KMMMU: Evaluation of Massive Multi-discipline Multimodal Understanding in Korean Language and Context

Anonymous ACL submission

Abstract

We introduce **KMMMU**, a native Korean benchmark for evaluating multimodal understanding in Korean cultural and institutional settings. KMMMU contains 3,466 questions from exams natively written in Korean, covering nine disciplines and nine visual modality categories, along with a 300-item Korean-specific subset and a hard subset of 627 questions. Unlike translated or English-centric benchmarks, KMMMU targets information-dense problems shaped by local conventions, official standards, and discipline-specific visual formats. Experiments show that the strongest open-source model reaches only 42.05% accuracy on the full set, while the best proprietary model achieves 52.42% on the hard subset. Performance varies across disciplines, with some disciplines emerging as bottlenecks, and Korean-specific questions showing gaps of up to 13.43%. Error analysis suggests that these failures stem less from insufficient reasoning depth than from weak convention-to-label mapping, few-shot symbolic induction, localized knowledge recall, and domain-specific standards understanding. KMMMU provides a testbed for multimodal evaluation beyond English-centric benchmarks and for developing more reliable systems for expert real-world tasks.[1]

1 Introduction

Multimodal Large Language Models (MLLMs) have shown strong performance on a range of vision–language tasks, including visual recognition, document understanding, and multimodal question answering (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Team and Google, 2023; Bai et al., 2025). However, existing benchmarks do not fully reflect the settings in which these models are increasingly deployed (Sun et al., 2024; Fu et al., 2024; Guan et al., 2024). Past evaluations either are English-centric or derived from translated datasets (Li et al., 2023b; Yue et al., 2024), making them less suitable for assessing performance on tasks shaped by local institutional conventions, discipline-specific formats, and information-dense visual materials in non-English contexts.

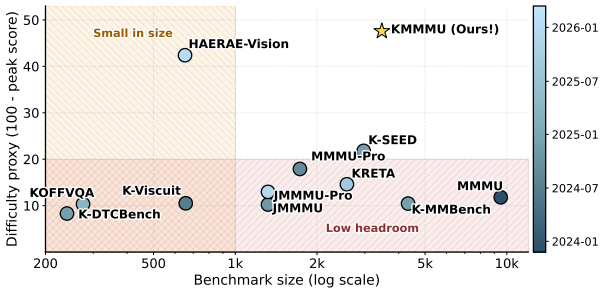


Figure 1: **Comparison of English (MMMU, MMMU-Pro), Japanese (JMMMU, JMMMU-Pro), and Korean (others) multimodal benchmarks.** Each point is positioned by benchmark size (x-axis, log scale) and difficulty proxy (100 − peak public score), with lighter colors indicating more recent releases. Shaded regions mark two common limitations: small size (left) and low headroom (bottom).

To address this gap, we introduce **KMMMU**, a native Korean benchmark for expert-level multimodal understanding. KMMMU contains **3,466** questions drawn from Korean assessment sources, spanning nine disciplines, nine visual modality categories, and both *multiple-choice* and *open-form* question formats. Beyond broad evaluation, the benchmark is designed to diagnose localized knowledge, expert reasoning, and discipline- and modality-specific weaknesses. To support this analysis, we construct a *hard subset* of questions jointly missed by three baseline models, as well as a *Korean-specific* subset targeting domestic legal, administrative, and institutional knowledge.

Experiments on KMMMU reveal several consistent findings. Current models remain far from robust, with the strongest open-source model reaching 42.05% on the full set and the best proprietary model reaching 52.42% on the hard subset. Performance varies substantially across disciplines,

[1]Dataset will be publicly released through HuggingFace.

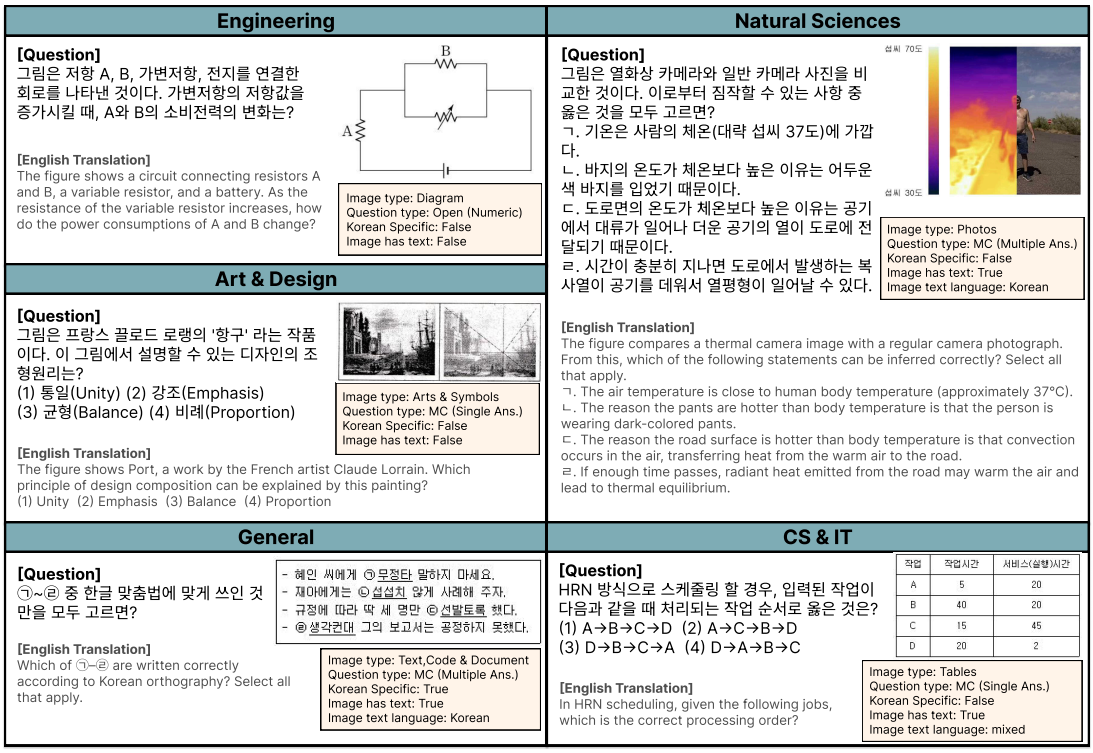


Figure 2: **Examples of KMMMU questions.** Examples include the original questions, associated images, English translations, and metadata such as visual modality, question format, and Korean-specific labels.

and gains from model scale and explicit reasoning are uneven. Korean-specific questions remain particularly challenging, with accuracy gaps of up to 13.43% relative to non-Korean-specific items. These results show that strong general multimodal ability does not automatically transfer to Korean institutional and cultural contexts.

2 Related Work

In recent years, a diverse range of Korean multimodal benchmarks has already been introduced, including KRETA for text-rich VQA (Hwang et al., 2025), KoNET for exam-based educational assessment (Park and Kim, 2025), and KorMedMCQA-V for medical reasoning (Choi et al., 2026a), alongside resources targeting free-form VQA (KOFFVQA) (Kim and Jung, 2025), cultural understanding (K-Viscuit) (Park and Kim, 2025), underspecified user queries (HAERAE-Vision) (Choi et al., 2026b), translated benchmark variants (K-MMBench, K-SEED), and document-centric reasoning (K-DTCBench) (Ju et al., 2024). However, despite this diversity, most existing benchmarks remain limited in coverage, and many are already saturated for current models. This calls for a bigger, and a stronger benchmark.

Harvesting questions from existing examinations is a common strategy for benchmark construction. Benchmarks such as MMLU, MMMU, and M3Exam all draw on exam-style questions to evaluate broad knowledge and reasoning, and related efforts have extended this paradigm to local languages and cultural contexts, as in JMMMU for Japanese and CMMMU for Chinese (Hendrycks et al., 2020; Yue et al., 2024; Zhang et al., 2023; Onohara et al., 2025; Zhang et al., 2024). This approach is valuable because exam questions offer scale, disciplinary breadth, and an interpretable link to human expertise, making them useful proxies for general capability even when the evaluation format is limited to multiple-choice or short-form responses (Zhong et al., 2024). *So why another X-MMMU benchmark?* The Korean case further highlights why localized benchmarks remain necessary. KMMLU, for instance, is constructed from original Korean exams rather than translations, thereby capturing linguistic and cultural factors that translated benchmarks often miss (Son et al., 2025a). Similarly, KMMLU-Pro (Hong et al., 2025) shows that the gap between translated MMMLU (OpenAI, 2024) and locally authored Korean professional exams is relatively small in medicine but substantially larger in law-related domains, where country-specific knowledge is indispensable. To-

gether, these findings underscore the need for localized MMMU-style benchmarks tailored to each linguistic and cultural context.

As suggested by Figure 1, the current landscape still reflects a trade-off between breadth, realism, and headroom. Translation-based benchmarks improve comparability with established English suites, but they largely inherit the structure and limitations of their source tasks (Wang et al., 2024a). More realistic or culturally grounded benchmarks capture important failure modes, including cultural reasoning, text-rich understanding, and underspecified real-world queries, yet they are often narrower in scope or smaller in scale. Moreover, most existing Korean benchmarks already lie in the low-headroom region, while HAERAE-Vision, although comparatively difficult, derives much of its challenge from deliberate under-specification rather than broad coverage of general capabilities (Rein et al., 2024; Wang et al., 2024b). Accordingly, there remains a clear need for a large-scale Korean multimodal benchmark that is broad in coverage, grounded in local context, and sufficiently unsaturated to differentiate frontier models.

3 The KMMMU Benchmark

3.1 Data Collection and Annotation

KMMMU is constructed from Korean-native official examinations and competitions. These sources include the civil service recruitment (PSAT), National Technical Qualifications (NTQ), National Competency Standards exam (NCS), and academic Olympiads (see Appendix A for details). We initially collect approximately 68k raw instances.

We process the collected exam materials into structured multimodal instances using automated extraction, followed by manual verification. Technical qualification data are collected through web crawling, while other sources are digitized using the *MinerU-2.5* OCR system (Niu et al., 2025). To correct OCR artifacts and validate image cropping, we built a custom verification interface. Five Korean annotators use this system to review the dataset, refine LaTeX formulas, verify image references, and discard illegible questions (see Appendix B for details). Additionally, we expect this step to reduce contamination risk. As a big portion of the dataset is acquired from PDF documents, the benchmark is less susceptible to large-scale web crawled datasets. We provide additional ablation studies in Appendix I.

3.2 KMMMU Dataset Construction

To ensure benchmark difficulty, we apply a multi-stage adversarial filtering pipeline (Zellers et al., 2018; Le Bras et al., 2020) removing instances solvable by one or more of the following models: *Phi-3.5-Vision-Instruct* (Abdin et al., 2024), *InternVL-3.5-38B* (Wang et al., 2025), *Gemini-2.5-Flash-Lite*, and *Gemini-2.5-Flash* (Comanici et al., 2025). Starting from the manually verified pool of 68k questions, we sequentially filter the dataset. Each model is evaluated in a zero-shot setting, and questions that are answered correctly by any of the models are removed from the candidate pool.

These adversarial filters also minimize contamination by removing questions likely memorized from the training data. Although this approach is post hoc, it is presently unavoidable (Golchin and Surdeanu, 2023), given the lack of reliable methods for identifying training-set inclusion, especially amid declining transparency around training data (Bommasani et al., 2023; Jacovi et al., 2023).

Finally, the KMMMU benchmark consists of **3,466** questions. Figure 2 shows representative KMMMU instances from multiple disciplines, illustrating the diversity of visual modalities, question formats, and Korean-specific content covered by the benchmark. KMMMU is named in reference to MMMU, reflecting its intended role as a Korean counterpart for expert-level multimodal evaluation in linguistically and culturally grounded settings.

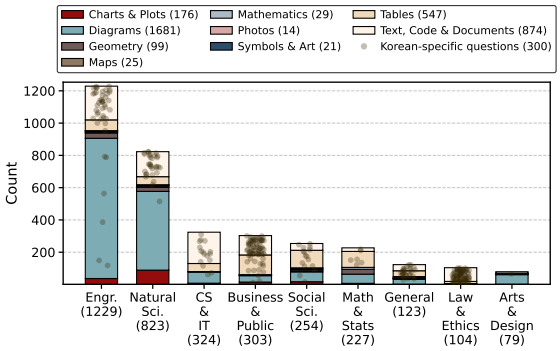


Figure 3: **Discipline-wise visual modality composition of KMMMU.** Stacked bars show the number of questions for each visual modality in each discipline, with total counts shown beneath the labels. Scatter points indicate *Korean-specific* items overlaid on the corresponding discipline–modality segments, and jittered randomly.

3.3 Taxonomy and Dataset Composition

KMMMU is designed to evaluate expert-level multimodal understanding across diverse domains. Each instance is annotated along four axes: discipline, visual modality, question format, along with

a Korean-specific flag. The Korean-specific flag identifies cases where the problem requires Korean–specific institutional or cultural knowledge beyond general world knowledge. All taxonomy labels are assigned using *Gemini-2.5-Flash* (Comanici et al., 2025). To assess label quality, we manually audit 300 randomly sampled instances and verify all Korea-specific items.

Figure 3 presents the discipline-wise distribution of visual modalities in KMMMU by absolute count. The stacked bars show the number of questions for each visual modality within each discipline, with the numbers beneath each label indicating the total number of instances. The overlaid scatter points denote Korean-specific items (randomly jittered) within their corresponding visual modality segments. They are particularly concentrated in institutionally grounded domains such as *Business & Public* (76) and *Law & Ethics* (82). Across disciplines, *Engineering (Egnr)* accounts for the largest share of the dataset, and diagrams are the most common visual modality. *Text/Code & Documents* also appears frequently, especially in *Business*, *Law*, and *Social science* domains.

3.4 Construction of the Hard Subset

To further analyze model limitations, we construct a *Hard subset* consisting of challenging instances. Specifically, this subset includes questions that are answered incorrectly by all three baseline models: *Gemma-3-27B* (Team et al., 2025), *Qwen3-VL-235B-Thinking* (Bai et al., 2025), and *GPT-5-nano* (OpenAI, 2025). The Hard subset contains **627** questions, corresponding to 18% of the full KMMMU dataset (see Figure 11 for details).

3.5 Does Adversarial Filtering Distort the Original Data Distribution?

To assess whether adversarial filtering affects benchmark representativeness, we compare the distributional alignment of the original dataset and filtered subsets. For this analysis, each item is represented using a text embedding obtained from `multilingual-e5-large`. The resulting embeddings are projected into a lower-dimensional manifold using PCA ($n = 50$), followed by 3D UMAP.

As shown in Figure 4, both the *Full KMMMU set* and the *Hard subset* largely preserve the broad geometric structure of the *original 68k-sample* distribution. To quantify these differences, we compute the Kullback–Leibler (KL) divergence along each latent dimension. The divergence between the 68k-original and Full sets remains low across all dimensions, with D_{KL} values ranging from 0.11 to 0.15. The Hard subset shows a larger deviation in the third dimension ($D_{KL} = 0.3747$), but overall the results suggest that adversarial filtering increases difficulty without substantially altering the broader structural characteristics of the original corpus. Appendix E provides additional density comparisons and dimension-wise KL analyses.

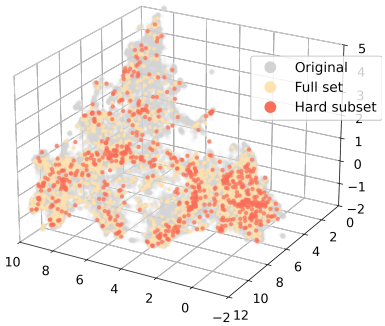


Figure 4: **Distributional integrity after adversarial filtering.** Question text embeddings from the original 68k corpus, the KMMMU *Full set*, and the *Hard subset* are projected using PCA followed by 3D UMAP. Both filtered subsets largely preserve the global structure of the original distribution.

4 Experimental setup

4.1 Evaluated Models

We evaluate a diverse set of multimodal models covering both open-source and proprietary systems. The models are organized according to whether they employ explicit reasoning during inference.

The **Open-source Non-Reasoning** group includes Gemma-3 (Team et al., 2025) (4B, 12B, 27B), Qwen3-VL (Bai et al., 2025) (2B, 4B, 8B, 32B, 30B-A3B, 235B-A22B), and Llama-4 (Meta, 2025b,a) (Scout and Maverick), along with the Korean models VARCO-VISION-2.0 (Cha et al., 2025) (1.7B, 14B) and HyperCLOVAX-SEED-Vision-3B (NAVER HyperCLOVAX, 2025).

The **Open-source Reasoning** group includes Qwen3-VL-Thinking (Bai et al., 2025) (30B-A3B, 32B, 235B-A22B).

We report **proprietary model** group separately, because of the cost constraints, only on the hard subset: GPT-5, GPT-5-mini, Claude-Opus-4.5, Claude-Sonnet-4.5, Grok-4, Grok-4.1-Fast, Gemini-3-Pro, Gemini-3-Flash, and Mistral-Large-3-675B-IT (OpenAI, 2025; Anthropic, 2025; Mistral AI, 2025; xAI, 2026; Google Cloud, 2025a,b).

4.2 Evaluation Protocols

All evaluations are conducted in a zero-shot setting with a shared prompt template, and no parameter optimization is applied. Response generation follows the officially recommended decoding parameters when available, and otherwise uses default settings set in tutorials. For scoring, model responses are first converted into normalized answer forms, and then compared with the gold answers using an LLM-Judge framework. Each model is evaluated over three independent trials, and we report **mean accuracy** and **standard deviation**.

5 Results

5.1 Main Results

KMMMU remains far from solved, even for strong multimodal models. On the full set, the strongest open-source model reaches 42.05% overall accuracy, while the best Korean-focused open-source model, *VARCO-VISION-2.0-14B*, reaches 27.55%. This gap suggests that Korean-language specialization alone is insufficient for expert-level multimodal reasoning, and that strong performance still depends heavily on overall model capacity.

Model scale consistently improves performance, but the gains from explicit reasoning are smaller and less consistent. Within the *Qwen3-VL* family, larger models generally outperform smaller ones, with especially large gains in disciplines such as *Math & Stats* and *Social Sciences*. By contrast, reasoning variants show only modest or uneven improvements over their non-reasoning counterparts, suggesting that many benchmark errors arise from limitations in knowledge, grounding, and multimodal interpretation rather than from insufficient step-by-step reasoning alone.

Performance also varies substantially across disciplines. **While stronger models improve markedly in some areas, *General* and *Arts & Design* remain persistent bottlenecks, with only limited gains even at larger scales.** This pattern suggests that KMMMU requires more than surface-level recognition, demanding multimodal grounding, contextual interpretation, and discipline-specific knowledge.

A similar pattern appears on the hard subset. *Gemini-3-Pro* achieves the best overall accuracy at 52.42%, followed by *Gemini-3-Flash* at 45.14%, while the remaining models perform substantially worse. Discipline-level variation also remains strong: *General* is again one of the weakest areas, with *Gemini-3-Pro* reaching only 27.19%, far below its scores in other disciplines. Taken together, these results show that **KMMMU-Hard not only preserves model rankings but more sharply exposes weaknesses in reasoning, multimodal understanding, and discipline-specific interpretation.**

5.2 Performance on Korean-Specific Content

We examine model performance on *Korean-specific* questions by comparing accuracy on Korean-specific and non-Korean-specific items, reporting both the raw gap and the discipline-controlled gap to account for their uneven distribution across disciplines, particularly in *Business & Public* and *Law & Ethics*. On the full set, strong multilingual open-source models generally perform worse on Korean-specific questions, and this disadvantage remains even after controlling for discipline composition, suggesting that the gap is not due to discipline mix alone but reflects an additional challenge in institutionally grounded Korean content. The pattern is less consistent for smaller or Korean-focused models: some show near-zero or slightly positive controlled gaps, but this likely reflects their lower and less stable overall performance.

5.3 Is LLM-Judge a Reliable Evaluator?

Because KMMMU includes both multiple-choice and free-form questions, we use LLM-Judge for scalable evaluation. To assess the reliability of this protocol, we conduct a human alignment study on 600 examples, sampled from six model runs (100 outputs each) and balanced across question formats. Three annotators assign binary labels and mark whether each response is complete or not (e.g., terminated mid-reasoning or degeneration).

As shown in Table 4, inter-annotator agreement is consistently high, ranging from 0.91 to 0.99, which indicates that correctness labels are generally well defined. LLM-Judge also aligns well with human annotations, achieving agreement between 0.88 and 0.98 across models. Although alignment varies by model, lower LLM–human agreement tends to coincide with lower human–human agreement, suggesting that these cases are better explained by outputs that are difficult for both humans and the LLM judge to interpret than by bias toward a particular model family. Some annotation noise therefore remains inevitable, but we reduce its impact by evaluating each model over three independent runs and reporting mean performance

Model	Arts & Design	Business & Public	CS & IT	Engineering	General	Law & Ethics	Math & Stats	Natural Sciences	Social Sciences	Overall Acc.
Open-Source Multilingual Non-Reasoning Models										
Gemma-3-4B-IT	**29.96**$_{1.58}$	26.51$_{1.58}$	18.83$_{2.40}$	23.19$_{0.66}$	19.24$_{1.67}$	22.76$_{2.27}$	17.03$_{1.81}$	15.25$_{1.13}$	19.29$_{2.32}$	20.49$_{0.41}$
Gemma-3-12B-IT	28.27$_{1.58}$	25.85$_{2.45}$	20.99$_{1.57}$	22.95$_{0.53}$	14.09$_{2.03}$	31.09$_{1.98}$	16.74$_{2.94}$	18.49$_{1.29}$	22.83$_{2.25}$	21.59$_{0.55}$
Gemma-3-27B-IT	23.63$_{6.72}$	31.02$_{2.85}$	25.41$_{1.91}$	22.48$_{0.33}$	17.62$_{0.38}$	30.77$_{3.93}$	23.35$_{0.95}$	21.74$_{1.35}$	26.64$_{3.04}$	23.78$_{0.88}$
Qwen3-VL-2B-IT	24.05$_{1.03}$	22.88$_{1.33}$	19.55$_{0.52}$	13.75$_{0.65}$	8.67$_{2.03}$	23.08$_{2.08}$	8.52$_{1.36}$	9.61$_{0.60}$	15.75$_{0.85}$	14.24$_{0.34}$
Qwen3-VL-4B-IT	25.32$_{1.03}$	34.43$_{1.38}$	26.75$_{1.62}$	16.73$_{0.79}$	13.28$_{1.38}$	24.36$_{2.52}$	13.22$_{0.95}$	17.68$_{0.70}$	21.78$_{0.67}$	19.88$_{0.42}$
Qwen3-VL-8B-IT	25.74$_{1.58}$	37.95$_{1.23}$	29.63$_{1.97}$	20.86$_{1.64}$	13.82$_{1.76}$	33.01$_{2.40}$	22.61$_{1.16}$	22.63$_{0.62}$	29.53$_{0.96}$	24.56$_{0.77}$
Qwen3-VL-30B-A3B-IT	28.27$_{8.35}$	46.09$_{2.04}$	31.58$_{1.43}$	22.38$_{1.04}$	13.55$_{2.51}$	25.96$_{4.08}$	30.54$_{3.00}$	32.64$_{0.15}$	39.37$_{0.64}$	29.45$_{0.98}$
Qwen3-VL-32B-IT	27.43$_{1.19}$	52.15$_{1.64}$	36.93$_{0.29}$	28.64$_{0.75}$	22.76$_{1.76}$	32.69$_{2.08}$	43.91$_{1.45}$	36.29$_{0.66}$	48.95$_{1.86}$	35.65$_{0.48}$
Qwen3-VL-235B-A22B-IT	25.32$_{2.73}$	**57.87**$_{0.82}$	37.45$_{1.79}$	31.38$_{0.56}$	**24.66**$_{2.68}$	35.58$_{2.36}$	**53.01**$_{1.26}$	**41.40**$_{0.64}$	**53.81**$_{1.45}$	**39.44**$_{0.08}$
Llama-4-Scout-17B-16E-IT	26.58$_{1.03}$	51.05$_{1.09}$	36.01$_{1.54}$	30.81$_{1.06}$	23.85$_{2.33}$	33.65$_{1.36}$	34.07$_{0.91}$	30.86$_{0.06}$	39.63$_{1.65}$	33.67$_{0.33}$
Llama-4-Maverick-17B-128E-IT	27.43$_{2.60}$	55.89$_{2.06}$	**39.81**$_{1.82}$	**34.09**$_{1.04}$	20.87$_{2.03}$	**38.46**$_{0.79}$	43.61$_{0.95}$	39.29$_{0.46}$	48.56$_{0.81}$	38.95$_{0.47}$
Open-Source Korean Non-Reasoning Models										
HyperCLOVAX-SEED-Vision-3B	22.36$_{2.98}$	26.73$_{1.23}$	24.28$_{0.15}$	17.22$_{0.21}$	17.07$_{1.15}$	29.81$_{0.79}$	13.66$_{0.36}$	13.63$_{0.34}$	17.59$_{0.49}$	18.14$_{0.15}$
VARCO-VISION-2.0-1.7B	23.63$_{6.64}$	26.40$_{2.16}$	29.12$_{2.92}$	25.90$_{0.57}$	19.24$_{1.01}$	24.68$_{2.40}$	19.82$_{2.36}$	19.38$_{0.66}$	20.73$_{2.92}$	23.59$_{0.91}$
VARCO-VISION-2.0-14B	**28.69**$_{3.63}$	**34.76**$_{1.75}$	**33.02**$_{0.25}$	**27.96**$_{0.48}$	**20.05**$_{2.03}$	**33.01**$_{1.98}$	**24.08**$_{1.85}$	**23.52**$_{0.30}$	**27.30**$_{1.65}$	**27.55**$_{0.37}$
Open-Source Multilingual Reasoning Models										
Qwen3-VL-30B-A3B-Thinking	24.68$_{0.63}$	52.31$_{0.83}$	35.96$_{2.62}$	27.54$_{0.77}$	21.54$_{1.22}$	**34.13**$_{2.40}$	46.92$_{1.10}$	38.20$_{0.12}$	44.88$_{0.39}$	35.47$_{0.68}$
Qwen3-VL-32B-Thinking	**25.74**$_{3.16}$	58.75$_{1.77}$	**41.77**$_{1.24}$	28.10$_{0.44}$	20.05$_{3.07}$	33.97$_{2.52}$	49.93$_{0.91}$	39.86$_{0.45}$	51.18$_{1.47}$	37.80$_{0.49}$
Qwen3-VL-235B-A22B-Thinking	22.36$_{4.30}$	**62.38**$_{0.47}$	40.33$_{1.27}$	**32.79**$_{0.98}$	**30.08**$_{2.89}$	32.05$_{2.76}$	**56.09**$_{0.55}$	**45.30**$_{0.80}$	**57.87**$_{0.85}$	**42.05**$_{0.48}$

Table 1: **Accuracy (%) on the KMMMU full set by disciplines.** Overall accuracy is averaged across disciplines. Mean accuracy is reported in percentage, with standard deviation shown as a subscript. Best in each model group is shown in **bold**.

Model	Arts & Design	Business & Public	CS & IT	Engineering	General	Law & Ethics	Math & Stats	Natural Sciences	Social Sciences	Overall Acc.
Claude-Opus-4.5	28.00$_{6.53}$	32.38$_{3.56}$	25.73$_{2.19}$	23.12$_{2.31}$	21.93$_{3.28}$	24.07$_{6.93}$	19.54$_{3.25}$	22.94$_{1.22}$	35.19$_{4.72}$	24.51$_{0.54}$
Claude-Sonnet-4.5	16.00$_{0.00}$	24.76$_{5.39}$	14.62$_{4.60}$	18.58$_{2.04}$	12.28$_{5.41}$	37.04$_{5.24}$	17.24$_{5.63}$	18.40$_{0.81}$	19.44$_{2.27}$	18.55$_{1.31}$
Gemini-3-Flash	48.00$_{0.00}$	42.86$_{7.00}$	48.54$_{1.65}$	46.52$_{1.32}$	21.93$_{3.28}$	46.30$_{9.44}$	37.93$_{0.00}$	46.10$_{1.84}$	56.48$_{3.46}$	45.14$_{0.98}$
Gemini-3-Pro	**50.67**$_{4.99}$	**61.90**$_{1.35}$	**57.31**$_{0.83}$	**54.04**$_{2.28}$	**27.19**$_{1.24}$	**46.30**$_{5.24}$	**48.28**$_{2.82}$	**52.60**$_{3.23}$	**58.33**$_{2.27}$	**52.42**$_{0.94}$
Mistral-Large-3-675B-IT	20.00$_{3.27}$	16.19$_{1.35}$	15.79$_{3.79}$	15.32$_{2.28}$	4.39$_{1.24}$	16.67$_{4.54}$	6.90$_{2.82}$	16.23$_{0.53}$	21.30$_{3.46}$	15.15$_{1.17}$
GPT-5	20.00$_{3.27}$	26.67$_{3.56}$	26.90$_{2.19}$	30.35$_{1.40}$	24.56$_{3.28}$	18.52$_{2.62}$	34.48$_{2.82}$	33.55$_{2.51}$	45.37$_{3.46}$	30.57$_{0.40}$
GPT-5-Mini	14.67$_{6.80}$	25.71$_{2.33}$	14.04$_{2.48}$	18.58$_{1.40}$	17.54$_{4.96}$	24.07$_{11.42}$	29.89$_{3.25}$	24.89$_{1.70}$	31.48$_{1.31}$	21.32$_{1.04}$
Grok-4	28.00$_{0.00}$	24.29$_{4.29}$	27.19$_{2.63}$	26.38$_{1.28}$	18.42$_{2.63}$	16.67$_{0.00}$	34.48$_{0.00}$	22.73$_{4.55}$	31.94$_{1.39}$	25.44$_{0.88}$
Grok-4.1-Fast	26.67$_{3.77}$	20.95$_{1.35}$	17.54$_{1.43}$	20.00$_{0.35}$	9.65$_{2.48}$	7.41$_{2.62}$	12.64$_{3.25}$	22.73$_{1.59}$	25.93$_{3.46}$	19.78$_{0.13}$

Table 2: **Accuracy (%) on the hard subset by disciplines.** Overall accuracy is averaged across disciplines. Mean accuracy is reported in percentage, with standard deviation shown as a subscript. The best result is shown in **bold**.

Model	Non Korean–spec	Korean –spec	Raw gap	Controlled gap
Open-Source Multilingual Models				
Qwen3-VL-8B-IT	24.33	27.11	+2.78	+4.01
Qwen3-VL-30B-A3B-IT	30.16	22.00	-8.16	-5.97
Qwen3-VL-30B-A3B-Thinking	36.32	26.67	-9.65	-7.99
Qwen3-VL-32B-IT	36.56	26.22	-10.33	-8.09
Qwen3-VL-32B-Thinking	38.70	28.33	-10.37	-10.00
Qwen3-VL-235B-A22B-IT	40.62	27.11	-13.51	-13.43
Qwen3-VL-235B-A22B-Thinking	43.18	30.22	-12.96	-11.17
Llama-4-Maverick-17B-IT	39.96	28.44	-11.51	-12.35
Llama-4-Scout-17B-IT	34.27	27.44	-6.83	-7.97
Open-Source Korean Models				
HyperCLOVAX-SEED-Vision-3B	17.57	24.22	**+6.66**	**+5.15**
VARCO-VISION-2.0-14B	27.70	26.11	-1.59	-4.18

Table 3: **Performance on Korean-specific questions.** Raw gap is the accuracy difference between Korean-specific and non-Korean-specific questions; controlled gap is the discipline-controlled difference. Negative values indicate worse performance on Korean-specific questions. The largest positive gap is bolded, and the largest negative gap is shown in red.

Model	H-H Agr.	LLM-H Agr.	No Answer
GPT-5-mini	95.0	93.7	1
VARCO-VISION-2.0-14B	97.0	94.8	4
Qwen3-VL-2B-IT	99.0	98.0	51
Qwen3-VL-30B-A3B-IT	97.0	95.9	18
Qwen3-VL-30B-A3B-Think	91.0	88.0	22
Qwen3-VL-32B-Think	93.0	92.5	31

Table 4: **Human alignment of LLM-Judge.** We report inter-annotator agreement (H-H Agr.), agreement between human annotations and LLM-Judge (LLM-H Agr.), and the no-answer rate on 100 sampled outputs per model.

and standard deviation. For more details on judging validation analysis, see Appendix G.2.

6 Error Analysis

In this section, we conduct targeted manual error analysis by reading through selected model generations. Our analyses examine paired reversals between *Qwen3-VL-32B-IT* and *Qwen3-VL-32B-Thinking*, Korean-specific comparisons between *Qwen3-VL-235B-A22B-IT* and *HyperCLOVAX-SEED-Vision-3B*, and representative bottleneck cases from persistently difficult disciplines, with additional reference to corresponding *Qwen3-VL-235B-A22B-Thinking* outputs where relevant.

Across inspected cases, we notice that, errors are not explained by reasoning depth alone. They more often reflect failures in answer completion, gaps in domain-specific and institutional knowledge, brittle category and label mapping, and weak rule induction on symbolic problems. Reasoning helps when the evidence is already available and the challenge

lies in answer organization or completion, but its benefits are limited when success depends on exact knowledge recall or subtle category distinctions.

6.1 Post-perceptual Effects of Reasoning

Different failure patterns across quantitative domains. Although reasoning improves overall performance in some quantitative disciplines, especially *Math & Stats* (43.91→49.93), its remaining failures follow different patterns across domains. To examine this, we sampled 25 items each from *Math & Stats*, *Engineering*, and *Natural Sciences* among questions answered correctly by *Qwen3-VL-32B-IT* but incorrectly by *Qwen3-VL-32B-Thinking*, as these disciplines show contrasting reasoning effects. The clearest pattern appears in *Math & Stats*. In our inspected sample, 72% (18/25) of these reversals were not caused by obviously worse intermediate reasoning, but by *answer finalization failure*. The thinking model often developed a partially correct or plausible solution path, but stopped before producing a fully resolved final answer. By contrast, reversals in *Engineering* and *Natural Sciences* more often reflected incorrect problem framing than incomplete finalization. In these cases, the thinking model sometimes appears to map partial visual or textual cues onto a familiar *device type, curve pattern, control category*, or *physical scenario* too early, and then elaborate that interpretation into a coherent but incorrect solution (see Appendix H.2 for detailed examples).

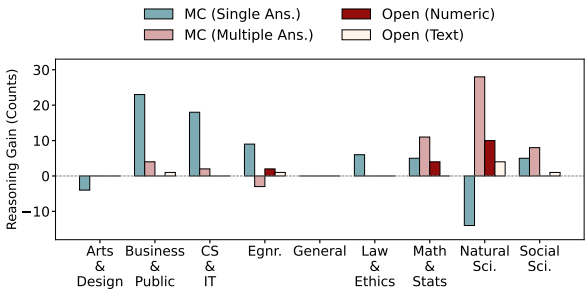


Figure 5: **Reasoning gain by discipline and question type in Qwen3-VL-32B (IT vs. Thinking).** Numbers in parentheses indicate the total number of questions in each category.

Reasoning gains in answer composition tasks. Among cases where *Qwen3-VL-32B-Thinking* succeeds and *Qwen3-VL-32B-IT* fails, the clearest gains appear on questions that require *answer composition*. This is especially visible for open-ended questions that ask for multiple requested outputs and for multiple-choice questions with multiple correct answers. In such cases, the *Thinking* variant does a notably better job in formatting their responses for questions requiring multiple outputs, while the *Instruct* variant often misses to do so, even after solving correctly. The non-reasoning model is more likely to provide a subset of the requested components whereas the reasoning model is more likely to preserve the required answer structure and return all necessary components.

This tendency is also reflected in the aggregate pattern across question types (Figure 5). Reasoning gains are large for multiple-answer and open-form questions, including numerical and text items. These findings suggest that explicit reasoning helps most with constraint tracking, structured decomposition, and complete answer assembly, so its benefits appear more in *output completeness* than in knowledge recovery, helping explain why the gains in the main results are uneven rather than uniform.

Modality-wise performance remains broadly similar, and inspected differences rarely come from one variant clearly reading or missing the image while the other does not (Appendix H.2, Figure 13).

6.2 Knowledge recall and category matching failure in Korean-specific questions

Interesingly, *Qwen3-VL-235B-A22B-IT* substantially outperforms *HyperCLOVAX-SEED-Vision-3B* on the full benchmark (39.44 vs. 18.14), the gap narrows considerably on Korean-specific questions. In a 300-item comparison, the two models achieve relatively similar performance, scoring 83/300 and 72/300, respectively. This reduced separation suggests that general reasoning ability provides limited advantage on Korean-specific items, many of which depend on regulation-specific knowledge or fine-grained administrative distinctions.

Figure 6 is a representative failure case concerning regulation-specific terminology. Korean law distinguishes 소형차 (small vehicle), defined by an engine displacement of 1000 to 1600 cc, from 승용차 (passenger vehicle), which refers to cars with up to 15 seats. However, *Qwen3-VL-235B-A22B-IT* appears to collapse both terms into the same English expression during intermediate reasoning, producing an incorrect answer. Similar patterns are reported by Son et al. (2025b), that multilingual models often translate inputs into a preferred language, introducing noise and reducing task performance. Overall, these errors suggest that Korean-specific failures arise more from localized knowledge than from image reading.

[Question - English Translation]
What is the minimum turning radius appropriate for a small vehicle specification? (Assume that the unit of specification is meters, and apply the Rules on the Standards for Road Structure and Facilities.)

(1) 6 m (2) 7 m (3) 12 m (4) 15 m

[Image]

폭	높이	길이	축간 거리	앞내민 길이	뒷내민 길이
2.0	2.8	6.0	3.7	1.0	1.3

[Error Analysis]
The model conflated small vehicle with passenger car and used the wrong regulatory value. Under the design vehicle specification table, these are separate categories with different minimum turning radii: 6.0 m for passenger cars and 7.0 m for small vehicles. Therefore, the correct answer is 7 m.

[Qwen3-VL-235B-A22B-IT Response - English Translation]
This question asks for the minimum turning radius appropriate for a small vehicle, and explicitly states that the Rules on the Standards for Road Structure and Facilities should be applied.
The given table shows the specifications of a small vehicle, including its height, length, wheelbase, front inner wheel track, and rear inner wheel track.

1. What is turning radius?
It refers to the minimum radius within which a vehicle can make a turn.
.. (truncated)
2. Applying the Rules on the Standards for Road Structure and Facilities
According to the Korean road-related regulations and the Rules on the Standards for Road Structure and Facilities issued by the Ministry of Land, Infrastructure and Transport, the minimum turning radius for a small vehicle is 6.
3. Relation to the table data
The values in the table (e.g., wheelbase 3.7 m, length 6.0 m, etc.) represent the actual dimensions of the vehicle, and in principle the turning radius could be calculated from them. However, since the question explicitly says to apply the Rules on the Standards for Road Structure and Facilities, the correct approach is to use the prescribed standard value rather than perform a separate calculation.
Therefore, according to the regulation, the minimum turning radius appropriate for a small vehicle is **6 m**.

Final answer: $\\boxed{(1) 6 m}$

Figure 6: **Example of a Korean-specific regulatory category mismatch.** *Qwen3-VL-235B-A22B-IT* reads the table correctly, but maps *small vehicle* to the wrong category and applies the wrong standard. This is a failure of institutional knowledge recall and lexical category matching, not OCR.

6.3 Disciplinary Bottlenecks

Among the subject groups, *Arts & Design* and *General* remain consistently difficult across models, suggesting bottlenecks that are not readily resolved by either scale or explicit reasoning. Error analysis indicates that the two categories are challenging for different reasons.

In *General*, many failures arise on linguistically oriented items sourced from the KLO (Korea Linguistic Olympiad) exam. Each of these problems require huge cognitive load to solve, mixing heterogeneous problem types such as linguistics and notation puzzles, Korean orthography and semantic change, dictionary ordering and some also requiring the model to infer a latent symbol-to-sound or symbol-to-word rule from a small set of examples. We observe that in most failures, models often capture only parts of the pattern, then produces a plausible but unsupported answer, which points to weak few-shot pattern induction.

In *Arts & Design*, by contrast, many items require recalling the exact expert label for a specialized visual convention. For example in Appendix H.3, while some models manage to correctly identify the visual ques of question, they fail to select the exact standardized term, especially when distinguishing between closely related expert categories such as *local* versus *partial projection*, *cutting line* versus *revolved-section line*, or similar notation symbols. As these tasks depend heavily on precise retrieval of domain-specific nomenclature, when the knowledge is absent models fail to solve, even with more parameters. Taken together, these results suggest two complementary directions for improvement. For *Arts & Design*, stronger performance may require pretraining on materials from niche domains, particularly sources that contain Korean-specific technical terminology and conventions. For *General*, gains may depend more on post-training with instruction data that imposes higher cognitive load, requiring models to coordinate multiple abilities, such as pattern induction, linguistic reasoning, and knowledge retrieval, within a single problem.

7 Conclusion

We introduce **KMMMU**, a native Korean benchmark for expert-level multimodal understanding in culturally and institutionally grounded settings. Across 3,466 carefully verified questions, KMMMU shows that current MLLMs remain far from robust on Korean real-world assessment materials. Our findings suggest that key failures arise less from insufficient reasoning depth than from weak convention-to-label mapping, few-shot symbolic induction, localized knowledge recall, and familiarity with domain-specific standards and terminology. These bottlenecks help explain the limited and uneven gains from reasoning, the persistent difficulty of Korean-specific content, and strong disciplinary variation in performance. We hope KMMMU to serve as a rigorous benchmark for evaluating expert-level Korean multimodal understanding and a practical testbed for developing more culturally grounded and institutionally aware MLLMs.

Limitations

Coverage and representativeness. Although KMMMU spans many disciplines, it is not a comprehensive model of all real-world multimodal use cases. The benchmark is exam-centric and emphasizes information-dense, structure-heavy visuals, so performance may not directly transfer to everyday perception, interactive settings, or non-exam domains.

Annotation noise and taxonomy subjectivity. The discipline, visual modality, and Korean-specific labels are generated from an LLM-assisted annotation pipeline, in which model-proposed labels are later consolidated by human annotators. This design improves scalability, but it also introduces a potential source of noise, since the initial model proposals may be imperfect and some category boundaries are inherently ambiguous. Although we audit a random subset and manually verify all Korean-specific items, some residual label noise is likely to remain, especially for fine-grained disciplines and multi-skill questions.

Uncertainty about data contamination. Data contamination remains an important concern for benchmarking, especially because model developers rarely disclose training data with enough granularity to enable direct verification. As a result, we cannot precisely determine whether some KMMMU items, source documents, or near-duplicate variants were included in pretraining corpora. Our construction choices provide only partial mitigation: many questions are digitized from official exam materials instead of being directly collected from web QA repositories, and the final benchmark retains only items unsolved by multiple strong models. The relatively low performance of current systems also suggests that widespread contamination is unlikely to fully explain the benchmark results. We include supplementary contamination analyses in the Appendix I, but they offer only indirect evidence. A more rigorous assessment would require substantially greater transparency about model training data than is currently available.

Evaluation noise for mixed-format answers. Because KMMMU includes both multiple-choice and free-form items, scalable evaluation relies on LLM-Judge, which can be sensitive to prompt design and answer formatting. Despite using deterministic decoding and spot-checks, some grading errors may remain, particularly when responses are verbose, underspecified, or unconventional in format.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Anthropic. 2025. Introducing Claude Opus 4.5. Anthropic Newsroom. Published: 2025-11-24. Accessed: 2026-01-05.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*.

Young-rok Cha, Jeongho Ju, SunYoung Park, Jong-Hyeon Lee, Younghyun Yu, and Youngjune Kim. 2025. Varco-vision-2.0 technical report. *arXiv preprint arXiv:2509.10105*.

Byungjin Choi, Seongsu Bae, Sunjun Kweon, and Edward Choi. 2026a. Kormedmcqa-v: A multimodal benchmark for evaluating vision-language models on the korean medical licensing examination. *arXiv preprint arXiv:2602.13650*.

Dasol Choi, Guijin Son, Hanwool Lee, Minhyuk Kim, Hyunwoo Ko, Teabin Lim, Ahn Eungyeol, Jungwhan Kim, Seunghyeok Hong, and Youngsook Song. 2026b. What users leave unsaid: Under-specified queries limit vision-language models. *arXiv preprint arXiv:2601.06165*.

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*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Google Cloud. 2025a. Gemini 3 Flash (Preview) | Generative AI on Vertex AI. Google Cloud Documentation. Release date: 2025-12-17. Accessed: 2026-01-05.

Google Cloud. 2025b. Gemini 3 Pro (Preview) | Generative AI on Vertex AI. Google Cloud Documentation. Release date: 2025-11-18. Accessed: 2026-01-05.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. 2025. From kmmlu-redux to kmmlu-pro: A professional korean benchmark suite for llm evaluation. *arXiv preprint arXiv:2507.08924*.

Taebaek Hwang, Minseo Kim, Gisang Lee, Seonuk Kim, and Hyunjun Eun. 2025. Kreta: A benchmark for korean reading and reasoning in text-rich vqa attuned to diverse visual contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33409–33420.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084.

Jeongho Ju, Daeyoung Kim, SunYoung Park, and Youngjune Kim. 2024. Varco-vision: Expanding frontiers in korean vision-language models. *Preprint*, arXiv:2411.19103.

Yoonshik Kim and Jaeyoon Jung. 2025. Koffvqa: An objectively evaluated free-form vqa benchmark for large vision-language models in the korean language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 575–585.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International conference on machine learning*, pages 1078–1088. Pmlr.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

Meta. 2025a. meta-llama/Llama-4-Maverick-17B-128E-Instruct. Hugging Face model card and weights. Model release date: 2025-04-05. Accessed: 2026-01-05.

Meta. 2025b. meta-llama/Llama-4-Scout-17B-16E. Hugging Face model card and weights. Model release date: 2025-04-05. Accessed: 2026-01-05.

Mistral AI. 2025. Mistral Large 3 (v25.12) | Mistral Docs. Mistral Documentation. Dated: 2025-12-02. Accessed: 2026-01-05.

NAVER HyperCLOVAX. 2025. HyperCLOVAX-SEED-Vision-Instruct-3B. Hugging Face model card and weights. Model release date: 2025-04-24 (as stated in repository license). Accessed: 2026-01-05.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

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 (Volume 1: Long Papers)*, pages 932–950.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). https://huggingface.co/datasets/openai/MMMLU.

OpenAI. 2025. Introducing GPT-5. OpenAI. Published: 2025-08-07. Accessed: 2026-01-05.

Sanghee Park and Geewook Kim. 2025. Evaluating multimodal generative ai with korean educational standards. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 671–688.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025a. 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 (Volume 1: Long Papers)*, pages 4076–4104.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Chanuk Lim, Srikant Panda, Minhyuk Kim, Nikunj Drolia, Dasol Choi, and 1 others. 2025b. Pushing on multilingual reasoning models with language-mixed chain-of-thought. *arXiv preprint arXiv:2510.04230*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

Gemini Team and Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

xAI. 2026. Models and Pricing (xAI API Documentation). xAI Developer Docs. Accessed: 2026-01-05.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 93–104.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, and 1 others. 2024. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv: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. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2299–2314.

A Data Sources and Collection Scope

KMMMU is collected from four high stakes sources in South Korea. We summarize the collection scope for each source below.

A.1 PSAT

We annotate ten years of past examinations from civil service recruitment tracks. The PSAT includes Language Logic, Data Interpretation, and Situational Judgment sections that assess logical reasoning and information integration.

A.2 National Technical Qualifications

We collect fifteen years of questions from 252 distinct certification exams, including Information Processing Engineer, Electric Engineer, and Fire Safety Manager. These exams cover a wide range of technical domains across industrial and engineering fields.

A.3 Olympiads

To incorporate academically challenging reasoning problems, we gather ten years of Olympiad questions spanning middle school, high school, and

university levels. The collected problems focus primarily on mathematics and science.

A.4 NCS

We include three years of National Competency Standards examinations covering all ten competency areas, such as Communication, Numeracy, and Problem Solving. These exams are used in recruitment for public sector organizations.

B Annotation and Quality Control Details

The construction of KMMMU uses a rigorous pipeline that combines automated processing with human verification to ensure high data fidelity.

B.1 Human Verification Interface

We utilized a custom built annotation tool to verify and correct the output of the OCR pipeline. Raw data digitized by MinerU-2.5 (Niu et al., 2025) often contained artifacts and formula errors. Figure 7 shows the interface where five Korean annotators reviewed the parsed content against the original PDF source. Annotators were instructed to

- Correct LaTeX formatting for mathematical formulas
- Verify that image references in the text matched the cropped images
- Discard questions where essential visual information was illegible or missing.

All five annotators are native Korean speakers with at least a bachelor's degree and prior experience in annotation or dataset curation. They are also familiar with AI-related workflows, which helped them reliably identify OCR artifacts, formula corruption, and image–text mismatches during verification.

B.2 Automatic Labeling and Taxonomy Consolidation

We annotate several auxiliary attributes to support analysis and stratified reporting, including discipline, visual modality type, question format and a Korean-specific flag. All taxonomy labels are assigned using Gemini-2.5-Flash. For each labeling job, the model is given the question text and its associated image, and outputs the most appropriate label.

We use an open labeling step that does not constrain predictions to a fixed label set. This reduces forced assignments when an instance does not cleanly match a predefined taxonomy. All label types are generated independently.

Manual audit and consolidation We conduct a manual audit by randomly sampling around 300 instances and reviewing the assigned labels. Based on the audited outputs, we consolidate the discipline taxonomy through human curation into 45 sub-discipline categories and 9 macro discipline categories.

Verification of Korean-specific cases Because false positives can inflate localization analyses, we manually verify all instances labeled as Korean-specific. We confirm that each positive case requires Korean-specific knowledge or context rather than general world knowledge expressed in Korean.

B.3 Adversarial Filtering Protocol

To ensure benchmark difficulty, we apply a multi-stage adversarial filtering pipeline (Zellers et al., 2018; Le Bras et al., 2020) that removes instances solvable by current multimodal models without advanced reasoning. Starting from a manually verified pool of approximately 68,000 questions, we apply the following procedure.

1. **Data cleaning and de-duplication.** We first remove samples with invalid image links and de-duplicate near-duplicate questions across exam years using image and text similarity checks.
2. **Model-based adversarial filtering.** We then sequentially filter the remaining candidate pool using four multimodal models: PHI-3.5-VISION-INSTRUCT (Abdin et al., 2024), INTERNVL-3.5-38B (Wang et al., 2025), GEMINI-2.5-FLASH-LITE, and GEMINI-2.5-FLASH (Comanici et al., 2025). Each model is evaluated in a zero-shot setting, and questions answered correctly at each stage are removed from the candidate pool.
3. **Final retention.** Only questions that remain unsolved after all four filtering stages are retained in the final benchmark.

The resulting KMMMU benchmark contains **3,466** curated questions.

C Korean-Specific Context

To provide a concrete illustration of KMMMU, Figure 8 presents a *Korean-Specific* instance from

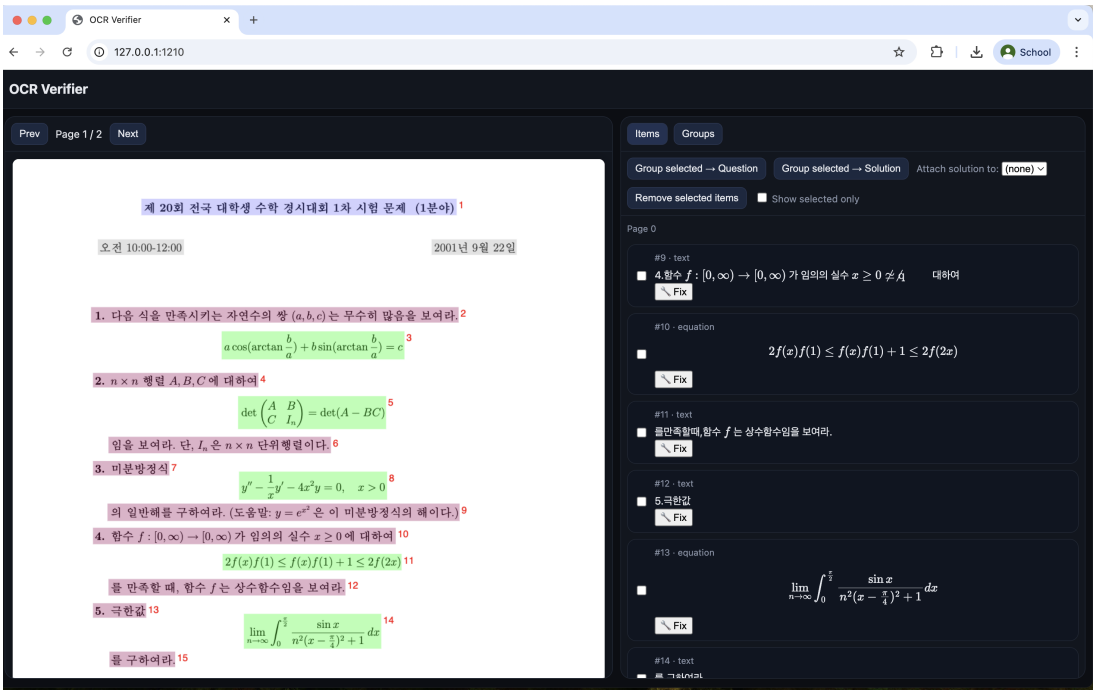


Figure 7: **Annotation tool interface used for OCR verification.** The tool displays the original PDF page on the left and the parsed text and images on the right, allowing annotators to correct OCR errors and validate image cropping in real time.

the benchmark. Unlike standard multimodal benchmarks, which often emphasize culturally invariant knowledge such as Physics or Mathematics, KMMMU includes a dedicated subset of questions that require localized knowledge grounded in Korean institutional and legal contexts. In this example, the input consists of an image containing regulation text and a corresponding question, and the model must interpret the visual text referring to the "extraction area slope criteria" in the specific context of South Korea's *Mountainous Districts Management Act* to identify the correct legal standard (Option 3). This example shows that solving such questions requires not only optical character recognition, but also grounded knowledge of Korean administrative law.

D Detailed Dataset Statistics

In this section, we provide a granular breakdown of the dataset composition. Beyond the overview (Table 5), we report (i) the distribution of fine-grained disciplin categories (Table 6), (ii) the question format distribution (Table 7), and (iii) the examples of visual modalities (sub-visual modalities, Figure 9)

D.1 Discipline Category Distribution

Table 6 details the frequency of questions across 45 fine-grained Discipline categories. The distribution reflects the emphasis on STEM (Science, Technology, Engineering, and Mathematics) fields, with *Physics*, *Civil Engineering*, and *Mechanical Engineering* constituting the largest portions. This heavy tail in engineering disciplines ensures that KMMMU serves as a robust benchmark for technical domain expertise.

Statistic	Count
Total Questions	3,466
Hard Subset	627
Discipline Categories	9
Sub-discipline Categories	45
Visual Modality Types	9
Question Types	5
Questions with in-image texts	2,383 (68.75%)
Korean-specific questions	300 (8.65%)

Table 5: **Dataset distribution outline.** We report counts and percentages for key attributes such as in-image text and Korean-specific content.

D.2 Question Format Distribution

Because KMMMU contains both multiple-choice and free-form items, the answer format affects evaluation difficulty and failure modes. Table 7 reports the distribution of question formats in the benchmark.

Korean Specific Question

[Original Image]

채취지역의 평균 경사도는 (①) 이하이어야 하고, 채취등을 완료한 후 절개사면의 기울기(비탈면의 높이에 대한 수평거리의 비율을 말한다)는 건축용 석재인 경우에는 (②)기준에 적합할 것. 다만, 채취등을 함으로써 절개사면 없이 평탄지로 될 수 있는 경우에는 그러하지 아니하다.

[Original Question]

문제: 산지관리법령산 대통령령이 정하는 토석채취허가기준 중 산지의 경사도에 관한 기준이다. ( )안에 알맞은 것은?

선택지:
(1) ① 45도, ② 1: 0.5 이하
(2) ① 45도, ② 1: 0.4 이하
(3) ① 35도, ② 1: 0.5 이하
(4) ① 35도, ② 1: 0.4 이하

[Image Context Translation]

Image Context: The average slope of the extraction area must be ① or less. Regarding the slope of the cut surface after extraction... it must comply with the criteria of ②

[English Translation]

Question: These are the criteria for soil and stone extraction permits determined by the Presidential Decree under the *Mountainous Districts Management Act*. What fits in the parentheses ( )?
Options:
(1) ① 45˚ , ② 1:0.5 or less
(2) ① 45˚ , ② 1:0.4 or less
(3) ① 35˚ , ② 1:0.5 or less
(4) ① 35˚ , ② 1:0.4 or less

Figure 8: **Data Card for a Korean-Specific Question.** The figure aggregates the raw inputs and their translations. **[Original Image]** The original visual input containing a text-rich regulation box. **[Original Question]** The original question text in Korean. **[Translation]** English translations for both the visual context and the question. Correctly answering this question requires retrieving specific legal provisions regarding slope limits for soil extraction permits in South Korea, demonstrating the benchmark's focus on localized expert knowledge.

D.3 Visual Modality Taxonomy

KMMMU includes a wide range of fine-grained visual types, including circuit, mechanical, and structural diagrams, document-style text images, tables, mathematical figures, charts, maps, symbols, and photographs. For analysis, we consolidate these fine-grained types into 9 macro-level visual modality categories. Technical diagrams constitute a particularly large portion of the dataset, reflecting KMMMU's emphasis on professional and schematic visual reasoning. Figure 9 presents representative examples of these fine-grained visual types before consolidation.

D.4 Question Type Taxonomy

Table 8 summarizes the distribution of answer formats within each macro subject. This table clarifies which subjects are dominated by multiple choice items versus numerical or descriptive responses. It also provides context for interpreting task-wise performance, since answer format affects both evaluation difficulty and failure modes.

E Additional Distributional Analysis

Figure 10 provides per-dimension density comparisons for the original 68k corpus, the KMMMU *Full set*, and the *Hard subset* in the 3D UMAP space. Across all three dimensions, the filtered subsets broadly preserve the major density peaks and overall multimodal structure of the original distribution. The *Full set* remains especially close to the original corpus, while the *Hard subset* shows a somewhat larger shift in parts of the latent space.

To quantify these differences, we compute the Kullback–Leibler (KL) divergence between the original distribution and each filtered subset along each UMAP dimension. For the *Full set*, the divergence remains low across all three dimensions ($D_{KL} = 0.1184$, 0.1459, and 0.1437 for Dimensions 1–3, respectively). The *Hard subset* shows similarly low divergence on Dimensions 1 and 2 (0.1081 and 0.1699), but a larger deviation on Dimension 3 (0.3747). Overall, these results are consistent with the main-text UMAP visualization: adversarial filtering increases difficulty while largely

Subject	Cnt	Subject	Cnt	Subject	Cnt
Physics	474	General Knowledge & Interdisciplinary	67	Economics	33
Civil & Structural Engineering	332	Industrial & Systems Engineering	64	Earth & Geological Sciences	33
Mechanical Engineering	223	Architecture & Urban Studies	57	Library, Archival & Information Science	30
Electrical & Electronics Engineering	221	Human Resources & Organizational Studies	56	Transportation & Logistics	29
Computer Science	164	Linguistics & Language Studies	54	Agriculture & Life Sciences	27
Statistics & Probability	149	Biology	43	Software & Programming	27
Mathematics	121	Safety, Risk & Reliability Engineering	42	Astronomy & Space Science	26
Business Administration & Management	104	Data Science & Analytics	41	Psychology	22
Environmental Science & Engineering	99	Finance & Accounting	40	Cognitive & Behavioral Sciences	20
Geography & Spatial Studies	95	Sociology & Social Sciences	36	Biomedical & Health Sciences	20
Manufacturing & Production Engineering	90	Chemical Engineering	76	Communication & Media Studies	17
Public Administration & Policy	89	Chemistry	75	Education & Pedagogy	16
Design & Visual Arts	88	Law & Legal Studies	74	Marketing & Consumer Studies	15
Chemical Engineering	76	Materials Science & Metallurgy	69	Ethics & Philosophy	12
Chemistry	75	Information Technology & Systems	69	Artificial Intelligence & Machine Learning	12

Table 6: **Distribution of sub-discipline categories in KMMMU**.

Question Type	Count	%
Multiple Choice (Single Answer)	2,831	81.68
Multiple Choice (Multiple Answers)	346	9.98
Open Format (Numerical / Calculation)	207	5.98
Open Format (Text)	67	1.93
Essay (Descriptive)	15	0.43
Total	3466	100

Table 7: **Question type distribution.**

preserving the broader distributional structure of the original corpus.

F Analysis of the Hard Subset

F.1 Distributional Characteristics of the Hard Subset

We analyze the structural composition of the hard subset to better understand the types of instances that contribute to systematic model failures.

We first examine the prevalence of Korean-specific items (Table 9). Korean-specific questions account for 8.65% of the full set (300/3,466), but 12.12% of the hard subset (76/627). This increase suggests that localized Korean content is somewhat overrepresented among harder examples. However, because the majority of hard-subset questions are still not Korean-specific, localization alone does not fully explain the difficulty of the subset.

F.2 Model Performance on Hard subset

Table 2 reports accuracy on the adversarially filtered hard subset, obtained by restricting full-set evaluation outputs to the retained hard-subset instances. Performance drops substantially relative to the full-set results across nearly all models, confirming that the hard subset is meaningfully more difficult.

Even the strongest model remains below 20% overall accuracy, with VARCO-VISION-2.0-1.7B achieving the highest overall score at 19.56%. This result suggests that adversarial filtering successfully removes many easier instances while preserving questions that remain challenging even for relatively strong multimodal systems.

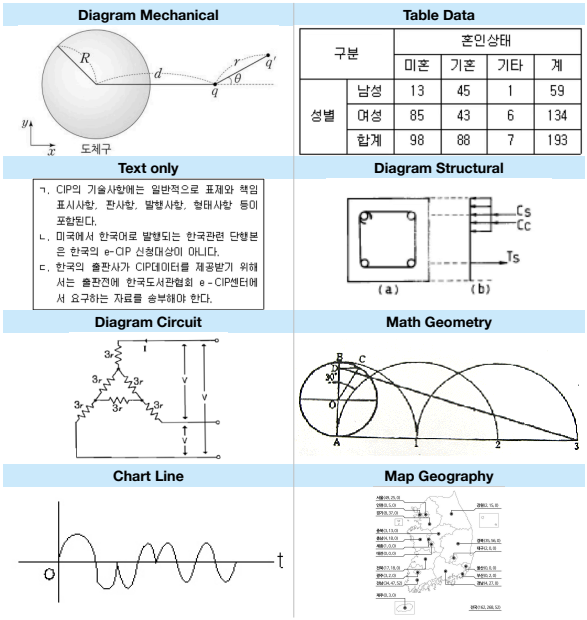


Figure 9: **Representative examples of fine-grained visual types in KMMMU.** Before consolidation into the final macro-level visual modality categories, the dataset included diverse fine-grained visual types, such as specialized engineering diagrams, document-style text images, and South Korean geographic maps.

Performance also varies considerably across disciplines. For example, *Law & Ethics* and *Arts & Design* remain difficult for most models, while *Engineering* and *Natural Sciences* still show modest separation among stronger systems. At the same time, reasoning models do not exhibit a consistent advantage over non-reasoning models on this subset. This pattern suggests that many hard-subset failures arise not simply from insufficient chain-of-

Macro Subject	Multiple Choice (Single Answer)	Multiple Choice (Multiple Answers)	Open Format (Numerical)	Open Format (Text)	Essay (Descriptive)
Engineering	1182	25	20	2	0
Natural Sciences	450	180	154	27	11
CS & IT	311	11	2	0	0
Business & Public	263	37	2	2	0
Math & Stats	161	36	27	3	0
Social Sciences	193	55	2	4	0
General/Interdisciplinary	89	0	0	30	4
Arts & Design	79	0	0	0	0
Law & Ethics	103	1	0	0	0

Table 8: **Dataset attributes by macro subject and answer type.**

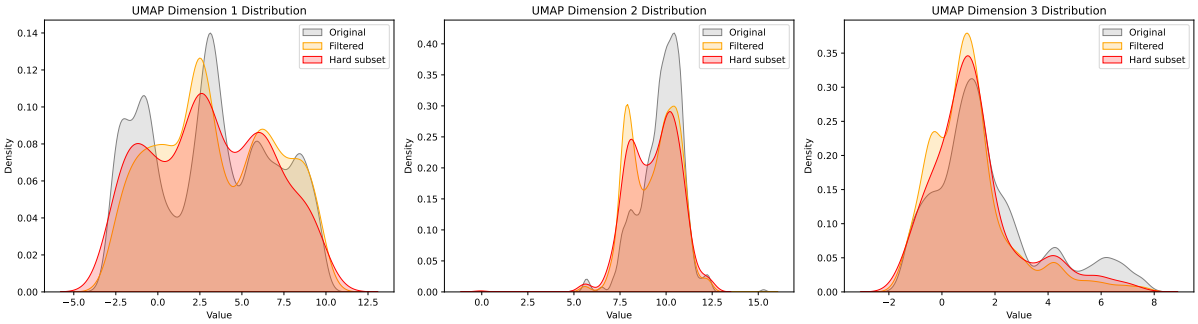


Figure 10: **Per-dimension density comparison after adversarial filtering.** Kernel density estimates over the three UMAP dimensions for the original 68k corpus, the KMMMU *Full set*, and the *Hard subset*. The filtered subsets broadly retain the major density peaks and multimodal trends of the original distribution, although the *Hard subset* shows a somewhat larger deviation in Dimension 3.

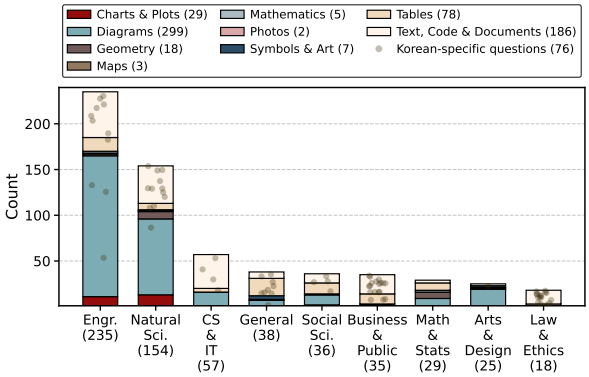


Figure 11: **Discipline-wise visual modality composition of KMMMU Hard Set.** Stacked bars show the number of ques- tions for each visual modality in each discipline, with total counts shown beneath the labels. Scatter points indicate Korean-specific items overlaid on the corre- sponding discipline–modality segments. The hard subset is concentrated in *Engineering* and *Natural Sciences*, similar to Full set.

thought depth, but from more persistent limitations in knowledge, grounding, visual interpretation, and answer execution.

Split	Korean Specific	Not Korean Specific	Ratio (%)
Full	300	3166	8.65
Hard	76	551	12.12

Table 9: **Number of Korean-specific questions each in Full set and Hard subset.**

G Reliability of LLM-Judge

G.1 Annotation Protocol

To validate the reliability of our evaluation pipeline, we conducted a manual annotation study using a custom annotation interface (Figure 12). Three annotators independently reviewed each sample with access to the question, associated image, gold answer, model response, and parsed answer.

For each sample, annotators first evaluated whether the parsed answer faithfully reflected the answer expressed in the original model response, labeling it as *match*, *mismatch*, or *no_answer*. Here, *mismatch* indicates that the parser failed to preserve the intended answer, while *no_answer* indicates that the model response itself did not contain a complete answer.

Annotators then assessed correctness with respect to the gold answer in two ways: once based on the full model response and once based on the

Model	Arts & Design	Business & Public	CS & IT	Engineering	General	Law & Ethics	Math & Stats	Natural Sciences	Social Sciences	Overall Acc.
Open-Source Multilingual Non-Reasoning Models										
Gemma-3-4B-IT	$22.67_{1.89}$	$21.90_{1.35}$	$9.94_{4.14}$	$19.01_{2.95}$	$14.04_{1.24}$	$18.52_{6.93}$	$10.34_{2.82}$	$10.82_{1.33}$	$24.07_{5.24}$	$16.06_{1.43}$
Gemma-3-12B-IT	$14.67_{3.77}$	$8.57_{4.04}$	$11.70_{4.60}$	$15.32_{1.84}$	$6.14_{3.28}$	$22.22_{0.00}$	$5.75_{4.30}$	$8.87_{1.70}$	$12.04_{7.29}$	$12.01_{0.72}$
Gemma-3-27B-IT	$16.44_{8.46}$	$11.63_{10.45}$	$15.11_{7.09}$	$14.04_{4.78}$	$14.04_{12.59}$	$16.20_{15.06}$	$9.51_{8.57}$	$12.61_{6.48}$	$16.24_{0.60}$	$13.72_{5.88}$
Llama-4-Scout-17B-16E-IT	$13.33_{3.77}$	$9.52_{1.35}$	$9.36_{2.19}$	$17.73_{0.53}$	$10.53_{2.15}$	$20.37_{2.62}$	$6.90_{2.82}$	$11.26_{2.72}$	$19.44_{4.54}$	$13.98_{0.53}$
Llama-4-Maverick-17B-128E-IT	$17.33_{3.77}$	$13.33_{1.35}$	$20.47_{1.65}$	$16.74_{1.06}$	$10.53_{2.15}$	$22.22_{4.54}$	$17.24_{4.88}$	$14.29_{1.91}$	$21.30_{3.46}$	$16.37_{1.08}$
Qwen3-VL-2B-IT	$25.33_{3.77}$	$9.52_{4.86}$	$12.87_{2.98}$	$11.49_{1.84}$	$3.51_{2.48}$	$12.96_{2.62}$	$4.60_{3.25}$	$6.28_{0.31}$	$6.48_{1.31}$	$9.73_{1.14}$
Qwen3-VL-4B-IT	$18.67_{1.89}$	$10.48_{3.56}$	$11.70_{0.83}$	$12.34_{0.35}$	$3.51_{1.24}$	$3.70_{2.62}$	$3.45_{2.82}$	$4.98_{0.61}$	$8.33_{2.27}$	$9.20_{0.08}$
Qwen3-VL-8B-IT	$20.00_{3.27}$	$14.29_{2.33}$	$16.37_{3.60}$	$14.04_{0.35}$	$5.26_{0.00}$	$5.56_{0.00}$	$6.90_{2.82}$	$8.01_{0.81}$	$14.81_{2.62}$	$11.96_{0.13}$
Qwen3-VL-30B-A3B-IT	$21.33_{8.22}$	$14.29_{4.04}$	$6.43_{2.19}$	$13.62_{0.92}$	$4.39_{2.48}$	$9.26_{6.93}$	$8.05_{1.63}$	$10.17_{0.61}$	$16.67_{6.00}$	$11.70_{1.21}$
Qwen3-VL-32B-IT	$14.67_{4.99}$	$12.38_{4.86}$	$14.62_{3.60}$	$13.90_{1.45}$	$7.02_{3.28}$	$11.11_{4.54}$	$9.20_{1.63}$	$11.04_{1.40}$	$21.30_{3.46}$	$12.92_{0.78}$
Qwen3-VL-235B-A22B-IT	$9.33_{4.99}$	$12.38_{2.69}$	$13.45_{1.65}$	$17.02_{0.92}$	$7.89_{4.30}$	$7.41_{5.24}$	$13.79_{2.82}$	$14.94_{1.06}$	$24.07_{2.62}$	$15.05_{0.87}$
Open-Source Korean Non-Reasoning Models										
HyperCLOVAX-SEED-Vision-3B	$9.33_{1.89}$	$11.43_{0.00}$	$16.37_{2.19}$	$15.46_{0.53}$	$4.39_{1.24}$	$11.11_{0.00}$	$8.05_{1.63}$	$8.66_{0.61}$	$14.81_{2.62}$	$12.23_{0.27}$
VARCO-VISION-2.0-1.7B	$20.00_{3.27}$	$23.81_{4.86}$	$22.81_{7.58}$	$24.40_{1.12}$	$3.51_{1.24}$	$24.07_{6.93}$	$13.79_{4.88}$	$15.15_{2.14}$	$16.67_{0.00}$	$19.56_{0.72}$
VARCO-VISION-2.0-14B	$17.33_{1.89}$	$21.90_{4.86}$	$19.88_{0.83}$	$20.43_{1.81}$	$7.02_{1.24}$	$27.78_{12.00}$	$12.64_{4.30}$	$11.47_{0.61}$	$16.67_{3.93}$	$16.96_{1.54}$
Open-Source Multilingual Reasoning Models										
Qwen3-30B-A3B-Thinking	$18.00_{2.00}$	$11.43_{0.00}$	$4.39_{2.63}$	$12.55_{0.21}$	$6.58_{1.32}$	$16.67_{0.00}$	$15.52_{1.72}$	$11.69_{0.65}$	$15.28_{4.17}$	$11.80_{0.48}$
Qwen3-32B-Thinking	$22.67_{1.89}$	$6.67_{4.86}$	$16.37_{1.65}$	$12.34_{1.04}$	$6.14_{1.24}$	$16.67_{4.54}$	$14.94_{4.30}$	$11.47_{3.84}$	$13.89_{6.00}$	$12.55_{0.53}$
Qwen3-VL-235B-A22B-Thinking	$4.00_{3.27}$	$6.67_{5.87}$	$3.51_{2.48}$	$7.80_{5.54}$	$4.39_{6.20}$	$7.41_{5.24}$	$6.90_{7.45}$	$9.31_{6.59}$	$13.89_{10.39}$	$7.66_{5.44}$

Table 10: **Accuracy (%) on the KMMMU hard subset by disciplines.** This table reports results recomputed by restricting full-set evaluation outputs to the adversarially filtered hard subset. Overall accuracy is averaged across all disciplines. Mean accuracy is reported in percentage, with standard deviation shown as a subscript. The best result for each discipline and overall accuracy is shown in bold.

Model	Human agr.	Human κ	Parsed agr.	Parsed κ	Response agr.	Response κ	No-answer rate
Qwen3-2B-IT	0.99	0.967	0.980	0.935	0.980	0.935	0.51
Qwen3-30B-IT	0.97	0.930	0.959	0.905	0.928	0.845	0.18
Qwen3-32B-Thinking	0.93	0.656	0.925	0.332	0.495	0.018	0.31
Qwen3-30B-Thinking	0.91	0.818	0.880	0.531	0.326	0.544	0.22
VARCO2-14B	0.97	0.917	0.948	0.855	0.929	0.805	0.04
GPT-5-Mini	0.95	0.898	0.937	0.872	0.892	0.785	0.01
Average	0.95	0.864	0.921	0.739	0.742	0.468	0.21

Table 11: **Human agreement and judge–human alignment across six model runs.** "Human agr." denotes pairwise human agreement, and "Human κ" the corresponding Cohen's κ. "Parsed" and "Response" report judge–human alignment under parsed-answer-based and full-response-based evaluation, respectively.

parsed answer, each labeled as *correct*, *incorrect*, or *no_answer*. This design allowed us to distinguish parsing failures from genuine model errors. In addition, annotators verified the consistency of the recorded question type and image type using *match*, *mismatch*, or *unsure*, and could provide free-form comments for ambiguous cases.

G.2 Human Alignment Results

The alignment study contains 600 examples drawn from six model runs (100 outputs each), balanced across question formats. Table 11 reports overall pairwise human agreement and judge–human alignment across these runs. Human agreement is consistently high, with pairwise agreement ranging from 0.91 to 0.99 and Cohen's κ ranging from 0.66 to 0.97. This indicates that the annotation task is generally well defined, although agreement becomes weaker for some reasoning-heavy outputs.

Overall, parsed-answer-based judging aligns substantially better with human labels than full-response judging. Averaged across the six runs, parsed-answer-based judging achieves 0.921 agreement and 0.739 Cohen's κ, compared with 0.742 agreement and 0.468 κ for full-response judging. This gap is especially pronounced for reasoning models, where long responses often contain partially correct intermediate reasoning without a clearly finalized answer.

Table 12 further decomposes judge–human alignment by response completeness. The advantage of parsed-answer-based judging is strongest on *no_answer* cases: averaged across runs, it reaches 0.98 accuracy, whereas full-response judging drops to 0.67. For this subset, we emphasize accuracy rather than Cohen's κ, since label imbalance is severe and κ becomes less stable and less informative.

Parsed-answer-based judging also remains stronger on answered cases. Across the six runs, it achieves 0.910 accuracy and 0.741 Cohen's κ, compared with 0.796 accuracy and 0.665 κ for

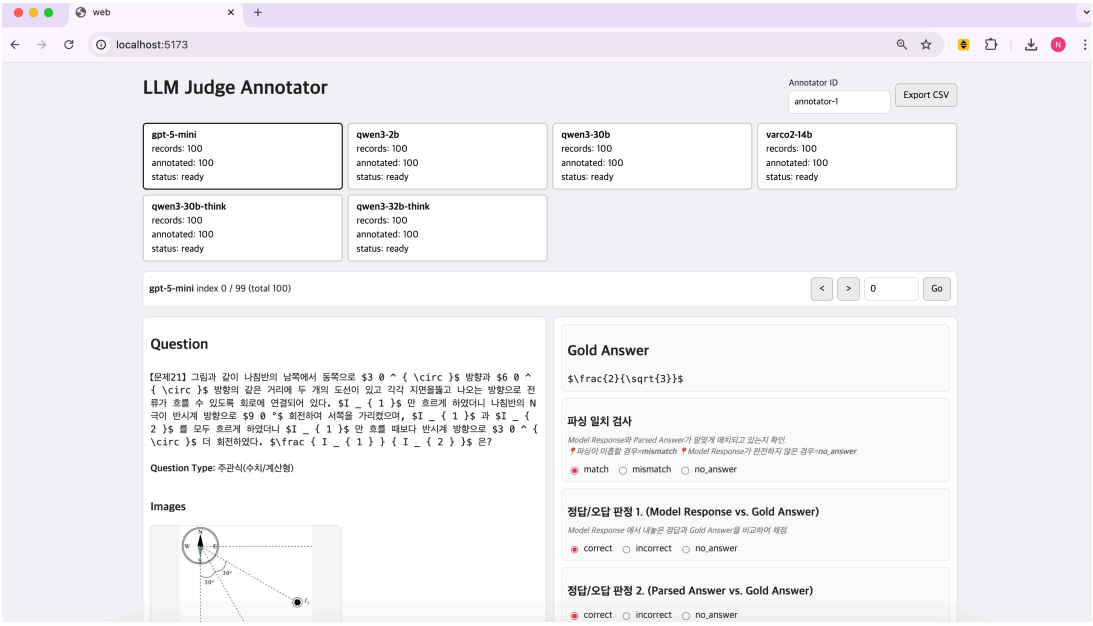


Figure 12: **Annotation interface for manual validation of LLM-Judge outputs.** For each sample, annotators review the question, image, gold answer, model response, and parsed answer, and record parsing consistency, correctness judgments, metadata consistency, and optional comments.

Model	Parsed acc. (*no answer*)	Response acc. (*no answer*)	Parsed acc. (answered)	Parsed κ (answered)	Response acc. (answered)	Response κ (answered)
Qwen3-2B-IT	0.96	0.96	1.000	1.000	1.000	1.000
Qwen3-30B-IT	0.94	0.72	0.962	0.920	0.975	0.948
VARCO2-14B	1.00	1.00	0.946	0.853	0.926	0.802
GPT-5-Mini	1.00	1.00	0.936	0.871	0.891	0.783
Qwen3-32B-Thinking	1.00	0.20	0.889	0.316	0.734	0.277
Qwen3-30B-Thinking	0.95	0.15	0.729	0.486	0.347	0.179
Average	0.98	0.67	0.910	0.741	0.796	0.665

Table 12: **Judge–human alignment broken down by response completeness.** Parsed-answer-based judging remains more robust on incomplete responses and generally aligns better with human labels on answered cases as well. For the *no_answer* subset, accuracy is more informative than Cohen's κ because of severe label imbalance.

full-response judging. Thus, the benefit of parsed-answer-based evaluation is not limited to incomplete outputs; it also improves alignment on responses that contain a final answer.

Manual inspection of disagreement cases suggests that many residual mismatches arise from answer-formatting and completion issues rather than broad evaluator failure. In multiple-choice questions, some models produce the content of the correct option rather than its explicit index, which can cause an otherwise correct response to be judged as incorrect. More broadly, disagreement is concentrated in cases where the response contains extended or partially correct reasoning but fails to end with a clearly finalized answer. Taken together, these results support our use of parsed-answer-based judging as the primary evaluation protocol, especially for long or reasoning-heavy model outputs.

H Error Analysis Details

H.1 Error Inspection Methodology

To investigate the mechanisms underlying the patterns in Tables 1–3, we conducted targeted manual error inspection over three focused subsets.

First, for paired reasoning comparison, we examined reversal cases between *Qwen3-VL-32B-IT* and *Qwen3-VL-32B-Thinking*. To analyze domain-specific reasoning effects, we sampled 25 items each from *Math & Stats*, *Engineering*, and *Natural Sciences* among questions answered correctly by *Qwen3-VL-32B-IT* but incorrectly by *Qwen3-VL-32B-Thinking*, yielding 75 inspected reversals in total.

Second, for Korean-specific failures, we analyzed incorrect outputs from a 300-item comparison set between *Qwen3-VL-235B-A22B-IT* and *HyperCLOVAX-SEED-Vision-3B*. We randomly sampled 25 incorrect cases from each model for qualitative inspection, focusing on recurring pat-

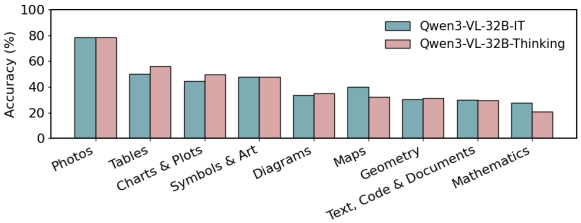


Figure 13: **Accuracy by visual modality for Qwen3-VL-32B-IT and Qwen3-VL-32B-Thinking.** Performance remains broadly similar across visual modality categories, suggesting that explicit reasoning does not systematically change raw visual evidence extraction. The main differences appear to arise after evidence extraction, such as in task framing, constraint tracking, and answer finalization.

terns of localized knowledge failure, regulatory category mismatch, and terminology grounding errors.

Third, to characterize persistent disciplinary bottlenecks, we additionally inspected representative failure cases from *Arts & Design* and *General*, focusing on *Qwen3-VL-235B-A22B-IT* with reference to corresponding *Qwen3-VL-235B-A22B-Thinking* outputs where relevant.

Each inspected case was reviewed by two authors, who examined the image, question, model output, and ground-truth answer. Disagreements were resolved through discussion.

H.2 Additional Qualitative Examples for Post-perceptual Reasoning Effects

Figure 13 reports performance by visual modality for *Qwen3-VL-32B-IT* and *Qwen3-VL-32B-Thinking*. Across most modality categories, the two variants remain broadly similar, with no consistent pattern indicating that explicit reasoning systematically improves raw visual evidence extraction.

To complement the aggregate modality comparison, we include a representative reversal case from *Natural Sciences* where *Qwen3-VL-32B-IT* answers correctly but *Qwen3-VL-32B-Thinking* fails in Figure 14. This example illustrates the broader pattern discussed in the main text: the difference does not arise from one variant clearly seeing the image while the other does not, but from how the extracted evidence is framed and translated into a final judgment.

In this case, the *Thinking* variant does not fail because it misses the basic visual structure or the relevant physical relation. Instead, it becomes anchored on an overly rigid conceptual rule and evaluates the option through that internal schema rather than the condition stated in the question itself. This qualitatively matches the pattern in our reversal inspection for *Natural Sciences*, where errors often stem from premature commitment to an incorrect problem frame rather than from missing visual evidence.

Figure 15 shows a representative *Engineering* reversal in which both variants initially misread the same FT diagram, but differ in whether they recover from that early structural error. The underlined spans mark the decision points that anchor each model's subsequent reasoning. For *Qwen3-VL-32B-IT*, the underlined text shows an initial misinterpretation followed by an explicit self-correction once the model recognizes that the top gate is an AND gate, allowing it to recover the correct answer. By contrast, *Qwen3-VL-32B-Thinking* remains committed to the mistaken assumption that the top event is connected to an OR gate. That early structural error then propagates through the entire derivation, leading the model to produce a fully consistent but fundamentally incorrect recovery analysis. This example matches the broader pattern we observed in *Engineering*: failures often arise from incorrect diagram-level structure interpretation, after which the model elaborates a coherent solution under the wrong logical frame.

H.3 Additional Qualitative Examples for Disciplinary Bottlenecks

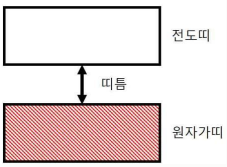
For disciplinary bottlenecks, we use *Qwen3-VL-235B-A22B-IT* as a consistent reference point for representative qualitative examples. In additional inspected cases, we also examined corresponding outputs from *Qwen3-VL-235B-A22B-Thinking*, and observed qualitatively similar failure patterns. The examples in this appendix illustrate recurring errors in exact convention-to-label mapping in *Arts & Design* and few-shot symbolic induction or terminology grounding in *General*.

Example 1: Expert category mismatch in *Arts & Design*. Figure 16 shows a representative *Arts & Design* failure in which the model captures the coarse spatial organization of an architectural plan but fails to assign the correct standardized category. Rather than identifying the plan type required by the question, the model overinterprets the stacked horizontal layout as evidence for an orthogonal or cross-shaped structure, and then reinforces this mistaken frame with a plausible but incorrect villa as-

[Question - English Translation]
The following figure shows the energy band structure of a certain solid. Select all correct statements about energy band structures and semiconductors from the options below.
ㄱ. In an insulator, the valence band is completely filled with electrons, and the band gap is larger than that of a semiconductor.
ㄴ. Even if a voltage greater than the band gap is applied, no current flows in an insulator.
ㄷ. Even if a band gap exists, the material is metallic if the Fermi level lies inside the conduction band.
ㄹ. When an electron transitions from the conduction band to the valence band, the frequency of the emitted light increases as the band gap becomes larger.

[Gold Answer] ㄱ,ㄷ,ㄹ

[Image]



[Qwen3-VL-32B-IT Response - English Translation]
... (truncated)

ㄴ. Even if a voltage greater than the band gap is applied, no current flows in an insulator.
❌ Incorrect.
This statement is wrong. Even if the applied voltage is greater than the band gap, current generally does not flow in an insulator, but if a very large voltage (e.g., the breakdown voltage) is applied, dielectric breakdown can occur and current may flow. In other words, the absolute statement that "no current flows even if a larger voltage is applied" is incorrect. Therefore, this statement is inaccurate.
ㄷ. Even if a band gap exists, the material is metallic if the Fermi level lies inside the conduction band.
✅ Correct.
A metal is characterized by a partially filled conduction band or by overlap between the valence and conduction bands, so that there is no band gap or only a very small one. However, if the Fermi level lies inside the conduction band, electrons can move easily, so the material has very high conductivity and behaves like a metal. This distinguishes it from semiconductors and insulators. Therefore, this statement is correct.
... (truncated)

Thus, as the band gap becomes larger, the emitted energy increases, and the frequency also increases.
Therefore, the correct statements are: ㄱ, ㄷ, ㄹ
Final answer: $\\boxed{ㄱ, ㄷ, ㄹ}$

[Qwen3-VL-32B-Thinking Response - English Translation]
Okay, let's tackle this problem. So, the question is about the energy band structure of a solid, specifically a semiconductor. The image shows two bands: the conduction band (전도띠) on top and the valence band (원자가띠) at the bottom, with a gap labeled ""띠틈"" between them.
... (truncated)

Option ㄱ: "절연체는 원자가띠에 전자가 완전히 채워져있고 띠틈이 반도체 보다 크다." (Insulators have the valence band completely filled with electrons, and the band gap is larger than in semiconductors.)
This seems right. Insulators have a large band gap (usually more than 3 eV), while semiconductors have a smaller one (like 1-2 eV). The valence band is full in both insulators and semiconductors at absolute zero, but the gap size is the key. So ㄱ is correct.
... (truncated)

Option ㄷ: "띠틈이 있더라도 페르미 준위가 전도띠 내부에 위치하면 금속이다." (Even if there is a band gap, if the Fermi level is inside the conduction band, it's a metal.)

Wait, metals don't have a band gap. In metals, the conduction band and valence band overlap, so there's no gap. If there's a band gap (as in semiconductors or insulators), then the Fermi level can't be inside the conduction band unless the gap is zero. So if there is a band gap, the Fermi level can't be inside the conduction band. So this statement is incorrect. Because if there's a band gap, then the conduction band is empty (at 0K), so the Fermi level would be in the gap. So ㄷ is wrong.
... (truncated)

ㄷ: If there's a band gap, and the Fermi level is in the conduction band, then it's a metal. But metals don't have a band gap. So if there's a band gap, it's not a metal. So ㄷ is wrong.
... (truncated)

So the correct answers are $\\boxed{ㄱ, ㄹ}$

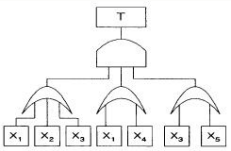
Figure 14: **Rigid conceptual framing in a *Natural Sciences* reversal.** *Qwen3-VL-32B-IT* correctly applies the relevant conductivity criterion, whereas *Qwen3-VL-32B-Thinking* overcommits to an overly rigid band-gap-based schema and rejects the crucial statement about the Fermi level in the conduction band.

[Question - English Translation]
In the system analyzed by the fault tree (FT) shown in the figure, all components corresponding to the basic events are currently failed.
If the repair team restores Component 1 through Component 5 in order, at the completion of which component repair does the system become operational?

(1) Component 2 (2) Component 3 (3) Component 4 (4) Component 5

[Gold Answer] Option 2

[Image]



[Qwen3-VL-32B-IT Response - English Translation]
The given FT has the following structure:
- The top event T (system failure) occurs through an AND gate over the three lower events.
- Each lower event is composed of an OR gate.

So, T occurs only when all three OR-gate outputs occur.
... (truncated)
The system becomes operational only if T does not occur.
... (truncated)
At first, the model says:
If even one of the three OR-gate outputs becomes normal, T does not occur and the system becomes normal.
Therefore, when Component 1 is repaired, the system becomes operational.
Then it revises its interpretation:
Ah! Because the top gate is an AND gate, T does not occur only when all three OR-gate outputs are normal.

T=OR1∧OR2∧OR3

So, for the system to be operational, OR1, OR2, and OR3 must all be normal.

It then tracks the repair sequence:
- After Component 1: OR1 = normal, OR2 = normal, OR3 = failed
  → T=normal∧normal∧failed=failed
- After Component 2: OR1 = normal, OR2 = normal, OR3 = failed
  → failed
- After Component 3: OR1 = normal, OR2 = normal, OR3 = normal
  → system operational

Final answer: $\\boxed{(2) Component 3}$

[Qwen3-VL-32B-Thinking Response - English Translation]
... (truncated)

The top event T is connected to an OR gate, and each of the three sub-gates is also an OR gate.
So, the system fails if any of the three sub-gates is true.
Therefore, the system is operational only when all three sub-gates are false.
... (truncated)

The conditions are:
1. First sub-gate (X1,X2,X3)(X1, X2, X3)(X1,X2,X3): false only if X1 = 0, X2 = 0, X3 = 0
2. Second sub-gate (X1,X4)(X1, X4)(X1,X4): false only if X1 = 0, X4 = 0
3. Third sub-gate (X3,X5)(X3, X5)(X3,X5): false only if X3 = 0, X5 = 0
   ... (truncated)

So, the system is operational only when all five components are restored.
It then tracks the repair sequence:
- After repairing X3: First sub-gate = 0, second = 1, third = 1
  → T = 1, system failed
- After repairing X4: First = 0, second = 0, third = 1
  → T = 1, system failed
- After repairing X5: All sub-gates = 0
  → T = 0, system operational
  ... (truncated)

The top event T is connected to an OR gate. So T occurs if any of the three sub-gates occur. For T not to occur, all three sub-gates must not occur.
... (truncated)

Final answer: $\\boxed{(4) Component 5}$

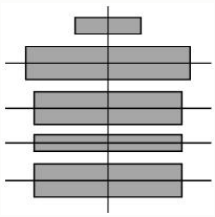
Figure 15: **Structural misinterpretation in an *Engineering* reversal (Fault Tree analysis).** *Qwen3-VL-32B-IT* partially corrects an early gate-level misinterpretation, whereas *Qwen3-VL-32B-Thinking* persists with an incorrect top-level gate reading and derives the wrong recovery point.

[Question - English Translation]
The following figure represents one of the three main plan types of the Italian terracedominant architectural style.
What is the name of this type, and which villa corresponds to it?

(1) Linear type — Villa Lante (2) Parallel type — Villa d'Este
(3) Orthogonal type — Medici Villa (4) Cross type — Isola Bella

[Gold Answer] Option **2**

[Image]



[Error Analysis]
The model does not fail to read the gross visual structure. Instead, it overinterprets the stacked horizontal layout as a cross/orthogonal plan and then reinforces this mistaken category with memorized architectural associations (e.g., linking the figure to Villa Medici). This example illustrates a failure of precise expert category mapping rather than low-level visual perception.

[Qwen3-VL-235B-A22B-Thinking Response - English Translation]
Okay, let's see. The question is about Italian villa architectural styles, specifically the Terracedominant style. It's asking for the name of the plan type shown in the image and the corresponding villa name.
..(truncated)

The image shows a central vertical axis with horizontal elements. If it's a cross shape, then it's Cross type, but the options have orthogonal as option 3. Wait, maybe the plan here is orthogonal, meaning the wings are at right angles to the main axis.

Wait, the Terracedominant style's three types: serial (rooms in a line), parallel (two parallel wings), and cross (wings at right angles). But I'm a bit fuzzy.
Wait, the Villa Medici in Fiesole is considered to have an orthogonal plan. Let me check my memory. The Villa Medici in Fiesole has a central block with wings extending to the sides, forming a sort of cross. Wait, no, maybe it's more of a U-shape or something else.
..(truncated)

Option 3 is orthogonal and Medici Villa. If the image shows a plan where the main axis is vertical and there are horizontal wings (so a cross shape), then it's orthogonal. The Medici Villa (Villa Medici) is often cited as an example of the orthogonal plan.

Upon checking, the three main types of Italian Renaissance villa plans are:

1. Serial (or linear) - rooms arranged in a line.
2. Parallel - two parallel wings.
3. Orthogonal (or cross) - with wings at right angles.

The Villa Medici in Fiesole is considered an example of the orthogonal plan. So if the image shows a cross shape (vertical and horizontal axes), then it's 직교형 and the villa is 메디치장. So option 3.
</think>

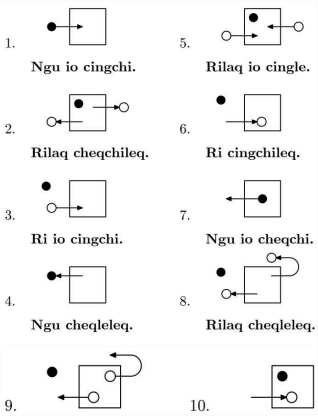
correct option: $\\boxed{(3) 직교형－메디치장}$

Figure 16: **Exact architectural category misclassification in *Arts & Design*.** The model recognizes the overall structure of the plan, but fails to map it to the correct standardized architectural category. Instead, it overcommits to an orthogonal-plan interpretation and supports it with a plausible but incorrect villa association.

[Question - English Translation]
The following are expressions in Shanghainese describing a series of situations.
Describe the following two situations shown in the pictures in Shanghainese (number 9 and 10).

[Gold Answer]
9. Rilaq io cheqle. 10. Ri cingleleq.

[Image]



[Error Analysis]
The model captures the coarse visual semantics of the scene, but fails to infer the exact symbolic rule needed to compose the correct Shanghainese sentence. Its prediction appears to be driven by nearest-example analogy rather than systematic rule induction, leading to a plausible but incorrect output.

[Qwen3-VL-235B-A22B-IT Response - English Translation]
First, let us analyze the patterns in the existing sentences and pictures.
Each sentence consists of two elements:
- subject (person/object): Ngu, Ri, Rilaq
- action/state: io, cheqchileq, cingchi, cheqleleq, etc.

And in the pictures:
- there are black dots (●) and white dots (○),
- arrows indicate direction or movement,
- and the square appears to represent a space or boundary.

Let us examine each sentence:
1. Ngu io cingchi. → The black dot enters the square (→)
   .. (truncated)
7. Ngu io cheqchi. → The black dot enters the square (→), while the white dot is outside
8. Rilaq cheqleleq. → The white dot exits the square (←), while the black dot is inside, and the white dot has a loop (circular) arrow

Now let us analyze Pictures 9 and 10:
Picture 9:
- The black dot (●) is outside the square on the left
- The white dot (○) is inside the square and has a loop arrow
- The white dot is circling inside the square
- The black dot has an arrow pointing left outside the square → that is, it appears to be moving away or going out

→ This is similar to No. 8, but in No. 8 the white dot goes out of the square, whereas in No. 9 the white dot is circling inside it.
   .. (truncated)
→ No. 9 is the white dot circling inside and then going out → since this is similar to No. 8, "Rilaq cheqleleq" is possible.
But in No. 8, the white dot loops inside the square and then exits left, and No. 9 shows the same action.
Difference: in No. 9, the black dot also has an arrow pointing left. → This seems to be additional information, and the sentence appears to be centered on the white dot.
→ Therefore, No. 9 is also presumed to be Rilaq cheqleleq.
Picture 10:
- The black dot (●) is inside the square
- The white dot (○) is inside the square and has an arrow coming in from right to left → that is, the white dot is entering the square
- The black dot is not moving

→ This is similar to No. 6: Ri cingchileq. → The white dot is inside, and the black dot is outside (in No. 6, the white dot has an arrow showing that it is entering)
   .. (truncated)
→ No. 3: Ri io cingchi. → the white dot enters, while the black dot is outside
In No. 10: the white dot enters, while the black dot is inside → this is different, but the sentence seems to focus on the white dot's action
→ Therefore, it is presumed to be Ri io cingchi.

Final conclusion: $\\boxed{9. Rilaq cheqleleq, 10. Ri io cingchi.}$

Figure 17: **Few-shot symbolic induction failure in *General*.** A Shanghainese-language item requiring the model to infer a latent mapping from a small set of diagram–expression pairs and apply it to new cases. Although *Qwen3-VL-235B-A22B-IT* produces a detailed step-by-step analysis, it fails to recover the full correspondence system and instead relies on partial surface analogies, leading to a plausible but incorrect answer.

[Question - English Translation]
Which of the following applies the following Romanization rule?

(1) 독도: Dok-do (2) 반구대: Ban-gudae
(3) 독립문: Dok-rip-mun (4) 인왕리: Inwang-ri

[Gold Answer] Option 3

[Image]

발음상 혼동의 우려가 있을 때에는 음절 사이에 붙임표(-)를 쓸 수 있다.

[Error Analysis]
The model demonstrates partial knowledge of the Romanization rule, but misapplies the official standard by selecting a familiar, plausible example rather than the option that most precisely satisfies the rule. This error reflects over-reliance on intuitive ambiguity resolution instead of exact rule-based discrimination.

[Qwen3-VL-235B-A22B-IT Response - English Translation]
The given rule is: "A hyphen (-) may be used between syllables when there is a possibility of confusion in pronunciation."
This means that when writing Korean in the Roman alphabet, a hyphen may be used if the boundary between syllables is unclear or if pronunciation could be ambiguous.
Looking at each option:
(1) 독도: Dok-do
→ 독도 consists of two syllables, and when dok and do are written together, there may be a risk of confusion in pronunciation. In practice, Dokdo is often written as Dokdo, but writing it as Dok-do is also permitted to prevent ambiguity. This conforms to the rule.
(2) 반구대: Ban-gudae
→ 반구대 has three syllables: ban, gu, and dae. However, gudae combines gu and dae, and it is hard to say that this creates a real risk of confusion between syllables. Rather, Ban-gu-dae would be a more accurate syllabic separation. Therefore, this spelling does not fully conform to the rule.

.. (truncated)
However, the question asks, **"Which of the following applies the rule?"** so we need to choose the option that most clearly reflects the intent of the rule, namely, the prevention of ambiguity.
Here, (1) Dokdo: Dok-do is the most representative example. In practice, Dokdo is also written as Dokdo, but Dok-do is officially allowed, and this reflects the purpose of preventing confusion.

Therefore, the most appropriate answer is: $\\boxed{(1) 독도: Dok-do}$

Figure 18: **Exact standards and rule-criterion misapplication in *General*.** An item testing the official Romanization rule for hyphen use. The model gives a broadly reasonable explanation of the rule, but selects the wrong option because it applies an approximate plausibility-based criterion rather than the exact condition required by the formal standard.

sociation. This is not a low-level perception failure: the model recognizes salient geometric structure, but fails at precise convention-to-label mapping among closely related expert categories. The case therefore illustrates a recurring bottleneck in *Arts & Design*, where errors arise not from missing the visual content altogether, but from overconfident misclassification of specialized visual conventions into the wrong technical label.

Example 2: Few-shot symbolic induction failure in *General*. As shown in Figure 17, this item requires the model to infer a latent mapping between diagram configurations and Shanghainese expressions from a small set of paired examples, and then apply that rule to unseen cases. The model produces a long, locally plausible analysis, but fails to recover the full underlying correspondence system. Instead, it partially matches surface patterns and then drifts into self-invented regularities, yielding answers that are structurally plausible but incorrect. This example illustrates a recurring *General* bottleneck in KMMMU: some items require few-shot rule induction from sparse symbolic evidence, not just fluent explanation or broad world knowledge.

Example 3: Exact standards and rule-criterion misapplication in *General*. Figure 18 shows a case that depends on precise application of an official Romanization rule, rather than broad linguistic plausibility alone. The model produces a reasonable explanation and discusses related principles, but selects the wrong option because it applies an approximate criterion instead of the exact standard required by the question. This pattern appears in multiple *General* items that test official terminology, orthographic conventions, or certification-style definitions: the model often gives a broadly sensible account, but misses the precise condition that determines correctness.

I Ablation Study

I.1 Evaluation of Image Dependency

Model	Original	Text-only
Gemini-3-Flash	$45.15_{0.98}$	$20.43_{0.48}$
GPT-5-Mini	$21.32_{0.33}$	$9.75_{0.85}$

Table 13: **Ablation results for image dependency.** We compare the average accuracy and standard deviation (Acc_{std}) of models on the original multimodal dataset versus the text-only baseline.

To verify that KMMMU functions as a genuinely multimodal benchmark, we measure how strongly performance depends on access to visual information. We conduct a text-only ablation in which models receive the textual question and answer options, but the associated image is removed.

Table 13 shows substantial performance drops for both GEMINI-3-FLASH and GPT-5-MINI under this setting. For GEMINI-3-FLASH, accuracy declines from 45.15% to 20.43%, a drop of 24.72 percentage points. For GPT-5-MINI, accuracy declines from 21.32% to 9.75%, a drop of 11.57 percentage points. These results indicate that many KMMMU questions cannot be solved reliably from text alone.

We also manually inspected the 60 cases answered correctly by GPT-5-MINI in the text-only setting. A non-trivial subset of visually accompanied items remains solvable without direct image access, but these cases often do not reflect genuine visual understanding. Instead, they typically fall into several recurring patterns: (i) the textual prompt already specifies most of the decisive constraints, making the image largely auxiliary; (ii) the answer can be inferred from strong domain priors or option elimination rather than visual grounding; (iii) the option structure, numerical form, or canonical diagram schema enables answer reconstruction without actual image reading; and (iv) in some quantitative science items, the core reasoning is already determined by symbolic conditions in the text, with the image serving mainly as contextual support.

Overall, these findings support the multimodal validity of KMMMU while also clarifying that visual accompaniment and strict image-essentiality are not identical. Although removing images causes large performance drops, some items remain text-solvable because they contain sufficient textual, structural, or prior-driven cues to permit correct answering without direct image use.

I.2 Data Contamination Analysis

To probe possible memorization, we run a prefix-completion test in which models receive the first 35% of a question together with its image and are asked to generate the remaining continuation. We restrict this analysis to questions longer than 150 tokens, since shorter exam-style items often begin with generic instructions that provide too little question-specific content for a meaningful reconstruction test.

Source	Model	Exactness$_{NH}$	Exactness$_{H}$	Refusal$_{NH}$	Refusal$_{H}$	Hallucination$_{NH}$	Hallucination$_{H}$
PSAT	Gemini-3-Flash	14.20	13.14	17.53	24.70	76.10	68.13
	Gemini-3-Pro	0.34	0.46	95.60	94.78	3.20	3.21
	GPT-5-Mini	0.00	0.00	99.60	100.00	0.00	0.00
NCS	Gemini-3-Flash	10.05	6.54	29.08	50.20	50.20	35.57
	Gemini-3-Pro	0.18	0.04	99.21	100.00	0.40	0.00
	GPT-5-Mini	0.65	0.45	95.65	100.00	1.19	1.19
NTQ	Gemini-3-Flash	15.70	18.58	19.66	15.61	61.54	76.84
	Gemini-3-Pro	0.33	2.03	93.86	79.15	4.44	16.92
	GPT-5-Mini	0.45	0.25	97.96	99.83	0.51	0.17
Olympiads	Gemini-3-Flash	10.48	10.91	23.12	37.56	55.58	49.48
	Gemini-3-Pro	0.13	0.38	97.92	97.41	1.56	2.33
	GPT-5-Mini	0.52	0.18	94.56	98.96	2.85	1.04

Table 14: **Prefix-completion analysis for potential data contamination.** Models are given the first 35% of each question together with the associated image, and asked to generate the remaining continuation. *Exactness* denotes a judge-assigned 0–100 faithfulness rating with respect to the reference continuation. *Refusal* and *Hallucination* denote judge-labeled failure modes, reported as percentages, under the No-hint (NH) and Hinted (H) settings.

We evaluate three frontier models under two settings. In the first, no additional metadata is provided. In the second, we provide the exam name and year as a potential memorization trigger.

We evaluate generated continuations using GEMINI-3-FLASH as a judge. For each continuation, the judge assigns (i) an exactness rating from 0 to 100 based on overlap with the reference continuation and preservation of key details, and (ii) categorical labels indicating refusal or hallucination. Thus, exactness reflects judge-rated reconstruction fidelity, whereas refusal and hallucination characterize distinct failure behaviors rather than the same measurement scale.

As shown in Table 14, GPT-5-MINI and GEMINI-3-PRO exhibit very high refusal rates across all four sources, typically accompanied by near-zero exactness ratings. This suggests that these models usually do not attempt faithful continuation under the prefix-completion setup.

GEMINI-3-FLASH attempts continuation more often, yielding lower refusal rates and somewhat higher exactness ratings than the other two models. However, its exactness remains low overall, and the hinted setting does not produce a consistent increase across sources. Moreover, hallucination rates remain high, indicating that many attempted continuations are low-fidelity generations rather than faithful reconstructions.

Taken together, these results do not provide strong evidence that benchmark performance is driven by simple memorization of question continuations. If contamination were a major driver under this setup, we would expect more consistently faithful reconstruction and clearer improvement when exam metadata is provided as a hint. Instead, the dominant pattern is either refusal or low-fidelity continuation.