss is computed only over the assistant answer tokens; prompt tokens corresponding to the system/user messages and image-conditioned input are masked from the loss.

Training is implemented with the HuggingFace Trainer. We use gradient checkpointing and save LoRA adapters separately for each base model. At evaluation time, the fine-tuned models are evaluated with the same input format, decoding setup, normalization, and COMET/METEOR scoring protocol as the zero-shot models.

The main hyperparameters are shown in Table 8.

C Additional Results

The additional results provide complementary breakdowns of model behavior across visual element type, difficulty, question modality, language, answer format, and lexical-overlap metrics. The visual-element analysis shows that tables are generally easier than diagrams and graphs, while Qwen3.5-27B and Gemini-3.1-Flash-Lite are the strongest open-weight and closed models, respectively, in most categories (Table 12). The difficulty-tagging models show that Gemini-3-Flash-Preview gives the strongest overall performance among the models used for difficulty estimation (Table 13). The zero-shot breakdowns show consistent drops on Text+Visual questions compared with Text-only questions, a large gap between Easy and Hard items, and relatively stable but language-dependent performance across the six languages (Tables 14, 15, and 16). The lexical-overlap zero-shot results report BLEU, ROUGE-L, and accuracy separately, supporting the same ranking trends without relying only on COMET, METEOR, or their mean score (Table 9). The repurposed OpenQA results show similar trends, with Gemini-3.1-Flash-Lite and Qwen3.5-27B remaining the strongest closed and open models across language and modality splits (Tables 17 and 18). The corresponding OpenQA lexical-overlap table further reports BLEU, ROUGE-L, and thresholded accuracy for this subset (Table 11). Finally, the MCQ results are substantially higher than the open-ended OpenQA results, but still vary by language and modality, with Gemini-3.1-Flash-Lite and Qwen3.5-27B again giving the best overall performance (Tables 19, 20, and 10). These auxiliary metrics help verify that the observed trends are not specific to a single evaluation formulation. This adds a metric-level robustness check.

Model	BLEU	ROUGE-L	Accuracy
Open Models			
InternVL3.5-14B	0.13	0.28	0.26
InternVL3.5-8B-IT	0.09	0.20	0.17
Qwen3.5-27B	**0.18**	**0.39**	**0.36**
Qwen3.5-9B	0.15	0.32	0.29
Gemma-3-12B-IT	0.06	0.16	0.11
Gemma-3-27B-IT	0.08	0.18	0.15
Llama-3.2-11B-Vision-IT	0.03	0.08	0.02
Closed Models			
Claude-Sonnet-4.6	0.19	0.41	0.40
GPT-5.4-Mini	0.16	0.34	0.33
Gemini-3.1-Flash-Lite	**0.23**	**0.44**	**0.46**

Table 9: Zero-shot OpenQA results on the test set using lexical-overlap and thresholded accuracy metrics. Accuracy denotes the proportion of predictions with Mean Score ≥ 0.5.

Model	Key	Answer	BLEU	ROUGE-L
Open-Models MCQ				
Qwen3.5-27B	**0.79**	**0.66**	**0.75**	**0.54**
Gemma-3-27B-IT	0.31	0.24	0.39	0.36
Closed-Models MCQ				
Gemini-3.1-Flash-Lite	**0.78**	**0.69**	**0.77**	**0.55**
GPT-5.4-Mini	0.50	0.40	0.52	0.42

Table 10: MCQ results using key accuracy, answer accuracy, and lexical-overlap metrics. Key Acc. measures prediction of the correct option key, while Answer Acc. measures prediction of the corresponding answer text.

Model	BLEU	ROUGE	Acc
Open-Models OpenQA			
Qwen3.5-27B	**0.15**	**0.23**	**0.32**
Gemma-3-27B-IT	0.09	0.15	0.18
Closed-Models OpenQA			
Gemini-3.1-Flash-Lite	**0.17**	**0.26**	**0.35**
GPT-5.4-Mini	0.13	0.19	0.26

Table 11: OpenQA results using lexical-overlap and thresholded accuracy metrics. Accuracy denotes the proportion of predictions with Mean Score ≥ 0.5.

Model	Diagram	Graph	Table	Other
Open-Models-Zeroshot				
InternVL3.5-14B	0.23	0.22	0.29	0.12
InternVL3.5-8B-Inst	0.16	0.15	0.22	0.09
Qwen3.5-27B	**0.28**	**0.31**	**0.47**	**0.41**
Qwen3.5-9B	0.17	0.16	0.38	0.30
Gemma-3-12B-IT	0.05	0.08	0.22	0.22
Gemma-3-27B-IT	0.08	0.11	0.22	0.23
Llama-3.2-11B-Vision-IT	0.02	0.04	0.04	0.00
Closed-Models-Zeroshot				
Claude-Sonnet-4.6	0.30	0.22	**0.56**	0.49
GPT-5.4-Mini	0.26	0.28	0.31	0.39
Gemini-3.1-Flash-Lite	**0.36**	**0.30**	0.53	**0.49**

Table 12: Accuracy breakdown by single visual-element type on the test set for open-weight and closed zero-shot models.

Model	COMET	METEOR	Mean
Difficulty Estimation Models			
Claude-Haiku-4.5	0.58	0.17	0.37
GPT-5-Mini	0.66	0.33	0.49
Gemini-3-Flash-Preview	**0.73**	**0.43**	**0.58**

Table 13: Zero-shot results for the models used in difficulty estimation, evaluated on the full set 2406 samples.

Task
You are given a pdf file - a test in English, Bulgarian or Croatian or another language. Do object detection for each question to find its ymin, xmin, ymax, xmax.
A question is defined as:
- All the text belonging to a question.
- All the possible choice answers (if any).
- Any diagrams, images, or figures that are part of the question.
- Any tables or charts that are part of the question.
- Any mathematical expressions or equations that are part of the question.
- Any special formatting (like bold or italics) that is part of the question.
- Absolutely ignore extration of multiple choice questions (MCQs). Only extract open-ended questions.

Validation Criteria:
- Ensure that the bounding boxes fully encompass the entire question, including all text, images, diagrams, tables, and any other relevant content.
- Ensure the bounding box encompasses multipart question properly. Multipart questions can be present in a given pdf as 1 A), 2 C) or 36.1, 36.2 or 6 i), 6 ii), 6 a) ii), etc.
- Multipart questions should only be extracted as one whole question, not as separate question for each part.
- Is the question fully formed? Are all choice answers, diagrams, table, charts, etc included? Is it answerable on its own?
- The bounding boxes should not include content from other questions or unrelated material on the page.
- Never extract MCQs (Multiple-choice questions)

Output Format:
Return ONLY a raw JSON list of objects. Do not wrap it in markdown code blocks (like ```json).
Use only numbers for page indexing, based on the pdf page.
Question_id - the actual question_id as it appears in the PDF.
- If there is more than one exam paper in the pdf, append '_paper_1', '_paper_2', etc.
- If there is only one exam paper in the pdf still append '_paper_1' in the end of the question_id.
- If the same paper contains the same question number, e.g., emumeration resets, consider the new enumeration as a new paper - _paper_2.

For bounding boxes, use the box_2d format: [ymin, xmin, ymax, xmax], with coordinates normalized between 0 and 1000.
Ignore questions that are spanning on more than one page. I am only interested in questions fully contained on a single page.
Example output:

```
[
    {"question_id": "1_paper_1", "page": 1, "box_2d": [112, 20, 244, 890]},
    {"question_id": "2_paper_1", "page": 3, "box_2d": [250, 20, 380, 890]},
    {"question_id": "3_paper_2", "page": 2, "box_2d": [386, 20, 516, 890]},
    {"question_id": "4_paper_2", "page": 1, "box_2d": [522, 20, 652, 890]},
    {"question_id": "5_paper_3", "page": 1, "box_2d": [658, 20, 788, 890]},
    {"question_id": "6_paper_3", "page": 1, "box_2d": [794, 20, 924, 890]}
]
```

Figure 4: Crop (bounding box) extraction prompt.

	Text+Visual			Text-only		
Model	**COMET**	**METEOR**	**Mean**	**COMET**	**METEOR**	**Mean**
Open-Models-Zeroshot						
InternVL3.5-14B	0.59	0.17	0.38	0.63	0.21	0.42
InternVL3.5-8B-Inst	0.54	0.13	0.34	0.54	0.13	0.33
Qwen3.5-27B	**0.65**	**0.24**	**0.44**	**0.67**	**0.28**	**0.48**
Qwen3.5-9B	0.61	0.19	0.40	0.65	0.24	0.44
Gemma-3-12B-IT	0.51	0.11	0.31	0.52	0.10	0.31
Gemma-3-27B-IT	0.52	0.12	0.32	0.51	0.13	0.32
Llama-3.2-11B-Vision-IT	0.41	0.06	0.24	0.44	0.08	0.26
Closed-Models-Zeroshot						
Claude-Sonnet-4.6	0.64	0.28	0.46	0.67	0.33	0.50
GPT-5.4-Mini	0.63	0.23	0.43	0.67	0.26	0.46
Gemini-3.1-Flash-Lite	**0.66**	**0.29**	**0.47**	**0.71**	**0.37**	**0.54**

Table 14: Performance breakdown by question type on the test set for open-weight and closed zero-shot models. Text-only contains only textual content, while Text+Visual includes diagrams, graphs, tables, or other figures. Mean is computed as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

```
### Task
You are given a PDF file - a test in English, Bulgarian, Croatian, or another language.
I need you to extract the correct answers for each question from the answer section/table that is
    typically provided at the end of the PDF or in a dedicated answers section.

### Important Instructions:
- The answers are usually within the same PDF file. If two PDF files are given, the first one is the
    test and the second one is the answers.
- READ the answers from the PDF document, do NOT generate or guess answers
- Extract only the final correct answers, not explanations or working steps
- Answers that require drawings or proofs or explanations only if the whole question part requires it.
    In those cases return $Drawing$ , $Proof$, $Explanation$, etc. for the whole question or just the
     part that requires it.
- Any text should be kept in the original language.
- Question_id - the actual question_id as it appears in the PDF.
    - If there is more than one exam paper in the pdf, append '_paper_1', '_paper_2', etc.
    - Multipart questions should only be present once in the json, not for each part.
    - If the same paper contains the same question number, e.g., emumeration resets, consider the
        new enumeration as a new paper - _paper_2.

### Answer Format Requirements:
- Use only numbers for question id, based on the numeric value of the question. If there is more than
    one exam paper in the pdf, append '_paper_1', '_paper_2', etc.
- The output format should be latex, unless it is pure text or just a number.
- All formulas should be in latex format - mathematical, chemical, biological, physics, etc. Escape "\"
    -> "\\" in latex expressions to ensure valid JSON.
- In case of Chinese PDFs make sure to escape latex style properly!
- Provide the page where the answer was found in the 'answer_page' field.

For Open/Text Questions:
- Convert mathematical expressions to LaTeX format when necessary (e.g., fractions, formulas).
- Convert chemical formulas to LaTeX when appropriate
- If a single question has multiple answers but is not Multi-part, return them in a single list element,e.
    g., "answers": ["$\\frac{3}{4}$, $\\sqrt{x + 2}$"] or "answers": ["A, B"] or "answers": ["$\\frac{
    e^{2x}(3-2x)}{(1-x)^2}$, Always decreasing"]

For Multi-part Questions (e.g., 37A, 37B, 2.1, 2.2, 2 A), 3 B), i), ii), etc):
- Return as list of elements: ["A", "B"] or ["$\\frac{3}{4}$", "$\\sqrt{x + 2}$"]

For Numerical Questions:
- Provide exact numerical values
- Include units if specified in the answer key

### Output Format:
Return ONLY a raw JSON object. Do not wrap it in markdown code blocks (like ```json).

[
    {"question_id": 1_paper_1, "answer_page": 12, "answers" : ["$\\frac{3}{4}$"]},
    {"question_id": 1_paper_2, "answer_page": 5, "answers" : ["57 cm", "5", "$\\sqrt{x + 2}$"]},
    {"question_id": 1_paper_2, "answer_page": 10, "answers" : ["oxygen", "molecule"]},
]
```

Figure 5: Answer extraction prompt.

Task
You are given a PDF file - a test in English, Bulgarian, Croatian, or another language.
For each question in the PDF, I need you to analyze and extract metadata about the question type and visual elements.

Information Required for Each Question:
1. Type of question:
   - "text" - only textual information
   - "image" - has visual elements
2. Question_id - string of the actual question_id as it appears in the PDF. Also append '_mcq' in the end if the question is a multiple-choice one. If there is more than one exam paper in the pdf, append '_paper_1', '_paper_2', etc.
   - If there is only one exam paper in the pdf still append '_paper_1' in the end of the question_id.
   - Multipart questions should only be present once in the json, not for each part, e.g., if a question 7 has parts A, B, C or i), ii), etc. it should only be present once in the json.
   - If the same paper contains the same question number, e.g., emumeration resets, consider the new enumeration as a new paper - _paper_2.
3. Visual Elements (1 or 0 for each category):
   - Diagram: Schematic or illustrative drawings used to represent physical systems, setups, objects, or spatial relationships
   - Graph: Visuals that represent quantitative relationships
   - Tables: Structured data presented in rows and columns
   - Other/Figure: Other visual elements that cannot be classified with the above categories

Information Required for the Entire PDF File:
1. Date of the exam (if present in the PDF, 'Unknown' if not)
2. Language of the exam
3. Grade level (if present in the PDF, 'Unknown' if not)
4. Source file name (provide the filename without extension)
5. Subject of the exam (if present in the PDF, 'Unknown' if not)

Output Format:
Return ONLY a raw JSON object. Do not wrap it in markdown code blocks (like ```json).

Figure 6: Metadata extraction prompt.

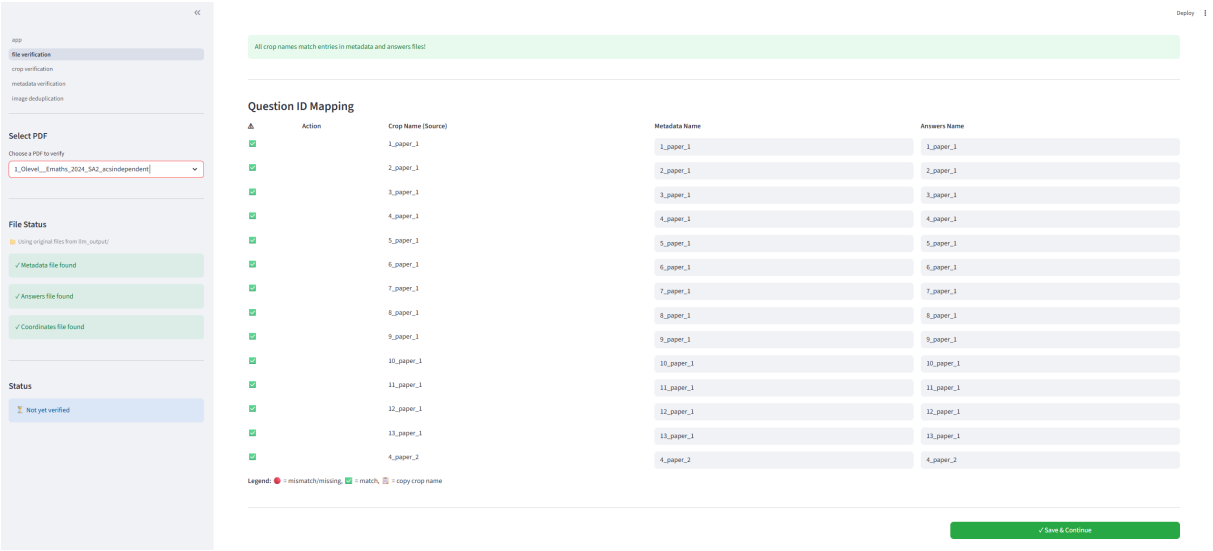


Figure 7: Example of the filename verification user interface.

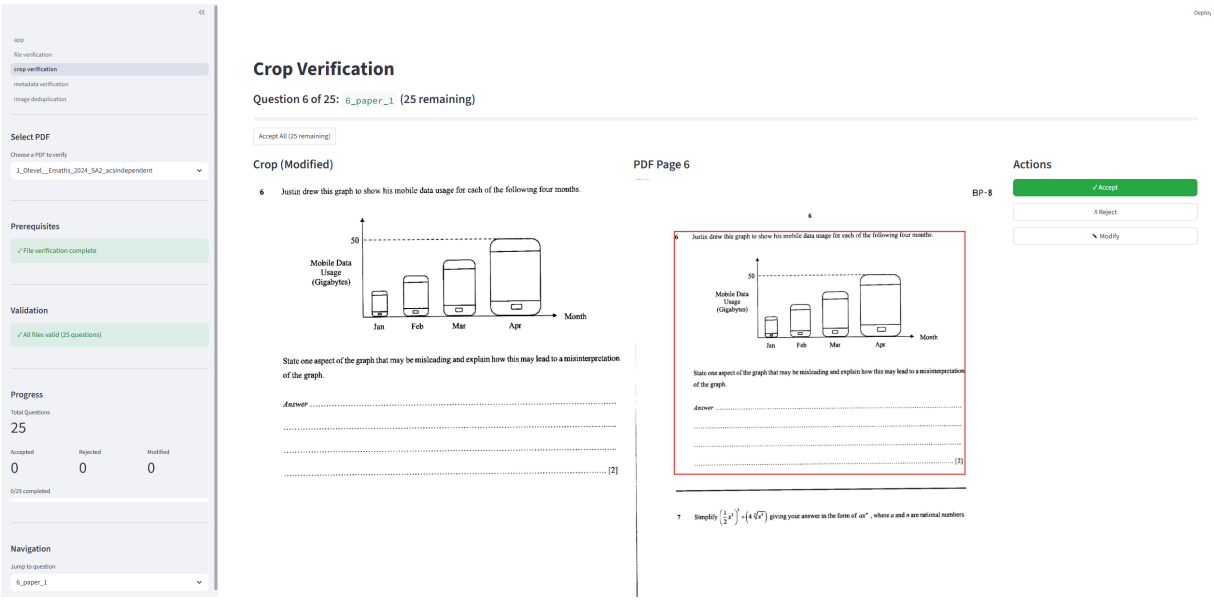


Figure 8: Example of the crop verification user interface.

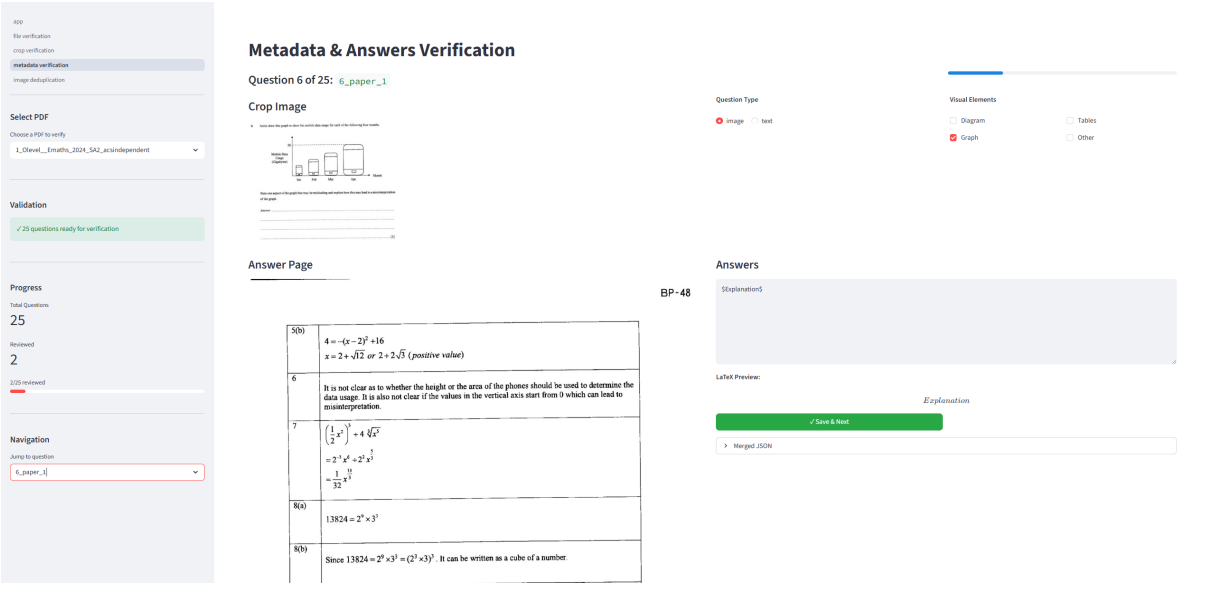


Figure 9: Example of the crop verification user interface.

	Easy			Hard		
Model	**COMET**	**METEOR**	**Mean**	**COMET**	**METEOR**	**Mean**
Open-Models-Zeroshot						
InternVL3.5-14B	0.74	0.30	0.52	0.54	0.13	0.33
InternVL3.5-8B-Inst	0.65	0.21	0.43	0.47	0.08	0.28
Qwen3.5-27B	**0.81**	**0.43**	**0.62**	**0.57**	**0.16**	**0.37**
Qwen3.5-9B	0.79	0.36	0.58	0.54	0.13	0.34
Gemma-3-12B-IT	0.63	0.17	0.40	0.45	0.06	0.25
Gemma-3-27B-IT	0.63	0.22	0.43	0.44	0.06	0.25
Llama-3.2-11B-Vision-IT	0.45	0.09	0.27	0.41	0.07	0.24
Closed-Models-Zeroshot						
Claude-Sonnet-4.6	0.79	0.49	0.64	0.58	0.20	0.39
GPT-5.4-Mini	0.79	0.40	0.59	0.57	0.16	0.37
Gemini-3.1-Flash-Lite	**0.84**	**0.53**	**0.69**	**0.60**	**0.22**	**0.41**

Table 15: Performance breakdown by difficulty on the test set for open-weight and closed zero-shot models. Easy and Hard indicate the assigned difficulty labels. Mean is computed as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

Model	**Bulgarian**	**Chinese**	**Croatian**	**English**	**Italian**	**Serbian**
Open-Models-Zeroshot						
InternVL3.5-14B	0.34	0.46	0.38	**0.49**	0.43	0.36
InternVL3.5-8B-Inst	0.26	0.42	0.35	0.36	0.38	0.31
Qwen3.5-27B	**0.45**	**0.46**	**0.48**	0.47	**0.47**	**0.45**
Qwen3.5-9B	0.40	0.43	0.44	0.45	0.43	0.42
Gemma-3-12B-IT	0.30	0.28	0.32	0.32	0.30	0.31
Gemma-3-27B-IT	0.31	0.28	0.34	0.32	0.34	0.32
Llama-3.2-11B-Vision-IT	0.22	0.23	0.25	0.31	0.24	0.24
Closed-Models-Zeroshot						
Claude-Sonnet-4.6	0.48	0.42	0.52	0.48	0.49	**0.51**
GPT-5.4-Mini	0.44	0.41	0.45	0.48	0.46	0.45
Gemini-3.1-Flash-Lite	**0.49**	**0.50**	**0.55**	**0.53**	**0.52**	**0.51**

Table 16: Performance breakdown by language on the test set for open-weight and closed zero-shot models. Values report Mean, computed as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

Model	**Bulgarian**	**Chinese**	**Croatian**	**English**	**Italian**	**Serbian**
Open-Models OpenQA						
Qwen3.5-27B	**0.50**	**0.36**	**0.45**	**0.52**	**0.48**	**0.41**
Gemma-3-27B-IT	0.41	0.27	0.38	0.45	0.38	0.34
Closed-Models OpenQA						
Gemini-3.1-Flash-Lite	**0.54**	**0.35**	**0.49**	**0.54**	**0.50**	**0.47**
GPT-5.4-Mini	0.52	0.33	0.42	0.47	0.44	0.39

Table 17: Performance breakdown by language for the repurposed OpenQA subset. Values report Mean, computed as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

	Text+Visual			Text-only		
Model	**COMET**	**METEOR**	**Mean**	**COMET**	**METEOR**	**Mean**
Open-Models OpenQA						
Qwen3.5-27B	**0.67**	**0.21**	**0.44**	**0.69**	**0.22**	**0.46**
Gemma-3-27B-IT	0.60	0.12	0.36	0.63	0.13	0.38
Closed-Models OpenQA						
Gemini-3.1-Flash-Lite	**0.68**	**0.23**	**0.46**	**0.71**	**0.27**	**0.49**
GPT-5.4-Mini	0.66	0.19	0.43	0.65	0.18	0.41

Table 18: Performance breakdown by question type for the repurposed OpenQA subset. Text-only contains only textual content, while Text+Visual includes visual elements. Mean is computed as $0.5 \times$COMET$+0.5 \times$METEOR.

Model	Bulgarian	Chinese	Croatian	English	Italian	Serbian
Open-Models MCQ						
Qwen3.5-27B	0.84	0.73	0.84	0.72	**0.86**	0.77
Gemma-3-27B-IT	0.66	0.41	0.64	0.63	0.70	0.55
Closed-Models MCQ						
Gemini-3.1-Flash-Lite	**0.85**	**0.68**	**0.88**	**0.76**	**0.87**	**0.83**
GPT-5.4-Mini	0.73	0.48	0.76	0.57	0.80	0.68

Table 19: MCQ performance breakdown by language. Values report Mean, computed on the full answer prediction as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.

	Text+Visual			Text-only		
Model	COMET	METEOR	Mean	COMET	METEOR	Mean
Open-Models MCQ						
Qwen3.5-27B	**0.92**	**0.67**	**0.80**	**0.91**	**0.66**	**0.78**
Gemma-3-27B-IT	0.78	0.35	0.56	0.82	0.44	0.63
Closed-Models MCQ						
Gemini-3.1-Flash-Lite	**0.91**	**0.66**	**0.78**	**0.94**	**0.73**	**0.84**
GPT-5.4-Mini	0.83	0.46	0.64	0.86	0.54	0.70

Table 20: MCQ performance breakdown by question type. Text-only contains only textual content, while Text+Visual includes visual elements. Mean is computed on the full answer prediction as $0.5 \times$ COMET $+ 0.5 \times$ METEOR.