VMMU: A Vietnamese Multitask Multimodal Understanding and Reasoning Benchmark

Vy Tuong Dang[*1], An Vo[*‡1,2,3], Emilio Villa-Cueva[2], Quang Tau[1], Duc Dm[1], Thamar Solorio[2], Daeyoung Kim[1]
[1]KAIST, [2]MBZUAI, [3]University of Michigan
[*]*Equal contribution*
{vydang,ducdm200158,kimd}@kaist.ac.kr
voan@umich.edu {emilio.villa,thamar.solorio}@mbzuai.ac.ae

Abstract

Vision language models are predominantly evaluated on English-centric multimodal benchmarks, leaving their behavior in other languages and cultures insufficiently understood. We introduce VMMU, a Vietnamese Multitask Multimodal Understanding and Reasoning benchmark of 2,548 multiple-choice questions across seven domains: Mathematics, Physics, Chemistry, Biology, Geography, Driving Test, and IQ Test. Every question requires jointly reasoning over Vietnamese text and non-text visual evidence (charts, diagrams, tables, traffic scenes), and is distributed as a single rendered page image. We evaluate 10 open-source and 5 closed-source VLMs on VMMU. Test-time compute improves performance from 57% to 78% accuracy, but it remains inferior to an expert baseline (99% accuracy). Extensive error analyses show that closed-source VLMs already achieve strong Vietnamese OCR performance, yet still struggle on VMMU. This suggests that the primary bottleneck is multimodal grounding and reasoning rather than text recognition or Vietnamese language understanding. Code and data are available at: vmmu-bench.github.io.

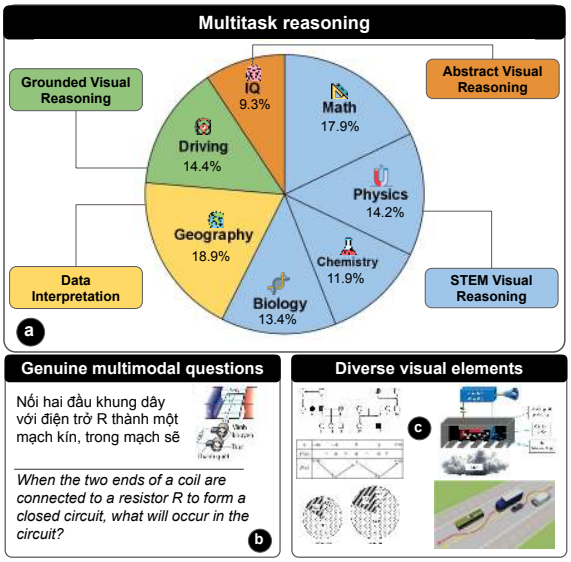


Figure 1: (a) VMMU covers seven domains and supports multitask multimodal reasoning. (b) Many items require joint interpretation of Vietnamese text and the accompanying image. (c) VMMU includes diverse visual evidence (e.g., diagrams, charts, tables, and traffic scenes).

1 Introduction

Vision language models (VLMs) have made rapid progress on multimodal benchmarks, demonstrating strong performance on complex visual and textual reasoning tasks in English (Yue et al., 2024, 2025; Liu et al., 2024; Lu et al., 2022, 2024). However, it remains unclear how well these models perform in low-resource languages when language understanding, visual grounding, and reasoning are all required to complete a task. In this work, we study this question in the context of Vietnamese, a widely spoken language with nearly 100 million speakers that remains significantly underrepresented in multimodal evaluation.

This underrepresentation matters because many tasks require understanding image-based documents: people frequently rely on screenshots or photos that include instructions with structured visual content. For example, learning materials often consist of textbook pages or PDFs that include questions alongside visual information (Fig. 1b). Similarly, people routinely encounter administrative and government documents (e.g., application forms, notices, and records) that combine dense text with visually structured elements such as tables, stamps, and form layouts. VLMs therefore have the potential to support Vietnamese users across a range of these settings, from helping parents explain homework solutions at home to enabling cit-

[‡]Work done while AV was at MBZUAI and KAIST.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 825–870
November 6-10, 2026 ©2026 Association for Computational Linguistics

izens to interpret administrative documents. To accomplish this, models must process image-based questions that integrate *Vietnamese text* with *structured visual content* (e.g., charts, diagrams, tables; Fig. 1c), which requires reliable optical character recognition (OCR) of Vietnamese, adequate grounding between textual instructions and visual evidence, and robust multimodal reasoning. Frontier VLMs have recently been reported to reach gold-medal performance on English mathematics and coding olympiads (Chervonyi et al., 2025; Luong and Lockhart, 2025; El-Kishky et al., 2025), suggesting they may be sufficiently capable to reliably support such applications. Whether this capability actually transfers to Vietnamese multimodal tasks, however, remains untested.

Several LLM/VLM benchmarks include partial or full Vietnamese coverage (Singh et al., 2025; Bui et al., 2025; Xuan et al., 2025), but they are either text-only or contain only text embedded in images (rather than true visual evidence such as graphs, tables, or charts). As a result, they do not require joint visual grounding and language reasoning. Consequently, prior benchmarks offer limited insight into VLM performance on *genuinely multimodal* Vietnamese questions, where non-text visual evidence is essential and must be integrated with the text to reach the correct answer. Importantly, English-only multimodal evaluation (Yue et al., 2024, 2025) is not sufficient to assess multilingual generalization. Vietnamese tasks exhibit differences in rendered text, labels, domain distributions, and conventions, limiting the usefulness of English benchmarks for diagnosing whether VLMs reliably ground and reason in Vietnamese settings.

To address this gap, we introduce VMMU: a Vietnamese Multitask Multimodal Understanding and Reasoning evaluation suite (Fig. 1). We define *multimodal questions* as those in which non-text visual evidence is necessary to solve the problem, so the model must integrate the question with visual elements. We filter out all samples where the image is not needed to answer the question, only retaining items that require understanding non-text visual evidence (e.g., diagrams, charts, tables).

VMMU consists of 2,548 carefully curated questions spanning seven domains (Fig. 1a): Mathematics, Physics, Chemistry, Biology, Geography, Driving Test, and IQ Test. The benchmark is constructed using a semi-automated pipeline with human validation to ensure multimodal integrity and question quality. Unlike prior datasets that simplify or discard visual information (Dao et al., 2023; Zhang et al., 2023), VMMU preserves visual content in its original form, enabling direct evaluation of multimodal reasoning. Our key findings are:

1. Open-source models average only 37% accuracy and non-thinking closed-source models 57% on VMMU, both below the average high-school student (66%). Test-time compute via thinking models improves to 78% but remains well below the expert reference (∼99%), so no group is reliable enough for tutoring or automatic checking (§4).

2. OCR is a major bottleneck for open-source VLMs but not for closed-source ones. Closed-source models achieve strong Vietnamese embedded-text recognition (mean BLEU 89%), while open-source models score substantially lower (mean BLEU 58%). Thus, open-source failures are partly attributable to OCR, whereas proprietary failures point more strongly to multimodal grounding and reasoning (§5.1).

3. Separating text from visual evidence improves reliability. Providing question and options as text alongside a visual crop improves every evaluated VLM by +7 points on average. This finding suggests that separating textual content from the image encourages models to rely more on visual reasoning while reducing distraction from text embedded in the image (§5.2).

4. Translating Vietnamese inputs into English consistently hurts performance for most models by -2 points on average, which shows that native-language multimodal evaluation cannot be replaced by translation, even for models trained predominantly on English (§5.3).

2 Related work

Multimodal Evaluation Recent multimodal benchmarks evaluate VLMs on increasingly diverse and structured visual media, including text-in-image, documents, and charts (Singh et al., 2019; Biten et al., 2019; Mathew et al., 2021; Masry et al., 2022), and broad multi-discipline reasoning that requires jointly processing text and visual evidence (Yue et al., 2024, 2025; Hao et al., 2025). These benchmarks typically provide the textual question and the visual evidence as separate inputs. **In contrast, VMMU's primary setting renders the question, answer options, and visual**

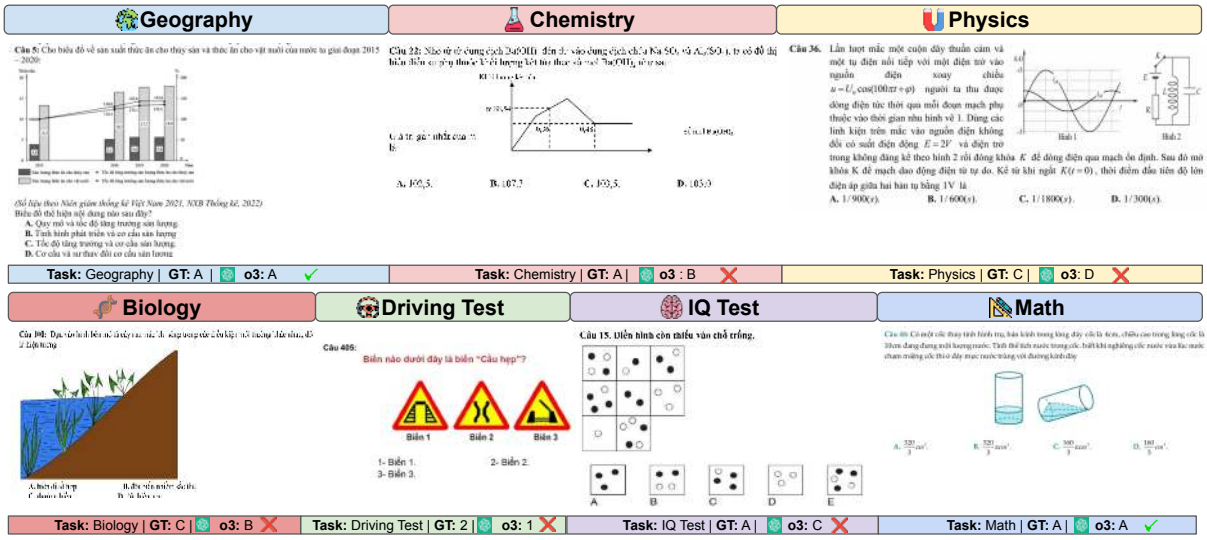


Figure 2: Sample questions from VMMU spanning seven domains. English translations are shown in Fig. 32.

evidence together within a single image, matching how Vietnamese exam materials commonly appear (Fig. 2). Beyond this unified-image setting, VMMU also contributes a genuine multimodal filter that removes text-only screenshots, and a suite of diagnostic variants (see Sec. 5) that isolate distinct failure modes rather than reporting only a single aggregate score. Previous benchmarks are also largely English-centric, and directly translating them does not capture Vietnamese-native layouts, in-image labels, or curriculum conventions, leaving limited understanding of how OCR, language effects, and multimodal reasoning interact in low-resource language settings such as Vietnamese.

Benchmark	#VN Q	#VN MM Q	Genuine MM
EXAMS-V (Das et al., 2024)	0	0	✗
SeaExam (Liu et al., 2025a)	1,745	0	✗
VNHSGE (Dao et al., 2023)	19,300	0	✗
VMLU (Bui et al., 2025)	10,880	0	✗
M3Exam (Zhang et al., 2023)	1,817	116	*partial*
VMMU **(Ours)**	**2,548**	**2,548**	✓

Table 1: Vietnamese coverage in representative benchmarks. #VN MM Q counts Vietnamese questions intended to be answered with an *image input*. **Genuine MM** marks questions requiring integration of Vietnamese text with *non-text visual evidence* (e.g., charts/diagrams/maps), excluding text-only screenshots.

Multilingual and Vietnamese Benchmarks Vietnamese text-only language models and evaluation benchmarks have advanced substantially (Nguyen and Nguyen, 2020; Bui et al., 2025), paralleling language-specific evaluation efforts for other non-English languages such as Korean (Son et al., 2025) and Kazakh (Togmanov et al., 2025). Non-English *multimodal* benchmarks, however, remain concentrated on a handful of languages such as Chinese (Zhang et al., 2024) and Japanese (Onohara et al., 2025; Miyai et al., 2025). Meanwhile, Vietnamese multimodal coverage in particular is sparse. Most existing Vietnamese benchmarks (Tab. 1) either contain no multimodal questions or rely on weakly multimodal formats (e.g., text-only screenshots), which prevents systematic evaluation of multimodal integration and related failure modes. For example, although VNHSGE (Dao et al., 2023) contains Vietnamese exam questions, it converts all images to text-only format during preprocessing. In contrast, VMMU provides Vietnamese questions where visual content is integral and is designed to support diagnostic analysis of OCR, multimodal reasoning, language effects, and answer-option biases.

3 The VMMU Benchmark

VMMU targets scenarios where Vietnamese language content is embedded within images and must be interpreted jointly with structured visual evidence (e.g., diagrams, charts, maps, tables; see Fig. 4).

3.1 Benchmark scope

VMMU contains 2,548 questions spanning seven domains: Mathematics, Physics, Chemistry, Biology, Geography, Driving Test, and IQ Test. Examples are illustrated in Fig. 2. These seven domains collectively probe three multimodal reasoning skills: (*i*) STEM and data interpretation (equations, diagrams, tables, plots): Math, Physics,

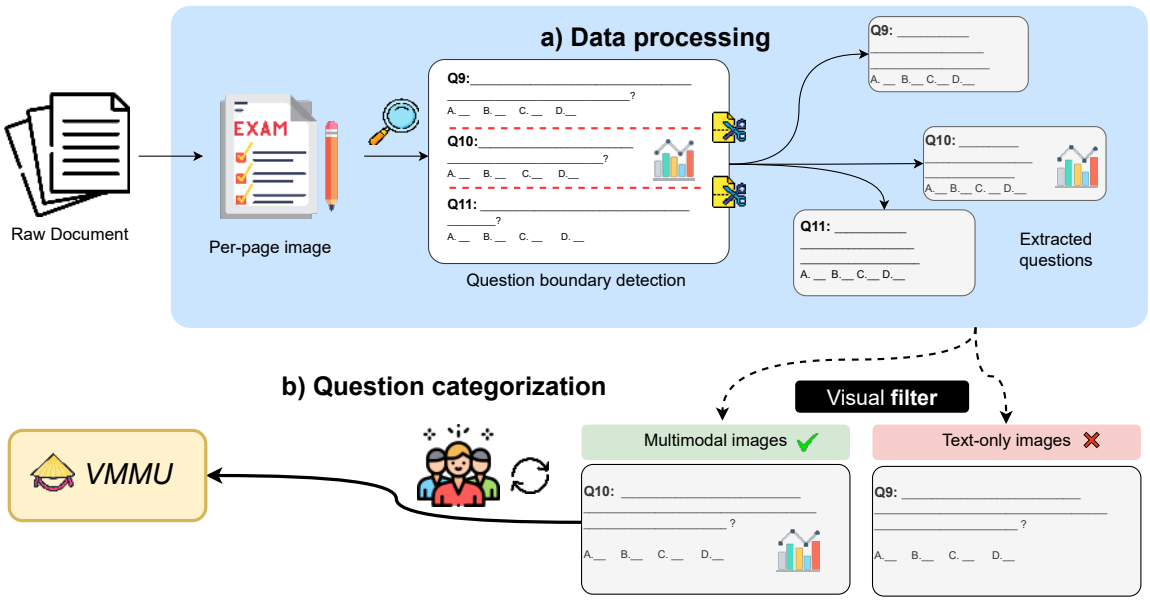


Figure 3: VMMU data curation pipeline. (a) Data processing renders PDFs as images and detects question boundaries. (b) Items are filtered for multimodal content and verified by native speakers.

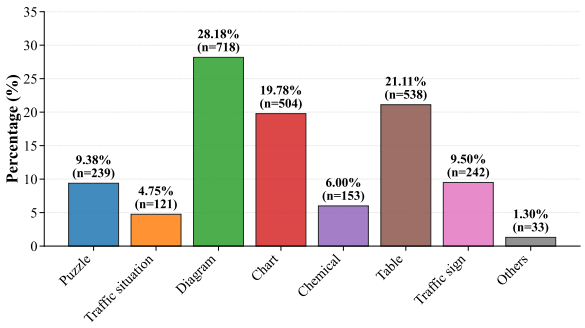


Figure 4: Distribution of visual evidence categories in VMMU.

Chemistry, Biology, Geography; (*ii*) rule-governed grounded reasoning (traffic signs and situations): Driving Test; and (*iii*) abstract visual reasoning (pattern completion and spatial logic): IQ Test. All questions are multiple-choice. A very small fraction ($\approx$1%) requires selecting multiple answers because the image contains more than one question. As the first large-scale Vietnamese multimodal benchmark of this kind, we intentionally prioritize curriculum-grounded sources (official exams and government materials) so that ground-truth answers, licensing, and question boundaries are unambiguous, which is a prerequisite for clean diagnostic analysis. Extending VMMU to in-the-wild multimodal scenarios (e.g., UI screenshots, Vietnamese cultural scenes) is a natural next step that requires different collection and verification pipelines, and we leave it to future work.

Tasks 1–5: STEM and data interpretation (Math, Physics, Chemistry, Biology, Geography) Questions for these five domains are collected from Vietnam's *High School Graduation Examination*, a nationwide standardized test used for graduation eligibility and university/college admission. Items span equations, scientific diagrams, Vietnamese-specific geographical charts, and maps. Samples are in Secs. E.1 to E.5.

Task 6: Rule-governed grounded reasoning (Driving Test) Vietnamese driving license examinations are standardized safety-critical assessments. These questions test traffic-rule knowledge and situational judgment from scenario-based images that require visual understanding and application of Vietnamese traffic regulations. Samples are in Sec. E.6.

Task 7: Abstract visual reasoning (IQ Test) IQ Test items are drawn from standardized pattern-recognition and spatial-logic puzzles, testing non-verbal reasoning independent of subject-matter knowledge. Samples are in Sec. E.7.

3.2 Data sources and curation pipeline

We collect question-answer pairs from official government sources and standardized national academic tests (Tab. 16), used under Article 25 of Vietnam's IP Law (non-commercial research; see Ethical Considerations). Our semi-automated

pipeline (see Fig. 3) renders exam PDFs to page images, detects question boundaries with Tesseract OCR (Smith, 2007), filters out text-only screenshots via geometric analysis of non-text visual structures, and finalizes every item through triple verification by three Vietnamese native speakers (co-authors). Full details are in Sec. D.

4 Main evaluation: Thinking models achieve the strongest performance

We evaluate VLMs on VMMU, where question text, options, and visual evidence are rendered together in a single image. Our goal is to characterize how current VLMs perform on Vietnamese multimodal questions, and what level of reliability they reach relative to humans.

4.1 Experiments

We evaluate 3 groups of models on the full VMMU benchmark with the prompt shown in Figs. 9 and 10: (1) 10 open-source VLMs spanning multiple families and sizes, (2) 3 non-thinking closed-source VLMs, and (3) 4 thinking variants of closed-source VLMs (including 2 recent thinking models, GPT-5 and Opus-4.6). Full experimental details are in Sec. B.

Model	Math	Phys	Chem	Bio	Geo	Drv	IQ	Mean
Human & Random baseline								
High schoolers (Average)	64.50	66.70	66.80	62.80	71.90	–	–	66.54
Human expert (Best)	98.00	100.0	100.0	100.0	100.0	–	–	99.60
Random baseline	25.00	24.66	24.38	24.67	25.00	33.24	24.46	25.92
Open-source VLMs								
Aya-Vision-8B	24.34	26.04	27.81	24.93	26.20	29.16	20.42	25.56
Aya-Vision-32B	23.90	24.10	19.54	20.53	14.55	34.60	20.00	22.46
Gemma-3-4B	26.32	17.45	22.19	22.29	27.65	40.60	21.67	25.45
Gemma-3-27B	47.37	30.47	38.74	30.79	47.40	43.87	35.00	39.09
Qwen-2.5-VL-32B	61.18	49.17	44.70	42.82	71.31	54.50	41.25	52.13
Qwen-2.5-VL-72B	58.33	49.86	48.34	40.47	64.66	53.95	46.67	51.75
Llama-4-Scout	59.65	44.88	45.03	38.71	69.02	49.05	37.08	49.06
Llama-4-Maverick	42.54	29.36	32.78	27.57	53.22	53.95	29.17	38.37
Mistral-Small-3.2-24B	34.65	30.28	28.48	28.15	30.98	39.78	30.83	31.88
Mistral-Medium-3	44.74	36.57	37.75	29.33	36.59	45.78	33.75	37.78
Model Mean	42.30	33.82	34.54	30.56	44.16	44.52	31.58	37.35
Non-thinking closed-source VLMs								
Gemini-2.5-Flash	82.46	67.04	78.15	63.05	85.24	71.39	52.08	71.34
Sonnet-4.0	64.25	41.08	53.31	44.87	48.44	58.04	44.17	50.58
GPT-4.1	46.27	43.21	44.37	44.87	69.85	66.21	46.25	51.58
Model Mean	64.33	50.44	58.61	50.93	67.84	65.21	47.50	57.83
Thinking closed-source VLMs								
GPT-5	91.23	76.45	85.10	72.73	89.40	75.20	58.75	78.41
Opus-4.6	83.33	67.31	79.47	71.26	90.85	73.84	56.25	74.62
o3	84.87	68.98	82.78	67.16	88.98	74.66	50.42	73.98
Gemini-3-Pro	**92.54**	**81.16**	**91.72**	**80.94**	**94.59**	**91.28**	**72.08**	**86.33**
Model Mean	87.99	73.48	84.77	73.02	90.96	78.75	59.38	78.34

Table 2: Thinking closed-source VLMs are substantially stronger than non-thinking closed-source VLMs across all seven domains of VMMU (78.34% vs. 57.83% mean), while open-source VLMs fall further behind (37.35%). Human numbers are estimated from the official score distributions of the Vietnamese national high school graduation exam (VnExpress, 2024).

4.2 Results

Results on open-source models Open-source VLMs achieve only 37.35% mean accuracy on VMMU (Tab. 2), drastically worse than the average Vietnamese high-school student (66.54%). Even the strongest open-source models (32B Qwen-2.5-VL-32B at 52.13% and 72B Qwen-2.5-VL-72B at 51.75%) remain well below average-student level on every task. Within some families, larger variants improve accuracy (Gemma, Mistral). However, scaling to larger models is not uniformly beneficial (e.g., Aya, Qwen and Llama). **Overall, open-source VLMs perform far below the average student**. We diagnose the main cause of this failure in Sec. 5.1.

Results on non-thinking closed-source models Non-thinking closed-source VLMs do better, but still fall short of the average student (57.83% mean vs. 66.54%; Tab. 2). Only Gemini-2.5-Flash exceeds the student-average reference (71.34%), and both Sonnet-4.0 and GPT-4.1 remain near 50%. **Non-thinking closed-source VLMs also underperform the average student**, so using a frontier proprietary VLM without thinking capabilities is not sufficient for VMMU.

Results on thinking closed-source models Enabling thinking in closed-source VLMs yields the largest single jump in our evaluation (78.34% mean vs. 57.83% for non-thinking proprietary; Tab. 2). Thinking models are the only group that exceeds the average-student threshold, with the strongest model (Gemini-3-Pro) reaching 86.33%. However, mean accuracy remains well below the best-test-taker reference (99.60%, Tab. 2), far from the reliability level that would be needed to deploy these models as tutors or for automated checking, even though frontier VLMs already perform at Olympiad-gold level on English-language tasks (Chervonyi et al., 2025; Luong and Lockhart, 2025; El-Kishky et al., 2025). Such a gap is expected given that VMMU questions are mainly STEM, driving, and IQ, which require multi-step reasoning rather than surface-level perception. **Thinking VLMs outperform the average-student baseline, but remain far from the expert level.**

Discussions Our conclusions are stable across runs: repeating the evaluation 5 times preserves model ranking and qualitative trends (Sec. A.3). The performance gap is not an artifact of answer parsing either: SOTA VLMs follow the required

output format reliably with near-perfect answer-extraction rates (Sec. A.5). We also find that non-thinking VLMs are substantially more over-confident and less well-calibrated than thinking VLMs (Sec. A.4). Shuffling multiple-choice options yields only small accuracy changes ($\Delta = -0.33\%$) and preserves model rankings, which shows that option-position bias does not drive our findings (Sec. A.1). Across models, performance varies substantially by domains: IQ is consistently hardest, while Geography and Math are comparatively easier (Tab. 2). On the other hand, even the strongest thinking closed-source model (Gemini-3-Pro at 86.33%) is roughly 13 points below the best-test-taker reference. Thus, current VLMs are not yet ready for automatic checking or tutoring on Vietnamese multimodal questions. The remainder of this section analyzes *why* models fail, in order to identify directions for improvement.

Model	BLEU (%)↑	F1 ↑	CER (%)↓	WER (%)↓
Open-source VLMs				
Aya-Vision-8B	2.72	0.20	229.30	253.09
Aya-Vision-32B	10.48	0.33	150.73	180.28
Gemma-3-4B	65.12	0.80	24.79	37.23
Gemma-3-27B	78.60	0.87	16.18	23.89
Qwen-2.5-VL-32B	77.50	0.88	19.05	24.17
Qwen-2.5-VL-72B	82.30	0.89	10.73	18.49
Llama-4-Maverick	70.59	0.81	48.77	53.91
Llama-4-Scout	38.95	0.54	217.92	211.96
Mistral-Medium-3	81.23	0.89	9.08	16.11
Mistral-Small-3.2-24B	72.28	0.84	19.39	28.41
Model Mean	57.98	0.70	74.59	84.75
SOTA closed-source VLMs				
Gemini-2.5-Flash	82.58	0.90	14.46	17.57
Sonnet-4.0	88.77	0.95	4.24	7.52
GPT-4.1	90.28	0.95	4.10	7.04
o3	94.28	0.97	4.09	5.25
Gemini-3-Pro	89.13	0.95	6.04	9.27
Model Mean	89.01	0.94	6.59	9.33

Table 3: OCR evaluation results across VLMs show a clear gap between open-source and SOTA closed-source models. This indicates that embedded-text OCR is a major bottleneck for open-source VLM performance.

5 Analyses and discussions

5.1 Open-source models fail at OCR while closed-source models do not

To understand the performance gap between open-source and closed-source models observed in Sec. 4, we analyze OCR capability, which is a prerequisite for any downstream multimodal reasoning over text-in-image inputs: a model that cannot reliably read the embedded Vietnamese text cannot be expected to answer correctly regardless of its reasoning ability.

Experiments We construct OCR ground truth via human-AI collaboration on a 210-question subset. We first ask o3 to transcribe all embedded text, including the prompt and answer options (see Figs. 11 and 12), then a human annotator verifies and edits the transcription to correct residual errors. We evaluate the same 10 open-source VLMs and 5 closed-source VLMs from Sec. 4 (see Sec. B for more details).

Results Across closed-source VLMs, OCR is consistently strong (mean BLEU 89.01%, F1 0.94%, CER 6.59%, WER 9.33%; Tab. 3). Reading Vietnamese text embedded in images is thus largely reliable for closed-source VLMs, so their downstream errors must instead come from multimodal grounding, i.e., locating the relevant visual evidence and reasoning over it, rather than from OCR failures. Therefore, we analyze these grounding/reasoning failures in the remaining subsections.

The picture is very different for open-source VLMs, which exhibit large and unstable OCR degradation. Several models collapse to unusable quality (e.g., Aya-Vision-8B and Llama-4-Scout both exceed 200% CER) and some fail to follow the OCR instruction entirely, hallucinating image content. Even the strongest open-source VLMs scores below the weakest proprietary one on every metric (Tab. 3). These findings suggest that **open-source models cannot reliably read the embedded Vietnamese text in the first place, and improving OCR and instruction-following is therefore a prerequisite for competitive open-source performance on** VMMU**.**

5.2 Separating text from visual evidence improves multimodal performance

Integrated question images require VLMs to simultaneously (i) read dense rendered text and (ii) locate and interpret the relevant visual evidence (e.g., charts, tables, maps, diagrams) within the same canvas. This can introduce distraction from text-heavy regions and make it harder to attend to the correct evidence, even when models are capable of reading the text (Yue et al., 2025). We test whether separating the textual prompt from the visual evidence improves reliability by making both components easier to process.

Experiments We create the separated setting (see Fig. 5) for the *full* VMMU benchmark by separating the textual questions and answer options from the visual evidence. Since SOTA VLMs already perform strongly on OCR tasks (see Sec. 5.1), we

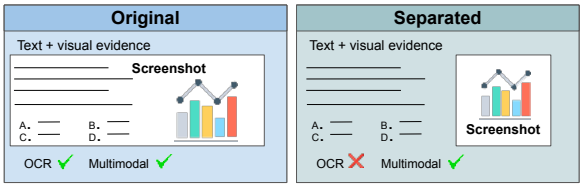


Figure 5: Comparison between the original and separated evaluation settings in VMMU. In the Original setting, the question text, answer options, and visual evidence are rendered together in a single image. In the separated setting, the question text and answer options are provided as plain text, while the visual evidence is provided as a separate image crop. In our evaluation, **only the content shown on the white canvas is passed to VLMs as image input**, and any text outside the white canvas is provided as plain text.

use o3 to automatically OCR-transcribe the textual elements for all 2,548 questions in VMMU, and then manually crop the required visual evidence regions into separate images. Due to evaluation cost, we evaluate the separated setting on a representative subset of models rather than the full model suite. This subset covers the main model groups in our study: open-source VLMs, non-thinking closed-source VLMs, and thinking closed-source VLMs. We use the prompts shown in Fig. 13.

The separated setting improves the performance of every model (+7.97 points on average; Tab. 4). The gains are largest for open-source VLMs (+9.65 points) and non-thinking closed-source VLMs (+8.56 points), while thinking closed-source VLMs improve only modestly (+2.10 points). This suggests that separating textual content from the image encourages models to rely more on visual reasoning while reducing distraction from text embedded in the image.

The benefit of the separated setting is a cleaner reasoning setup. Specifically, models no longer need to locate and transcribe text while simultaneously grounding the relevant visual evidence. Open-source VLMs benefit the most, likely because they are more affected by OCR and layout complexity, while non-thinking closed-source VLMs also gain substantially from reduced text-image interference. Thinking closed-source VLMs appear more robust to the original image format, so their smaller gains are consistent with partial saturation. Overall, **separating the textual questions and answer options from the visual evidence reduces text-in-image interference and makes grounded reasoning more reliable**, especially for open-source and non-thinking models.

Model	Original	Separated (Δ)
Open-source VLMs		
Gemma-3-27B	39.09	47.98 (+8.89)
Qwen-2.5-VL-72B	51.75	52.30 (+0.55)
Llama-4-Scout	49.06	52.55 (+3.49)
Llama-4-Maverick	38.37	52.07 (+13.70)
Mistral-Small-3.2-24B	31.88	45.82 (+13.94)
Mistral-Medium-3	37.78	55.08 (+17.30)
Model Mean	41.32	50.97 (+9.65)
Non-thinking closed-source VLMs		
Gemini-2.5-Flash	71.34	73.70 (+2.36)
Sonnet-4.0	50.58	64.44 (+13.86)
GPT-4.1	51.58	61.05 (+9.47)
Model Mean	57.83	66.40 (+8.56)
Thinking closed-source VLMs		
o3	73.98	76.59 (+2.61)
Gemini-3-Pro	86.33	87.92 (+1.59)
Model Mean	80.16	82.26 (+2.10)
Overall Mean	52.89	60.86 (+7.97)

Table 4: Separating textual content from visual evidence improves accuracy across all model groups. Open-source and non-thinking closed-source VLMs benefit the most, while thinking closed-source VLMs show smaller gains, suggesting that their performance is closer to saturation under the original image format.

5.3 English translation does not improve performance

A natural hypothesis is that performance is limited by language understanding, and that translating the prompt and answer options into English could improve results for models trained predominantly on English data (Mondshine et al., 2025; Etxaniz et al., 2024; Liu et al., 2025b).

Experiments We use the same subset of 210 ques-

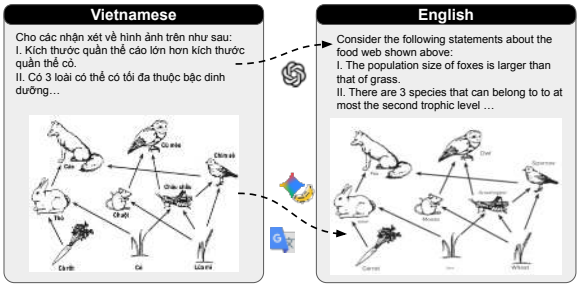


Figure 6: Our Vietnamese-to-English translation pipeline: o3 translates the question text, Google Translate (with Nano Banana 2 as fallback) translates the text in the image.

tions from Sec. 5.1, evaluated in the separated setting from Sec. 5.2. For the text component (i.e., questions and answer options), we use o3 to translate the Vietnamese text into English (see Fig. 6). For the text embedded within the visual evidence crop, we first apply Google Translate's image translation feature (Google, 2026). For cases where the translation is incorrect (6.19%), as identified through human verification, we use Nano Banana 2 (Google, 2025) to translate the image directly.

We evaluate this translation setting only on closed-source VLMs due to evaluation cost and because these models are the strongest models in our study. This provides a favorable test of the translation hypothesis: if translating Vietnamese inputs into English were beneficial, it should improve the performance of highly capable multilingual VLMs. We then pair the translated text with the translated visual evidence crop as input to each VLM to measure performance changes attributable to translation. Both the plain-text component and the text within the visual evidence crop are manually verified by native-speaker annotators to ensure translation quality. Only 6.19% of visual-crop translations required correction, which we report as a reliability signal for the translated condition.

Model	Vietnamese	English (Δ)
Gemini-2.5-Flash	82.38	76.19 (-6.19)
Sonnet-4.0	67.14	69.05 (+1.91)
GPT-4.1	67.14	66.19 (-0.95)
o3	81.43	78.10 (-3.33)
Gemini-3-Pro	92.38	88.57 (-3.81)
Model Mean	78.09	75.62 (-2.47)

Table 5: Translating Vietnamese questions into English reduces accuracy by -2.47 points on average. This suggests that English translation is not a reliable shortcut for VMMU, even for VLMs heavily trained on English data.

Results English translation consistently reduces the performance of most models (-2.47% accuracy; Tab. 5 and Fig. 32). We attribute this drop to two factors. First, translation introduces semantic drift, particularly for Vietnam-specific terminology in domains like traffic regulations, geography, and chemistry nomenclature, where direct English equivalents may not preserve the original meaning. Second, the consistent degradation across all models suggests that SOTA VLMs already possess sufficient Vietnamese understanding such that translation acts as a lossy intermediate step rather than a helpful shortcut.

This finding goes against the common assumption that translating low-resource language input to English improves performance (Mondshine et al., 2025; Etxaniz et al., 2024; Liu et al., 2025b), and suggests that **high-performance Vietnamese multimodal reasoning requires native-language training rather than translation-based workarounds**.

6 Conclusion

In this paper, we introduce VMMU, a Vietnamese multitask multimodal benchmark consisting of 2,548 questions across seven domains. Our evaluation shows that current VLMs still struggle with Vietnamese multimodal reasoning. The dominant source of error is *multimodal grounding and downstream reasoning*, rather than Vietnamese text recognition: closed-source models achieve reliable OCR performance, yet still make substantial errors on VMMU (Sec. 5.1). The benchmark remains challenging, although performance improves with test-time compute through thinking models and with the separated reformulation (Secs. 4 and 5.2).

Our additional analyses show that removing visual evidence leaves accuracy above chance, but this residual performance reflects standard exam test-taking behavior, such as option elimination and domain knowledge, rather than benchmark contamination. We further find that reliance on such priors accounts for roughly 42% of errors on average, where models commit to a prior answer despite the visual evidence being present (Secs. A.6 to A.8).

The remaining failures are also language-dependent. In particular, even when both the prompt and visual evidence are translated into English, accuracy consistently decreases (Sec. 5.3). This is because translation introduces semantic drift for Vietnam-specific terminology in traffic, geography, and chemistry, and adds a lossy intermediate step that strong VLMs do not need when they already handle Vietnamese reliably. These findings show that Vietnamese-native evaluation is necessary for measuring grounded multimodal reasoning in the settings users actually face. Future work should develop stronger joint text-vision reasoning under dense layouts and Vietnamese-centric instruction tuning that targets grounding and reasoning beyond what English-centric resources provide.

Limitations

This work focuses on evaluation and failure analysis rather than proposing a new training or fine-tuning method, as VMMU is the first benchmark of its kind for Vietnamese multimodal reasoning and large-scale Vietnamese multimodal training resources are not yet available. We hope VMMU will encourage the community to build larger native Vietnamese multimodal datasets for training and adaptation of VLMs in low-resource languages.

Ethical Considerations

Parts of this work were conducted via human-AI collaboration, where the human contributors are co-authors of this paper. This involvement posed minimal risk, and all participating authors agreed to contribute. The benchmark is curated from materials that are publicly published by governmental agencies and public educational institutions in Vietnam. All questions are used under Article 25 of Vietnam's Intellectual Property Law, which permits non-commercial research use while copyright remains with the original creators. Prior Vietnamese benchmarking efforts have followed similar practices (Bui et al., 2025; Dao et al., 2023). Furthermore, we will remove or replace any items upon request. We used large language models to assist with annotating parts of the dataset, with all outputs subsequently verified and corrected when needed by human annotators. We also used LLM-based tools for light editing and coding assistance.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00573160), and the 'Advanced GPU Utilization Support Program' funded by the Government of the Republic of Korea (Ministry of Science and ICT).

We also thank Rada Mihalcea (University of Michigan), Patrick Vibild, Hoang Minh Son (KAIST), and Duc-Vu Ngo for support, feedback and discussions of the earlier results. AV was supported by the Hyundai Motor Chung Mong-Koo Global Scholarship, and VTD was supported by Cohere Lab's Research Grant.

References

Anthropic. 2026. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6.

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

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4290–4300. IEEE.

Cuc Thi Bui, Nguyen Truong Son, Trang Van Truong, Viet Lam Phung, Pham Nhut Huy, Hoang Anh Le, Quoc Huu Van, Phong Nguyen-Thuan Do, Van Le Tran Truc, Duc Thanh Chau, and Le-Minh Nguyen. 2025. VMLU benchmarks: A comprehensive benchmark toolkit for vietnamese llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11495–11515. Association for Computational Linguistics.

Yuri Chervonyi, Trieu H. Trinh, Miroslav Olsák, Xiaomeng Yang, Hoang H. Nguyen, Marcelo Menegali, Junehyuk Jung, Junsu Kim, Vikas Verma, Quoc V. Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alphageometry2. *J. Mach. Learn. Res.*, 26:241:1–241:39.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Xuan-Quy Dao, Ngoc-Bich Le, The-Duy Vo, Xuan-Dung Phan, Bac Bien Ngo, Van-Tien Nguyen, Thi-My-Thanh Nguyen, and Hong Phuoc Nguyen. 2023. VNHSGE: vietnamese high school graduation examination dataset for large language models. *CoRR*, abs/2305.12199.

Rocktim Jyoti Das, Simeon Emilov 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7768–7791. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 Pro: the frontier of vision AI. https://blog.google/technology/developers/gemini-3-pro-vision/.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera Gilaberte, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, o3 contributors, Oleg Mürk, and 5 others. 2025. Competitive programming with large reasoning models. *CoRR*, abs/2502.06807.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in english? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 550–564. Association for Computational Linguistics.

Google. 2025. Nano banana 2. https://gemini.google/overview/image-generation/. Gemini AI image generator. Accessed: 2026-04-19.

Google. 2026. Google Translate – image translation. https://translate.google.com/?op=images. Accessed: 2026-04-19.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? EMMA: an enhanced multimodal reasoning benchmark. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025a. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6119–6136. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025b. Is translation all you need? A study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 9594–9614. 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*, volume 15064 of *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. 2024. 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.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Thang Luong and Edward Lockhart. 2025. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. Google DeepMind Blog.

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

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, and 30 others. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

MetaAI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

MistralAI. 2025a. Medium is the new large. https://mistral.ai/news/mistral-medium-3.

MistralAI. 2025b. Mistral Small 3. https://mistral.ai/news/mistral-small-3.

Atsuyuki Miyai, Shota Onohara, Jeonghun Baek, and Kiyoharu Aizawa. 2025. Jmmmu-pro: Image-based

japanese multi-discipline multimodal understanding benchmark via vibe benchmark construction. *Preprint*, arXiv:2512.14620.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond english: The impact of prompt translation strategies across languages and tasks in multilingual llms. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 1331–1354. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1037–1042. Association for Computational Linguistics.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. JMMMU: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 932–950. Association for Computational Linguistics.

OpenAI. 2025a. GPT-5 is here. https://openai.com/gpt-5/.

OpenAI. 2025b. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/.

OpenAI. 2025. Openai o3 and o4-mini system card. Technical report, OpenAI. Comprehensive system card detailing the capabilities, safety, and evaluation results for OpenAI o3 and o4-mini models.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2006–2017. Association for Computational Linguistics.

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

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18761–18799. Association for Computational Linguistics.

R. Smith. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633. IEEE Computer Society.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: measuring massive multitask language understanding in korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4076–4104. Association for Computational Linguistics.

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. 2025. Kazmmlu: Evaluating language models on kazakh, russian, and regional knowledge of kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14403–14416. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. 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 15894–15939. Association for Computational Linguistics.

VnExpress. 2024. Phổ điểm 9 môn thi tốt nghiệp THPT 2024 . https://vnexpress.net/pho-diem-9-mon-thi-tot-nghiep-thpt-2024-4770524.html.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025a. Vision language models are biased. *CoRR*, abs/2505.23941.

An Vo, Mohammad Reza Taesiri, Daeyoung Kim, and Anh Totti Nguyen. 2025b. B-score: Detecting biases in large language models using response history. In *Forty-second International Conference on Machine*

Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025. OpenReview.net.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. 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. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15134–15186. Association for Computational Linguistics.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, and 4 others. 2024. CMMMU: A chinese massive multi-discipline multimodal understanding benchmark. *CoRR*, abs/2401.11944.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Appendix for: VMMU: A Vietnamese Multitask Multimodal Understanding and Reasoning Benchmark

Contents

A Additional findings **14**
- A.1 Option-position bias does not drive our findings . . . 14
- A.2 Image-required questions are harder than text-sufficient ones . . 14
- A.3 Multiple runs do not change model ranking or our main conclusions . 14
- A.4 Non-thinking VLMs are overconfident . . . 15
- A.5 Proprietary and open-source VLMs follow the answer format reliably . . . 15
- A.6 Accuracy in the absence of visual evidence . . . 16
- A.7 No-image accuracy reflects test-taking heuristics, not contamination 16
- A.8 Over-reliance on priors causes failures when visual evidence conflicts 17

B Models and access details **19**

C Details of the prompts **20**

D Dataset curation details **22**
- D.1 Overview . . . 22
- D.2 Data source . . . 22
- D.3 Collection and cleanup . . . 22
- D.4 PDF rendering to page images . . 22
- D.5 Question boundary detection and extraction . . . 22
- D.6 Automatic modality classification (text-only vs. genuine multimodal) 22
- D.7 Human verification and ground-truth finalization . . . 23
- D.8 Release license . . . 23

E Qualitative results **25**
- E.1 Math: Qualitative results . . . 25
- E.2 Physics: Qualitative results . . . 27
- E.3 Chemistry: Qualitative results . . 29
- E.4 Biology: Qualitative results . . . 31
- E.5 Geography: Qualitative results . . 33
- E.6 Driving Test: Qualitative results . 35
- E.7 IQ Test: Qualitative results . . . 38
- E.8 Qualitative results for English translation . . . 41
- E.9 Qualitative results for removing visual evidence . . . 42

A Additional findings

A.1 Option-position bias does not drive our findings

Our earlier results report gaps between thinking and non-thinking VLMs (Sec. 4) and between original and separated settings (Sec. 5.2). A natural concern is whether these differences are artifacts of multiple-choice positional biases, where models over-prefer certain option letters or benefit from consistent option ordering (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024). We test this by re-running the separated setting with shuffled answer options.

Experiments Starting from the separated setting in Sec. 5.2, we randomly shuffle the order of answer options for each question while keeping the prompt text and visual evidence crop unchanged.

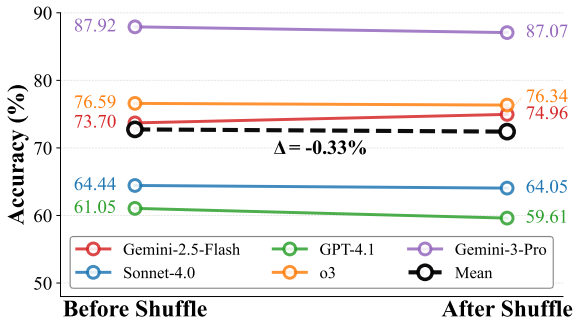


Figure 7: Accuracy before and after shuffling answer-option order. Changes are small ($\Delta = -0.33\%$) and model ranking is unchanged, showing limited sensitivity to option order.

Results Randomly shuffling the multiple-choice answer options produces only small accuracy changes across VLMs ($\Delta = -0.33\%$; Fig. 7), and the relative ranking of models is unchanged. The gaps reported in Secs. 4 and 5.2 are therefore not explained by option-ordering heuristics and reflect genuine differences in multimodal reasoning.

A.2 Image-required questions are harder than text-sufficient ones

Sec. 5.2 shows that *separated* setting improves accuracy, but this gain could be misattributed if a non-trivial portion of the benchmark is solvable from text alone. To confirm that VMMU primarily evaluates multimodal reasoning, we explicitly measure how often the image is necessary for answering.

Experiments We label each question as *image-required* or *text-sufficient* based on whether the correct answer can be determined from the textual prompt and options alone. We use o3 as a judge to make an initial decision for all 2,548 questions, then apply human verification to ensure labeling quality. This process yields 2,363 image-required questions (92.74%) and 185 text-sufficient questions (7.26%).

Model	Image required		Text sufficient	
	Original	Separated	Original	Separated
Gemini-2.5-Flash	69.92	72.75 (+2.83)	92.47	91.78 (-0.69)
Sonnet-4.0	49.82	63.22 (+13.40)	71.38	88.06 (+16.68)
GPT-4.1	51.31	60.30 (+8.99)	71.47	82.06 (+10.59)
o3	72.49	75.48 (+2.99)	95.26	93.70 (-1.56)
Gemini-3-Pro	**85.62**	**87.39** (+1.77)	**97.57**	**96.89** (-0.68)
Mean	65.83	71.83 (+6.00)	85.63	90.50 (+4.42)

Table 6: Separating the question text from the visual evidence (separated setting) improves performance on both image-required and text-sufficient questions (+4-6 points). Gains are largest for non-thinking VLMs (, ,), especially on image-required questions where separated setting yields sizable improvements over original setting. Thinking VLMs (,) benefit less and can slightly drop on text-sufficient questions, suggesting they already perform near saturation.

Results In Tab. 6, image-required questions are substantially harder than text-sufficient ones in both original (65.83% vs. 85.63%) and separated settings (71.83% vs. 90.50%). The separated setting yields comparable gains on both image-required and text-sufficient subsets (+4-6 percentage points). Despite this gain, accuracy on image-required questions on separated setting remains far from saturated (71.83%; Tab. 6). Grounded multimodal reasoning, therefore, remains the primary bottleneck.

A.3 Multiple runs do not change model ranking or our main conclusions

Since VLM decoding can be stochastic, we verify that our conclusions are not sensitive to a single evaluation run.

Experiments We sample 210 questions from 7 domains (30 questions per domain) and repeat evaluation for each proprietary VLM across 5 independent runs.

Model	Accuracy (↑) on 5 runs
Gemini-2.5-Flash	75.70 ± 3.64
Sonnet-4.0	54.67 ± 4.16
GPT-4.1	60.19 ± 4.03
o3	79.62 ± 4.03
Gemini-3-Pro	89.43 ± 1.73
Mean	71.92 ± 3.52

Table 7: Accuracy is stable across 5 independent runs (mean 71.92 ± 3.52), and model ranking is unchanged.

Results Accuracy is stable across runs (mean 71.92 ± 3.52), and the relative ordering remains unchanged: Gemini-3-Pro is best (89.43±1.73), followed by o3 (79.62±4.03) and Gemini-2.5-Flash (75.70±3.64), with GPT-4.1 (60.19±4.03) and Sonnet-4.0 (54.67±4.16) lower (Tab. 7). Thus, run-to-run variance does not affect our main conclusions.

A.4 Non-thinking VLMs are overconfident

Beyond accuracy, reliability depends on whether a model's expressed confidence matches its correctness. Overconfident errors are particularly problematic when VLMs are used for decision-making. We therefore measure self-reported confidence and calibration on VMMU.

Model	Confidence score (%) ↑	ECE score (%) ↓
Gemini-2.5-Flash	93.22	33.76
Sonnet-4.0	76.08	28.46
GPT-4.1	**96.76**	48.20
o3	56.72	**17.66**
Mean	80.69	32.02

Table 8: Non-thinking VLMs (, ,) report high self-confidence yet show substantially worse calibration (higher ECE), whereas the thinking VLM () is less confident but best calibrated.

Experiments We compute confidence and calibration on a single VMMU evaluation run (the same outputs in Sec. 4). Following Vo et al. (2025b), after the VLM selects an answer option, we ask in a second turn for a *verbalized confidence* score $p_i \in [0, 1]$ for its chosen option. We report the mean of these self-reported scores as the **confidence score**. We compute expected calibration error (ECE) directly from these per-example confidence scores by binning predictions into M confidence bins $\{B_m\}_{m=1}^{M}$:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \quad (1)$$

where $\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$ and $\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} p_i$. Lower ECE indicates better calibration.

Results Non-thinking VLMs (, ,) report very high confidence (e.g., GPT-4.1: 96.76%; Gemini-2.5-Flash: 93.22%), while the thinking model o3 reports much lower confidence (56.72) despite achieving the highest accuracy (Tab. 7). This mismatch is reflected in calibration: GPT-4.1 has the highest confidence but also the worst ECE (48.20), whereas o3 has the best ECE (17.66). Overall, these results show that non-thinking VLMs tend to be overconfident, while thinking VLMs are more conservative and better calibrated.

A.5 Proprietary and open-source VLMs follow the answer format reliably

End-to-end accuracy can be underestimated if models frequently produce outputs that are not parseable into a valid multiple-choice option. We therefore measure whether evaluation failures could be explained by answer-formatting issues rather than reasoning.

Experiments Using the same settings as the main evaluation (Sec. 4), we compute the TRUE extraction rate: the percentage of model responses from which a valid answer option can be extracted under our post-processing rules. We report this rate for all 5 proprietary VLMs and all 10 open-source VLMs.

Model	Extraction rate (%) ↑
Open-source VLMs	
Aya-Vision-8B	100.00
Aya-Vision-32B	100.00
Gemma-3-4B	94.10
Gemma-3-27B	97.47
Qwen-2.5-VL-32B	99.10
Qwen-2.5-VL-72B	99.66
Llama-4-Scout	94.50
Llama-4-Maverick	77.14
Mistral-Small-3.2-24B	93.98
Mistral-Medium-3	94.79
Model Mean	95.07
SOTA proprietary VLMs	
Gemini-2.5-Flash	98.09
Sonnet-4.0	**99.94**
GPT-4.1	99.66
o3	99.82
Gemini-3-Pro	99.14
Model Mean	99.33

Table 9: Both proprietary (mean 99.33%) and open-source (mean 95.07%) VLMs follow the required multiple-choice answer format reliably, so formatting is not a meaningful source of failures. The one notable exception is Llama-4-Maverick (77.14%).

Results Proprietary VLMs achieve near-perfect extraction rates (mean 99.33%; Tab. 9), and open-source VLMs are also reliable overall (mean 95.07%), with 9 of 10 above 93%. The only substantial outlier is Llama-4-Maverick (77.14%); all other open-source models exceed 93% and

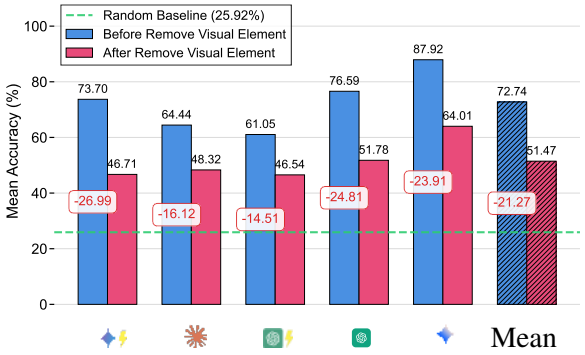


Figure 8: Accuracy with separated setting versus after removing the visual evidence. Accuracy drops by -21.27 percentage points on average, but remains above random chance (51.47% vs. 25.92%). This residual is decomposed in Sec. A.7, where we show it reflects domain knowledge and option-elimination heuristics rather than contamination.

several (8B Aya-Vision-8B, 32B Aya-Vision-32B) achieve 100%. Thus, answer-format compliance is not a meaningful contributor to the remaining failures for either group.

A.6 Accuracy in the absence of visual evidence

Having confirmed that answer extraction (Sec. A.5) and calibration (Sec. A.4) do not explain the errors, we now examine whether visual evidence is actually essential for solving VMMU. A common concern with multiple-choice multimodal benchmarks is that some accuracy may be recoverable without looking at the image, either through option-elimination heuristics, domain knowledge, or (in the worst case) contamination. To measure this residual, we evaluate all proprietary VLMs without visual evidence. A companion analysis (Sec. A.7) then decomposes what drives the residual and shows that it reflects expected test-taking behavior rather than contamination.

Experiments Starting from separated setting, we remove the visual evidence crop and provide only the text question and answer options (see Fig. 14).

Results Removing visual evidence significantly reduces accuracy for all VLMs (-21.27 percentage points; Fig. 8), but performance does not collapse to chance (51.47% vs. 25.92%). Notably, Gemini-3-Pro still achieves 64.01% accuracy without visual evidence. Qualitatively, remaining correct answers are typically produced by option elimination, domain knowledge, or incidental guessing (Figs. 33, 36 and 37), mirroring how human students answer multiple-choice exam items without full information. In contrast to option shuffling (Sec. A.1), which leaves accuracy essentially unchanged, this text-only ablation exposes the extent of such test-taking shortcuts. Sec. A.7 quantifies the contribution of these shortcuts and rules out contamination as an explanation.

A.7 No-image accuracy reflects test-taking heuristics, not contamination

The 51.47% mean accuracy observed when visual evidence is removed (Sec. A.6) could be misread as evidence that VMMU is solvable from text alone or that the benchmark is contaminated. We perform three diagnostic checks, plus a hard-subset robustness analysis, that rule out this interpretation and attribute the residual to expected test-taking behavior.

Experiments Starting from the no-image outputs of Sec. A.6, we manually categorize each response (human verification with LLM assistance) for two representative VLMs (Sonnet-4.0, Gemini-3-Pro) into 4 types: (i) *heuristic / domain-based* answers that eliminate implausible options or apply subject knowledge; (ii) *random guessing*; (iii) *self-reported memorization*, where the model explicitly claims to recall the question from exam repositories; and (iv) *non-image-required* items. We then (a) report response-type breakdowns (Tab. 10), (b) measure accuracy *within* each response type (Tab. 11), and (c) manually verify self-reported memorization claims and re-compute accuracy after filtering confirmed matches (Tab. 12).

Model	Heuristic/domain	Random	Self-memorization	Non-image-required
Sonnet-4.0	63.11	29.08	0.00	7.26
Gemini-3-Pro	77.83	4.67	9.93	7.26

Table 10: Response-type breakdown (%) on no-image outputs. Answers are dominated by domain knowledge and option-elimination heuristics, while self-reported memorization is rare.

Results We find that **(1) responses are dominated by heuristics, not memorization.** When the image is removed, 63.11% (Sonnet-4.0) and 77.83% (Gemini-3-Pro) of responses are driven by domain knowledge and option elimination (see Figs. 33, 34 and 36) or random (Fig. 37), while self-reported memorization (see Fig. 35) accounts for 0.00% and 9.93% respectively (Tab. 10). **(2) Self-reported memorization does not reliably help.** Even when Gemini-3-Pro explicitly claims to recall a question, it answers correctly only 54.15% of the time (Tab. 11), showing that such claims are noisy and are not equivalent to having memorized the cor-

rect answer. **(3) Verified contamination is small and does not affect conclusions.** Manually verifying all 253 self-reported memorization claims yields only 56 exact matches with the original exam source (2.1% of VMMU). Removing these items changes mean accuracy by 0.06% (Tab. 12), within the run-to-run variance reported in Sec. A.3.

Model	Heuristic/domain	Random	Self-memorization	Non-image-required
Sonnet-4.0	52.30	36.30	–	77.84
Gemini-3-Pro	67.73	25.21	54.15	94.05

Table 11: Accuracy (%) within each response type. Even when Gemini-3-Pro self-reports memorization, it answers correctly only 54.15% of the time, indicating these self-reports are noisy and do not reliably map to memorized ground-truth answers.

Model	Before filtering	After filtering (Δ)
Gemini-2.5-Flash	71.34	71.48 (+0.14)
Sonnet-4.0	50.58	50.44 (-0.14)
GPT-4.1	51.58	51.71 (+0.13)
o3	73.98	74.03 (+0.05)
Gemini-3-Pro	86.33	86.43 (+0.10)
Mean	**66.76**	**66.82** (+0.06)

Table 12: Overall accuracy (%) on before and after removing the 56 items (2.1% of VMMU) whose self-reported memorization was verified against the original exam source. Mean accuracy changes by only 0.06 pp, within run-to-run variance (Sec. A.3).

Model	Full set (2,548)	Hard subset (2,051) (Δ)
Gemini-2.5-Flash	71.34	66.92 (-4.42)
Sonnet-4.0	50.58	46.35 (-4.23)
GPT-4.1	51.58	46.74 (-4.84)
o3	73.98	69.94 (-4.04)
Gemini-3-Pro	86.33	84.15 (-2.18)
Mean	**66.76**	**62.82** (-3.94)

Table 13: Accuracy (%) on the full benchmark versus on the hard subset that conservatively removes the 497 questions all 5 proprietary VLMs answer correctly in the no-image condition. Accuracy drops for every model while preserving their relative ranking, confirming the benchmark remains challenging after the filter.

Hard-subset robustness. As a stricter robustness check, we construct a hard subset by conservatively removing every question that *all five* proprietary VLMs answer correctly in the no-image condition (n=497). On the remaining 2,051 questions, mean accuracy drops from 66.76% to 62.82% (-3.94 percentage points), with per-model drops ranging from 2.18 to 4.84 pp (Tab. 13). Crucially, the relative ranking of models is preserved (Gemini-3-Pro > o3 > Gemini-2.5-Flash > GPT-4.1 > Sonnet-4.0), and the benchmark remains far from saturated.

Taken together, these analyses indicate that above-chance no-image accuracy reflects expected exam-style test-taking behavior (option elimination, domain knowledge, occasional guessing), not contamination or a failure of the benchmark's multimodal design.

A.8 Over-reliance on priors causes failures when visual evidence conflicts

Beyond sustaining above-chance no-image accuracy, reliance on priors also has a behavioral cost: when the image contradicts the model's prior expectations, the model may anchor on the prior and ignore the visual evidence (Vo et al., 2025a).

Experiments Using the per-question predictions from the no-image condition (Sec. A.6) and the separated setting (Sec. 5.2), we define a *sticky-prior error* as a question where (i) the prediction is incorrect and (ii) it matches the no-image prediction. Such cases indicate that providing the visual evidence did not change the model's answer, consistent with the model committing to a text-only prior and disregarding the image. We report the sticky-prior error rate per model, both as a fraction of all questions and as a fraction of all errors.

Model	Separated error rate	Sticky-prior errors	Sticky/All errors
Gemini-2.5-Flash	26.30	8.21	33.47
Sonnet-4.0	35.56	14.39	42.64
GPT-4.1	38.95	15.18	41.90
o3	23.41	9.57	44.59
Gemini-3-Pro	12.08	5.54	**47.16**
Mean	**27.26**	**10.58**	**41.95**

Table 14: Sticky-prior failure analysis (%). *Sticky-prior errors* are errors whose predicted option equals the no-image prediction, indicating that providing the image did not move the model off its prior. On average, 41.95% of all errors are sticky-prior, i.e., nearly half of the remaining errors are cases where the visual evidence did not change the model's answer.

Results Across the 5 proprietary VLMs, **41.95% of all errors are sticky-prior errors on average** (Tab. 14): nearly half of the remaining failures are cases where providing the image did not move the model off its prior, consistent with the visual input being effectively ignored. Notably, the strongest model (Gemini-3-Pro, with the lowest overall error rate at 12.08%) has the *highest* sticky-prior share (47.16%), indicating that even top VLMs are disproportionately susceptible to prior over-

reliance when they do fail. Gemini-2.5-Flash is the only model below 40% (33.47%), suggesting its errors are more diverse and less dominated by prior commitment. The memorization excerpt in Sec. A.7 is a clear qualitative instance: the model commits to a prior answer inferred from familiar exam sources and does not revise it when reading the figure. Similar patterns recur in domains with strong conventional expectations (e.g., standard diagrams in physics and chemistry). This finding is completely aligned with (Vo et al., 2025a), which showed that VLMs are heavily biased due to relying on knowledge priors that even ignore visual evidence.

Two implications follow. First, improving VMMU performance is not solely a matter of better perception. It also requires models that downweight priors when the image carries contradicting evidence. Second, the no-image ablation (Sec. A.6) should be read not as evidence of contamination, but as a diagnostic of how strong the text-only prior is, with Tab. 14 translating that prior into a measurable rate of behavioral failure under visual conflict.

B Models and access details

Model	Model Name	Temperature	Platform
	Open-source VLMs		
Aya-Vision-8B (Üstün et al., 2024)	c4ai-aya-vision-8b	0.0	Cohere
Aya-Vision-32B (Üstün et al., 2024)	c4ai-aya-vision-32b	0.0	Cohere
Gemma-3-4B (Mesnard et al., 2024)	gemma-3-4b-it	0.0	Google AI Studio
Gemma-3-27B (Mesnard et al., 2024)	gemma-3-27b-it	0.0	Google AI Studio
Mistral-Small-3.2-24B (MistralAI, 2025b)	mistral-small-3.2-24b-instruct	0.0	OpenRouter
Mistral-Medium-3 (MistralAI, 2025a)	mistral-medium-3	0.0	OpenRouter
Qwen-2.5-VL-32B (Bai et al., 2025)	qwen/qwen2.5-vl-32b-instruct	0.0	OpenRouter
Qwen-2.5-VL-72B (Bai et al., 2025)	qwen/qwen2.5-vl-72b-instruct	0.0	OpenRouter
Llama-4-Maverick (MetaAI, 2025)	llama-4-maverick	0.0	OpenRouter
Llama-4-Scout (MetaAI, 2025)	llama-4-scout	0.0	OpenRouter
	Non-thinking proprietary VLMs		
Gemini-2.5-Flash (Comanici et al., 2025)	gemini-2.5-flash	0.0	Google AI Studio
Sonnet-4.0	claude-sonnet-4-20250514	0.0	Anthropic
GPT-4.1 (OpenAI, 2025b)	gpt-4.1-2025-04-14	0.0	OpenAI
	Thinking proprietary VLMs		
GPT-5 * (OpenAI, 2025a)	gpt-5-2025-08-07	N/A	OpenAI
Opus-4.6 * (Anthropic, 2026)	claude-opus-4-6	N/A	Anthropic
o3 * (OpenAI, 2025)	o3-2025-04-16	N/A	OpenAI
Gemini-3-Pro (DeepMind, 2025)	gemini-3.0-pro	N/A	Google AI Studio/OpenRouter

*reasoning_effort: medium (default thinking mode setting)

Table 15: VLMs configuration and platform details.

C Details of the prompts

Original setting prompt

English prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}.
Vietnamese prompt: Trả lời câu hỏi trắc nghiệm trong hình. Ghi đáp án bạn chọn trong dấu ngoặc nhọn với chữ cái tương ứng, ví dụ: {Z}.

Figure 9: The Vietnamese and English prompts used for multiple-choice question answering tasks.

Original setting prompt (Multiple answers)

English prompt: Answer the multiple-choice questions in the image. List your choices in curly brackets using option letters, e.g. {Y, Z}.
Vietnamese prompt: Trả lời các câu hỏi trắc nghiệm trong hình. Liệt kê các đáp án bạn chọn trong dấu ngoặc nhọn, sử dụng chữ cái, ví dụ: {Y, Z}.

Figure 10: The Vietnamese and English prompts used for question answering tasks on images containing multiple questions.

OCR prompt

English prompt: Perform OCR on the image and return the result in this JSON format:

{
"Text": "Extract only the textual content, excluding figures and tables. Use Unicode math symbols (e.g. $\surd$, π) and '/' for fractions (e.g. 1/5).",
"Image description": "If figures or tables are present, describe them in Vietnamese; otherwise, leave blank."
}

Vietnamese prompt: Thực hiện OCR trên hình ảnh và trả về kết quả theo định dạng JSON sau:

{
"Text": "Chỉ trích xuất phần văn bản, bỏ qua hình vẽ và bảng biểu. Sử dụng ký hiệu toán học Unicode (e.g. $\surd$, π) và '/' cho phân số (ví dụ: 1/5).",
"Image description": "Nếu có hình vẽ hoặc bảng biểu, mô tả chúng bằng tiếng Việt; nếu không có, để trống."
}

Figure 11: The Vietnamese and English prompts used for OCR tasks on individual question images in the dataset.

OCR prompt (Multiple Answers)

English prompt: Perform OCR on the image and return the result in this JSON format:

{
"Text": "Extract only the textual content, excluding figures and tables. Use Unicode math symbols (e.g. $\sqrt{}$, π) and '/' for fractions (e.g. 1/5).",
"Image description": "If figures or tables are present, describe them in Vietnamese; otherwise, leave blank."
}

Vietnamese prompt: Thực hiện OCR trên hình ảnh và trả về kết quả theo định dạng JSON sau:

{
"Text": "Chỉ trích xuất phần văn bản, bỏ qua hình vẽ và bảng biểu. Sử dụng ký hiệu toán học Unicode (e.g. $\sqrt{}$, π) và '/' cho phân số (ví dụ: 1/5).",
"Image description": "Nếu có hình vẽ hoặc bảng biểu, mô tả chúng bằng tiếng Việt; nếu không có, để trống."
}

Figure 12: The prompts used for OCR tasks on images containing two questions in the dataset.

Separated setting prompt

English prompt: [Question content]
Based on the information in the image, answer the multiple-choice question above. Write your chosen answer in double curly brackets with the corresponding letter, for example: {{Z}}.
Vietnamese prompt: [Nội dung câu hỏi]
Dựa vào thông tin trong hình để trả lời câu hỏi trắc nghiệm ở trên. Ghi đáp án bạn chọn trong dấu ngoặc nhọn kép với chữ cái tương ứng, ví dụ: {{Z}}.

Figure 13: The Vietnamese and English prompts used for question answering tasks on separated setting questions.

Separated setting prompt (after removing visual elements)

English prompt: [Question content] Based on the information provided, answer the multiple-choice question above. In all cases, you must select exactly one valid answer using the corresponding letter, even when data is insufficient or certain inference cannot be made. Write your chosen answer in double curly brackets with the corresponding letter, for example: {{Z}}.
Vietnamese prompt: [Nội dung câu hỏi]
Dựa vào thông tin được cung cấp để trả lời câu hỏi trắc nghiệm ở trên. Trong mọi trường hợp, bạn phải chọn đúng một đáp án hợp lệ bằng chữ cái tương ứng, kể cả khi thiếu dữ liệu hoặc không thể suy luận chắc chắn. Ghi đáp án bạn chọn trong dấu ngoặc nhọn kép với chữ cái tương ứng, ví dụ: {{Z}}.

Figure 14: The Vietnamese and English prompts used for question answering tasks when image is required but not provided in the experiment.

D Dataset curation details

D.1 Overview

We curate VMMU with a semi-automated pipeline that converts heterogeneous public documents into individual question images, then applies human verification for final quality control (Fig. 3). The pipeline has 3 goals: (i) reliably extract question boundaries (including multi-page items), (ii) retain only *genuine* multimodal questions (where non-text visual evidence is necessary), and (iii) produce consistent artifacts for evaluation (original and separated settings).

D.2 Data source

We collect question sets from publicly available repositories and official releases (Tab. 16).

Tasks 1-5: Mathematics, Physics, Chemistry, Biology, Geography We systematically collected questions using Selenium web scraping from public repositories. Our dataset comprises questions from 3 primary sources: (1) Official Exam Tests from the Vietnamese National High School Graduation Examination, (2) Official Sample Tests issued by the Ministry of Education and Training (MOET), and (3) Mock Exam Tests from Provincial Departments of Education.

Task 6: Driving Test The Vietnamese Ministry of Transport (MOT) maintains official question banks containing 250 questions for A1 licenses (motorcycles) and 600 questions for B2 licenses (automobiles). To achieve A1 level requires a minimum of 21/25 correct answers (84%), while B2 level requires 32/35 correct answers (91.4%) and must not make errors on critical questions. From this broader collection of 850 total questions, we selected 367 multimodal multiple-choice questions that contain visual elements integrated with Vietnamese text. Each selected question presents traffic scenarios through illustrations, accompanied by Vietnamese text describing the situation and 2-4 multiple-choice options.

Task 7: IQ Test We collected 240 IQ test questions from https://vndoc.com through manual screenshot capture. These questions span various cognitive domains (e.g., spatial reasoning, pattern completion, logical sequences, numerical reasoning, abstract thinking). Each question presents 4-8 multiple-choice options embedded within the image.

D.3 Collection and cleanup

Crawling and normalization We use a Selenium-based crawler to download documents in .pdf/.doc/.docx. All Word files are converted to PDF to standardize downstream processing.

Removing trailing solutions/answer keys Many documents append answer keys or solution explanations after the questions. We drop these sections with a lightweight PDF segmentation step (pypdf): we scan pages for high-precision termination cues (e.g., "HẾT"; *THE END*) and fallback cues marking auxiliary sections (e.g., "BẢNG ĐÁP ÁN"; *ANSWER KEY*, and "LỜI GIẢI CHI TIẾT"; *DETAILED SOLUTION*). We keep pages from the beginning up to the first matched cue.

D.4 PDF rendering to page images

We render each PDF page to a PNG image using pdf2image at 300 DPI. Each page image stores document/page identifiers and pixel dimensions in a metadata file to support traceability during cropping.

D.5 Question boundary detection and extraction

Detecting question starts We run Tesseract OCR (Smith, 2007) to obtain line-level text and bounding boxes. We then detect question starts via Vietnamese layout markers using regex patterns (e.g. "Câu\s?\d+[:.]", where "Câu" means *Question*). We additionally detect group directives (e.g., "Dùng dữ kiện cho câu 1–5"; *Use the given data for Questions 1–5*) to preserve shared context for question groups.

Cropping and multi-page stitching Given detected markers, we construct a crop plan that maps each question to a page region. If a question continues onto the next page, we stitch the cropped fragments vertically (cv2.vconcat) to form a single question image.

D.6 Automatic modality classification (text-only vs. genuine multimodal)

To filter out weakly multimodal items (text-only screenshots), we classify each extracted question image based on the presence of *non-text visual structures* (e.g., charts/diagrams/tables/maps). We binarize the image (grayscale + adaptive thresholding), detect connected components via contours, and apply simple geometric rules to separate text-like blocks from non-text graphics. If at least one

Domain(s)	Source	Platform
Mathematics, Physics, Chemistry, Biology, Geography	Official exam tests from the Vietnamese National High School Graduation Examination. Official sample tests issued by the Ministry of Education and Training (MOET). Mock exam tests released by Provincial Departments of Education.	https://thuvienhoclieu.com https://vndoc.vn
Driving Test	Official question banks and regulation-aligned materials from the traffic police department and the Ministry of Transportation (MOT).	https://daotaohoclaixeoto.com https://tracuuphapluat.info
IQ Test	Manual collection from publicly available sources.	https://vndoc.com

Table 16: VMMU data sources by subject group.

sufficiently large, non-text-like region is detected, we label the item as multimodal, otherwise, it is labeled text-only.

D.7 Human verification and ground-truth finalization

Automated extraction is followed by human verification to ensure correct crop boundaries, correct modality labels, and correct answer keys.

Review interface We built a review tool (React + TailwindCSS) that displays extracted question images alongside the original source document for side-by-side inspection (Fig. 15). Annotators can: (i) accept/reject the multimodal label, (ii) flag cropping issues for manual repair, and (iii) record the correct answer.

Cross-verification For final answer keys, 3 native speakers independently verify each label. We accept an answer only under unanimous agreement, and disagreements are resolved via discussion and re-checking against the original document.

All annotators are co-authors of this paper, aged between 20 and 30, and provided informed consent for their participation and the intended use of their annotations.

D.8 Release license

VMMU is released under the Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0, https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial use with attribution. Copyright remains with the original exam-material creators, as noted in the Ethical Considerations.

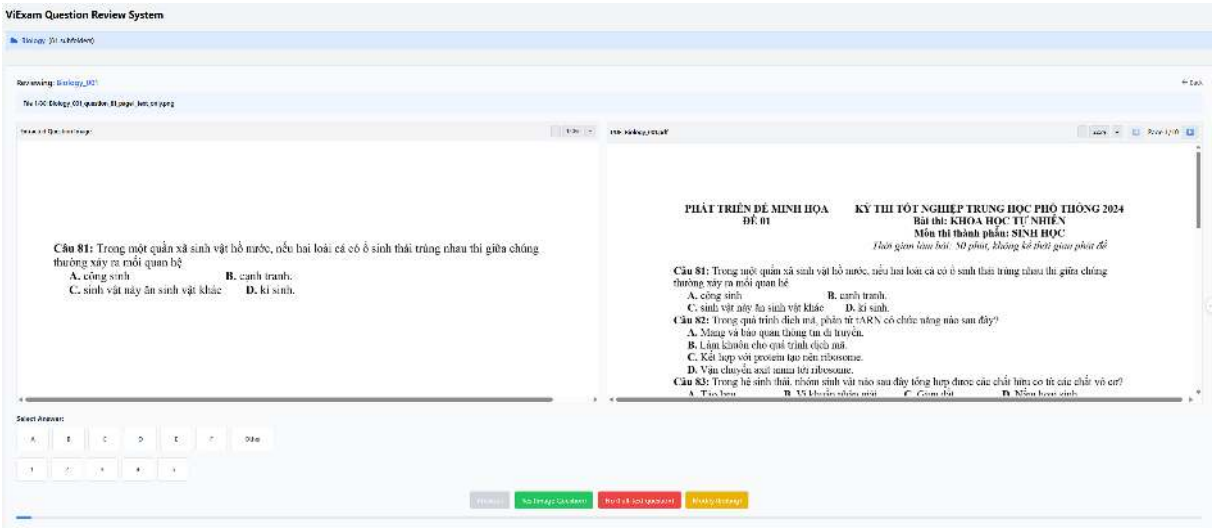
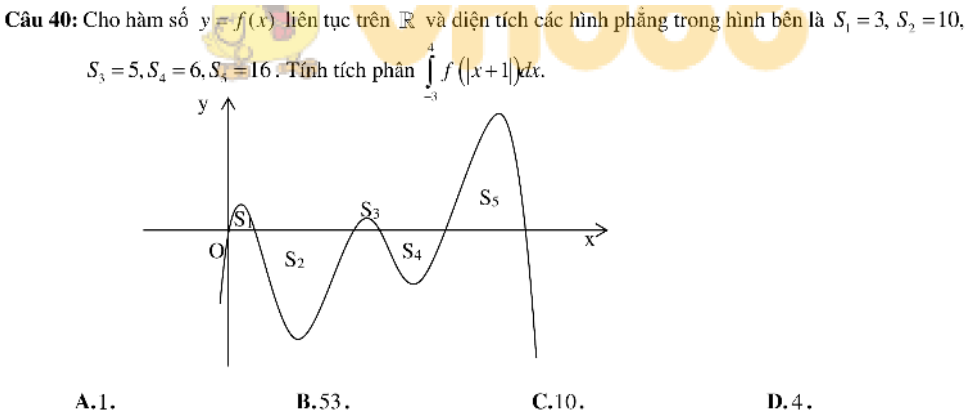


Figure 15: VMMU question review system for human verification and ground-truth annotation. The interface shows an automatically extracted question crop (left) alongside the original source document rendered in-browser (right) for side-by-side checking. Annotators validate whether visual evidence is required, flag cropping errors, and record the correct answer choice.

E Qualitative results

E.1 Math: Qualitative results

Math



Question: Given a continuous function $y = f(x)$ on $\mathbb{R}$ with the graph shown and the areas labeled in the figure $S_1 = 3,\ S_2 = 10,\ S_3 = 5,\ S_4 = 6,\ S_5 = 16$. Compute $\int_{-3}^{4} f(|x+1|)\ dx$.

A. 1 B. 53 C. 10 D. 4

Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}

Ground truth: D ✓

Gemini-2.5-Flash: {5} ✗

Sonnet-4.0: {C} ✗

GPT-4.1: {B} ✗

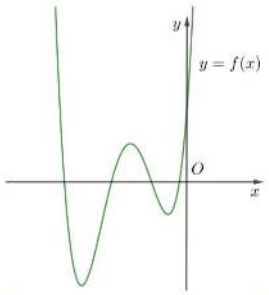
o3: {B} ✗

Gemini-3-Pro: {A} ✗

Figure 16: VLMs fail to correctly evaluate an integral defined by a graphical function combined with an absolute-value transformation. Notably, Gemini-2.5-Flash also fails to produce an answer in the required format.

Math

Câu 50: Cho hàm số $F(x)$ có $F(0)=0$. Biết $y=F(x)$ là một nguyên hàm của hàm số $y=f(x)$ đồ thị như hình vẽ. Số điểm cực trị của hàm số $G(x)=\left|F\left(x^6\right)-x^3\right|$ là



A. 4. **B.** 5. **C.** 6. **D.** 3.

Question: Let $F(x)$ satisfy $F(0) = 0$. Suppose $y = F(x)$ is an antiderivative of $y = f(x)$ whose graph is shown. The number of local extrema of the function $G(x) = \left|F(x^6) - x^3\right|$ is
A. 4 B. 5 C. 6 D. 3
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: D ✓

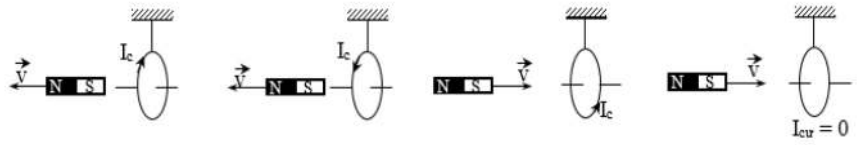
Gemini-2.5-Flash: {1} ✗
Sonnet-4.0: {B} ✗
GPT-4.1: {C} ✗
o3: {B} ✗
Gemini-3-Pro: {A} ✗

Figure 17: VLMs fail to correctly count the local extrema of an absolute-value composite function defined through an antiderivative and a provided graph. Gemini-2.5-Flash additionally fails to comply with the required answer format.

E.2 Physics: Qualitative results

Physics

Câu 29[TH]: Hình vẽ nào dưới đây biểu diễn **đúng** chiều dòng điện cảm ứng trong vòng dây?



Hình 1 **Hình 2** **Hình 3** **Hình 4**

A. Hình 1. **B.** Hình 2. **C.** Hình 3. **D.** Hình 4.

Question: Which diagram below correctly represents the direction of the induced current in the loop?

A. Figure 1 B. Figure 2 C. Figure 3 D. Figure 4

Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}

Ground truth: A ✓

Gemini-2.5-Flash: {C} ✗

Sonnet-4.0: {D} ✗

GPT-4.1: {B} ✗

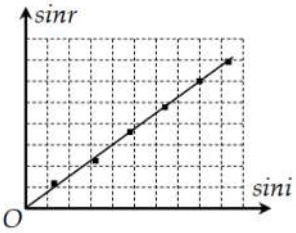
o3: {B} ✗

Gemini-3-Pro: {C} ✗

Figure 18: VLMs struggle with electromagnetic induction: identifying the correct direction of induced current from the motion of a magnet relative to a loop.

Physics

Câu 29: Để đo chiết suất n của một chất lỏng trong suốt X, một bạn đã chiếu chùm sáng laze hẹp vào một cái bể (có chứa chất lỏng X) với góc tới i, sau đó tiến hành đo góc khúc xạ r. Thay đổi góc tới i để thu được góc khúc xạ r khác nhau. Hình bên là đồ thị mô tả mối quan hệ giữa sini và sinr. Biết độ chia nhỏ nhất trên các trục của đồ thị có giá trị bằng nhau. Giá trị trung bình của n thu được trong thí nghiệm trên là



A. 1,33. **B.** 1,50. **C.** 1,25. **D.** 1,62.

Question: To measure the refractive index n of a transparent liquid X, a narrow laser beam is incident on a container (filled with liquid X) at angle i, and the refraction angle r is measured. Varying i yields different r. The plot shows the relationship between $\sin i$ and $\sin r$, and the smallest grid spacing on both axes is the same. The average value of n obtained from the experiment is
A. 1.33 B. 1.50 C. 1.25 D. 1.62
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: A ✓

Gemini-2.5-Flash: {C} ✗
Sonnet-4.0: {C} ✗
GPT-4.1: {B} ✗
o3: {B} ✗
Gemini-3-Pro: {B} ✗

Figure 19: VLMs struggle with interpreting refraction graphs, estimating the refractive index of a liquid from a plot of $\sin i$ versus $\sin r$.

E.3 Chemistry: Qualitative results

Chemistry

Câu 27: Kết quả thí nghiệm của các dung dịch X, Y, Z, T với thuốc thử được ghi ở bảng sau:

Thuốc thử / Chất	X	Y	Z	T
Dung dịch $AgNO_3/NH_3$, đun nhẹ	Không có kết tủa	Ag↓	Không có kết tủa	Ag↓
$Cu(OH)_2$, lắc nhẹ	$Cu(OH)_2$ không tan	Dung dịch xanh lam	Dung dịch xanh lam	Dung dịch xanh lam
Nước brom	Mất màu nước brom và có kết tủa trắng xuất hiện	Mất màu nước brom	Không mất màu nước brom	Không mất màu nước brom

Các dung dịch **X**, **Y**, **Z**, **T** lần lượt là

A. Phenol, axit fomic, saccarozơ, glucozơ. **B.** Anilin, glucozơ, glixerol, saccarozơ.
C. Anilin, glucozơ, glixerol, fructozơ. **D.** Phenol, glucozơ, glixerol, axit fomic.

Question: The experimental results for solutions X, Y, Z, T with different reagents are shown in the table. Identify the substances corresponding to X, Y, Z, T in order.
A. Phenol, formic acid, sucrose, glucose. B. Aniline, glucose, glycerol, sucrose. C. Aniline, glucose, glycerol, fructose. D. Phenol, glucose, glycerol, formic acid.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: B ✓

Gemini-2.5-Flash: {C} ✗
Sonnet-4.0: {D} ✗
GPT-4.1: {A} ✗
o3: {D} ✗
Gemini-3-Pro: {C} ✗

Figure 20: VLMs fail to reason over a multi-reagent qualitative analysis table to identify unknown organic compounds.

Chemistry

Câu 9: Sơ đồ pin Galvani tổng quát được kí hiệu như sau:

(anode) X | X^{x+} || Y^{y+} | Y **(cathode)** hay đơn giản là pin Galvani **X – Y**.

Cho biết sức điện động chuẩn của các pin sau:

Pin điện hóa	**X – Y**	**M – Y**	**M – Z**
Sức điện động chuẩn (V)	0,20	0,60	0,30

Sự sắp xếp nào sau đây đúng với tính khử của các kim loại?

A. Y < X < Z < M. **B.** X < Y < M < Z. **C.** M < Z < Y < X. **D.** X < Y < Z < M.

Question: A galvanic cell is denoted by (anode) $X \mid X^{x+} \parallel Y^{y+} \mid Y$ (cathode), i.e., the cell $X - Y$. Given the standard EMFs of the following cells: $X - Y = 0.20\,\text{V}$, $M - Y = 0.60\,\text{V}$, and $M - Z = 0.30\,\text{V}$. Which ordering correctly describes the reducing strength of the metals?
A. $Y < X < Z < M$. B. $X < Y < M < Z$. C. $M < Z < Y < X$. D. $X < Y < Z < M$.

Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}

Ground truth: C ✓

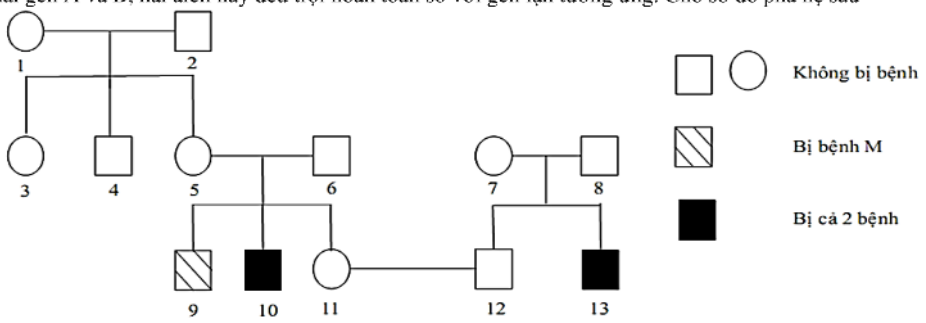
Gemini-2.5-Flash: {A} ✗
Sonnet-4.0: {B} ✗
GPT-4.1: {D} ✗
o3: {A} ✗
Gemini-3-Pro: {A} ✗

Figure 21: VLMs fail to correctly reason over a pedigree involving X-linked recessive mutations and genetic recombination, resulting in incorrect evaluation of multiple statements.

E.4 Biology: Qualitative results

Biology

Câu 38. Ở người, bệnh M và bệnh N là hai bệnh do đột biến gen lặn nằm ở vùng không tương đồng trên nhiễm sắc thể giới tính X, khoảng cách giữa 2 gen là 40 cM. Người bình thường mang hai gen A và B, hai alen này đều trội hoàn toàn so với gen lặn tương ứng. Cho sơ đồ phả hệ sau



Biết rằng không phát sinh các đột biến mới ở tất cả các cá thể trong phả hệ. Có bao nhiêu phát biểu sau đây là đúng?
I. Tất cả những người nam trong phả hệ đều bị bệnh.
II. Có thể xác định được chính xác kiểu gen của 9 người.
III. Nếu người số 13 có vợ không bị bệnh nhưng bố của vợ bị cả hai bệnh thì xác suất sinh con gái bị bệnh là 35%.
IV. Xác suất sinh con đầu lòng bị cả hai bệnh của cặp vợ chồng số 11- 12 là 4,5%.
A. 1. **B.** 2. **C.** 3. **D.** 4.

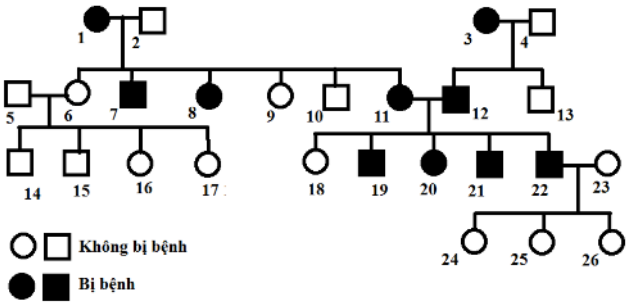
Question: In humans, diseases M and N are caused by recessive mutations in two non-homologous regions on the X chromosome, with the distance between the two genes being 40 cM. Normal individuals carry alleles A and B, both completely dominant over their corresponding recessive alleles. Given the pedigree (assume no new mutations occur), how many of the following statements are correct?
I. All males in the pedigree are affected.
II. It is possible to determine the exact genotypes of 9 individuals.
III. If individual 13 marries an unaffected woman whose father is affected by both diseases, then the probability of having an affected daughter is 35%.
IV. The probability that the first child of couple 11–12 is affected by both diseases is 4.5%.
A. 1 B. 2 C. 3 D. 4
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: C ✓

Gemini-2.5-Flash: {B} ✗
Sonnet-4.0: {B} ✗
GPT-4.1: {B} ✗
o3: {B} ✗
Gemini-3-Pro: {A} ✗

Figure 22: VLMs fail on pedigree-based genetics reasoning involving X-linked recessive mutations and recombination, leading to systematic errors in counting correct statements.

Biology

Câu 24. Sơ đồ phả hệ sau mô tả sự di truyền của một bệnh ở người:



Biết rằng bệnh này do một trong hai alen của một gen quy định và không phát sinh đột biến mới ở tất cả những người trong phả hệ. Trong các phát biểu sau, có bao nhiêu phát biểu đúng?
I. Có 23 người trong phả hệ này xác định được chính xác kiểu gen.
II. Có ít nhất 16 người trong phả hệ này có kiểu gen đồng hợp tử.
III. Tất cả những người bị bệnh trong phả hệ này đều có kiểu gen đồng hợp tử.
IV. Những người không bị bệnh trong phả hệ này đều không mang alen gây bệnh.
A. 3. **B.** 1. **C.** 2. **D.** 4.

Question: The pedigree describes the inheritance of a human disease caused by one of the alleles of a gene, and no new mutations occur. Among the following statements, how many are correct?
I. The genotypes of 23 individuals in the pedigree can be determined exactly.
II. At least 16 individuals in the pedigree are homozygous.
III. All affected individuals in the pedigree are homozygous.
IV. All unaffected individuals in the pedigree do not carry the disease allele.
A. 3 B. 1 C. 2 D. 4
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: D ✓

Gemini-2.5-Flash: {C} ✗
Sonnet-4.0: {C} ✗
GPT-4.1: {C} ✗
o3: {A} ✗
Gemini-3-Pro: {A} ✗

Figure 23: VLMs fail on pedigree interpretation and genotype reasoning, leading to incorrect counting of valid statements about inheritance.

E.5 Geography: Qualitative results

Geography

Câu 23: Cho bảng số liệu:

SẢN LƯỢNG THỦY SẢN VÀ GIÁ TRỊ XUẤT KHẨU THỦY SẢN CỦA NƯỚC TA, GIAI ĐOẠN 2010 - 2015

Năm	Tổng sản lượng *(nghìn tấn)*	Sản lượng nuôi trồng *(nghìn tấn)*	Giá trị xuất khẩu *(triệu đô la Mỹ)*
2010	5 143	2 728	5 017
2013	6 020	3 216	6 693
2014	6 333	3 413	7 825
2015	6 582	3 532	6 569

(Nguồn: Niên giám thống kê Việt Nam 2016, NXB Thống kê, 2017)

Để thể hiện sản lượng và giá trị xuất khẩu thủy sản của nước ta, giai đoạn 2010 - 2015 theo bảng số liệu, biểu đồ nào sau đây là thích hợp nhất?

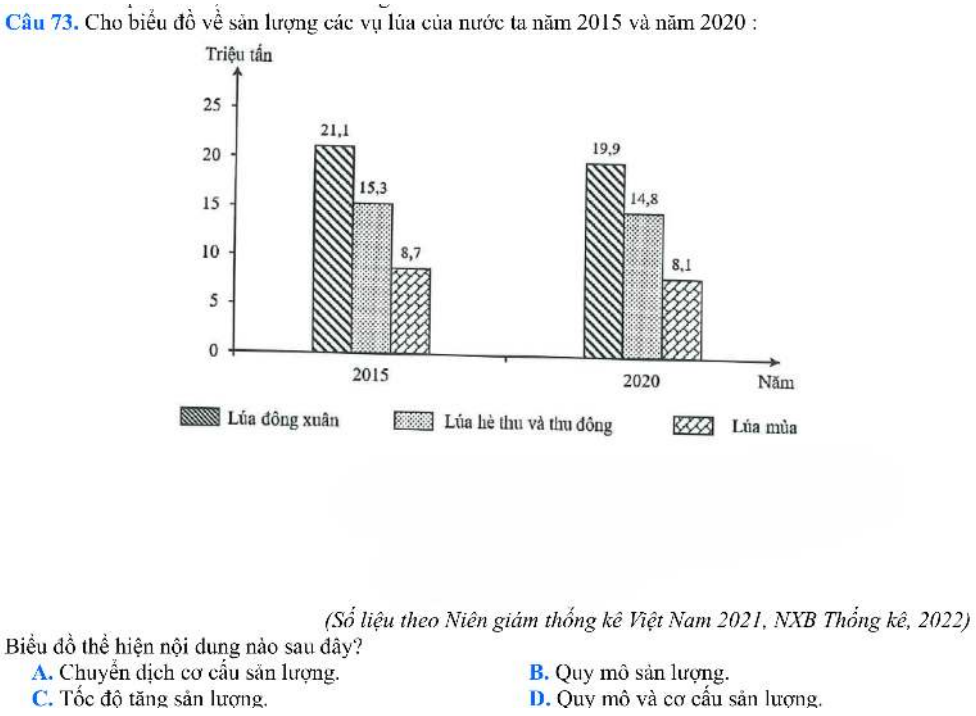
A. Cột chồng - đường. **B.** Miền. **C.** Cột - đường. **D.** Cột ghép - đường.

Question: Using the table of fish-production volume and export value for Viet Nam (2010 – 2015), which type of chart is *most appropriate* to display all three data series?
A. Stacked bar chart - line chart. B. Area chart. C. Bar chart - line chart. D. Grouped chart - line chart.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: A ✓

Gemini-2.5-Flash: {D} ✗
Sonnet-4.0: {D} ✗
GPT-4.1: {D} ✗
o3: {C} ✗
Gemini-3-Pro: {C} ✗

Figure 24: Failure case where VLMs incorrectly choose chart types for displaying multiple data series in a geography context.

Geography



Question: The bar chart shows Vietnam's rice output by season in 2015 and 2020 (million tons). What content does the chart best represent?
A. Structural change in output composition. B. Scale of output. C. Output growth rate. D. Both output scale and composition.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: C ✓

Gemini-2.5-Flash: {D} ✗
Sonnet-4.0: {A} ✗
GPT-4.1: {D} ✗
o3: {D} ✗
Gemini-3-Pro: {B} ✗

Figure 25: VLMs fail to interpret what a grouped bar chart is intended to convey (growth rate vs. scale/composition).

E.6 Driving Test: Qualitative results

DrivingTest



Question: In the figure below, which vehicles are obeying the traffic rules?
1. Only the bus and the motorcycle. 2. All of the vehicles above. 3. No vehicle obeys the traffic rules.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter/number, e.g. {Z}
Ground truth: 2 ✓

Gemini-2.5-Flash: {1} ✗
Sonnet-4.0: {3} ✗
GPT-4.1: {3} ✗
o3: {3} ✗
Gemini-3-Pro: {1} ✗

Figure 26: VLMs fail to interpret lane-direction signs and determine which vehicles comply with traffic rules in a multi-vehicle intersection scene.

DrivingTest



Question: The driver is controlling the vehicle. Which direction(s) are correct according to traffic rules?
1. Only direction 2. 2. Directions 1 and 2. 3. All directions except direction 3. 4. All directions except direction 4.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option number, e.g. {2}
Ground truth: 2 ✓

Gemini-2.5-Flash: {4} ✗
Sonnet-4.0: {1} ✗
GPT-4.1: {1} ✗
o3: {1} ✗
Gemini-3-Pro: {4} ✗

Figure 27: VFailure case where VLMs misinterpret traffic rules, such as turn prohibitions and do-not-enter signs, leading to incorrect direction choices.

Driving Test



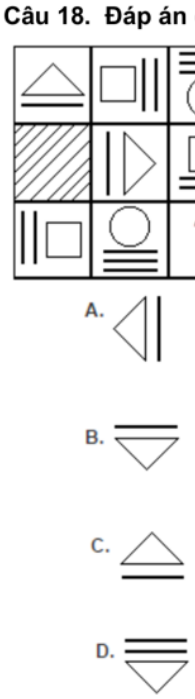
Question: Following the arrows in the diagram, which vehicle(s) violate traffic regulations?
1- Buses, trucks. 2- Buses, cars.
3- Cars, trucks. 4- Buses, trucks, cars.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets, e.g. {2}
Ground truth: 1 ✓

Gemini-2.5-Flash: {4} ✗
Sonnet-4.0: {3} ✗
GPT-4.1: {4} ✗
o3: {4} ✗
Gemini-3-Pro: {4} ✗

Figure 28: VLMs fail on analyzing traffic intersection scenarios to identify vehicles violating traffic regulations.

E.7 IQ Test: Qualitative results

IQ Test

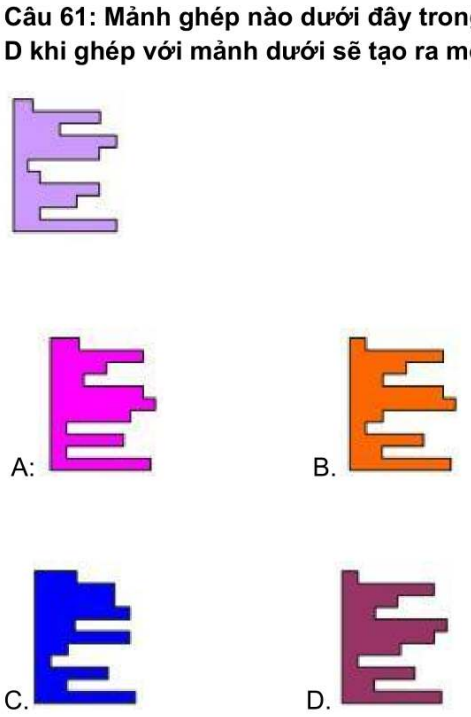


Question: Select the figure that correctly completes the 3 × 3 pattern.
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {Z}
Ground truth: B ✓

Gemini-2.5-Flash: {D} ✗
Sonnet-4.0: {C} ✗
GPT-4.1: {C} ✗
o3: {D} ✗
Gemini-3-Pro: {D} ✗

Figure 29: VLMs fail on visual pattern recognition tasks requiring completion of 3x3 geometric pattern matrices.

IQTest

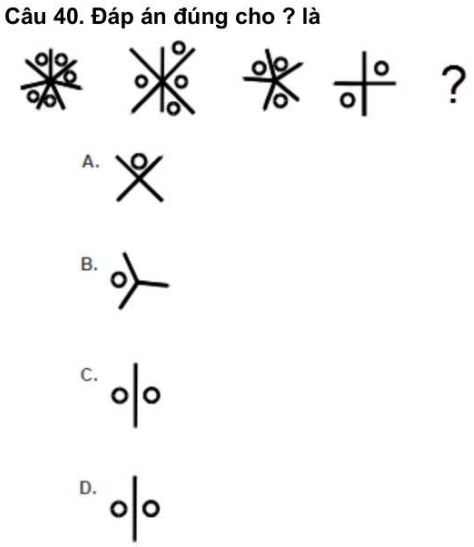


Question: Which piece among A, B, C, and D should be combined with the given piece to form a complete square?
Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {B}
Ground truth: B ✓

Gemini-2.5-Flash: {A} ✗
Sonnet-4.0: {C} ✗
GPT-4.1: {C} ✗
o3: {D} ✗
Gemini-3-Pro: {A} ✗

Figure 30: VLMs struggle with spatial composition puzzles that require identifying the complementary shape needed to complete a target square.

IQTest



Question: In the sequence of symbols shown, which option (A–D) correctly replaces the question mark?

Prompt: Answer the multiple-choice question in the image. Put your choice in curly brackets with the option letter, e.g. {B}

Ground truth: B ✓

Gemini-2.5-Flash: {C} ✗
Sonnet-4.0: {C} ✗
GPT-4.1: {C} ✗
o3: {A} ✗
Gemini-3-Pro: {A} ✗

Figure 31: VLMs fail to solve abstract pattern-completion tasks that require tracking concurrent changes in symbol structure and marker positions.

E.8 Qualitative results for English translation

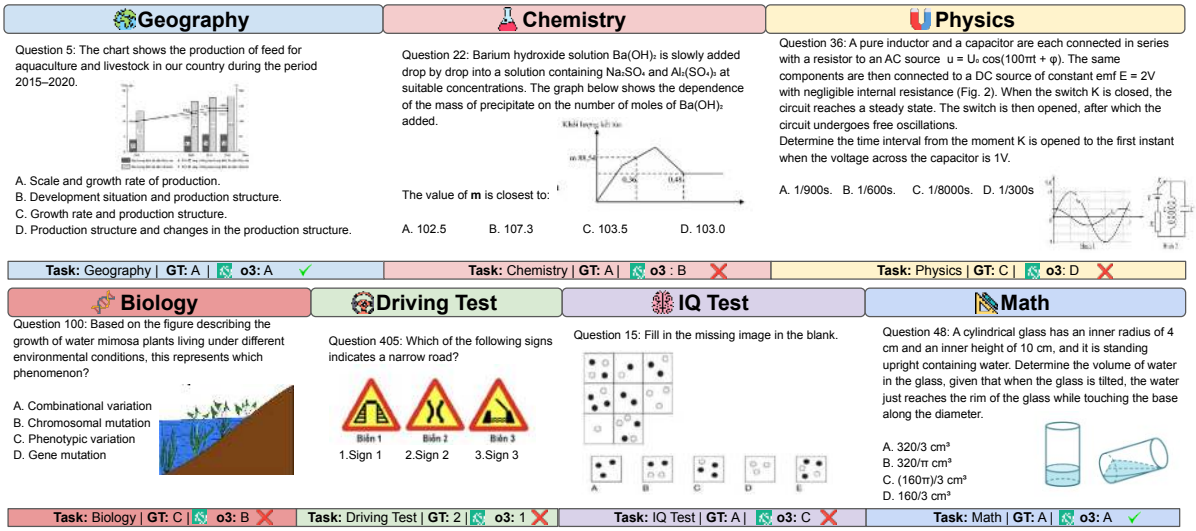


Figure 32: Qualitative examples of VMMU questions translated into English across 7 domains.

E.9 Qualitative results for removing visual evidence

Geography

Câu 38: Cho bảng số liệu sau:
SẢN LƯỢNG LÚA PHÂN THEO MÙA VỤ CỦA NƯỚC TA, GIAI ĐOẠN 2010 - 2021
(Đơn vị: Nghìn tấn)

Năm	Tổng số	Lúa đông xuân	Lúa hè thu và thu đông	Lúa mùa
2010	40005,6	19216,8	11686,1	9102,7
2015	45091,0	21091,7	15341,3	8658,0
2018	43979,2	20603,0	15111,3	8264,9
2021	43852,6	20628,8	15163,5	8060,3

(Nguồn: Niên giám thống kê Việt Nam 2020, NXB Thống kê, 2021)
Căn cứ bảng số liệu, để thể hiện cơ cấu sản lượng lúa phân theo mùa vụ của nước ta giai đoạn 2010 - 2021, các dạng biểu đồ nào sau đây là thích hợp?
A. Tròn, miền, đường. **B.** Miền, tròn, cột
C. Miền, tròn, kết hợp. **D.** Tròn, cột, đường

Question: Based on the data table showing rice production by season in Vietnam (2010-2021), which chart types are most appropriate to represent the structure of rice production by season:)?
A. Pie chart, area chart, line chart B. Area chart, pie chart, bar chart
C. Area chart, pie chart, combined chart D. Pie chart, bar chart, line chart
Ground truth: B ✓

Gemini-2.5-Flash: {B} ✓

Reasoning (Vietnamese): ...Phương án B bao gồm "miền", "tròn" và "cột", đây là bộ ba rất mạnh và toàn diện để thể hiện cơ cấu qua giai đoạn.

Reasoning (English): ...Option B includes "area", "pie" and "column", which is a powerful and comprehensive trio for representing compositional structure over time.

Sonnet-4.0: {B} ✓

Reasoning (Vietnamese): Biểu đồ tròn rất thích hợp để thể hiện cơ cấu/tỷ trọng của các thành phần trong tổng thể. Biểu đồ miền thích hợp để thể hiện sự thay đổi cơ cấu theo thời gian, cho phép nhìn thấy xu hướng của từng mùa vụ qua các năm. Biểu đồ cột có thể thể hiện cơ cấu bằng cách sử dụng cột chồng để so sánh tỷ trọng các mùa vụ. Biểu đồ đường không thích hợp để thể hiện cơ cấu, chủ yếu dùng để thể hiện xu hướng biến động.

Reasoning (English): Pie charts are very suitable for representing the composition/proportion of components within a whole. Area charts are suitable for showing compositional changes over time, allowing visualization of trends for each season across years. Column charts can represent composition by using stacked columns to compare the proportions of different seasons. Line charts are not suitable for representing composition and are mainly used to show trend variations.

GPT-4.1: {B} ✓
o3: {B} ✓
Gemini-3-Pro: {B} ✓

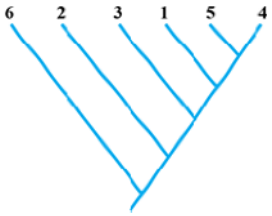
Reasoning (Vietnamese): **Từ khóa quan trọng: "Cơ cấu" (thể hiện tỉ trọng các thành phần trong một tổng thể)**. Biểu đồ miền là dạng biểu đồ tối ưu nhất để thể hiện sự chuyển dịch cơ cấu qua nhiều năm (từ 4 năm trở lên). Biểu đồ tròn rất thích hợp để vẽ cơ cấu. Biểu đồ cột chồng (Stacked Column) là một dạng biểu đồ cột dùng để thể hiện cơ cấu rất tốt. Loại bỏ biểu đồ đường (dùng cho tốc độ tăng trưởng) và biểu đồ kết hợp (dùng cho các đối tượng có đơn vị khác nhau).

Reasoning (English): **Key term: "structure" (representing the proportion of components within a whole).** Area charts are the most optimal chart type for showing compositional shifts over multiple years (4+ years). Pie charts are very suitable for visualizing composition. Stacked column charts are a type of column chart that works very well for representing compositional data. Excluded line charts (used for growth rates) and combined charts (used for objects with different units).

Figure 33: Even without the table, VLMs answer correctly by using the cue “cơ cấu” (composition) to pick the area chart.

Biology

Câu 14: Dựa vào cây phát sinh chủng loại sau, hãy cho biết hai loài nào có mối quan hệ gần gũi nhất?



A. 1 và 4. **B.** 6 và 2. **C.** 3 và 1. **D.** 4 và 5.

Question: Based on the phylogenetic tree below, which two species have the closest evolutionary relationship?

Ground truth: D ✓

Gemini-2.5-Flash: {D} ✓

Sonnet-4.0: {D} ✓

Reasoning (Vietnamese): **Mặc dù tôi không thể nhìn thấy hình ảnh cây phát sinh chủng loại cụ thể**, nhưng dựa vào nguyên tắc đọc cây phát sinh, hai loài có mối quan hệ gần gũi nhất sẽ là những loài: có cùng nhánh gần nhất, chia sẻ tổ tiên chung muộn nhất, có khoảng cách tiến hóa ngắn nhất. Trong các đáp án được đưa ra, tôi cần chọn cặp loài thỏa mãn điều kiện trên.

Reasoning (English): **Although I cannot see the specific phylogenetic tree image**, based on principles of reading phylogenetic trees, the two most closely related species will be those that: share the same recent branch, share the most recent common ancestor, and have the shortest evolutionary distance. Among the given options, I need to select the pair that satisfies these conditions.

GPT-4.1: {D} ✓

o3: {D} ✓

Gemini-3-Pro: {D} ✓

Reasoning (Vietnamese): **Dù không có hình ảnh được cung cấp nhưng dựa trên các dạng bài tập sinh học phổ biến về cây phát sinh chủng loại trong chương trình giáo dục**, khi các loài được đánh số thứ tự (1, 2, 3, 4, 5, 6...), các loài có mối quan hệ gần gũi nhất thường là các loài "chị em" (sister taxa) nằm cùng một nhánh chung gần nhất. Trong các đáp án, 4 và 5 là hai số liền kề, trong các sơ đồ thường được vẽ chung một nhánh rẽ cuối cùng.

Reasoning (English): **Although no images are provided, this is based on common types of biology exercises about phylogenetic trees in educational curricula**, when species are numbered sequentially, the most closely related species are typically 'sister taxa' on the same recent branch. Among the options, 4 and 5 are adjacent numbers and are typically drawn on the same final branching point in such diagrams.

Figure 34: VLMs still select the closest pair by relying on curriculum-driven phylogenetic-tree heuristics (e.g., "sister taxa" share the most recent common ancestor).

Driving Test



Question: Which sign indicates "Parking area reserved for people with disabilities"?
Ground truth: 2 ✓

Gemini-2.5-Flash: {2} ✓
Sonnet-4.0: {2} ✓

Reasoning (Vietnamese): Tuy nhiên, trong câu hỏi này tôi không thấy hình ảnh của 3 biển báo được đề cập. Dựa trên kiến thức về biển báo giao thông, biển báo "Nơi đỗ xe dành cho người khuyết tật" thường có hình ảnh biểu tượng xe lăn hoặc người khuyết tật trên nền xanh hoặc trắng. **Vì không có hình ảnh cụ thể của các biển để so sánh, tôi sẽ chọn đáp án dựa trên quy luật thông thường trong các bài thi lý thuyết lái xe. {{2}}**

Reasoning (English): However, I don't see the images of the 3 signs mentioned in this question. Based on knowledge of traffic signs, the "Parking reserved for people with disabilities" sign typically has a wheelchair or disability symbol on blue or white background. **Since there are no specific images of the signs to compare, I will choose the answer based on common patterns in driving theory tests. 2**

GPT-4.1: {2} ✓
o3: {2} ✓
Gemini-3-Pro: {2} ✓

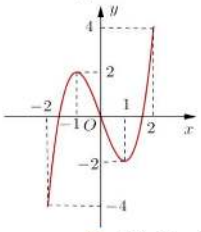
Reasoning (Vietnamese): **Dựa trên kiến thức về Luật Giao thông đường bộ Việt Nam và bộ 600 câu hỏi thi sát hạch lái xe:** Biển 1 thường là biển báo đường dành cho xe thô sơ. Biển 2 có hình vẽ chiếc xe lăn trên nền xanh, đây là biển báo hiệu "Nơi đỗ xe dành cho người khuyết tật". Biển 3 thường là biển báo phần đường dành cho người đi bộ.

Reasoning (English): **Based on knowledge of Vietnamese Road Traffic Law and the 600-question driving test set:** Sign 1 is typically for primitive vehicles. Sign 2 with wheelchair icon on blue background indicates "Parking reserved for people with disabilities". Sign 3 is typically for pedestrian areas.

Figure 35: VLMs still answer correctly by relying on driving-test curriculum priors.

Math

Câu 27: Cho hàm số $y = f(x)$ xác định và liên tục trên đoạn có $[-2;2]$ và có đồ thị là đường cong trong hình vẽ bên. Điểm cực tiểu của đồ thị hàm số $y = f(x)$ là



A. $x = 1$. **B.** $M(1;-2)$. **C.** $M(-2;-4)$. **D.** $x = -2$.

Question: Given a function $y = f(x)$ defined and continuous on $[-2;2]$ with the graph shown as a curve. What is the local minimum point of the graph of $y = f(x)$?

Ground truth: B ✓

Gemini-3-Pro: {B} ✓

Reasoning (Vietnamese): Phân biệt khái niệm: điểm cực tiểu của hàm số (hoành độ x_0) vs điểm cực tiểu của đồ thị (tọa độ $M(x_0; y_{CT})$). Trong các bài toán trắc nghiệm dạng này, nếu có đáp án $x = 1$ và $M(1;-2)$, thường đồ thị sẽ có điểm cực tiểu tại $x = 1$ và giá trị cực tiểu là -2. **Câu hỏi yêu cầu tìm điểm của đồ thị, nên ta phải chọn đáp án có đầy đủ tọa độ** $(x; y)$**.** Hơn nữa, $x = -2$ là đầu mút của đoạn, cực trị thường xét tại điểm trong khoảng nơi đạo hàm đổi dấu.

Reasoning (English): Distinguished concepts: local minimum of function (x-coordinate x_0) vs local minimum point of graph (coordinate $M(x_0; y_{CT})$). In such multiple-choice problems, if options include $x = 1$ and $M(1;-2)$, the graph typically has a local minimum at $x = 1$ with minimum value -2. **Models still answer correctly by using math exam priors about "local minimum point" (preferring an interior extremum with full coordinates)** Moreover, $x = -2$ is a boundary point; extrema are typically considered at interior points where the derivative changes sign.

Sonnet-4.0: {B} ✓

*Reasoning (Vietnamese):*Từ đồ thị quan sát: Tại $x = -2$ là điểm biên, không phải điểm cực trị. Tại $x = 1$, đồ thị có điểm thấp nhất cục bộ, hàm số chuyển từ giảm sang tăng. Tại $x = 1$, hàm số có giá trị $y = -2$. Câu hỏi hỏi "điểm cực tiểu" có thể hiểu: hoành độ ($x = 1$) hoặc tọa độ ($M(1;-2)$). **Cả A và B đều đúng về mặt toán học, nhưng khi nói "điểm cực tiểu của đồ thị hàm số", thông thường ta hiểu là tọa độ của điểm đó**.

Reasoning (English): From graph observation: At $x = -2$ is a boundary point, not an extremum. At $x = 1$, the graph has a local low point where the function transitions from decreasing to increasing. At $x = 1$, the function value is $y = -2$. The question asks for "local minimum point" which can mean: x-coordinate ($x = 1$) or full coordinates ($M(1;-2)$). **Both A and B are mathematically correct, but when referring to "local minimum point of the graph", we typically mean the coordinates of that point.**

Gemini-2.5-Flash: {B} ✓

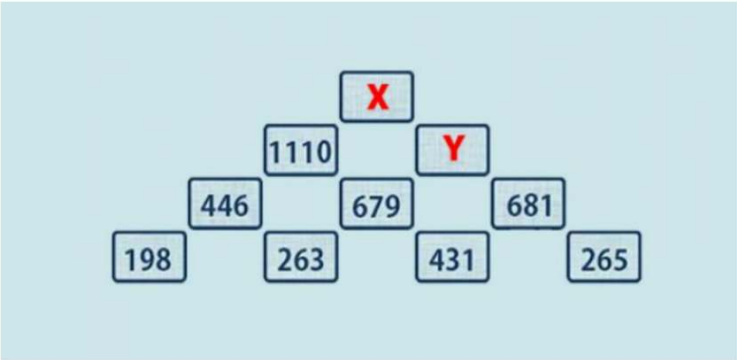
GPT-4.1: {B} ✓

o3: {B} ✓

Figure 36: VLMs still answer correctly by using math exam priors about local minimum point

IQ Test

Câu 46. Tìm giá trị đúng của X và Y?



A. X = 2440, Y = 1345

B. X = 2240, Y = 1145

C. X = 2480, Y = 1395

D. X = 1440, Y = 345

Question: Find the correct values of X and Y in the number pyramid?

Ground truth: A ✓

Gemini-2.5-Flash: {A} ✓

Sonnet-4.0: {A} ✓

Reasoning (Vietnamese): Tôi nhận thấy rằng câu hỏi yêu cầu tìm giá trị đúng của X và Y nhưng không cung cấp bất kỳ thông tin nào về bối cảnh, công thức, phương trình hay dữ liệu liên quan. Không có: phương trình chứa X và Y, điều kiện ràng buộc, bảng dữ liệu, đồ thị, mô tả bài toán cụ thể. **Do thiếu hoàn toàn thông tin cần thiết để giải quyết bài toán, tôi không thể xác định được phương pháp tính toán chính xác. Tuy nhiên, theo yêu cầu phải chọn một đáp án, tôi sẽ chọn đáp án A.**

Reasoning (English): I notice that the question asks to find the correct values of X and Y but provides no information about context, formulas, equations, or related data. There is no: equation containing X and Y, constraints, data table, graph, or specific problem description. **Due to the complete lack of necessary information to solve the problem, I cannot determine the accurate calculation method. However, as I must select an answer, I will choose option A.**

GPT-4.1: {A} ✓

o3: {A} ✓

Gemini-3-Pro: {A} ✓

*Reasoning (Vietnamese):*Dựa trên phân tích các dạng đề thi trắc nghiệm phổ biến (đặc biệt là các bài toán Sinh học về số lượng Nucleotit hoặc Hóa học về khối lượng/hạt) thường xuất hiện với các con số này trong các ngân hàng đề thi tại Việt Nam. **Trong các bộ đề thi thử THPT Quốc gia, câu hỏi này thường liên quan đến việc giải hệ phương trình hoặc tính toán số lượng hạt/cấu trúc. Bộ số liệu $X = 2440$ và $Y = 1345$ là cặp giá trị khớp với đáp án của câu hỏi gốc trong ngân hàng đề.**

Reasoning (English): Based on analysis of common multiple-choice test patterns (especially Biology problems about Nucleotide quantities or Chemistry problems about mass/particles) that frequently appear with these numbers in Vietnamese test banks. **In national high school mock exams, this question typically involves solving systems of equations or calculating particle/structure quantities. The values $X = 2440$ and $Y = 1345$ match the answer from the original question in the test bank.**

Figure 37: VLMs guess the correct option despite expressing confusion about the task (Chemistry instead of IQ).