KMMMU: A Korean Massive Multi-discipline Multimodal Understanding Benchmark

Anonymous ACL submission

Abstract

The evaluation of expert-level multimodal capability remains constrained by a reliance on English-centric corpora and translated approximations that fail to capture the nuances of localized professional reasoning. To address this limitation, we introduce **KMMMU** which is a native Korean benchmark constructed exclusively from high-stakes professional and academic examinations. The dataset comprises 3,466 verified image-text pairs spanning diverse disciplines and task types. These questions necessitate the interpretation of information-dense visual inputs including technical diagrams and administrative tables where visual evidence is indispensable for the solution. While scaling and inference-time compute improve logical reasoning, our results suggest that they do not fully mitigate limitations in structural perception or cultural grounding. The observed disparity between processing text-rich documents and abstract diagrams points to ongoing challenges in structural visual reasoning. We hope that KMMMU serves as a useful diagnostic resource to address these fine-grained visual and institutional blind spots in future multimodal systems.

1 Introduction

Multimodal Large Language Models have demonstrated remarkable progress in visual recognition and document understanding (Alayrac et al., 2022; Kim et al., 2022; OpenAI, 2023; Team and Google, 2023; Li et al., 2023a; Liu et al., 2023; Bai et al., 2023; OpenAI, 2025; Bai et al., 2025; Google Cloud, 2025). However, current benchmarking methodologies lag behind the complex environments where these systems are increasingly deployed (Sun et al., 2024; Fu et al., 2024; Guan et al., 2024) . Most existing evaluations rely on English-centric datasets or translated variants of general knowledge tasks (Li et al., 2023b; Yue et al., 2024). This approach understates two critical challenges in real-world deployment which are expert-level reasoning under high-stakes assessment formats and deep localization that requires more than simple linguistic fluency (Liu et al., 2024; Li et al., 2023b; Lu et al., 2023; Huang et al., 2023; Son et al., 2025a). Consequently, high scores on current leaderboards often fail to reflect a model's true proficiency in specific cultural and professional contexts (Zhou et al., 2023; Gudibande et al., 2023).

To contribute to addressing this gap, we introduce KMMMU, a native Korean benchmark designed to evaluate expert-level multimodal understanding. Unlike datasets derived from translation, KMMMU is constructed from sources including the Public Service Aptitude Test, National Technical Qualifications, academic Olympiads, and National Competency Standards in South Korea. These sources present problems that necessitate reasoning over information-dense visuals such as technical diagrams, administrative tables, and legal schematics. The dataset comprises 3,466 quality-assured questions curated through a hybrid pipeline of automated digitization and multi-stage human verification.

One distinguishing aspect of KMMMU is its emphasis on institutional grounding. We include a Korean-specific subset that targets domestic legal frameworks, administrative procedures, and cultural conventions, which helps separate general reasoning from localization demands. We also define a hard subset from consensus failures of strong baselines to focus analysis on challenging instances.

Evaluations of proprietary and open-source models suggest that scaling and inference-time reasoning improve some logical tasks but do not fully resolve perceptual challenges. Even advanced models struggle with structural visual parsing in diagrams and with grounded knowledge needed for localized expert problems. We hope KMMMU will serve as a diagnostic benchmark for culturally grounded and professionally relevant multimodal systems.

2 Related Works

Prior multimodal evaluation primarily addressed general visual question answering (Goyal et al., 2017; Hudson and Manning, 2019) or specific skills such as chart and document understanding (Masry et al., 2022; Methani et al., 2020; Mathew et al., 2021; Singh et al., 2019). Recent initiatives have introduced expert-oriented benchmarks to approximate professional problem solving (Lu et al., 2023, 2022; Zhang et al., 2024). We position KMMMU within this landscape to address the specific requirements of Korean expert settings (Table 1).

Global Expert Benchmarks MMMU (Yue et al., 2024) and its extensions like CMMMU (Zhang et al., 2024) and JMMMU (Onohara et al., 2025) evaluate broad reasoning across academic disciplines. While these benchmarks establish a rigorous standard for expert evaluation, they generally rely on globally oriented content and do not assess cultural or institutional grounding. KMMMU diverges from this approach by utilizing native sources from the Public Service Aptitude Test and national technical qualifications. This design ensures the evaluation captures the distinct legal and administrative conventions of South Korea rather than relying on translation.

Korean-Centric Benchmarks Existing Korean resources often dissociate professional expertise from multimodal capabilities. Benchmarks such as KOFFVQA (Kim and Jung, 2025) focus on general scene understanding but lack the complexity required to test advanced reasoning. Conversely, KMMLU (Son et al., 2025a) and KMMLU-Pro (Hong et al., 2025) provides a robust assessment of expert knowledge but remains restricted to a text-only format. Although the recent KO-VLM suite (Marker-Inc-Korea, 2024) improves realism through industrial document parsing, it prioritizes information extraction over complex logic. KMMMU bridges this gap by presenting high-level problems that demand multi-step quantitative reasoning on original visual formats. It integrates an explicit Korean-specific flag to quantify localization and serves as a complementary expert-level benchmark in the field.

3 The KMMMU Benchmark

3.1 Data Collection and Annotation

KMMMU is constructed from high stakes national examinations and academic competitions in South Korea. We select these sources to organize problems that require advanced reasoning and domain knowledge rather than surface level pattern matching. The benchmark draws from four primary sources including PSAT, national technical qualifications, Olympiads, and NCS. Source specific collection scope and examples are provided in Appendix A.

Annotation and quality control. We combine automated extraction with manual verification. Technical qualification exam data are collected through web crawling, while other sources are digitized using the MinerU-2.5 OCR system (Niu et al., 2025). Five Korean annotators review each instance to verify the question text, associated images, and explanatory materials. Details on the verification interface and the filtering protocol are provided in Appendix B. The resulting dataset contains 3,466 curated questions.

3.2 Taxonomy and Dataset Composition

KMMMU is designed to evaluate expert-level multimodal understanding across diverse domains and problem formats. To support analysis beyond overall accuracy, we annotate each instance along four axes: subject, task, visual modality, and question format, together with a Korean-specific flag.

Labels and quality control. We assign each instance to one of 45 fine-grained subjects aggregated into 9 macro disciplines, and label the primary competency required (task), the dominant visual form, and the answer format. We mark items that require institutional or culturally grounded knowledge specific to South Korea as Korean-specific. We obtain initial subject and task labels using GEMINI-2.5-FLASH-LITE, and visual modality labels and the Korean-specific flag using GEMINI-2.5-FLASH, followed by human consolidation. We validate label quality via a manual audit of roughly 300 randomly sampled instances and manually verify all positive cases for the Korean-specific flag. Additional implementation details are provided in Appendix B.2.

Distributions. Figure 1 shows broad coverage across professional disciplines. KMMMU is dominated by structure-heavy inputs such as diagrams, document-style visuals, and tables (Table 2), and many instances contain in-image text that requires combining text extraction with multimodal reasoning. KMMMU also includes 299 Korean-

Benchmark	Year	Language	Subjects	Expert	Multimodal	Native KO	KO-Spec
Global Expert Multimodal Benchmarks							
MMMU (Yue et al., 2024)	2023	EN	30	✓	✓	✗	✗
CMMMU (Zhang et al., 2024)	2024	ZH	39	✓	✓	✗	✗
JMMMU (Onohara et al., 2025)	2025	JA	30	✓	✓	✗	✗
Korean-Centric Benchmarks							
KOFFVQA (Kim and Jung, 2025)	2023	KO	General	✗	✓	✓	Low
KMMLU (Son et al., 2025a)	2024	KO	45	✓	✗	✓	High
KMMMU (Ours)	**2025**	**KO**	**45**	✓	✓	✓	**High**

Table 1: **Comparison of KMMMU with the MMMU series and existing Korean-centric benchmarks.** *Expert* denotes college-level or professional expertise requirements. *Native KO* indicates construction from original Korean sources rather than translation. *KO-Spec* refers to the inclusion of South Korean-specific legal, administrative, or cultural context.

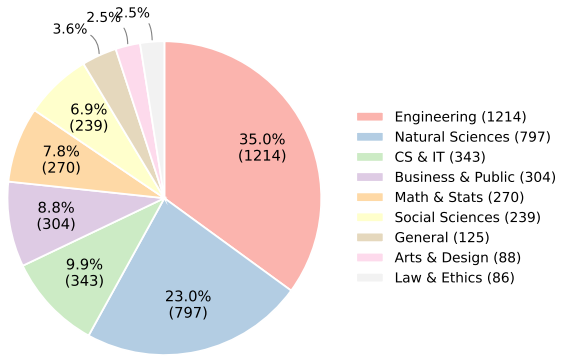


Figure 1: **Macro subject type distribution.**

Visual Modality	Count	%
Diagram	1,751	50.52
Text, Code & Documents	869	25.07
Tables	549	15.84
Charts & Plots	151	4.36
Mathematics	88	2.54
Maps	23	0.66
Symbols & Art	22	0.64
Photographs	13	0.38

Table 2: **Visual modality categories.**

specific items that target localization and institutional grounding, concentrated in domains such as law and public administration. Detailed task and question-format distributions are provided in Appendix D.

4 Experimental setup

4.1 Evaluated Models

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

Non-reasoning models. This group consists of instruction-tuned multimodal models that generate answers in a single pass without an explicit reasoning mode. Our evaluation covers open-source multilingual models including GEMMA-3 (Team et al., 2025) in the 4B, 12B, and 27B variants, QWEN3-VL (Bai et al., 2025) in the 2B, 4B, 8B, 32B, 30B-A3B, and 235B-A22B variants, as well as the LLAMA-4 (Meta, 2025b,a) family with the Scout and Maverick variants. We also include Korean-focused multimodal models, VARCO-VISION-2.0 (Cha et al., 2025) with 1.7B and 14B parameters and HYPERCLOVAX-SEED-VISION-3B (NAVER HyperCLOVAX, 2025).

Reasoning models. We additionally evaluate models that are designed to perform explicit multi-step reasoning prior to producing a final answer. This group consists of QWEN3-VL-THINKING (Bai et al., 2025) variants with 30B-A3B, 32B, and 235B-A22B parameters, as well as the Korean reasoning model KO-REASON-G3-12B (Son et al., 2025b). These models are configured to internally allocate additional computation for structured reasoning, which is relevant for problems requiring multi-step visual and domain-specific analysis.

Frontier models. We report results on this subset separately, including closed-source frontier models like GPT-5, CLAUDE-OPUS-4.5, GEMINI-3-PRO, etc (OpenAI, 2025; Anthropic, 2025; Google Cloud, 2025).

Hard subset. We define a hard subset to stress-test failure cases that persist under strong baselines.

Specifically, we select examples that are answered incorrectly by both GEMMA-3-27B and GEMINI-2.5-FLASH-LITE. We report results on this subset separately, including closed-source frontier models, to better expose differences that are muted on the full split.

4.2 Evaluation Protocols

For all models, All evaluations were conducted in a zero-shot setting. Response generation was also performed using either the officially recommended decoding parameters provided by the model developers or the default settings when no explicit recommendations were available. All models were evaluated using the same prompt template across all experiments, and no additional prompt-specific tuning or parameter optimization was applied.

The primary metric is mean accuracy. We performed 3 independent trials for each model, and report the mean accuracy and the standard deviation.

5 Results

5.1 Main Results

Table 3 reports accuracy by macro subject and overall average. Two patterns emerge. First, stronger models benefit from scale, but scale alone is not sufficient to close the gap on KMMMU. Second, enabling reasoning mode produces broad improvements that are most salient when questions require multi step computation or constraint satisfaction rather than direct recognition. This suggests that KMMMU contains a substantial fraction of items where thei limiting factor is not only perception but also the ability to execute and verify intermediate reasoning steps.

Open-source Korean models. Korean focused open source models remain behind the strongest multilingual open source baselines. This gap is consistent across subjects, which indicates that improved Korean coverage alone does not fully address the benchmark demands. Instead, competitive performance appears to require both robust visual grounding and broad expert level knowledge across domains.

Closed-source Frontier models. We additionally evaluate frontier models on the hard subset.

Results reported in Table 4 show clearer separation among frontier models, suggesting that the subset concentrates cases where shallow visual parsing is insufficient and correctness depends on more reliable grounding and reasoning.

Cross-subject variability. Macro subject breakdown reveals that some domains are more stable across model choices, while others exhibit larger dispersion. This dispersion is consistent with differences in localization and institutional grounding requirements, which are more prominent in BUSINESS & PUBLIC and LAW & ETHICS. In contrast, subjects that are supported by structured visual evidence tend to be less sensitive to the model choice. We further quantify these patterns through macro task analysis in Appendix E.

5.2 Performance on Korean-Specific Content

To test whether Korean-specific knowledge introduces an additional bottleneck, we stratify items by `is_korean_specific` and compare accuracy on `True` versus `False` items. Appendix Table 18 reports the counts and ratios of Korean-specific items in the Full and Hard splits.

Localization as a remaining barrier. Across stronger models, accuracy drops on the Korean-specific subset even when performance on non-Korean-specific items is high. This suggests that these items are not simply harder samples from the same distribution, but require additional localization, including institutional conventions, administrative context, and culturally grounded constraints.

Interpreting small or positive gaps. Some weaker models exhibit near-zero or even positive differences between `True` and `False`. This does not necessarily indicate better localization. When overall accuracy is low, both subsets can approach a similar error rate, which compresses the gap and can occasionally reverse its sign. For this reason, the stratified gap is most informative for models that perform reliably on the non-Korean-specific subset. We provide per-model stratified results in Appendix Table 21.

5.3 Is LLM-as-a-Judge a Consistent Evaluator?

Because KMMMU includes both multiple-choice and free-form questions, we use LLM-as-a-judge for scalable evaluation. We run a controlled study to assess prompt sensitivity under a fixed set of model outputs.

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	$25.00_{4.10}$	$20.18_{1.00}$	$11.76_{1.35}$	$12.14_{0.58}$	$10.40_{1.60}$	$18.99_{4.70}$	$2.84_{0.21}$	$7.49_{1.57}$	$11.58_{0.64}$	$11.41_{0.50}$
Gemma-3-12B-IT	$30.30_{2.62}$	$25.99_{2.48}$	$19.14_{1.98}$	$15.60_{1.08}$	$9.33_{1.22}$	$31.01_{1.78}$	$10.00_{1.70}$	$14.09_{1.50}$	$16.60_{2.56}$	$16.68_{0.42}$
Gemma-3-27B-IT	$22.73_{4.10}$	$24.89_{2.56}$	$20.31_{3.22}$	$17.93_{1.84}$	$14.13_{4.41}$	$21.71_{8.25}$	$15.19_{2.06}$	$17.40_{1.19}$	$19.53_{3.25}$	$18.63_{0.36}$
Qwen3-VL-2B-IT	$24.62_{3.99}$	$27.19_{1.98}$	$20.41_{0.87}$	$13.95_{0.48}$	$9.07_{0.46}$	$24.03_{0.67}$	$11.48_{3.03}$	$11.59_{0.38}$	$17.29_{1.35}$	$15.59_{0.23}$
Qwen3-VL-4B-IT	$23.11_{1.74}$	$37.06_{1.33}$	$25.95_{2.49}$	$17.11_{0.95}$	$13.60_{1.39}$	$27.52_{2.93}$	$15.31_{1.40}$	$18.95_{0.78}$	$23.71_{2.72}$	$20.75_{0.32}$
Qwen3-VL-8B-IT	$24.62_{1.74}$	$40.24_{1.81}$	$28.57_{1.82}$	$21.33_{1.65}$	$17.60_{0.00}$	$31.01_{3.74}$	$24.94_{2.52}$	$23.96_{1.07}$	$30.54_{0.72}$	$25.42_{0.53}$
Qwen3-VL-32B-IT	$27.27_{3.01}$	$52.63_{2.16}$	$36.73_{1.05}$	$28.91_{0.94}$	$24.27_{1.85}$	$37.21_{6.98}$	$48.52_{4.55}$	$40.07_{0.97}$	$46.72_{2.38}$	$37.08_{0.81}$
Qwen3-VL-30B-A3B-IT	$25.76_{8.53}$	$50.00_{2.30}$	$33.33_{3.03}$	$26.44_{1.57}$	$20.80_{2.12}$	$31.78_{4.08}$	$41.85_{2.25}$	$38.31_{2.13}$	$37.10_{1.05}$	$33.77_{0.44}$
Qwen3-VL-235B-A22B-IT	$23.48_{2.37}$	$57.79_{1.00}$	$36.73_{1.46}$	$31.58_{1.19}$	$28.27_{3.95}$	$35.66_{1.34}$	$55.06_{1.83}$	$44.12_{1.02}$	$49.37_{3.27}$	$40.10_{0.31}$
Llama-4-Scout-17B-16E-IT	$25.38_{3.28}$	$52.19_{0.68}$	$35.86_{1.54}$	$30.20_{1.26}$	$27.47_{3.23}$	$32.17_{0.67}$	$34.57_{1.07}$	$32.12_{0.50}$	$39.61_{3.36}$	$33.95_{0.54}$
Llama-4-Maverick-17B-128E-IT	$28.79_{2.37}$	$56.91_{2.05}$	$39.46_{1.50}$	$33.66_{1.67}$	$25.07_{3.23}$	$37.98_{2.93}$	$44.32_{2.04}$	$39.98_{1.01}$	$47.70_{2.74}$	$39.20_{0.38}$
Open-Source Korean Non-Reasoning Models										
HyperCLOVAX-SEED-Vision-3B	$21.59_{2.27}$	$20.83_{0.95}$	$18.46_{0.17}$	$11.15_{0.50}$	$12.00_{1.39}$	$21.71_{1.34}$	$10.25_{1.40}$	$9.37_{0.59}$	$16.32_{0.84}$	$13.16_{0.13}$
VARCO-VISION-2.0-1.7B	$18.18_{7.10}$	$12.72_{3.81}$	$15.06_{1.38}$	$12.52_{1.93}$	$6.67_{2.01}$	$11.63_{4.03}$	$5.93_{3.21}$	$7.90_{2.29}$	$11.16_{2.72}$	$11.03_{2.39}$
VARCO-VISION-2.0-14B	$24.24_{6.46}$	$28.73_{2.01}$	$26.63_{2.37}$	$24.44_{0.26}$	$13.87_{2.31}$	$27.52_{1.34}$	$15.19_{0.74}$	$18.90_{1.71}$	$25.80_{3.14}$	$22.82_{0.16}$
Reasoning Models										
Qwen3-VL-30B-A3B-Thinking	$34.09_{15.75}$	$64.25_{4.93}$	$49.76_{10.21}$	$49.81_{7.87}$	$50.40_{9.09}$	$41.86_{9.93}$	$76.05_{5.91}$	$58.43_{6.64}$	$67.78_{6.54}$	$55.76_{7.53}$
Qwen3-VL-32B-Thinking	$28.41_{3.41}$	$62.83_{2.37}$	$49.66_{0.84}$	$47.50_{0.81}$	$44.80_{3.49}$	$36.82_{3.55}$	$72.84_{1.83}$	$56.92_{0.51}$	$67.50_{1.89}$	$53.73_{0.49}$
Qwen3-VL-235B-A22B-Thinking	$23.86_{3.41}$	$64.04_{1.16}$	$45.87_{1.02}$	$45.77_{1.26}$	$49.60_{3.67}$	$36.05_{1.16}$	$75.43_{0.93}$	$58.89_{1.28}$	$65.27_{1.45}$	$53.39_{0.19}$
KO-REAson-G3-12B-1009	$30.68_{1.97}$	$53.40_{0.68}$	$47.42_{2.15}$	$55.82_{1.42}$	$53.07_{1.22}$	$32.56_{4.03}$	$76.67_{1.28}$	$57.05_{1.19}$	$63.46_{1.69}$	$55.90_{0.33}$

Table 3: **Accuracy (%) on the full set by macro subject with overall accuracy.** Overall accuracy is averaged across all subjects. Mean accuracy is reported in percentage, with standard deviation shown as a subscript.

Model	Arts & Design	Business & Public	CS & IT	Engineering	General	Law & Ethics	Math & Stats	Natural Sciences	Social Sciences	Overall Acc.
Gemini-3-Pro	$69.01_{0.29}$	$57.84_{3.40}$	$69.30_{0.76}$	$64.10_{1.27}$	$64.24_{1.32}$	$63.46_{1.92}$	$60.78_{6.79}$	$82.99_{0.59}$	$74.16_{0.99}$	$72.13_{1.64}$
Gemini-3-Flash	$60.27_{0.15}$	$52.94_{2.94}$	$61.40_{2.74}$	$54.21_{2.77}$	$55.93_{1.32}$	$47.44_{4.84}$	$54.90_{3.40}$	$73.81_{1.56}$	$65.17_{2.46}$	$66.67_{0.95}$
GPT-5	$51.50_{0.88}$	$28.43_{4.49}$	$51.75_{0.76}$	$39.56_{2.91}$	$43.51_{2.25}$	$45.51_{1.11}$	$37.25_{3.40}$	$76.87_{2.12}$	$59.55_{4.54}$	$61.75_{1.89}$
GPT-5-Mini	$42.80_{0.33}$	$17.65_{7.78}$	$53.95_{4.56}$	$31.14_{2.29}$	$31.65_{2.22}$	$32.05_{4.44}$	$29.41_{10.19}$	$70.75_{1.56}$	$52.81_{0.99}$	$49.73_{3.41}$
Claude-Opus-4.5	$42.58_{0.63}$	$34.31_{6.12}$	$56.14_{3.31}$	$35.16_{1.10}$	$32.87_{1.17}$	$38.46_{1.92}$	$35.29_{5.88}$	$64.97_{1.18}$	$45.82_{2.81}$	$53.55_{3.41}$
Claude-Sonnet-4.5	$34.00_{0.83}$	$21.57_{1.70}$	$55.70_{5.32}$	$24.54_{3.86}$	$25.02_{1.13}$	$22.44_{4.84}$	$41.18_{5.88}$	$56.46_{4.25}$	$36.20_{2.49}$	$42.62_{3.28}$
Grok-4	$42.42_{1.62}$	$29.41_{5.09}$	$42.11_{2.63}$	$36.63_{2.29}$	$37.63_{3.09}$	$33.33_{6.75}$	$19.61_{3.40}$	$57.82_{1.18}$	$49.69_{0.78}$	$44.26_{2.84}$
Grok-4.1-Fast	$32.54_{1.10}$	$27.45_{1.70}$	$29.82_{3.31}$	$19.05_{1.68}$	$28.85_{1.28}$	$25.00_{5.09}$	$17.65_{5.88}$	$37.41_{1.18}$	$45.57_{2.19}$	$26.23_{4.34}$
Mistral-Large-3-675B-Instruct-2512	$23.39_{1.99}$	$22.55_{6.79}$	$35.53_{3.48}$	$21.25_{1.68}$	$20.54_{2.60}$	$19.87_{2.94}$	$15.69_{3.40}$	$24.49_{2.04}$	$23.22_{2.34}$	$32.79_{4.92}$

Table 4: **Accuracy (%) on the hard subset by macro subject with overall accuracy.** Overall accuracy is averaged across all subjects. Results are reported as mean accuracy, with standard deviation shown as a subscript.

Evaluation protocol. We grade `hard` outputs from three evaluated systems using two judge models with deterministic decoding (temperature 0). For each judge, we compare three prompt templates (P1–P3) and report accuracy on the full `hard` split ($N = 1{,}053$). Prompt templates and additional details are provided in Appendix G.

Prompt-induced variation. Table 5 shows that accuracy can vary noticeably across prompts even when evaluated outputs are held fixed. This effect is strongest for GEMINI-2.5-FLASH-LITE, indicating that in mixed-format settings, measured accuracy can be sensitive to prompt-specific parsing and extraction behavior. GEMINI-3-FLASH exhibits smaller variation, suggesting more stable grading under prompt changes. Prompts with explicit answer extraction can help when responses contain long reasoning traces, but they may also introduce additional failure modes when extraction is imperfect.

Takeaway. Overall, these results suggest that LLM-as-a-judge can be useful for mixed-format evaluation, but prompt choice meaningfully affects reliability. In a small manual audit, P2 aligned most closely with human judgments, while P3 occasionally over-accepted answers when the extraction step was imperfect.

6 Error Analysis

To identify the cognitive bottlenecks of current MLLMs on KMMMU, we categorized failure cases into five distinct types. *Knowledge Shortage* aggregates failures where the model lacks relevant domain rules or cannot correctly map the problem statement to the appropriate knowledge. *Reasoning Error* covers coherent attempts that follow an incorrect inference chain. *Visual Perception Error* captures misreading or missing visual evidence in diagrams, tables, or documents. Finally, *Incomplete Response* and *Hallucination* represent output validity issues including truncated answers or unsupported fabrications.

Evaluated Model Outputs (hard)	Gemini-3-Flash (Judge)				Gemini-2.5-Flash-Lite (Judge)			
	P1	P2	P3	Range	P1	P2	P3	Range
Llama-4-Maverick-17B-128E-Instruct	25.7%	26.3%	26.0%	0.6	11.2%	19.9%	30.3%	19.1
Qwen3-VL-30B-A3B-Thinking	50.3%	56.6%	60.5%	10.2	37.5%	28.8%	53.4%	24.6
Gemini-3-Pro	68.1%	68.9%	68.6%	0.8	38.5%	52.9%	76.6%	38.1

Table 5: **Prompt sensitivity of LLM-as-a-judge on mixed-format evaluation.** Accuracy is reported for three judge prompts: **P1** minimal binary verdict, **P2** binary verdict with answer-format matching examples, and **P3** parse-then-judge that explicitly extracts the final answer before emitting the verdict. We also report **Range** (max−min across P1–P3; percentage points) to summarize prompt-induced variation.

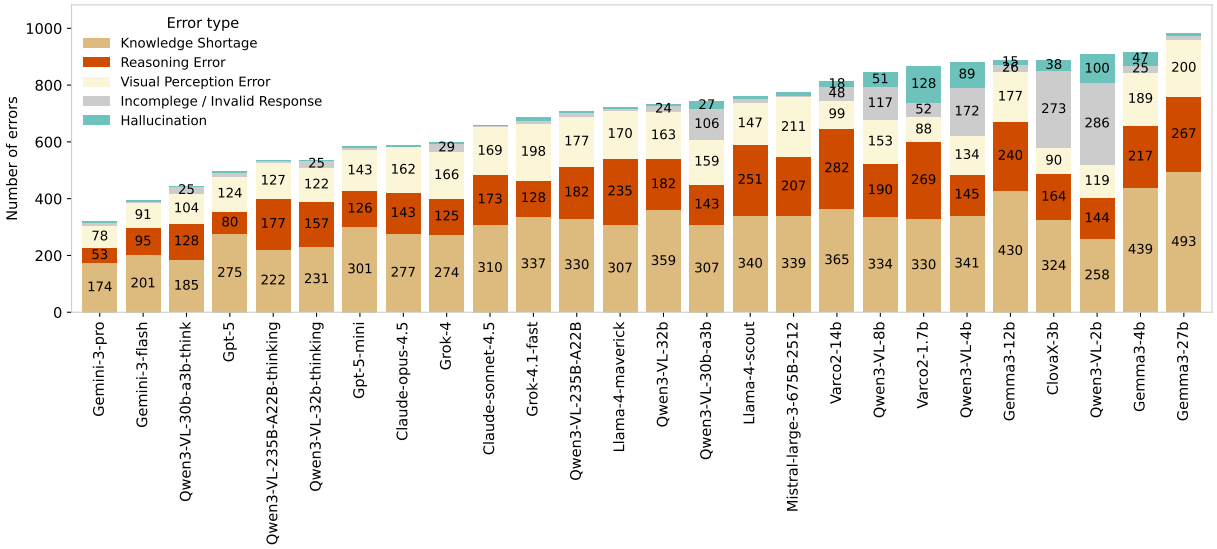


Figure 2: **Error composition on the KMMMU hard split.** We report the number of failures in each error category for each model.

6.1 The Shift in Cognitive Bottlenecks

Figure 2 illustrates that failure modes evolve as model capability increases. To determine which error types best predict performance ranking, we performed a regression analysis detailed in Appendix H.

For open source models, the primary bottleneck appears to be basic generation stability. Regression analysis identifies *Incomplete Response Rate* as the strongest predictor of ranking ($\beta = -0.09, p < 0.001$) followed closely by *Knowledge Shortage* ($\beta = -0.08, p < 0.001$). This indicates that many failures occur before the model can attempt complex multimodal reasoning. In this regime, *Visual Perception Error* has a smaller statistical impact ($\beta = -0.035$) which is consistent with a survivorship effect where instability and missing knowledge dominate first.

For frontier models, the landscape changes significantly. As output validity failures decrease, bottlenecks concentrate on substantive task difficulty. While *Knowledge Shortage* remains the top predictor ($\beta = -0.05, p < 0.001$) reflecting the expert domains of KMMMU, *Visual Perception Error* rises to become the second most critical bottleneck ($\beta = -0.045, p < 0.001$). This shift implies that once stability issues are resolved, fine grained visual understanding becomes the limiting factor for further gains.

6.2 Perceptual Bottlenecks and Structural Reasoning

To localize this perceptual limitation, we measured the visual perception error rate across different modalities as summarized in Table 6. The results reveal a stark contrast between text anchored modalities and structurally abstract ones. Models demonstrate high proficiency on text rich inputs such as *Text, Code & Documents* and *Tables* with minimal visual perception errors. This indicates that frontier models have effectively solved OCR based perception. However, performance degrades sharply on abstract visual inputs like *Charts & Plots* and *Diagrams*. Failures in these categories suggest that the

Visual Modality	Visual Perception Error (%)	Count (N)
Charts & Plots	39.11%	381
Diagrams	21.87%	4,386
Symbols & Art	19.67%	61
Mathematics	16.38%	232
Maps	11.63%	43
Photos	8.82%	34
Tables	5.48%	1,753
Text, Code & Documents	3.35%	2,360
Others	2.50%	40

Table 6: **Visual Perception Error Rate by Modality in Frontier Models.**

Model	Text, Code & Docs	Tables	Diagrams	Charts & Plots
Gemini-3-Pro	**41.7%**	**18.8%**	**30.4%**	**27.9%**
Gemini-3-Flash	48.9%	27.2%	40.4%	32.6%
GPT-5	66.7%	25.2%	49.1%	51.2%
Claude-4.5-Sonnet	73.1%	39.1%	70.8%	69.8%

Table 7: **Error rate comparison by visual modality among frontier models.** We report error rates on the hard split for major modalities.

challenge lies not in transcription but in preserving relational structure such as connectivity and layout dependent constraints.

This structural gap explains the performance advantage of GEMINI-3-PRO shown in Table 7. While most frontier models struggle with schematic inputs, GEMINI-3-PRO maintains significantly lower error rates on diagrams. This stability suggests stronger robustness to relational visual representations which effectively bridges the bottleneck between visual encoding and downstream reasoning.

6.3 Impact of Reasoning mode

We evaluated the impact of reasoning mode by comparing the QWEN3-VL series in standard instruction tuned (IT) form versus reasoning mode. Table 8 shows that reasoning mode yields large accuracy gains across model scales. Our analysis indicates that the thinking process primarily mitigates output stability issues and knowledge deficits. Comparing QWEN3-VL-30B variants reveals that incomplete responses are nearly eliminated and knowledge shortages are reduced. However, visual perception errors comprise a larger share of the remaining failures in reasoning models. This implies that while reasoning can refine logic and knowledge application, it cannot hallucinate correct visual details from a misperceived image embedding.

Base Model	Instruct (IT)	Thinking	Gain (Δ)
Qwen3-VL-30B-A3B	33.95%	55.76%	**+21.81%**
Qwen3-VL-32B	33.77%	60.28%	**+26.51%**
Qwen3-VL-235B	40.10%	58.87%	**+18.77%**

Table 8: **Accuracy gains from enabling reasoning mode in QWEN3-VL.** We compare instruction tuned models with their reasoning mode variants on the same evaluation setting.

6.4 Lessons from Error Patterns

Our error analysis points to three areas where progress appears most needed. First, the modality gap suggests that transcription oriented pretraining alone may not be sufficient for structural visual reasoning, motivating objectives that emphasize explicit structure extraction. Second, the persistence of knowledge shortages even in frontier models indicates that linguistic proficiency is not the same as institutional grounding, and that stronger coverage of Korean administrative and legal contexts may be important. Finally, the limited impact of reasoning mode on perceptual errors implies that long-form reasoning can drift away from the original visual evidence, suggesting a need for mechanisms that encourage iterative visual re-attention during inference.

7 Ablation Study

7.1 Evaluation of Image-Dependency

Model	Original	Text-only
Gemini-3-Flash	$64.91_{0.11}$	$37.14_{0.48}$
GPT-5-Mini	$47.63_{0.33}$	$23.18_{0.29}$

Table 9: **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 the validity of KMMMU as a multimodal benchmark, we evaluate the degree to which models rely on visual information to solve the problems. A robust visual reasoning benchmark should contain questions that are difficult or impossible to answer without access to the corresponding images. We conduct a text only baseline experiment where the models receive the textual question and options but the visual input is entirely omitted. Table 9 presents the performance gap between the original multimodal setting and the text only setting for GEMINI-3-FLASH and GPT-5-MINI. The results

indicate that the removal of images leads to a significant degradation in accuracy for both models. GEMINI-3-FLASH shows an average accuracy drop of 27.77% while GPT-5-MINI exhibits a decrease of 24.45%. This substantial performance decline demonstrates that KMMMU effectively captures the necessity of visual information. The models cannot rely solely on textual context or internal knowledge to solve a large majority of the tasks. The high performance gap confirms that our benchmark successfully measures the ability of models to integrate and reason over multimodal inputs.

7.2 Data Contamination Analysis

To assess the risk of memorization, we run a prefix completion test in which models receive the first 35% of a question and generate the remaining continuation. We apply this test to questions longer than 150 tokens and evaluate three frontier models under two settings. The first setting provides no additional metadata. The second setting provides the exam name and year as a potential memorization trigger.

We evaluate reconstructions using an LLM as a judge. We use GEMINI-3-FLASH to score exactness on a 0 to 100 scale based on lexical overlap, key detail preservation, and overall faithfulness to the reference question. We also report refusal and hallucination rates to characterize behavior under incomplete prompts.

Across all sources, GPT-5-MINI and GEMINI-3-PRO mostly refuse to complete the continuations, with refusal rates typically above 95%. This behavior is consistent with safety and policy training that discourages reproducing exam materials. GEMINI-3-FLASH attempts completions more often, but its exactness remains low and does not systematically increase in the hinted setting. These trends suggest that observed benchmark performance is unlikely to be explained by simple memorization. Full results by source, model, and hint setting are reported in Appendix Table 26.

8 Conclusion

We introduced **KMMMU**, a native Korean benchmark designed to evaluate expert-level multimodal understanding in real-world assessment settings. KMMMU is built from South Korean examinations and competitions and comprises 3,466 carefully verified questions with information-dense visuals, broad subject coverage, and rich annotations for subject, task, and visual modality. To better disentangle general capability from localization, we additionally curated a Korean-specific subset that requires institutional grounding, and we constructed a hard subset to concentrate frontier-level failure cases.

Our evaluations across open-source, Korean-focused, and frontier multimodal models reveal a clear capability gradient, but also a consistent set of remaining bottlenecks. While larger models and reasoning mode generally improve accuracy on problems that demand multi-step computation or constraint satisfaction, they do not fully resolve perceptual challenges. Error analyses indicate that once output stability improves, failures increasingly concentrate on *structural* visual understanding and grounded domain knowledge, with diagrams and plots remaining particularly error-prone.

Taken together, these results position KMMMU as a diagnostic benchmark for the next stage of multimodal research: models that robustly preserve relational structure in complex visuals and that exhibit institutionally grounded reasoning in localized professional contexts. We hope KMMMU will serve as a reference point for developing culturally aware and professionally competent multimodal systems, and for evaluating progress beyond English-centric or translation-based testbeds.

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 (e.g., diagrams and documents), so performance may not directly transfer to everyday perception, interactive settings, or non-exam domains.

Annotation noise and taxonomy subjectivity. Subject, task, and modality labels are ultimately based on human consolidation of initial model-proposed annotations, and some categories admit ambiguous boundaries. While we audit a random subset and manually verify all Korean-specific positives, residual label noise is likely, especially for fine-grained subjects and multi-skill questions.

Evaluation noise for mixed-format answers. Because KMMMU includes both multiple-choice and free-form items, scalable evaluation relies on

LLM-as-a-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

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.

Jinze Bai, Shuai Bai, Shusheng Yang, and 1 others. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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.

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*.

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.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

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*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

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.

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*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Marker-Inc-Korea. 2024. Ko-vlm benchmark: A comprehensive evaluation dataset for korean vision-language models. https://github.com/Marker-Inc-Korea/KO-VLM-Benchmark. Accessed: 2026-01-04.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

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. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

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*.

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*.

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.

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*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

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 3 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

B.2 Automatic Labeling and Taxonomy Consolidation

We annotate several auxiliary attributes to support analysis and stratified reporting, including subject, task, image type, and a Korean-specific flag. Initial subject and task labels are produced by GEMINI-2.5-FLASH-LITE. Image-type labels and the Korean-specific flag are produced by 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 subject taxonomy through human curation into 45 sub-subject categories and 9 macro subject categories. Task labels are also consolidated through human review into a final set of 11 task types.

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.

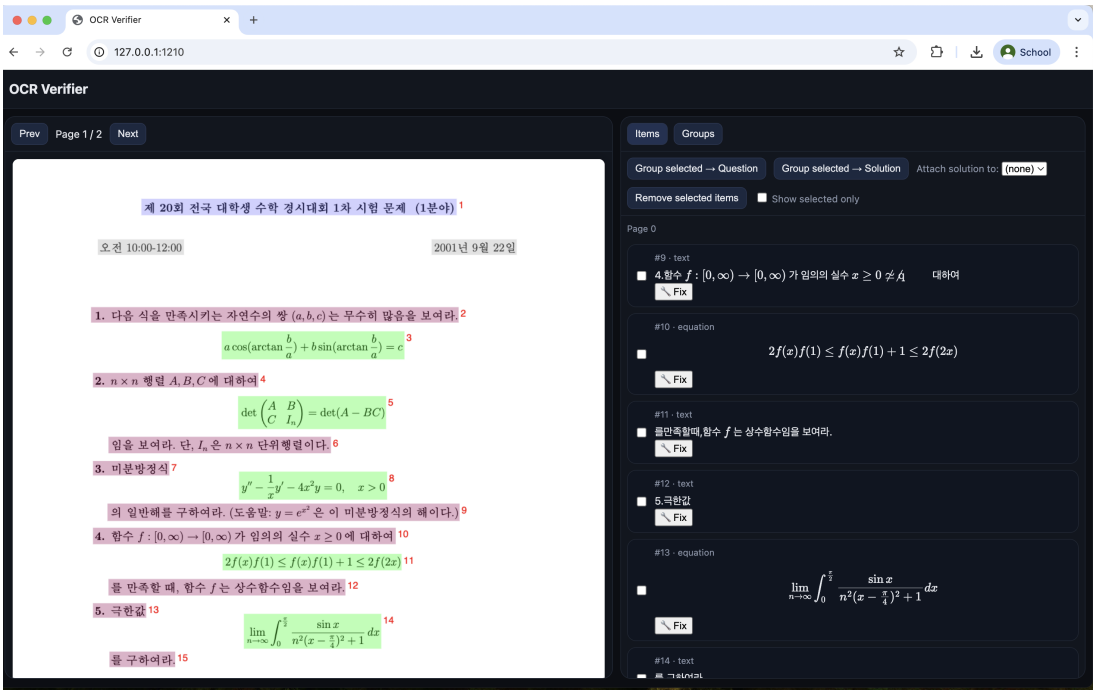


Figure 3: **Annotation tool interface used for OCR verification.** The tool displays the original PDF page on the left and the parsed text and images on the right, allowing annotators to correct OCR errors and validate image cropping in real time.

B.3 Adversarial Filtering Protocol

To ensure the benchmark evaluates frontier level capabilities, we implemented a multi stage adversarial filtering process designed to remove trivial samples that can be solved by existing mid tier models. We follow the rejection strategy below.

1. **De duplication.** We removed duplicate questions across different exam years using perceptual hashing for images and n gram overlap for text.

2. **Model based adversarial filtering.** We employed a cascade of multimodal models including INTERNVL-3.5-38B (Wang et al., 2025), GEMINI-2.5-FLASH-LITE, and GEMINI-2.5-FLASH (Comanici et al., 2025). If any of these models answered a question correctly with high confidence, the instance was excluded. This ensures that KMMMU focuses on the challenging tail of the distribution.

We initially collected approximately 68,000 raw samples. After removing invalid image links and duplicates, we apply multi stage adversarial filtering with frontier multimodal models to exclude trivial items and questions solvable without visual reasoning. The resulting dataset contains 3,466 curated questions.

C Korean-Specific Context

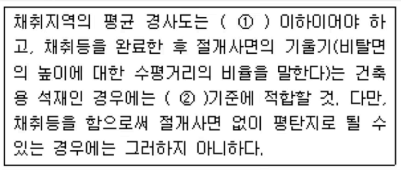
To provide a concrete understanding of the KMMMU benchmark, we present a detailed visualization of a "Korean-Specific" instance in Figure 4. Standard multimodal benchmarks often rely on universal knowledge (e.g., physics, basic math) that is invariant across cultures. In contrast, KMMMU includes a dedicated subset of questions that require localized knowledge. Figure 4 shows a sample data card from the dataset. The input consists of an image containing a specific regulation text and the corresponding question. As shown in the translation panels, the model must interpret the visual text ("extraction area slope criteria") specifically within the context of the South Korean *Mountainous Districts Management Act* to identify the correct legal standards (Option 3). This necessitates that the model possesses not only optical character recognition capabilities but also grounded knowledge of Korean administrative laws.

D Detailed Dataset Statistics

In this section, we provide a granular breakdown of the dataset composition. Beyond the overview (Ta-

Korean Specific Question

[Original Image]



[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 4: **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.

ble 10), we report (i) the distribution of fine-grained subject categories (Table 11), (ii) the macro task distribution (Table 12), (iii) the question format distribution (Table 13), and (iv) the taxonomy of visual modalities (Table 14).

Statistic	Count
Total Questions	3,466
Macro Subject Categories	9
Sub-Subject Categories	45
Macro Visual Modality Types	8
Macro Task Types	11
Question Types	5
Questions with in-image texts	2,383 (68.75%)
Questions w/o in-image texts	1,083 (31.25%)
Korean-specific questions	299 (8.63%)

Table 10: **Dataset overview.** We report counts and percentages for key attributes such as in-image text and Korean-specific content.

D.1 Subject Category Distribution

Table 11 details the frequency of questions across 45 fine-grained subject 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.

D.2 Macro Task Distribution

Table 12 summarizes the distribution of macro task labels. This breakdown clarifies what primary competency is most frequently required by KMMMU (e.g., concept application, data and visual interpretation, and quantitative reasoning), and supports task-wise analyses in later sections.

D.3 Question Format Distribution

Because KMMMU contains both multiple-choice and free-form items, the answer format affects eval-

uation difficulty and failure modes. Table 13 reports the distribution of question formats in the benchmark.

D.4 Visual Modality Taxonomy

Table 14 presents our hierarchical visual taxonomy. We grouped fine-grained visual types into 8 supercategories. *Diagrams* (including circuit, mechanical, and structural diagrams) are the most dominant modality (49.9%), reflecting the dataset's focus on professional schematics interpretation. To provide visual references for these categories, we display representative examples in Figure 5.

D.5 Question Type Taxonomy

Table 15 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.

D.6 Analysis of the Hard Subset

To further analyze model performance on the most challenging instances, we identified a "Hard Subset" ($N = 1,053$). This subset consists of questions that require complex multi-step reasoning and were frequently answered incorrectly during our pilot evaluation with open-source models. Table 17 shows the distribution of tasks within this hard subset. Notably, *Data and Visual Interpretation* and *Concept Application* dominate this subset, suggesting that the primary difficulty for current models lies not in simple knowledge recall, but in the synthesis of visual data with domain concepts. We also report the prevalence of Korean-specific items by split in Table 18. The Hard set contains 104 Korean-specific questions (9.92%), which is slightly higher than the Full set (8.63%), indicating that localization-heavy items are somewhat overrepresented among challenging instances.

E Macro-task Performance Breakdown

This appendix provides extended quantitative results that are omitted from the main text due to space constraints. We report macro-task performance breakdowns and dataset composition statistics to support replication and alternative aggregations.

Table 19 reports per-model accuracy by macro task on the full split. This breakdown complements the macro-subject analysis by separating items that may share subject coverage but require different operations such as information extraction, multi-step reasoning, or quantitative computation. Values are reported as mean and standard deviation over three runs.

Table 20 reports the same macro-task breakdown on the hard subset. The hard subset magnifies failure modes that survive after removing easier instances, so the task-wise profile can shift even when the model ranking is preserved.

F Korean-Specific Performance Breakdown

This appendix reports a per-model breakdown of performance on Korean-specific content. We partition the benchmark using the binary flag `is_korean_specific` and compute accuracy separately on `True` and `False` subsets. We also report the gap in percentage points as $\Delta = \mathrm{Acc}_{\texttt{True}} - \mathrm{Acc}_{\texttt{False}}$. All values are reported as mean and standard deviation across three runs.

F.1 Subset Size

Table 18 reports the number of Korean-specific items in the full and hard splits. Because the `True` subset is smaller, the variance of $\mathrm{Acc}_{\texttt{True}}$ can be higher, and Δ should be interpreted together with both subset sizes. Because the `True` subset is relatively small, both $\mathrm{Acc}_{\texttt{True}}$ and Δ can have higher variance than their `False` counterparts. We therefore interpret Δ together with subset sizes and the reported standard deviations.

F.2 Per-model Results

Table 21 reports per-model accuracy on Korean-specific content for the full split. Models are sorted by Δ to highlight which systems degrade most on localized content. A small or positive Δ is not sufficient evidence of localization strength when the overall accuracy is near a shared error floor. When Δ is small, differences can fall within run-to-run variation, especially on the smaller `True` subset. Accordingly, we treat large negative Δ as stronger evidence of a localization bottleneck than near-zero or positive values at low absolute accuracy.

Table 22 reports the same breakdown for the hard split. The hard split can also change the prevalence of Korean-specific items, so we report both split-level sizes (Table 18) and per-model breakdowns for transparency.

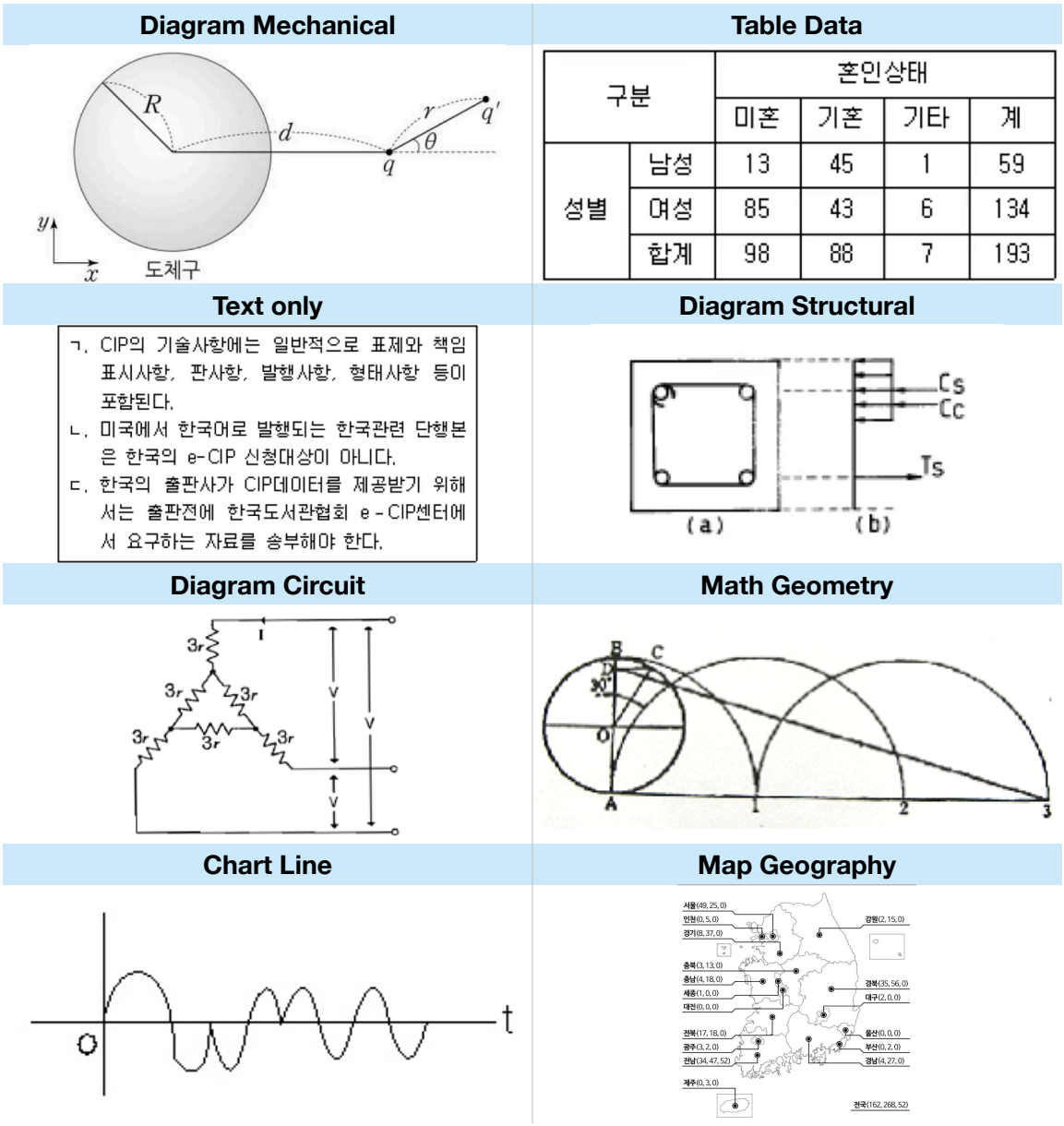


Figure 5: **Representative examples for each visual modality category.** KMMMU requires models to interpret various distinct visual formats, including specialized engineering diagrams and South Korean geographic maps.

G Additional Analysis on Prompt Sensitivity of LLM-as-a-Judge

We conduct a lightweight manual check to compare judge prompts against human verification. P2 matches human judgments most consistently. While P3 can reduce some format-related ambiguity by extracting a final answer, we observe cases where imperfect extraction leads to overly permissive verdicts, suggesting that explicit parsing can introduce additional failure modes.

Prompt-induced variation. Table 5 shows that measured accuracy can vary substantially across prompts even when the evaluated outputs are held fixed. The effect is most pronounced for GEMINI-2.5-FLASH-LITE, indicating sensitivity to prompt-specific parsing and extraction behavior in mixed-format grading. In contrast, GEMINI-3-FLASH exhibits smaller variation, suggesting more stable grading under prompt changes.

Implications for evaluation. These results highlight that reported accuracy reflects not only model behavior but also the reliability of the grading protocol. Accordingly, prompt choice should be validated with spot-checks when grading mixed-format outputs.

G.1 LLM-as-a-Judge Prompt Templates

We adopt an LLM-as-a-judge protocol compatible with mixed-format questions. Table 23 lists the three templates used in our experiments, and the main text quantifies their prompt sensitivity under fixed evaluated outputs. Table 23 lists the three prompt templates used in our experiments.

H Detailed Regression Analysis

To rigorously quantify the impact of distinct failure modes on overall performance, we conducted an Ordinary Least Squares (OLS) regression analysis. The dependent variable was the overall accuracy of each model, while the independent variables were the error rates for the five error categories categorized in Section 6. We standardized all independent variables to unit variance to allow for a direct comparison of the resulting β coefficients. A negative coefficient indicates that an increase in that specific error type correlates with a decrease in overall accuracy, identifying it as a performance bottleneck.

Table 24 presents the standardized coefficients for both open-source and frontier model groups. For open-source models, the analysis identifies *Incomplete Response* as the most statistically significant predictor of performance ranking ($\beta = -0.088$, $p < 0.001$). This is closely followed by *Knowledge Shortage* ($\beta = -0.081$, $p < 0.001$). These figures suggest that the primary constraints for these models lie in generation stability and domain knowledge availability. Conversely, *Visual Perception Error* exhibits a relatively smaller impact ($\beta = -0.035$), consistent with a survivorship bias where models frequently fail due to instruction following or knowledge deficits before visual perception becomes the deciding factor.

In the frontier model regime, the regression coefficients indicate a fundamental shift in critical challenges. The impact of *Incomplete Response* diminishes to a negligible level ($\beta = -0.007$), confirming that generation stability issues are largely resolved. However, *Knowledge Shortage* remains the dominant predictor ($\beta = -0.052$) reflecting the high difficulty of the expert domains in KMMMU. Most notably, *Visual Perception Error* rises in relative importance to become the second most critical bottleneck ($\beta = -0.045$). This statistical evidence supports the conclusion that expert-level multimodal proficiency is increasingly bounded by fine-grained visual understanding once foundational capabilities are established.

I Per Model Error Rates by Image Sub Category

This section provides a fine grained view of model behavior by reporting error rates for frequent image sub categories on the hard split. We include sub categories with at least 10 samples to reduce instability from small counts. For each sub category, we report the number of evaluated instances, the number of errors, and the resulting error rate. This breakdown is intended to support qualitative interpretation of the main text rather than to rank models on rare categories.

J Data Contamination Analysis Details

To assess the risk of memorization, we conduct a prefix completion test in which a model is given the first 35% of a question and asked to reconstruct the remaining 65%. We run this experiment on questions longer than 150 tokens to ensure that the continuation is non-trivial. We evaluate three frontier models under two settings. The first is a no-hint setting. The second is a hinted setting that provides the exam name and year, which can serve as a potential trigger for recall.

Evaluation with LLM-as-a-Judge We score reconstructions using an LLM-as-a-Judge framework based on GEMINI-3-FLASH. The judge compares the reconstructed continuation against the reference question and outputs an exactness score on a 0 to 100 scale. The rubric emphasizes lexical overlap, preservation of key details such as numbers, entities, and constraints, and overall faithfulness to the reference. The judge is instructed to be strict and to award high scores only when the reconstruction closely matches the original text. In addition, the judge flags whether the model refused and whether it hallucinated unsupported details.

Results and interpretation Table 26 reports mean exactness, median exactness, refusal rate, and hallucination rate for each source and hint setting. Two patterns are consistent across sources. First, GPT-5-MINI and GEMINI-3-PRO exhibit very high refusal rates, often above 95%, and near-zero exactness in both settings. This behavior is consistent with training that discourages reproducing exam materials, which reduces the practical risk of direct leakage. Second, GEMINI-3-FLASH attempts completions more frequently and achieves higher mean exactness, but the scores remain low overall and do not reliably increase in the hinted

setting. In some sources, hinted prompts reduce exactness. Together, the absence of consistently high exactness scores and the lack of systematic improvement under hints suggest that benchmark performance is unlikely to be explained by simple memorization.

Subject Category	Count
Physics	474
Civil & Structural Engineering	332
Mechanical Engineering	223
Electrical & Electronics Engineering	221
Computer Science	164
Statistics & Probability	149
Mathematics	121
Business Administration & Management	104
Environmental Science & Engineering	99
Geography & Spatial Studies	95
Manufacturing & Production Engineering	90
Public Administration & Policy	89
Design & Visual Arts	88
Chemical Engineering	76
Chemistry	75
Law & Legal Studies	74
Materials Science & Metallurgy	69
Information Technology & Systems	69
General Knowledge & Interdisciplinary	67
Industrial & Systems Engineering	64
Architecture & Urban Studies	57
Human Resources & Organizational Studies	56
Linguistics & Language Studies	54
Biology	43
Safety, Risk & Reliability Engineering	42
Data Science & Analytics	41
Finance & Accounting	40
Sociology & Social Sciences	36
Economics	33
Earth & Geological Sciences	33
Library, Archival & Information Science	30
Transportation & Logistics	29
Agriculture & Life Sciences	27
Software & Programming	27
Astronomy & Space Science	26
Psychology	22
Cognitive & Behavioral Sciences	20
Biomedical & Health Sciences	20
Communication & Media Studies	17
Education & Pedagogy	16
Marketing & Consumer Studies	15
Ethics & Philosophy	12
Artificial Intelligence & Machine Learning	12
Energy & Power Systems	11
Sports & Exercise Science	4

Table 11: **Distribution of Subject Categories in KMMMU**

Macro Task	Count	%
Concept Application	948	27.35
Data & Visual Interpretation	852	24.58
Quantitative Reasoning	516	14.89
Knowledge Recall	347	10.01
Language Understanding	183	5.28
Geometry & Spatial Reasoning	155	4.47
Information Extraction & Classification	134	3.87
Logic & Rule Reasoning	119	3.43
Optimization & Planning	107	3.09
Code & Debugging	58	1.67
Other	47	1.36

Table 12: **Macro task distribution.**

Question Type	Count	%
Multiple Choice (Single Answer)	2,851	82.25
Multiple Choice (Multiple Answers)	347	10.01
Short Answer (Numerical / Calculation)	176	5.08
Short Answer (Text)	69	2.00
Descriptive	23	0.66

Table 13: **Question type distribution.**

Super-category	Sub-category	Count
Diagrams (1,731)	diagram_mechanical	726
	diagram_structural	391
	diagram_circuit	330
	diagram_flowchart	103
	diagram_network	67
	diagram_geologic	43
	diagram_chemical	42
	diagram_biological	26
Text, Code & Documents (869)	text_only	456
	text_blank	276
	document_scan	105
	text_code	32
Tables (549)	table_data	508
	table_blank	40
Mathematics (88)	math_geometry	65
	math_function	23
Charts & Plots (151)	chart_line	109
	chart_bar	18
	chart_scatter	13
	chart_pie	5
	Others (mixed, box, radar)	6
Maps (23)	map_geographic	8
	map_topographic	7
	map_weather	5
	map_astronomy	3
Symbols & Art (22)	symbol_pictogram	16
	painting	8
Photographs (13)	photo_general	11
	Others (astronomy, organism)	2

Table 14: **KMMMU Dataset Visual Modality Taxonomy.** We organize fine-grained visual sub-categories into a set of semantically coherent super-categories to support structured analysis of multimodal reasoning. The dataset is rich in technical diagrams and document-based visuals.

Macro Subject	Multiple Choice (Single Answer)	Multiple Choice (Multiple Answers)	Short Answer (Text)	Descriptive	Short Answer (Numerical)
Engineering	1142	39	5	2	26
Natural Sciences	486	170	15	14	112
CS & IT	315	13	6	2	7
Business & Public	265	31	4	0	4
Math & Stats	175	64	9	1	21
Social Sciences	206	28	4	0	1
General	90	1	25	4	5
Arts & Design	88	0	0	0	0
Law & Ethics	84	1	1	0	0

Table 15: **Dataset attributes by macro subject and answer type.**

Table 16: Macro Subject Distribution in the KMMMU Hard Set

Macro Subject	Count
Engineering	357
Natural Sciences	267
Math & Stats	98
CS & IT	91
Business & Public Admin	76
Social Sciences	61
General/Interdisciplinary	52
Arts & Design	34
Law & Ethics	17
Total	**1,053**

Macro Task	Count	%
Data & Visual Interpretation	299	28.40
Concept Application	274	26.02
Quantitative Reasoning	123	11.68
Knowledge Recall	112	10.63
Language Understanding	67	6.36
Logic & Rule Reasoning	44	4.17
Optimization & Planning	39	3.70
Information Extraction & Classification	35	3.32
Geometry & Spatial Reasoning	33	3.13
Code & Debugging	15	1.42
Other	12	1.13

Table 17: **Macro task distribution on the hard subset.** Percentages are computed over the hard subset ($N = 1{,}053$).

Split	Korean Specific	Not Korean Specific	Ratio (%)
Full	299	3167	8.63
Hard	104	949	9.92

Table 18: **Number of Korean-specific questions each in Full and Hard sets.**

Model	Concept Application	Data & Visual Interpretation	Quantitative Reasoning	Knowledge Recall	Language Understanding	Geometry & Spatial Reasoning	Info Extraction & Classification	Logic & Rule Reasoning	Optimization & Planning	Code & Debugging	Other	Overall Acc.
clovaX-3b	$13.47_{0.64}$	$11.11_{0.38}$	$8.14_{0.67}$	$17.48_{0.33}$	$19.67_{1.89}$	$7.74_{1.71}$	$23.13_{2.59}$	$20.45_{1.94}$	$13.08_{2.47}$	$15.52_{1.72}$	$9.93_{2.46}$	$13.16_{0.13}$
gemma3-12b	$20.25_{0.11}$	$16.74_{1.36}$	$8.85_{0.62}$	$22.38_{0.60}$	$13.30_{0.63}$	$9.25_{0.74}$	$18.91_{2.62}$	$22.97_{3.97}$	$14.63_{1.95}$	$14.94_{3.59}$	$9.22_{1.23}$	$16.68_{0.42}$
gemma3-27b	$20.96_{0.34}$	$18.54_{0.41}$	$13.18_{1.21}$	$24.88_{1.92}$	$14.39_{2.28}$	$10.11_{1.49}$	$23.38_{3.11}$	$18.49_{2.22}$	$16.20_{1.95}$	$22.99_{6.97}$	$18.44_{2.46}$	$18.63_{0.36}$
gemma3-4b	$14.24_{1.02}$	$8.84_{1.42}$	$5.49_{0.99}$	$21.23_{1.09}$	$14.03_{2.70}$	$4.52_{0.65}$	$13.18_{0.43}$	$12.04_{2.57}$	$6.85_{1.43}$	$12.07_{4.56}$	$8.51_{3.69}$	$11.41_{0.50}$
koreason	$47.78_{2.11}$	$60.33_{1.23}$	$76.68_{2.20}$	$38.42_{1.30}$	$36.07_{2.89}$	$75.48_{0.65}$	$46.77_{1.14}$	$51.54_{3.88}$	$66.36_{0.00}$	$47.70_{1.99}$	$75.89_{4.43}$	$55.90_{0.33}$
llama-4-maverick	$37.06_{0.22}$	$43.70_{0.24}$	$41.54_{1.29}$	$39.87_{1.09}$	$30.60_{4.13}$	$32.26_{1.29}$	$44.78_{1.29}$	$35.01_{0.97}$	$38.01_{2.70}$	$39.66_{1.72}$	$23.40_{3.69}$	$39.20_{0.38}$
llama-4-scout	$31.72_{1.64}$	$35.80_{0.89}$	$34.24_{1.43}$	$36.12_{0.44}$	$26.41_{1.38}$	$30.97_{1.29}$	$46.52_{1.14}$	$34.73_{3.40}$	$34.58_{5.84}$	$28.74_{4.98}$	$32.62_{4.91}$	$33.95_{0.54}$
qwen3-2b	$17.58_{0.70}$	$14.55_{1.24}$	$9.11_{1.03}$	$22.19_{1.04}$	$15.48_{2.21}$	$8.39_{1.94}$	$26.37_{1.14}$	$21.85_{2.22}$	$12.46_{1.43}$	$9.77_{3.59}$	$8.51_{0.00}$	$15.59_{0.23}$
qwen3-30b-a3b	$30.59_{1.92}$	$39.32_{0.47}$	$33.91_{0.67}$	$33.53_{0.73}$	$28.23_{0.32}$	$17.85_{0.37}$	$45.77_{3.02}$	$35.57_{4.31}$	$39.56_{2.70}$	$28.74_{5.54}$	$25.53_{2.13}$	$33.77_{0.44}$
qwen3-30b-a3b-think	$47.15_{7.12}$	$63.03_{7.43}$	$69.12_{8.62}$	$44.19_{6.98}$	$38.98_{5.05}$	$66.88_{10.65}$	$50.75_{4.89}$	$54.34_{11.16}$	$70.72_{8.98}$	$45.98_{11.22}$	$60.99_{0.83}$	$55.76_{7.53}$
qwen3-32b	$33.30_{0.76}$	$43.15_{1.30}$	$39.53_{1.08}$	$35.16_{1.80}$	$27.14_{2.07}$	$29.46_{2.98}$	$41.79_{4.16}$	$37.54_{2.95}$	$41.12_{4.07}$	$36.78_{4.34}$	$31.21_{6.14}$	$37.08_{0.81}$
qwen3-32b-thinking	$44.59_{2.09}$	$61.46_{0.95}$	$67.12_{0.40}$	$39.58_{1.48}$	$37.16_{0.95}$	$67.53_{3.31}$	$51.99_{0.86}$	$50.42_{2.52}$	$68.54_{3.00}$	$48.85_{4.34}$	$60.28_{4.43}$	$53.73_{0.49}$
qwen3-4b	$21.52_{0.46}$	$19.29_{0.95}$	$15.70_{0.70}$	$29.39_{2.29}$	$19.85_{0.83}$	$10.54_{0.99}$	$32.09_{1.29}$	$25.77_{2.70}$	$16.20_{1.08}$	$31.61_{7.18}$	$12.77_{3.69}$	$20.75_{0.32}$
qwen3-8b	$25.07_{1.39}$	$25.27_{2.66}$	$21.06_{1.07}$	$32.76_{2.04}$	$23.50_{1.97}$	$16.99_{0.37}$	$36.57_{2.99}$	$31.37_{1.28}$	$23.99_{3.89}$	$28.74_{2.63}$	$16.31_{1.23}$	$25.42_{0.53}$
qwen3-VL-235B-A22B-thinking	$45.25_{0.18}$	$61.27_{0.41}$	$65.96_{2.50}$	$36.70_{2.58}$	$39.53_{3.01}$	$64.52_{2.23}$	$48.01_{2.62}$	$53.50_{1.28}$	$71.96_{0.93}$	$46.55_{1.72}$	$58.87_{3.25}$	$53.39_{0.19}$
qwen3-VL-235b-A22B-Instruct	$34.77_{1.68}$	$47.81_{0.82}$	$45.54_{1.51}$	$35.64_{0.88}$	$31.88_{3.01}$	$28.39_{3.92}$	$44.53_{1.55}$	$39.78_{1.75}$	$45.17_{2.35}$	$35.06_{2.63}$	$34.75_{6.84}$	$40.10_{0.31}$
varco2-1.7b	$13.99_{2.35}$	$9.00_{3.18}$	$5.75_{1.48}$	$15.85_{2.59}$	$11.66_{3.11}$	$10.32_{3.92}$	$14.18_{3.00}$	$10.08_{5.51}$	$9.03_{5.63}$	$14.37_{1.00}$	$4.26_{0.00}$	$11.03_{2.39}$
varco2-14b	$24.30_{0.60}$	$19.48_{0.31}$	$21.96_{1.48}$	$26.13_{3.21}$	$20.22_{3.94}$	$24.95_{0.99}$	$30.85_{2.15}$	$26.05_{3.85}$	$16.20_{0.54}$	$27.59_{1.72}$	$19.86_{2.46}$	$22.82_{0.16}$

Table 19: **Accuracy (%) by macro task with overall accuracy.** Values are reported as mean_{std} across three runs.

Model	Concept Application	Data & Visual Interpretation	Quantitative Reasoning	Knowledge Recall	Language Understanding	Geometry & Spatial Reasoning	Info Extraction & Classification	Logic & Rule Reasoning	Optimization & Planning	Code & Debugging	Other	Overall Acc.
Gemini-3-Pro	$67.40_{3.04}$	$74.69_{1.84}$	$71.82_{1.24}$	$59.52_{1.86}$	$70.15_{1.49}$	$57.58_{3.03}$	$60.95_{1.65}$	$62.88_{1.31}$	$78.63_{1.48}$	$60.00_{6.67}$	$75.00_{0.00}$	$69.01_{0.29}$
Gemini-3-Flash	$61.07_{0.56}$	$64.55_{2.74}$	$65.85_{3.73}$	$58.33_{2.73}$	$49.25_{4.48}$	$44.44_{4.63}$	$53.33_{7.19}$	$50.76_{4.73}$	$64.96_{3.92}$	$48.89_{3.85}$	$55.56_{12.73}$	$60.27_{0.15}$
GPT-5	$48.30_{2.43}$	$61.43_{0.97}$	$58.54_{1.41}$	$36.90_{3.14}$	$42.79_{1.72}$	$31.31_{3.50}$	$39.05_{7.19}$	$46.97_{3.47}$	$67.52_{1.48}$	$42.22_{3.85}$	$58.33_{8.33}$	$51.50_{0.88}$
GPT-5-mini	$40.27_{1.80}$	$53.07_{1.17}$	$49.59_{1.63}$	$25.89_{1.55}$	$33.83_{4.31}$	$23.23_{6.31}$	$28.57_{8.57}$	$35.61_{5.72}$	$61.54_{4.44}$	$40.00_{6.67}$	$47.22_{9.62}$	$42.80_{0.33}$
Claude-Opus-4.5	$38.69_{0.97}$	$51.51_{1.53}$	$46.88_{2.05}$	$30.65_{1.36}$	$38.81_{2.99}$	$28.28_{1.75}$	$35.24_{3.30}$	$34.09_{6.82}$	$53.85_{2.56}$	$51.11_{7.70}$	$41.67_{8.33}$	$42.58_{0.63}$
Claude-Sonnet-4.5	$27.13_{1.69}$	$43.92_{0.39}$	$38.21_{2.44}$	$28.57_{4.46}$	$24.38_{2.28}$	$23.23_{4.63}$	$31.43_{4.95}$	$29.55_{2.27}$	$38.46_{6.78}$	$46.67_{6.67}$	$27.78_{9.62}$	$34.00_{0.83}$
Grok-4	$39.78_{3.81}$	$47.71_{0.70}$	$53.12_{3.67}$	$33.33_{2.73}$	$31.84_{4.80}$	$29.29_{1.75}$	$38.10_{4.36}$	$35.61_{5.72}$	$48.72_{5.13}$	$48.89_{3.85}$	$50.00_{14.43}$	$42.42_{1.62}$
Grok-4.1-fast	$35.40_{3.12}$	$30.21_{0.51}$	$47.70_{0.47}$	$23.21_{0.89}$	$23.88_{5.38}$	$31.31_{1.75}$	$26.67_{3.30}$	$21.21_{1.31}$	$41.03_{2.56}$	$24.44_{3.85}$	$50.00_{8.33}$	$32.54_{1.10}$
Mistral-Large-3-675B-Instruct-2512	$22.14_{1.87}$	$24.30_{2.04}$	$23.58_{2.93}$	$23.21_{3.09}$	$20.40_{5.65}$	$22.22_{9.74}$	$28.57_{2.86}$	$18.94_{1.31}$	$29.06_{6.45}$	$24.44_{3.85}$	$30.56_{4.81}$	$23.39_{1.99}$

Table 20: **Accuracy (%) on the hard subset by macro task.** Values are reported as mean_{std} over three runs.

Model	**#True**	**Acc$_{\texttt{True}}$ (%)**	**#False**	**Acc$_{\texttt{False}}$ (%)**	**Δ (pp)**
KO-REAson-G3-12B-1009	299	$32.66_{2.04}$	3167	$58.01_{0.17}$	-25.35
Qwen3-32B-Thinking	299	$31.77_{1.16}$	3167	$55.83_{0.45}$	-24.05
Qwen3-VL-235B-A22B-Thinking	299	$32.44_{0.58}$	3167	$55.35_{0.30}$	-22.91
Qwen3-VL-30B-A3B-Thinking	299	$35.45_{7.91}$	3167	$57.62_{7.50}$	-22.17
Qwen3-VL-235B-A22B-IT	299	$26.98_{1.65}$	3167	$41.34_{0.26}$	-14.36
Llama-4-Maverick-17B-128E-IT	299	$28.32_{0.39}$	3167	$40.29_{0.40}$	-11.98
Qwen3-VL-32B-IT	299	$26.31_{3.35}$	3167	$38.12_{0.82}$	-11.81
Qwen3-VL-30B-A3B-IT	299	$23.52_{0.19}$	3167	$34.73_{0.45}$	-11.21
Llama-4-Scout-17B-16E-IT	299	$27.20_{0.39}$	3167	$34.69_{0.62}$	-7.49
VARCO-VISION-2.0-14B	299	$22.30_{1.58}$	3167	$22.99_{0.07}$	-0.70
Qwen3-VL-2B-IT	299	$15.27_{1.51}$	3167	$15.66_{0.38}$	-0.39
Qwen3-VL-4B-IT	299	$20.62_{2.92}$	3167	$20.86_{0.34}$	-0.24
Gemma-3-27B-IT	299	$18.84_{3.58}$	3167	$18.66_{0.15}$	+0.18
Qwen3-VL-8B-IT	299	$25.98_{3.37}$	3167	$25.46_{0.52}$	+0.51
VARCO-VISION-2.0-1.7B	299	$11.71_{3.48}$	3167	$11.00_{2.29}$	+0.71
Gemma-3-12B-IT	299	$21.85_{0.70}$	3167	$16.25_{0.45}$	+5.60
HyperCLOVAX-SEED-Vision-3B	299	$18.73_{1.86}$	3167	$12.69_{0.30}$	+6.04
Gemma-3-4B-IT	299	$17.28_{1.72}$	3167	$10.88_{0.64}$	+6.40

Table 21: **Per-model accuracy on Korean-specific content (Full split).** Values are reported in **percentage** using the notation mean_{std} across 3 runs. Δ is in percentage points (pp): $\Delta = \text{Acc}_{True} - \text{Acc}_{False}$, and models are sorted by Δ in ascending order (more negative indicates larger degradation on Korean-specific items).

Model	**#True**	**Acc$_{\texttt{True}}$**	**#False**	**Acc$_{\texttt{False}}$**	**Δ (pp)**
Gemini-3-Pro	104	$58.01_{1.47}$	949	$70.09_{0.53}$	-12.08
Gemini-3-Flash	104	$52.88_{1.67}$	949	$61.02_{0.28}$	-8.13
GPT-5	104	$33.01_{1.11}$	949	$53.50_{0.92}$	-20.48
GPT-5-mini	104	$31.41_{1.11}$	949	$43.96_{0.28}$	-12.55
Grok-4	104	$30.45_{2.00}$	949	$43.64_{1.84}$	-13.20
Grok-4.1-fast	104	$19.55_{1.11}$	949	$33.83_{1.13}$	-14.28
Claude-Opus-4.5	104	$35.58_{2.54}$	949	$43.36_{0.43}$	-7.78
Claude-Sonnet-4.5	104	$32.37_{2.22}$	949	$34.22_{0.74}$	-1.84
Mistral-Large-3-675B-Instruct-2512	104	$26.28_{2.42}$	949	$22.95_{1.95}$	+3.33

Table 22: **Per-model accuracy on Korean-specific content (Hard split).** Mean$\pm$std across 3 runs. Δ is in percentage points (pp): $\Delta = \text{Acc}_{\texttt{True}} - \text{Acc}_{\texttt{False}}$.

#1

```
You will be given a question, gold answer, and a model response.
Decide whether the model response matches the gold answer.
- Output EXACTLY ONE LINE (first line only): [TRUE] or [FALSE]
Do not output anything else.
```

#2

```
You will be given a question, gold answer, and a model response.
Decide whether the model response matches the gold answer.

Step 1) Determine question type:
- If the question contains explicit options/choices (e.g., ①②③④⑤, A/B/C/D, 1-5, ㄱ-ㄹ),
treat it as MULTIPLE-CHOICE.
- Otherwise treat it as FREE-FORM.

Step 2) For MULTIPLE-CHOICE:
- Extract the model's FINAL selected option only.
- Normalize equivalent formats:
- ①②③④⑤ == 1 2 3 4 5
- "4번", "정답은 ④", "answer: (4)" all mean option 4
- "A", "(A)", "Option A" all mean option A

Step 3) For FREE-FORM:
- Accept paraphrases that preserve the same meaning.
- The response must contain the same final conclusion as the gold answer.

==================================
- Output EXACTLY ONE LINE: [TRUE] or [FALSE]
Do not output anything else.
```

#3

```
You will be given a question, gold answer, and a model response.
Decide whether the model response matches the gold answer.
First, decide question type:
- MULTIPLE-CHOICE if explicit options exist, else FREE-FORM.
Second, extract the model's final answer:
- Extract the model's FINAL answer and normalize formats (①=1, "4번"=4, etc.).
- Mark TRUE when the normalized final option equals the gold option.
- If the response is partially correct, missing a required item, or contains any
contradiction, output [FALSE].
- Do not reward verbosity. Do not infer missing information.

- Output EXACTLY ONE LINE:
Final Answer: extracted final answer, Decision: [TRUE] or [FALSE]
Do not output anything else.
```

Table 23: Judge prompt templates used for model evaluation.

Error Type	**Open-Source Models** (β)	**Frontier Models** (β)
Incomplete Response	**-0.088** ($p < 0.001$)	-0.007 ($p < 0.001$)
Knowledge Shortage	**-0.081** ($p < 0.001$)	**-0.052** ($p < 0.001$)
Reasoning Error	-0.052 ($p < 0.001$)	-0.042 ($p < 0.001$)
Visual Perception	-0.035 ($p < 0.001$)	**-0.045** ($p < 0.001$)
Hallucination	-0.037 ($p < 0.001$)	-0.003 ($p = 0.002$)

Table 24: **OLS Regression Results: Predictors of Model Accuracy.** Values represent standardized beta coefficients (β). A larger negative value indicates that the error type is a stronger cause of performance degradation.

Image Sub-Category	Model	Count	Error	Error Rate (%)
chart_line	Gemini-3-Pro	32	11	34.38
	Gemini-3-Flash	32	10	31.25
	GPT-5	32	16	50.00
	Claude-Opus-4.5	32	20	62.50
	Grok-4	32	23	71.88
diagram_circuit	Gemini-3-Pro	109	41	37.61
	Gemini-3-Flash	109	49	44.95
	GPT-5	109	55	50.46
	Claude-Opus-4.5	109	77	70.64
	Grok-4	109	73	66.97
diagram_flowchart	Gemini-3-Pro	22	9	40.91
	Gemini-3-Flash	22	8	36.36
	GPT-5	22	12	54.55
	Claude-Opus-4.5	22	14	63.64
	Grok-4	22	13	59.09
diagram_mechanical	Gemini-3-Pro	242	59	24.38
	Gemini-3-Flash	242	88	36.36
	GPT-5	242	104	42.98
	Claude-Opus-4.5	242	135	55.79
	Grok-4	242	119	49.17
diagram_network	Gemini-3-Pro	19	8	42.11
	Gemini-3-Flash	19	10	52.63
	GPT-5	19	10	52.63
	Claude-Opus-4.5	19	12	63.16
	Grok-4	19	13	68.42
diagram_structural	Gemini-3-Pro	70	24	34.29
	Gemini-3-Flash	70	28	40.00
	GPT-5	70	44	62.86
	Claude-Opus-4.5	70	50	71.43
	Grok-4	70	40	57.14
document_scan	Gemini-3-Pro	19	4	21.05
	Gemini-3-Flash	19	5	26.32
	GPT-5	19	6	31.58
	Claude-Opus-4.5	19	10	52.63
	Grok-4	19	9	47.37
table_data	Gemini-3-Pro	188	32	17.02
	Gemini-3-Flash	188	50	26.60
	GPT-5	188	46	24.47
	Claude-Opus-4.5	188	64	34.04
	Grok-4	188	87	46.28
text_blank	Gemini-3-Pro	104	49	47.12
	Gemini-3-Flash	104	54	51.92
	GPT-5	104	84	80.77
	Claude-Opus-4.5	104	82	78.85
	Grok-4	104	80	76.92
text_only	Gemini-3-Pro	117	51	43.59
	Gemini-3-Flash	117	60	51.28
	GPT-5	117	75	64.10
	Claude-Opus-4.5	117	73	62.39
	Grok-4	117	85	72.65

Table 25: **Per model error rates by image sub category on the hard split.** We report the number of evaluated instances, the number of errors, and the resulting error rate. Only sub categories with at least 10 instances are shown to reduce noise from small counts.

Source	Model	Hinted	Mean Acc	Std	Median	Refusal %	Halluc %
PSAT	Gemini-3-Flash	False	14.20	13.39	15.0	17.53	76.10
		True	13.14	14.62	10.0	24.70	68.13
	Gemini-3-Pro	False	0.34	2.05	0.0	95.60	3.20
		True	0.46	2.03	0.0	94.78	3.21
	GPT-5-Mini	False	0.00	0.00	0.0	99.60	0.00
		True	0.00	0.00	0.0	100.00	0.00
NCS	Gemini-3-Flash	False	10.05	12.68	5.0	29.08	50.20
		True	6.54	11.02	0.0	50.20	35.57
	Gemini-3-Pro	False	0.18	2.54	0.0	99.21	0.40
		True	0.04	0.63	0.0	100.00	0.00
	GPT-5-Mini	False	0.65	3.99	0.0	95.65	1.19
		True	0.45	2.01	0.0	100.00	1.19
National Technical Qualification	Gemini-3-Flash	False	15.70	17.33	10.0	19.66	61.54
		True	18.58	16.97	15.0	15.61	76.84
	Gemini-3-Pro	False	0.33	2.31	0.0	93.86	4.44
		True	2.03	6.31	0.0	79.15	16.92
	GPT-5-Mini	False	0.45	4.13	0.0	97.96	0.51
		True	0.25	1.72	0.0	99.83	0.17
Olympiads	Gemini-3-Flash	False	10.48	12.45	5.0	23.12	55.58
		True	10.91	14.73	5.0	37.56	49.48
	Gemini-3-Pro	False	0.13	1.34	0.0	97.92	1.56
		True	0.38	2.46	0.0	97.41	2.33
	GPT-5-Mini	False	0.52	3.12	0.0	94.56	2.85
		True	0.18	1.72	0.0	98.96	1.04

Table 26: **Data contamination analysis via prefix completion.** We provide the first 35% of the question and measure the Exactness Score (0 to 100). Refusal and Hallucination rates are reported in percentages. Results are compared between the No-hint and Hinted settings across four data sources.