# Can Vision-Language Models Pass the PISA Test? A Multilingual Multimodal Benchmark from Global Student Assessments

**Anonymous ACL submission**

## Abstract

Vision-language models have demonstrated remarkable progress in multimodal reasoning. However, existing benchmarks remain limited in terms of high-quality, multilingual examples because *(i)* many benchmarks rely on synthetically generated content by large language models (LLMs) which may not reflect the complexity of real human assessments, and *(ii)* are restricted to English, limiting multilingual evaluation. To fill this gap, we introduce PISABENCH, a multilingual benchmark derived from the Programme for International Student Assessment (PISA), a standardized international test used to evaluate student competencies across more than eighty countries. We collect, preprocess and translate these expert-curated questions such that our final dataset contains 732 multimodal questions covering six languages. Finally, we evaluate 19 state-of-the-art vision-language models (both closed- and open-source) from eight model families and find that only four achieve an average test accuracy above 60%. We support our empirical findings with additional ablations, including contamination and error type analysis. We release PISABENCH under Apache 2.0 and the evaluation framework to the community.[1]

## 1 Introduction

Recent large language models (LLMs) and vision-language models (VLMs) have demonstrated remarkable abilities on tasks such as commonsense question answering (Sun et al., 2025; Chen et al., 2024; Toroghi et al., 2024), mathematical reasoning (Wang et al., 2026; Parashar et al., 2026; Dong et al., 2025) or PhD-level science questions (Rein et al., 2023). These capabilities are typically assessed using benchmark datasets designed to evaluate such capabilities.

However, current multimodal benchmarks exhibit several important shortcomings: *(i)* Many benchmarks rely on synthetically generated content from LLMs rather than human-authored tests, limiting the quality and reliability of the evaluation (Maheshwari et al., 2024; Gill et al., 2025); *(ii)* most benchmarks are predominantly in English, limiting the ability to identify language-specific biases in model reasoning (Dai et al., 2026); and *(iii)*, many benchmarks fail to cover and measure vision-specific skills such as spatial or geometric reasoning as commonly assessed in educational settings (Bean et al., 2025).

**The PISA Framework.** To address these systemic gaps, we identify a unique opportunity in the Programme for International Student Assessment (PISA). Developed by the OECD[2], PISA is a standardized framework designed to measure the functional literacy and problem-solving competencies of 15-year-old students globally (OECD, 2025). Unlike most vision-language benchmarks, PISA tasks are grounded in established psychometric principles and undergo rigorous expert review to ensure construct validity.

Crucially, these assessments are intrinsically multimodal; solving them requires the simultaneous integration of textual instructions with complex visual stimuli such as charts, maps, and geometric diagrams. Furthermore, because PISA is intended for international comparison across more than eighty countries, the source material is engineered to be linguistically and culturally invariant. This provides a robust foundation for evaluating the multilingual capabilities of VLMs without the confounding effects of translation artifacts or English-centric bias.

This multimodal complexity is illustrated in Figure 1, which depicts a spatial reasoning task: a student must deduce the total number of dots on a pile of dice from a shifted perspective, a task requiring precise visuospatial integration.

[1] Removed for double-blind review.

[2] https://www.oecd.org/pisa/

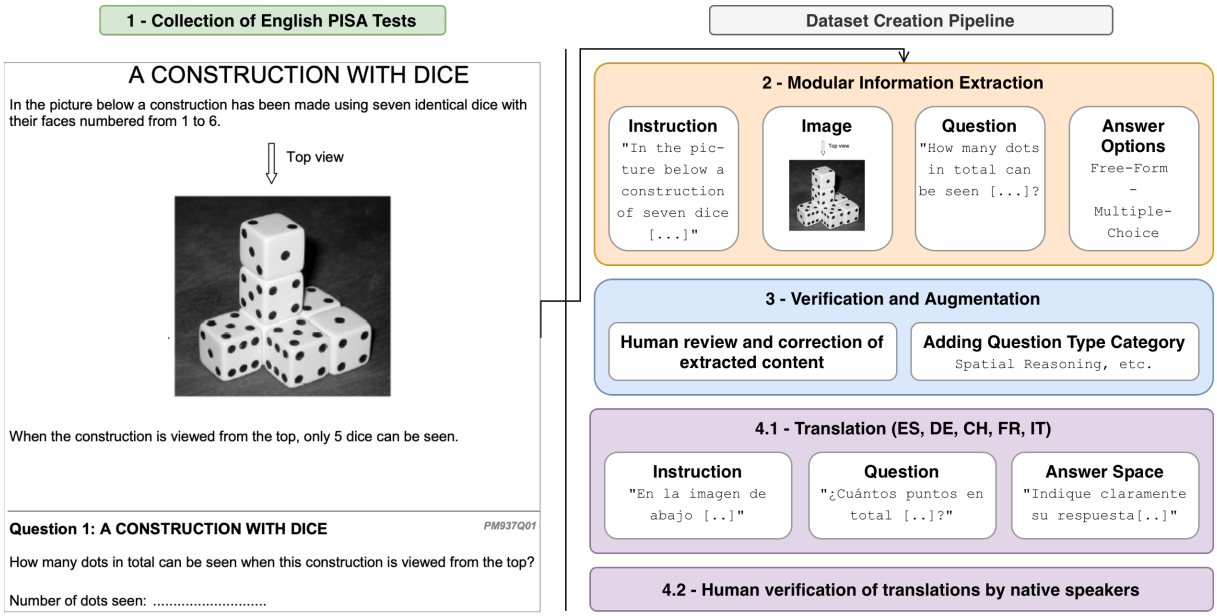



Figure 1: Overview of the dataset construction pipeline. We *(i)* collect tasks from the OECD PISA tests, *(ii)* decompose them into modular components (instruction, image, question, and answer options), *(iii)* verify, augment, and, if necessary, correct the extracted content, and *(iv)* translate them into five target languages (ES, DE, CH, FR, IT) and verify translations with native speakers.

**The PISABENCH Benchmark.** To leverage these structural advantages, we introduce PISABENCH, a curated multimodal benchmark systematically derived from the PISA assessment archives. While PISA provides the raw material, adapting these assessments for VLM evaluation requires rigorous filtering and restructuring to ensure suitability for automated inference. To this end, we extract and modularize instructions, questions, and visual assets, enriching each instance with fine-grained metadata and question-type categorizations.

To ensure the integrity of the benchmark, we manually verify that each example is self-contained and solvable without exogenous knowledge, discarding any items that rely on external context. Furthermore, to facilitate multilingual evaluation, we extend the initial English source questions into five additional languages (Spanish, German, Chinese, French, and Italian). The resulting dataset comprises 732 high-quality test questions (122 unique per language), providing a standardized measure of applied multimodal reasoning across diverse linguistic contexts.

We evaluate 19 state-of-the-art vision-language models from eight model families on PISABENCH and find that only four models achieve an average test accuracy above 60% across all languages. Furthermore, we observe that performance degrades in smaller models (<20 billion parameters) and on non-English splits. Given that the underlying PISA source material has been publicly accessible for several years, we include ablations to test for data contamination—investigating model performance in the absence of visual input—and provide a comparative analysis against established benchmarks such as MMLU (Hendrycks et al., 2021) and MMMU-Pro (Yue et al., 2025).

**Contributions.** We summarize our contributions as follows:

1. We create and release PISABENCH, a multilingual, parallel benchmark of 732 high-quality examples covering six languages, derived from the PISA tests, including question type categorization.

2. We conduct extensive evaluation of state-of-the-art vision-language models, revealing significant discrepancies across languages and question types.

3. We open-source our evaluation framework, enabling future research to easily replicate and evaluate vision language models using PISABENCH.

## 2 Dataset Construction

We construct PISABENCH through a four-stage pipeline illustrated in Figure 1: We first collect

the source corpus from English PISA tests and then extract information in a modular format. This is followed by a quality assurance phase involving human verification and, finally, a multilingual translation stage.

### 2.1 Stage 1: Collection of PISA Tests

We initially obtained assessment data from the 2009 and 2012 PISA cycles. These materials are publicly released by the OECD for research purposes, and we utilized the official documents to derive our initial English corpus.

To select suitable items, human annotators applied three selection criteria: *(i) completeness*, ensuring each example includes all necessary instructions and components; *(ii) clarity*, excluding items with ambiguous problem statements; and *(iii) multimodality*, requiring that the visual component is essential to the task logic. This process yielded a core set of 122 English assessment items.

### 2.2 Stage 2: Modular Information Extraction

We standardize the corpus by converting each item into a structured, modular format. As illustrated in Figure 1 (cf. 2 - Modular Information Extraction), annotators manually extracted the following fields:

- **Instruction**: The contextual description introducing the overall task.
- **Image**: The corresponding visual stimulus, such as figures or geometric diagrams.
- **Question**: The primary problem statement.
- **Answer Options**: The response candidates, categorized as free-form or multiple-choice.
- **Gold Answer**: The ground-truth solution.

Note that PISA often utilizes "unit-based" designs where multiple questions refer to a shared scenario (e.g., “Given your previous answer, now answer the following [..]”). To ensure each item is independent, we decoupled these sequences. Where a question referred to previous answers, we prepended the necessary context and instructions to make the item fully self-contained.

**Multiple Choice vs. Free-Form Answer.** PISA questions utilize both free-form and multiple-choice formats. To facilitate reproducible, automated evaluation via log-likelihood-based ranking (Gao et al., 2024; Ploner et al., 2025), we used GPT-4o to generate plausible distractors for free-form questions and rephrased prompts to a consistent multiple-choice template.

**Type Categorization.** To support fine-grained error analysis, we labeled each question with a reasoning category using GPT-4o. Each task was assigned to one of four domains: *(i)* spatial and geometric reasoning, *(ii)* quantitative reasoning, *(iii)* graph and pattern analysis, or *(iv)* text and diagram understanding. These labels were subsequently validated by human annotators during the quality control stage.

### 2.3 Stage 3: Quality Control

Following the modular extraction, two independent human annotators reviewed all materials to ensure high data fidelity. The verification process focused on confirming that each question is: *(i)* solvable solely through the provided image and instruction; *(ii)* faithful to the original content and intent; and *(iii)* supported by non-trivial answer options. If an item failed these criteria, it was manually corrected or regenerated until it passed two independent verification steps.

### 2.4 Stage 4: Translation into Target Languages

In the final stage, we translate the quality assured dataset into five additional languages: Spanish, German, French, Italian, and Chinese using GPT-4o. This results in a six-way parallel multilingual, multimodal benchmark. In the following sections, we analyze the translation quality based on evaluations by native speakers. The translation prompt is shown in Appendix B.

## 3 Translation Validation

As described in Section 2.4, we translated the English source material into five languages. We validate translation quality using three complementary approaches: human review, WMT23 COMET-KIWI (Rei et al., 2023) (reference-based), and GEMBA-MQM (Kocmi and Federmann, 2023) using GPT-4o (reference-free).

### 3.1 Human Validation

Native speakers manually review a random subset of 50 translated items including instruction, questions, and answer options. Each reviewer evaluates *(i)* completeness, ensuring no content is omitted, and *(ii)* correctness, ensuring meaning is preserved. Errors are classified as critical, major, or minor

| | Error-Free (%) | | Critical (%) | | Major (%) | | Minor (%) | |
|---|---|---|---|---|---|---|---|---|
| Language | GPT-4o | Human | GPT-4o | Human | GPT-4o | Human | GPT-4o | Human |
| Chinese | 66.3 | 86.0 | 17.2 | 0.0 | 3.3 | 0.0 | 10.7 | 12.0 |
| German | 64.6 | 86.0 | 15.5 | 0.0 | 8.2 | 4.0 | 9.0 | 8.0 |
| French | 71.7 | 88.0 | 12.3 | 0.0 | 6.6 | 0.0 | 7.4 | 6.0 |
| Italian | 75.2 | 82.0 | 12.3 | 0.0 | 2.5 | 0.0 | 8.2 | 18.0 |
| Spanish | 77.0 | 76.0 | 8.2 | 0.0 | 4.1 | 6.0 | 9.0 | 18.0 |

Table 1: Comparison of translation validation results: We report error-free translations, critical, major and minor mistakes (all in %) between GEMBA-MQM (GPT-4o) and human validation on 50 samples.

issues. We present the results in Table 1. All translations achieve an error-free rate of at least 76% with no critical issues across any language. Spanish has the highest fraction of major errors at 6%, and minor issues range between 6–18%. Importantly, no critical translation errors are found, indicating that core semantic content is faithfully preserved.

### 3.2 Automatic Validation

To complement the human validation, we evaluate all translations using neural metrics. COMET-KIWI is a multilingual transformer trained to predict translation quality scores between 0 and 1, comparing machine-translated output against a reference. GEMBA-MQM uses GPT-4o to detect translation errors without reference translations, classifying them into the same three categories as our human validation. We reuse the hyperparameters and few-shot examples provided by the authors (Kocmi and Federmann, 2023).

#### 3.2.1 WMT23 COMET-KIWI Metric

We present the COMET-KIWI scores in Figure 2. Boxes are consistently above 0.7, with Italian and Spanish achieving the highest medians (above 0.8), followed by Chinese, German, and French. Some outliers fall below 0.5, notably French with scores down to 0.3. Overall, translation quality achieves a sufficient level across all languages.

#### 3.2.2 GEMBA-MQM Metric

We show the GEMBA-MQM results in Table 1. Error-free rates range from 66.3% to 76.9%, with critical errors between 8.19% (Spanish) and 17.21% (Chinese), major errors between 2.46% (Italian) and 8.19% (German), and minor errors between 7.38% (French) and 10.66% (Chinese). Italian and Spanish show the highest error-free rates. The authors of GEMBA-MQM note that the metric should be used with caution in academic comparisons due to its reliance on proprietary, black-box models.

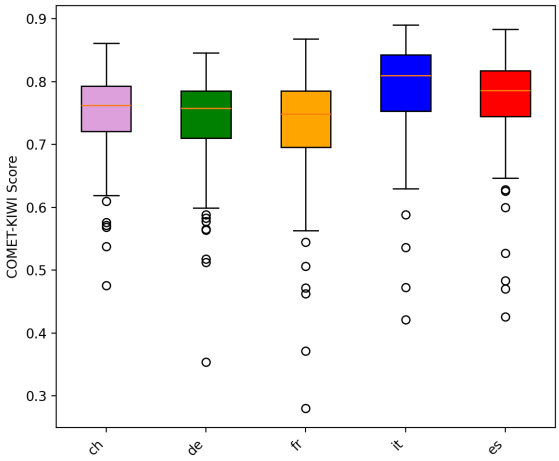



Figure 2: Distribution of WMT23 COMET-KIWI scores for each target language.

#### 3.2.3 Reconciling Human and Automatic Evaluation

A notable discrepancy emerges between the two evaluations: human annotators identify no critical errors, whereas GEMBA-MQM flags rates up to 17.21%. After manually reviewing the flagged items, we find that these largely correspond to stylistic deviations—such as lexical choices or syntactic reorderings—rather than semantic distortions or omissions. Human annotators, who assess translations holistically with domain context, consistently judge these items as meaning-preserving. This is consistent with prior observations that LLM-based metrics penalize surface-level variation more harshly than human judges (Kocmi and Federmann, 2023). We therefore interpret the GEMBA-MQM results as a conservative lower bound on translation quality.

In summary: *(i)* COMET-KIWI confirms strong semantic alignment (medians above 0.7); *(ii)* hu-

man validation confirms the absence of critical errors; and *(iii)* while GEMBA-MQM yields more conservative estimates, manual inspection attributes the gap to stylistic sensitivity rather than meaning-altering errors. Collectively, these results provide sufficient evidence that translation quality supports reliable cross-lingual evaluation.

# 4 Evaluation Results

## 4.1 Experimental Setup

In this section, we evaluate open-weight vision-language models (VLMs) on PISABENCH to assess the difficulty of our benchmark. We first evaluate on the original English split to establish a baseline performance. We then continue with evaluating all translated versions of our benchmark. Lastly, we conduct an error analysis as a supporting ablation. We also performed a contamination analysis, which is shown in Table 4.

**Evaluation Protocol.** For each example in our dataset, we provide the model with the instruction, question text, and the associated image. We apply the LLM-as-a-judge protocol with GPT-4o as the evaluator; thus, all models are tasked to generate a free-form textual answer, which is compared to the extracted gold reference.

**Models.** We consider the following models for our experiments: Qwen2.5-VL (3B – 72B) (Bai et al., 2025), Qwen3-VL (Team, 2025), Gemma-3 (4B – 27B) (Team et al., 2025), LLaVA (7B – 34B) (Liu et al., 2023), and Idefics3 (8B) (Laurençon et al., 2024) to cover different architectural concepts, pretraining objectives and model sizes. We further include GPT-4o and Claude-3.5-Haiku (Anthropic, 2025) as proprietary models.

## 4.2 Results

Table 2 summarizes the results on PISABENCH. The best performing model across languages is GPT-4o, achieving, e.g., 71.0 accuracy on English, 67.2 accuracy on Chinese, or 69.7 on French. The other proprietary model we investigated is the smaller Haiku model of the Claude family, which achieves an average score of 61.7, performing slightly worse (-1.6pp.) than the best-performing open-source model, Qwen2.5-VL (72B). Qwen2.5-VL (72B) is the only open-source model that is close to proprietary models, only being -1.6pp. accuracy behind GPT-4o. The worst-performing model is the SmolVLM2-500M-Video-Instruct, with an average accuracy of 20.5%, which is lower than a random guess (25%). Based on the results, we have three main observations.

**English Evaluation.** We observe that performance on the English split varies considerably across model sizes. The Qwen2.5-VL (72B) model, for instance, achieves +22.6pp. accuracy compared to the 3B counterpart. This observation also holds for other model families considered (Gemma3 (27B) vs. Gemma3 (4B): +25.3, LLaVA (34B) vs. LLaVA (7B): +12.9). Further, we find that dense models perform better, as seen in the comparison between Qwen3-VL 8B (58.9) and its sparse 30B counterpart with 3B active parameters (57.3). However, we also observe one outlier with Qwen2.5-VL (32B) which does not achieve better performance than its smaller 7B-version. Although this may seem obvious, a family hierarchy is considered a metric for a strong benchmark (Qian et al., 2026). The central premise is that well-constructed questions should induce performance differences that align with established capability hierarchies among models, such that more capable models systematically outperform less capable ones on those questions.

**Multilingual Evaluation.** For easier comparison, we included the column $\Delta_{\text{non-EN}}$, which shows the absolute difference between English results and the average across all translated splits. We observe that most results on the multilingual splits are lower than for the English one (See Table2). For instance, Qwen2.5-VL (72B) achieves 69.4 accuracy on the English split but only 58.2 on the German split. This trend also holds for the LLaVA model family and Idefics3. We note that these models are not explicitly trained on multilingual data. However, there are also a few outliers, e.g., SmolVLM2-256-Video-Instruct achieves +5.3pp. better scores with non-English examples. Only when evaluating on the Chinese split, we observe lower scores compared to the English split, with 31.1 accuracy. This observation indicates that, for the translated questions, comprehensive knowledge of the language is needed to answer them correctly, underlining the multilingual relevance of PISABENCH.

# 5 Ablations

## 5.1 Comparison to Related Benchmarks

In Table 3, we compare PISABENCH with commonly known benchmarks MMLU (Hendrycks

| MODEL | EN | DE | FR | IT | ES | CH | AVG | $\Delta_{\text{non-EN}}$ |
|---|---|---|---|---|---|---|---|---|
| *Proprietary Models* | | | | | | | | |
| GPT-4o | **71.0±4.1** | **68.9±4.2** | **69.7±4.2** | **65.6±4.3** | **64.8±4.3** | **67.2±4.3** | 67.8 | -3.8 |
| Claude-3.5-Haiku | 62.9±4.3 | 56.6±4.5 | 64.8±4.3 | 59.8±4.4 | 61.5±4.4 | 64.8±4.3 | 61.7 | -1.4 |
| *Qwen2.5 VL Family* | | | | | | | | |
| Qwen2.5-VL-3B-Instruct | 46.8±4.5 | 41.0±4.5 | 42.6±4.5 | 43.4±4.5 | 41.0±4.5 | 40.2±4.4 | 42.5 | -5.1 |
| Qwen2.5-VL-7B-Instruct | 52.4±4.5 | 48.4±4.5 | 56.6±4.5 | 54.1±4.5 | 46.7±4.5 | 47.5±4.5 | 50.9 | -1.8 |
| Qwen2.5-VL-32B-Instruct | 51.6±4.5 | 44.3±4.5 | 46.7±4.5 | 45.1±4.5 | 39.3±4.4 | 44.3±4.5 | 45.2 | -7.7 |
| Qwen2.5-VL-72B-Instruct | 69.4±4.1 | 58.2±4.5 | 60.7±4.4 | 64.8±4.3 | 63.1±4.4 | 63.9±4.3 | 63.3 | -7.2 |
| *Qwen3 VL Family* | | | | | | | | |
| Qwen3-VL-4B-Instruct | 48.4±4.5 | 50.0±4.5 | 51.6±4.5 | 49.2±4.5 | 45.1±4.5 | 52.5±4.5 | 49.5 | +1.3 |
| Qwen3-VL-8B-Instruct | 58.9±4.4 | 52.5±4.5 | 57.4±4.5 | 48.4±4.5 | 55.7±4.5 | 55.7±4.5 | 54.8 | -4.9 |
| Qwen3-VL-30B-A3B-Instruct | 57.3±4.4 | 48.4±4.5 | 50.8±4.5 | 50.8±4.5 | 44.3±4.5 | 50.0±4.5 | 50.3 | -8.4 |
| *Gemma Family* | | | | | | | | |
| gemma-3-4b-it | 45.2±4.5 | 35.2±4.3 | 36.9±4.4 | 36.9±4.4 | 36.1±4.3 | 38.5±4.4 | 38.1 | -8.4 |
| gemma-3-12b-it | 58.1±4.4 | 52.5±4.5 | 50.8±4.5 | 51.6±4.5 | 48.4±4.5 | 54.1±4.5 | 52.6 | -6.6 |
| gemma-3-27b-it | 60.5±4.4 | 63.9±4.3 | 61.5±4.4 | 63.9±4.3 | 61.5±4.4 | 54.1±4.5 | 60.9 | +0.5 |
| *SmolLM Family* | | | | | | | | |
| SmolVLM2-256M-Video-Instruct | 33.9±4.3 | 36.9±4.4 | 41.8±4.5 | 41.8±4.5 | 44.3±4.5 | 31.1±4.2 | 38.3 | +5.3 |
| SmolVLM2-500M-Video-Instruct | 29.8±4.1 | 23.0±3.8 | 13.1±3.1 | 17.2±3.4 | 19.7±3.6 | 20.5±3.7 | 20.6 | -11.2 |
| SmolVLM2-2.2B-Instruct | 38.7±4.4 | 23.8±3.9 | 25.4±3.9 | 29.5±4.1 | 34.4±4.3 | 30.3±4.2 | 30.4 | -10.0 |
| *LLaVA Family* | | | | | | | | |
| llava-1.5-7b-hf | 30.6±4.1 | 29.5±4.1 | 32.8±4.3 | 36.1±4.3 | 31.1±4.2 | 29.5±4.1 | 31.6 | +1.2 |
| llava-1.5-13b-hf | 35.5±4.3 | 32.8±4.3 | 27.0±4.0 | 31.1±4.2 | 28.7±4.1 | 33.6±4.3 | 31.5 | -4.8 |
| llava-v1.6-34b-hf | 43.5±4.5 | 36.9±4.4 | 36.9±4.4 | 34.4±4.3 | 38.5±4.4 | 41.8±4.5 | 38.7 | -5.8 |
| *Others* | | | | | | | | |
| Idefics3-8B-Llama3 | 47.6±4.5 | 42.6±4.5 | 36.9±4.4 | 38.5±4.4 | 42.6±4.5 | 36.9±4.4 | 40.9 | -8.1 |

Table 2: Accuracy (%) across languages for each model. Best per language in bold. Avg = mean across languages. $\Delta_{\text{non-EN}}$ = Avg(non-EN) - EN in percentage points. CH = Chinese, DE = German, EN = English, ES = Spanish, FR = French, IT = Italian. Output correctness decided by LLM (GPT-4o)

et al., 2021) and MMMU (Yue et al., 2025). We observe that our benchmark is significantly more difficult compared to MMLU for the models investigated, e.g., we find a performance of 68.5 on MMLU using Qwen2.5-VL (7B), whereas the corresponding performance on PISABENCH is only 52.4. However, we note that MMLU is a text-only benchmark, emphasizing that modalities other than text is more difficult for current language models.

When considering the results for MMMU, a corresponding multimodal benchmark, we observe that the results on our benchmark are in a similar range. For instance, we find that Qwen2.5-VL (72B) achieves 70.2 on MMMU and 69.4 on PISABENCH, and similarly Qwen3-VL (30B-A3B) achieves 59.0 on MMMU and 57.3 on PISABENCH. Despite similar accuracy ranges, the two benchmarks are complementary: MMMU tests college-level domain knowledge, whereas PISABENCH emphasizes applied visuospatial reasoning over factual recall. Notably, scaling behavior also differs—Qwen2.5-VL (32B) gains 5.8 points over the 7B variant on MMMU but performs slightly worse on PISABENCH.

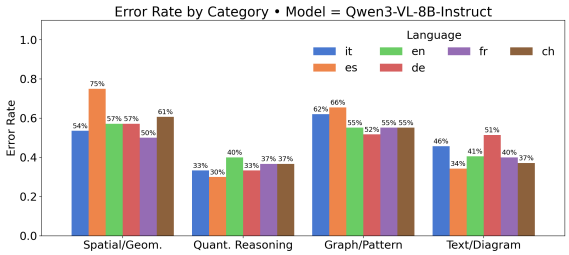
### 5.2 Error Analysis

As described in Section 2.2, we labeled each question with one of the following question type categories: spatial and geometric reasoning (23%), quantitative reasoning (24%), graph and pattern analysis (23%), and text and diagram understanding (29%). In this section, we investigate the errors the models have made to better understand the failure areas.

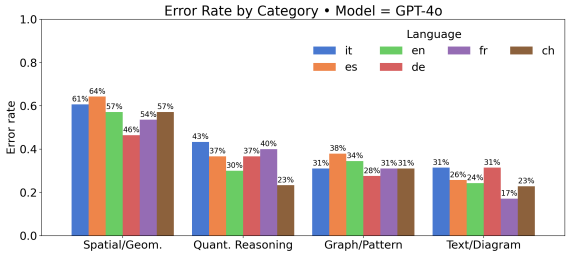
We show results for Qwen3-VL (8B) in Figure 3a and for GPT-4o in Figure 3b. For Qwen3-VL (8B), graph & pattern analysis yields the highest error rates across most languages, reaching up to 66% for Spanish. Spatial & geometric reasoning is similarly challenging, with Spanish exhibiting the highest error rate at 75% and Chinese at 61%. Quantitative reasoning produces the lowest error rates overall (30–40%), while text & diagram understanding falls in between, with most languages around 37–46% but German reaching 51%. Notably, Spanish shows pronounced weaknesses in spatial & geometric reasoning (75%) and graph & pattern analysis (66%), whereas it achieves one of the lowest error rates in text & diagram understand-

| MODEL | MMLU | MMMU | PISA$_{en}$ |
|---|---|---|---|
| *Qwen2.5 VL Family* | | | |
| Qwen2.5-VL-3B-Instruct | 66.3 | 46.1 | 46.8 |
| Qwen2.5-VL-7B-Instruct | 68.5 | 52.4 | 52.4 |
| Qwen2.5-VL-32B-Instruct | 75.2 | 58.2 | 51.6 |
| Qwen2.5-VL-72B-Instruct | 83.0 | 70.2 | 69.4 |
| *Qwen3 VL Family* | | | |
| Qwen3-VL-4B-Instruct | 64.6 | 54.0 | 48.4 |
| Qwen3-VL-8B-Instruct | 66.6 | 54.8 | 58.9 |
| Qwen3-VL-30B-A3B-Instruct | 66.3 | 59.0 | 57.3 |
| *Gemma Family* | | | |
| gemma-3-4b-it | 52.4 | 39.8 | 45.2 |
| gemma-3-12b-it | 72.1 | 48.7 | 58.1 |
| gemma-3-27b-it | 74.6 | 52.0 | 60.5 |
| *LLaVA Family* | | | |
| llava-1.5-7b-hf | 25.4 | 33.1 | 30.6 |
| llava-1.5-13b-hf | 25.4 | 35.7 | 35.5 |
| llava-v1.6-34b-hf | 25.4 | 50.2 | 43.5 |
| *Others* | | | |
| Idefics3-8B-Llama3 | 23.1 | 46.6 | 47.6 |

Table 3: Accuracy by benchmark for each model. For PISABENCH, the English split is shown.




(a) Qwen-3-VL-8B-Instruct




(b) GPT-4o

Figure 3: Comparison of error rates across languages for GPT-4o and Qwen-3-VL-8B-Instruct.

ing (34%). English, despite being the source language, does not consistently outperform the translated splits; for instance, it records 55% in graph & pattern analysis and 57% in spatial & geometric reasoning, comparable to or higher than several non-English languages.

These trends broadly hold when looking at the results for GPT-4o; however, all error rates are significantly lower on average, except for spatial & geometric reasoning, which remains persistently difficult. Particularly for quantitative reasoning and graph & pattern analysis, we observe substantially lower error rates. For example, the error rate in quantitative reasoning in Italian drops from 33% to 30%, while the error rate in graph & pattern analysis drops from 55% to 31% in French. The relative resilience of spatial & geometric reasoning errors across both models suggests that this category poses a fundamental challenge for current VLMs, independent of model scale.

### 5.3 Contamination Analysis

In this section, we investigate to what extent models may have memorized parts of the benchmark during pretraining, as our data source is a publicly available resource. To do so, we conduct a contamination analysis using two settings in which we evaluate the model with and without the images. Our hypothesis here is that if a model can correctly answer questions without the visual context, it indicates prior exposure to the same or similar PISA tests during pretraining, which would undermine the validity of the benchmark. We conduct the experiment using the original English split and all translated versions of our benchmark.

The results, depicted in Table 4, show a substantial decrease in accuracy across all languages and models when images are removed, indicating our benchmark does not suffer from pretraining contamination. For example, for Qwen2.5-VL-7B-Instruct, average accuracy decreases from 50.9% to 34.8%, and for GPT-4o, it drops from 67.8% to 34.9%.

This consistent decline demonstrates that models depend heavily on visual input to answer correctly. If prior exposure to our dataset were present, models would be expected to achieve comparable accuracy in the text-only setting by recalling question–answer pairs from pretraining. Instead, the large performance gap between the two settings confirms that models have not been trained on the benchmark content and that their predictions require visual reasoning.

These findings, together with the consistently low to moderate overall scores, indicate that PISABENCH exhibits low contamination and provides a reliable measure of multimodal reasoning. A likely explanation for the low contamination level is that each input question was reframed and partially adjusted to fit a multiple-choice format, making it more difficult for models to reproduce

| | Accuracy (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| Model | CH | DE | EN | ES | FR | IT | Avg. |
| *With Images* | | | | | | | |
| Qwen2.5-VL-7B-Instruct | 52.4 | 48.4 | 56.6 | 54.1 | 46.7 | 47.5 | 50.9 |
| GPT-4o | 71.0 | 68.9 | 69.7 | 65.6 | 64.8 | 67.2 | 67.8 |
| *Without Images* | | | | | | | |
| Qwen2.5-VL-7B-Instruct | 36.2 | 29.2 | 33.6 | 34.5 | 38.9 | 36.2 | 34.8 |
| GPT-4o | 34.4 | 28.6 | 37.1 | 35.2 | 39.3 | 34.4 | 34.8 |

Table 4: Contamination test results comparing model performance with and without access to images.

memorized answers from pretraining data.

## 6 Related Work

**LVLM Benchmarks.** Recent progress in large vision-language models (LVLMs) has been closely tied to the development of evaluation benchmarks. Early benchmarks primarily assessed visual perception and image understanding (Fu et al., 2025; Liu et al., 2024a, 2023; Meng et al., 2025), often restricting evaluation to multiple-choice or short-form VQA tasks. More recently, these benchmarks have been extended to more general areas such as cognition and reasoning, for example, benchmarks such as M3Exam (Zhang et al., 2023), SOK-Bench (Wang et al., 2024a), MathVista (Lu et al., 2024), or VL-ICL-Bench (Zong et al., 2025). Further, benchmarks like MMDU (Liu et al., 2024b), adopt open-ended questions combined with LLM-as-a-judge evaluation, while MMMU-pro (Yue et al., 2025) propose unifying text and images into a single visual representation.

**Multilingual Benchmarks.** In the domain of multilingual benchmarks, authors often begin with English datasets and translate them into other languages, such as XNLI (Conneau et al., 2018) or XCOPA (Ponti et al., 2020). More recent works utilize multilingual language models to translate the original test set, such as HumanEval-XL (Peng et al., 2024) or mHumanEval (Raihan et al., 2025). However, while this approach provides broad coverage, it may propagate cultural or linguistic biases through the translation using language models (Shi et al., 2023). P-MMEval (Zhang et al., 2025) and BenchMAX (Huang et al., 2025) address this issue by using parallel corpora to fairly assess cross-lingual capabilities, disentangling cultural knowledge from a language model's translation ability.

**Multilingual LVLM Benchmarks.** Many multilingual LVLM benchmarks evaluate the general natural language and image understanding capabilities of vision-language models such as xGQA (Pfeiffer et al., 2022), GEM (Su et al., 2021), and MaXM (Changpinyo et al., 2023). More recent benchmarks extend this line of work to reasoning and cognition, including M3Exam (Zhang et al., 2023), EXAM-V (Das et al., 2024), or M5-VGR (Schneider and Sitaram, 2024). Others focus on culture-specific reasoning, such as ALM-bench (Vayani et al., 2024) and CVQA (Romero et al., 2024), or domain-specific reasoning, such as medical reasoning in WorldMedQA-V (Matos et al., 2025). To mitigate the challenges of translating datasets into target languages, parallel benchmarks such as M4U (Wang et al., 2024b), PM4Bench (Gao et al., 2025), and XT-VQA (Yu et al., 2024) have been proposed.

## 7 Conclusion

In this paper, we introduce PISABENCH, a multilingual and multimodal benchmark designed to evaluate vision-language models on human-authored tests based on the international PISA study of the OECD. We translated the original English set using GPT-4o and verified translation accuracy with native speakers. Further, we enrich our dataset with question type categories that enable detailed analysis of failure areas.

In our experiments, we find that state-of-the-art vision-language models struggle to achieve high accuracy across all languages. We further observe significant gaps on our non-English splits, highlighting the need for better approaches to multilingual and multimodal reasoning. At last, we find particularly high error rates in geometric and spatial reasoning, indicating that this area is still challenging for current state-of-the-art VLMs.

## Limitations

While PISABENCH provides a valuable resource for evaluating multilingual multimodal reasoning, it also comes with several limitations. First, we do not perform the translations ourselves; instead, we rely on human annotators to verify and correct automatically generated translations. Although this process ensures sufficient quality for evaluation, it may not capture all subtle linguistic nuances, particularly in languages with complex morphology or idiomatic expressions.

Second, our benchmark uses accuracy rates as our main evaluation metric. However, the underlying PISA tests are designed to compare 15-year-old students across countries. As a result, the official test metrics are PISA scores, calculated using item response theory. We were only able to get the difficulty scores for a subset of our questions, such that we could only estimate the PISA scores in Appendix C.

Third, although the dataset consists of high-quality, human-authored examples, its size remains relatively modest. This makes PISABENCH particularly suitable as a resource-friendly evaluation benchmark, but less suitable for extensive and fine-grained error categorization or large-scale model training. Future extensions could address this limitation by expanding the dataset with additional PISA examples or complementary educational resources.

Fourth, we evaluated the translation quality using the GEMBA-MQM method with GPT-4o. Although this is the best-performing approach, we see a limitation when evaluating automatically translated data from GPT-4o using the same model. Research shows that language models recognize their own generations (Panickssery et al., 2024). However, we strengthen our evaluation with the COMET-KIWI evaluation and a manual evaluation. The evaluation using COMET-KIWI shows an overall good evaluation quality. Moreover, the manual evaluation is the most reliable analysis of the translation quality, which shows an overall good performance.

Finally, our evaluation protocol currently relies on LLM-as-judge evaluation. While this evaluation approach exploits the generation capabilities of LLMs, they may still miss subtler reasoning errors or reward overgeneralized answers. More robust evaluation methods, such as rubric-based scoring or task-specific human assessment, could further strengthen conclusions drawn from PISABENCH.

## References


Anthropic. 2025. Claude 3.5 Haiku. A lightweight large-language-model from Anthropic.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. *Preprint*, arXiv:2511.04703.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore. Association for Computational Linguistics.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaan Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, and Bo Ai. 2024. LLM-based multi-hop question answering with knowledge graph integration in evolving environments. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14438–14451, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Xinbang Dai, Huikang Hu, Yongrui Chen, Jiaqi Li, Rihui Jin, Yuyang Zhang, Xiaoguang Li, Lifeng Shang, and Guilin Qi. 2026. Elaipbench: A benchmark for expert-level artificial intelligence paper understanding. *Preprint*, arXiv:2510.10549.

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.

Harry Dong, Bilge Acun, Beidi Chen, and Yuejie Chi. 2025. Scalable llm math reasoning acceleration with low-rank distillation. *Preprint*, arXiv:2505.07861.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junyuan Gao, Jiahe Song, Jiang Wu, Runchuan Zhu, Guanlin Shen, Shasha Wang, Xingjian Wei, Haote Yang, Songyang Zhang, Weijia Li, Bin Wang, Dahua Lin, Lijun Wu, and Conghui He. 2025. Pm4bench: A parallel multilingual multi-modal multi-task benchmark for large vision language model. *Preprint*, arXiv:2503.18484.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Alexander Gill, Abhilasha Ravichander, and Ana Marasovic. 2025. What has been lost with synthetic evaluation? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9902–9945, Suzhou, China. Association for Computational Linguistics.

Ronald K Hambleton and Hariharan Swaminathan. 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A comprehensive multilingual evaluation suite for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16751–16774, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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. 2024a. Mmbench: Is your multi-modal model annbsp;all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024b. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *Preprint*, arXiv:2409.11968.

João Matos, Shan Chen, Siena Kathleen V. Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis Filipe Nakayama, José María Millet Pascual-Leone, Guergana K Savova, Hugo Aerts, Leo Anthony Celi, An-Kwok Ian Wong, Danielle Bitterman, and Jack Gallifant. 2025. WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7218–7231, Albuquerque, New Mexico. Association for Computational Linguistics.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2025. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

OECD. 2025. Pisa: Programme for international student assessment. `https://www.oecd.org/en/about/programmes/pisa.html`. Accessed: 2025-10-17.

Tomoya Okubo. 2022. Theoretical considerations on scaling methodology in pisa. Technical Report 282, Organisation for Economic Co-operation and Development (OECD), Directorate for Education and Skills.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. 2026. Inference-time computations for LLM reasoning and planning: A benchmark and insights. *Submitted to Transactions on Machine Learning Research*. Under review.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394, Torino, Italia. ELRA and ICCL.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Max Ploner, Jacek Wiland, Sebastian Pohl, and Alan Akbik. 2025. LM-pub-quiz: A comprehensive framework for zero-shot evaluation of relational knowledge in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 29–39, Albuquerque, New Mexico. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Qi Qian, Chengsong Huang, Jingwen Xu, Changze Lv, Muling Wu, Wenhao Liu, Xiaohua Wang, Zhenghua Wang, Zisu Huang, Muzhao Tian, Jianhan Xu, Kun Hu, He-Da Wang, Yao Hu, Xuanjing Huang, and Xiaoqing Zheng. 2026. Benchmark$^2$ : $Systematicevaluationofllmbenchmarks$. *Preprint*, arXiv:2601.03986.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. mHumanEval - a multilingual benchmark to evaluate large language models for code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11432–11461, Albuquerque, New Mexico. Association for Computational Linguistics.

Georg Rasch. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*, expanded edition edition. University of Chicago Press, Chicago. Originally published in 1960 by the Danish Institute for Educational Research.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: culturally-diverse multilingual visual question answering benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Lin Su, Nan Duan, Edward Cui, Lei Ji, Chenfei Wu, Huaishao Luo, Yongfei Liu, Ming Zhong, Taroon Bharti, and Arun Sacheti. 2021. GEM: A general evaluation benchmark for multimodal tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2594–2603, Online. Association for Computational Linguistics.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Yuan Wu, Qi Liu, and 15 others. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Comput. Surv.*, 57(11).

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.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Armin Toroghi, Willis Guo, Mohammad Mahdi Abdollah Pour, and Scott Sanner. 2024. Right for right reasons: Large language models for verifiable commonsense knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6601–6633, Miami, Florida, USA. Association for Computational Linguistics.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, and 1 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024a. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13384–13394.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024b. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and Yang Yu. 2026. A survey on large language models for mathematical reasoning. *ACM Comput. Surv.*, 58(8).

Xinmiao Yu, Xiaocheng Feng, Yun Li, Minghui Liao, Ya-Qi Yu, Xiachong Feng, Weihong Zhong, Ruihan Chen, Mengkang Hu, Jihao Wu, and 1 others. 2024. Cross-lingual text-rich visual comprehension: An information theory perspective. *arXiv preprint arXiv:2412.17787*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Hao-Ran Wei, Fei Huang, Bowen Yu, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. P-MMEval: A parallel multilingual multitask benchmark for consistent evaluation of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4809–4836, Suzhou, China. Association for Computational Linguistics.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2025. Vl-icl bench: The devil in the details of multimodal in-context learning. *Preprint*, arXiv:2403.13164.

## A Extraction Prompt

Figure 6 shows the prompt used for our information extraction step, especially for the data augmentation part, to ensure consistency and self-containment of each question.

## B Translation Prompt

We translate all English examples of our dataset using GPT-4 into our five target languages: Spanish, German, Chinese, French, and Italian. We use OpenAI's platform to generate the translation and depict the prompt used in Figure 5.

## C Approximating Model PISA Scores

PISA test scores are not reported as solve ratios but are instead measured on the PISA scale using item response theory (IRT) (Hambleton and Swaminathan, 2013; Okubo, 2022). To enable a direct comparison between model performance and student ability on the PISA scale, we estimate an approximate PISA-equivalent score using a Rasch model formulation (Rasch, 1980). We were able to find the official PISA difficulty parameters for a subset of questions in PISABENCH. We treat the model's binary correctness outcomes as responses to Rasch-parameterized items and infer the latent ability parameter $\theta$ by maximizing the Bernoulli log-likelihood under the logistic Rasch model. This inferred ability is then mapped to the PISA scale, providing an interpretable, though imperfect, score, as such scores are not widely publicly available. However, it helps us to establish a link between model accuracy on real assessment items and the corresponding human performance level.

We present the PISA of vision-language models in Table 5. Since only a limited number of benchmark items include published PISA item parameters, the resulting estimates rely on a comparatively small sample. As such, these PISA-equivalent scores should be interpreted as *indicative trend signals* rather than firm psychometric measurements. To ensure numerical stability, we constrain the Rasch ability parameter to $[-3, 3]$, which corresponds to roughly the 200-800 point range of the PISA scale. Under these bounds, a model answering every available item correctly attains the upper-limit score of 800, which is achieved by GPT-4o in the reading area. While this provides a useful approximation for positioning models along the PISA scale, a more directly comparable metric would be the per-item student accuracy distributions reported by PISA (i.e., the proportion of students solving each question correctly), which, unfortunately, are not publicly available for all released items. We therefore include the Rasch-based scores for transparency and illustrative comparison with human performance, while emphasizing that they should not be interpreted as evidence that language models currently match or exceed student proficiency on the full PISA assessment. Due to limited PISA scores assigned to our corresponding samples, we aggregated over years.

To enable a direct comparison between model performance and student ability on the PISA scale, we estimate an equivalent PISA score using a Rasch model formulation. For the subset of benchmark items where official PISA difficulty estimates were available, we treat the model's binary correctness outcomes as responses to Rasch-parameterized items and compute the latent ability parameter $\theta$.

## D Annotation Guidelines

Figure 5 shows an example of the annotation guidelines for Chinese. We explained what critical, major, and minor issues are. We also explicitly noted that the annotators should filter offensive and discriminatory language.

## Annotation Guidelines

The review is **not** about evaluating translations in terms of style or quality in detail. We solely want to determine whether the translated texts are **comprehensible** and whether **clear errors** are present. This means: only clear **semantic or syntactic errors** should be flagged.

**Table Structure.** Each row has **eight columns**: `question_id` (sequential ID, ignorable); `en_question` (English question); `en_choices` (four answer options, format: `["A1","A2","A3","A4"]`); `ch_question` (Chinese question); `ch_choices` (Chinese answer options, same format); `translation issue` (dropdown, required); `error category` (dropdown, required); `comment` (free-text, optional).

**Step 1 – "translation issue":**

- **No Issue** → Completely correct, no errors.
- **Partial Issue** → Minor errors: only `ch_question` *or* `ch_choices` affected; slight inaccuracies that do not entirely prevent comprehension.
- **Issue** → Serious errors: both `ch_question` *and* `ch_choices` wrong, or translations grossly misleading.

**Step 2 – "error category"** (set **None** if No Issue):

- **Wrong Context** → Linguistically correct but meaning does not fit context.
- **Syntax Error** → Grammatical/linguistic errors impairing readability.
- **Translation Error** → Clear mistranslation (wrong word or meaning).
- **Incomplete** → Translation incomplete; parts missing or cut off.

**Comments.** Use the `comment` field for justifications, recurring errors, or any offensive/discriminatory language encountered.

Figure 4: Review guidelines provided to annotators for evaluating translations, using Chinese as an example.

| Language | EN | | DE | | FR | | IT | | ES | | CH | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Math | Reading | Math | Reading | Math | Reading | Math | Reading | Math | Reading | Math | Reading | Math | Reading |
| Model | | | | | | | | | | | | | | |
| GPT-4o | 576 | 800 | 559 | 800 | 542 | 800 | 559 | 800 | 542 | 800 | 576 | 800 | 559 | 800 |
| Claude-3-5-Haiku | 632 | 800 | 559 | 650 | 677 | 604 | 593 | 604 | 559 | 604 | 576 | 604 | 599 | 644 |
| Qwen2.5-VL-3B-Instruct | 444 | 656 | 444 | 650 | 494 | 604 | 426 | 604 | 461 | 604 | 426 | 604 | 449 | 620 |
| Qwen2.5-VL-7B-Instruct | 526 | 718 | 526 | 566 | 542 | 800 | 526 | 800 | 542 | 650 | 494 | 566 | 526 | 683 |
| Qwen2.5-VL-32B-Instruct | 526 | 610 | 444 | 650 | 461 | 713 | 408 | 800 | 444 | 604 | 444 | 650 | 454 | 671 |
| Qwen2.5-VL-72B-Instruct | 593 | 718 | 559 | 800 | 542 | 800 | 612 | 800 | 542 | 800 | 576 | 800 | 570 | 786 |
| Qwen3-VL-4B-Instruct | 526 | 656 | 494 | 650 | 559 | 604 | 542 | 713 | 478 | 650 | 510 | 650 | 518 | 653 |
| Qwen3-VL-8B-Instruct | 526 | 800 | 526 | 650 | 542 | 650 | 494 | 713 | 494 | 800 | 526 | 650 | 518 | 710 |
| Qwen3-VL-30B-A3B-Instruct | 576 | 718 | 461 | 604 | 494 | 713 | 444 | 800 | 426 | 532 | 494 | 713 | 482 | 680 |
| gemma-3-4b-it | 461 | 456 | 426 | 501 | 461 | 470 | 461 | 470 | 444 | 532 | 478 | 470 | 455 | 483 |
| gemma-3-12b-it | 526 | 800 | 542 | 800 | 526 | 713 | 510 | 650 | 526 | 650 | 526 | 800 | 526 | 735 |
| gemma-3-27b-it | 576 | 656 | 542 | 650 | 576 | 650 | 576 | 604 | 542 | 650 | 494 | 566 | 551 | 629 |
| llava-1.5-7b-hf | 366 | 429 | 387 | 470 | 341 | 604 | 408 | 650 | 426 | 532 | 426 | 501 | 392 | 531 |
| llava-1.5-13b-hf | 408 | 512 | 408 | 470 | 366 | 470 | 341 | 532 | 366 | 470 | 387 | 501 | 379 | 492 |
| llava-v1.6-34b-hf | 408 | 718 | 408 | 650 | 408 | 650 | 387 | 713 | 408 | 713 | 461 | 650 | 413 | 682 |
| SmolVLM2-256M-Video-Instruct | 387 | 429 | 494 | 373 | 510 | 373 | 510 | 470 | 526 | 440 | 461 | 373 | 481 | 409 |
| SmolVLM2-500M-Video-Instruct | 313 | 337 | 387 | 408 | 341 | 281 | 387 | 332 | 408 | 373 | 341 | 440 | 362 | 361 |
| SmolVLM2-2.2B-Instruct | 461 | 456 | 387 | 332 | 366 | 373 | 387 | 408 | 444 | 440 | 387 | 440 | 405 | 408 |
| Idefics3-8B-Llama3 | 494 | 610 | 444 | 650 | 478 | 470 | 461 | 532 | 461 | 566 | 387 | 501 | 454 | 554 |
| *Test Takers* | | | | | | | | | | | | | | |
| 2000 | 493 | 504 | 490 | 484 | 517 | 505 | 457 | 487 | 476 | 493 | - | - | 487 | 495 |
| 2003 | 483 | 495 | 503 | 491 | 511 | 496 | 466 | 476 | 485 | 481 | - | - | 490 | 488 |
| 2006 | 474 | - | 504 | 495 | 496 | 488 | 462 | 469 | 480 | 461 | - | - | 483 | 478 |
| 2009 | 487 | 500 | 513 | 498 | 497 | 497 | 483 | 486 | 483 | 481 | 600 | 556 | 511 | 503 |
| 2012 | 481 | 498 | 514 | 508 | 495 | 505 | 485 | 490 | 484 | 488 | 613 | 570 | 512 | 510 |
| 2022 | 465 | 504 | 475 | 480 | 474 | 474 | 471 | 482 | 473 | 474 | 552 | 510 | 485 | 487 |

Table 5: Approximate PISA-equivalent model scores compared with national averages for the United States (EN), Germany, Italy, Spain, France, and China from the past years. We compute the PISA scores for LLMs on a subset of questions for which we were able to obtain difficulty parameters. Thus, the resulting LLM scores should be interpreted as approximate indicators rather than validated test scores.

**Prompt Template for Translations**

You are an expert translator from English to {lang} for educational assessments. You are given a multiple-choice question (MCQ) in English with its answer Translate the question and all answer options into <lang>, BUT keep all units, symbols, labels, and tokens exactly as in the original (no localization), so the text aligns with English annotations present in the associated image.

Requirements:
1. **Literal alignment with the source**:
- Do NOT convert or localize measurement units or quantities (e.g., keep 'miles per hour', 'mph', 'inches', 'pounds', 'Fahrenheit' exactly as written).
- Preserve all numbers, formulas, variables, dates, and proper nouns exactly (do not change numeric formatting: keep 3.5, not 3,5).
- Preserve abbreviations, acronyms, and labels verbatim if they may appear in the image (e.g., 'mph', 'NYC', axis labels, map keys).
2. **Natural but faithful {lang}**:
- Translate the surrounding prose naturally into lang but do NOT alter difficulty or meaning.
- Do NOT add hints, explanations, or paraphrases that change the task.

3. **Preserve the answer key**:
- Ensure the translated correct option remains the correct answer in {lang}.
- Do NOT reorder options or change their content beyond the literal translation.

4. **Output format**: Return ONLY the translated JSON in this exact structure, as you received it. Do not add any extra fields, comments, or explanations. If a term cannot be translated without breaking alignment with the image, keep it in English verbatim.

Figure 5: Prompt template used for generating translations using GPT-4o.

Prompt Template for Question Extraction from PDFs

You are an expert educational content creator and a skilled test designer. Your task is to analyze the provided PDF document, which contains educational content, exercises, and questions. Your goal is to perform the following actions for every question or exercise you find in the document: 1. **Extract the Core Questions**: Identify the primary questions or problem statements.
2. **Standalone Questions**: Ensure each question is self-contained, providing all necessary context for understanding without needing to refer back to the document. Assume only tables and images are provided separately if referenced.
3. **Reformulate into a Multiple-Choice Question (MCQ)**: Convert the question into a standard multiple-choice format with exactly four options (A, B, C, D).
4. **Generate a Correct Answer**: Identify the correct solution from the context of the document and assign it to one of the options. This can include a reference image if the question is visual in nature. Assume that the image is available to the test-taker, but refer to it in the question text.
5. **Generate Plausible Distractors**: Create three plausible but incorrect answer choices (distractors) that are relevant to the topic but are definitively wrong. The distractors should be well-formed and not nonsensical.
6. **Identify Scoring**: If the original question has a scoring system (e.g., "Worth 10 points"), you must extract and include this information. If no score is present, state "Score: Not specified."
7. **Translate if needed**: If the document is not in English, translate the question and options into English while preserving the original meaning. Apply also to the task name if it is not in English.
8. **Category Identification**: Identify whether the question falls under "math" or "reading" and specify a more detailed subcategory if possible (e.g., geometry, algebra for math; locate info, interpret text for reading).
The final output must be a JSON array, where each object represents one multiple-choice question. A PDF can contain more than one question. If there is more than one question, then identify each and produce one JSON object per question.

Figure 6: Prompt used for augmenting the original English test questions, e.g., required when multiple-choice answer options are not available or the question is not self-contained.