Evaluating Multimodal Generative AI with Korean Educational Standards

Sanghee Park[*]
NAVER Cloud AI
parksangheeeee@gmail.com

Geewook Kim[*][†]
NAVER Cloud AI
KAIST AI
gwkim.rsrch@gmail.com

Abstract

This paper presents the Korean National Educational Test Benchmark (KoNET), a new benchmark designed to evaluate Multimodal Generative AI Systems using Korean national educational tests. KoNET comprises four exams: the Korean Elementary General Educational Development Test (KoEGED), Middle (KoMGED), High (KoHGED), and College Scholastic Ability Test (KoCSAT). These exams are renowned for their rigorous standards and diverse questions, facilitating a comprehensive analysis of AI performance across different educational levels. By focusing on Korean, KoNET provides insights into model performance in less-explored languages. We assess a range of models—open-source, open-access, and closed APIs—by examining difficulties, subject diversity, and human error rates. The code and dataset builder will be made fully open-sourced at https://github.com/naver-ai/KoNET.

1 Introduction

The advancement of Large Language Models (LLMs) has spurred the integration of sophisticated generative AI systems into various applications (OpenAI, 2023). Recent developments combining LLMs with computer vision have resulted in powerful Multimodal LLMs (MLLMs) (Liu et al., 2023, 2024b; Laurençon et al., 2024b,a). However, questions remain about the true intelligence of these systems, especially their ability to generalize across novel tasks similar to human cognition.

Current benchmarks predominantly focus on English, overlooking the linguistic diversity worldwide and offering limited insights into low-resource languages like Korean. Moreover, many benchmarks do not compare AI performance to that of humans, making it difficult to precisely measure AI proficiency. Some benchmarks are also less connected to real-world application scenarios, hindering the applicability of MLLMs.

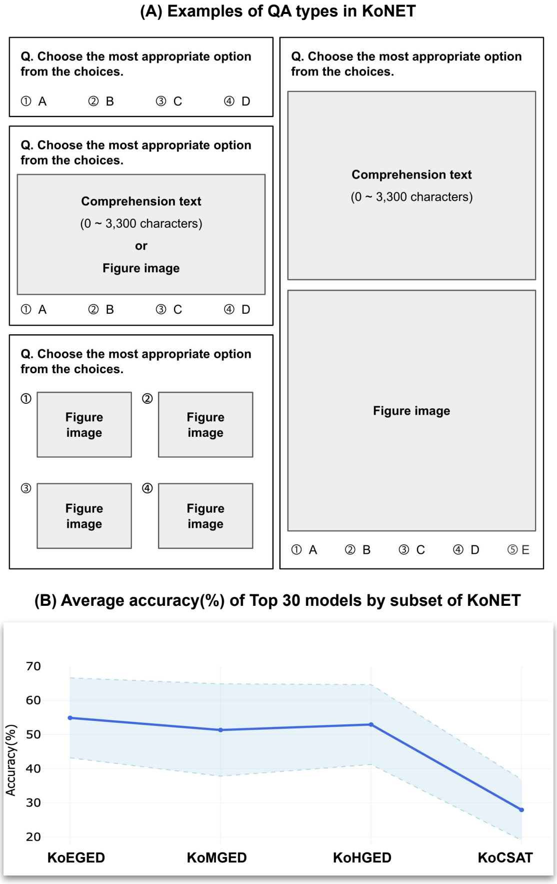


Figure 1: **Examples and Performance Overview of KoNET.** (a) Illustration of mathematics problem examples, highlighting the increased complexity and difficulty as the educational level progresses. (b) Demonstration of how the accuracy of contemporary AI models decreases with more advanced curricula. A detailed analysis is provided in Section 4.

To address these challenges, we introduce KoNET, a benchmark dataset leveraging four key Korean educational tests (refer to Figure 1). Each

[*] Sanghee Park and Geewook Kim contributed equally to this work and share first authorship.
[†] Corresponding author.

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
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Statistic	KoEGED	KoMGED	KoHGED	KoCSAT
Images	400	540	540	897
Questions	400	540	540	897
*K-QA	62 (15.5%)	65 (12.0%)	62 (11.5%)	57 (6.4%)
†TC-QA	123 (30.8%)	249 (46.1%)	284 (52.6%)	388 (43.3%)
‡MC-QA	215 (53.8%)	226 (41.9%)	194 (35.9%)	452 (50.3%)
Subjects	10	11	11	41
Choices	4 (100.0%)	4 (100.0%)	4 (100.0%)	5 (98.8%)
Avg word	29.9	42.7	48.0	113.0
Max word	106	362	410	786
Avg Char	113.0	167.2	193.6	475.9
Max Char	417	1,408	1,678	3,300
#choice	4	4	4	5

Table 1: **Key statistics of the KoNET benchmark.** *K-QA: Knowledge QA, †TC-QA: Text Comprehension QA, and ‡MC-QA: Mutimodal Comprehension QA.

Bench	Lang.	#Q	#I	#choice	*D	†H
AI2D	En	3,088	3,088	= 4 (100.0%)	✗	✗
ScienceQA	En	4,240	2,017	⩽ 5 (100.0%)	✗	✗
MMMU	En	900	1,900	⩽ 9 (94.1%)	✓	✗
Mathvista	En	1,000	1,000	⩽ 8 (53.4%)	✓	✗
KoNET (ours)	Ko	2,377	2,377	⩽ 5 (99.5%)	✓	✓

Table 2: **Comparison of Multiple-Choice QA Public Benchmarks.** *D indicates that difficulty levels are provided for each question, and †H denotes that human error rate data is available for certain items.

exam—KoEGED, KoMGED, KoHGED, and KoCSAT—provides detailed analyses of question difficulty, enabling nuanced evaluation of AI capabilities. Notably, KoCSAT includes data on the percentage of incorrect responses per item among examinees (human error rate), facilitating thorough comparisons of model behaviors with human performance. This benchmark allows for direct comparisons to human performance and underscores essential competencies crucial for AI-driven educational technologies, offering potential real-world applicability in the AI tutoring market.

Our key contributions include:

1. The introduction of KoNET, a comprehensive benchmark for evaluating Multimodal Generative AI Systems via Korean educational tests.

2. A thorough evaluation of various open-source, open-access, and closed API models.

3. Insights through multiple analytical frameworks, examining the relationship between human and model error rates.

2 Related Work

Text Benchmarks. MMLU (Hendrycks et al., 2021) assesses general language proficiency, while GSM8K (Cobbe et al., 2021), CS-Bench (Song et al., 2024), and SciBench (Wang et al., 2024b) focus on math, computer science, and science skills. These offer a focused evaluation of AI capabilities within educational contexts.

Multimodal Benchmarks. SEEDBench (Li et al., 2024) and MMStar (Chen et al., 2024a) provide general multimodal evaluations. Notably, there are educationally focused benchmarks such as ScienceQA (Lu et al., 2022) and MathVista (Lu et al., 2024), which assess AI's ability with scientific and mathematical content. Further, MMMU (Yue et al., 2024a) provides diverse subject evaluations, including Art and Medicine, while AI2D (Kembhavi et al., 2016) examines diagram interpretation in grade school science.

Korean Benchmarks. Korean benchmarks are limited, but efforts like K-MMLU (Son et al., 2024) and Ko-H5 (Park et al., 2024) have emerged. In multimodal contexts, KVQA (Kim et al., 2019) and CVQA (Romero et al., 2024) focus on VQA and cultural understanding. Despite the advances, there is a notable absence of Korean educational benchmarks, particularly in the multimodal domain. No existing frameworks comprehensively evaluate AI's educational performance across various school subjects within a Korean context.

3 Proposed Benchmark: KoNET

To offer a robust evaluation framework that facilitates comprehensive comparisons with human educational levels, we converts questions from Korea's national educational tests into a multimodal VQA format. Table 1 presents key statistics of KoNET, while Table 2 shows its main contributions.

3.1 Education System and Qualification Exams in Korea

Education is core to societal progress in Korea, with a structured system consisting of 6 years in elementary, 3 in middle, 3 in high school, and 4 in university or 2-3 in junior college (Centre, 2020).

The **General Educational Development (GED)** exams assess basic academic knowledge for individuals who have not completed formal schooling, granting qualifications equivalent to traditional graduation upon passing. The **College Scholastic Ability Test (CSAT)**, also known as "Suneung," is instrumental for college admissions and is recognized for its difficulty and ability to distinguish academic excellence.

3.2 Construction of KoNET

KoNET is constructed by parsing publicly available official PDFs from the Korea Institute of Curriculum and Evaluation[1]. The GED tests include all questions from the first and second sessions of 2023, with each exam comprising 20 or 25 multiple-choice questions per subject, with four options provided for each question. The CSAT incorporates questions from various subjects conducted in 2023, with a range of 20 to 45 questions each. While most are multiple-choice, some subjects have subjective questions. For the CSAT, human error rates are available for a selective subset of 327 questions. This subset reflects the challenges and complexities of these questions, as human error rate data is disclosed primarily for items with higher difficulty levels. Each data sample in KoNET is represented by a single image. More details are in Appendix A.

4 Experiment and Analysis

4.1 Setup

To thoroughly test contemporary models, we use 18 open-source LLMs, 20 open-source MLLMs, 4 closed-source LLMs, and 4 closed-source MLLMs, covering a range of sizes and complexities.

Response Generation. We employ the Chain-of-Thought (CoT) (Wei et al., 2022) as some KoNET problems requires complex reasoning. We use the OCR API[2], specialized for Korean, to translate image content for LLM models lacking vision capabilities. MLLMs use OCR as supplementary information. The ablations on CoT prompting and OCR are in Section 4. The CoT prompts used in this study are in Appendix B. In this study, we ensured a consistent evaluation environment for LLMs and MLLMs across multiple benchmarks, including KoNET, MMMU, and MathVista, using a unified prompt structure and input format. Recent multimodal benchmarks like MMMU-Pro (Yue et al., 2024b) and EXAMS-V (Das et al., 2024) embed all necessary information within images, requiring MLLMs to extract and interpret content directly. KoNET follows this approach, incorporating both questions and answer choices into images, eliminating the need for explicit question and option placeholders (Figure 4). LLMs do not receive direct textual inputs but can infer information via OCR-extracted text. Furthermore, KoNET includes problems where answer choices are images rather than text, requiring MLLMs to rely on visual reasoning. This design enables a more realistic assessment of multimodal comprehension and reasoning abilities.

Evaluation. We utilize the LLM-as-a-Judge approach (Zheng et al., 2023) with GPT-4o (OpenAI, 2023) to verify correctness. This method eliminates the need for manually parsing each model output, thereby minimizing potential errors.

4.2 Main Results

Table 3 outlines the main results, comparing KoNET performance with benchmarks like MathVista and ScienceQA. It also details subset performances for KoNET's components—elementary, middle, high school, and college exams.

Key insights include a general performance improvement with larger model sizes. Notably, there's a significant gap between closed-source APIs and open-source models, especially for KoNET, indicating open-source models lack tuning for Korean domains. Closed-source APIs likely excel due to Korea-targeted business strategies.

Models experience increased difficulty with advancing levels in the Korean curriculum, evident in subset performances. Complexity rises significantly at each educational stage, particularly in KoCSAT, highlighting the rigorous nature of these questions aligned with real-world standards.

The EXAONE-3.0-7.8B-Instruct model, a sovereign AI model specifically designed for the Korean language (bilingual in English and Korean), achieved a K-NET score of 45.5, significantly outperforming other models of similar size (7–8B). This suggests that benchmarks centered solely on English may not accurately assess AI performance in non-English or East Asian language environments. For instance, in the KoHGED (high school education exam), a question was based on the classic literary work Yongbieocheonga (Songs of the Dragons Flying to Heaven), a historical text from Korea's Joseon Dynasty published in 1445. This work is part of the standard curriculum in Korean education. Models lacking an understanding of the cultural context struggled to interpret the question and failed to provide the correct answer. In contrast, the EXAONE-3.0-7.8B-Instruct model successfully derived the correct response, demonstrating how linguistic and cultural specificity significantly impacts AI performance. No-

[1] https://www.kice.re.kr

[2] https://www.ncloud.com/product/aiService/ocr

Model	Size (B)	*Previous Benchmarks*				*Proposed KoNET Benchmarks*				KoNET
		Mathvista	ScienceQA	AI2D	MMMU	KoEGED	KoMGED	KoHGED	KoCSAT	
Open Source LLM										
Qwen2-0.5B-Instruct (Yang et al., 2024)	0.5	4.9	29.8	20.2	4.5	17.8	19.6	16.7	12.8	16.0
Qwen2-1.5B-Instruct (Yang et al., 2024)	1.5	2.8	32.6	19.6	6.1	25.8	20.6	22.0	14.3	19.2
gemma-2-2b-it (Team et al., 2024)	2.0	1.0	30.0	24.7	9.8	30.0	30.7	32.4	16.5	25.3
Phi-3-mini-4k-instruct (Abdin et al., 2024)	3.8	5.1	31.4	26.1	14.1	37.0	37.0	37.4	18.1	29.5
Phi-3.5-mini-instruct (Abdin et al., 2024)	3.8	5.5	34.9	26.8	10.9	29.0	28.0	23.5	14.6	21.8
Yi-1.5-6B-Chat (Young et al., 2024)	6.0	5.2	33.8	25.6	14.2	39.2	36.7	36.1	19.7	30.2
Mistral-7B-Instruct-v0.3(Jiang et al., 2023)	7.0	7.6	36.7	34.2	20.5	36.5	29.4	34.4	16.5	26.5
Qwen2-7B-Instruct (Yang et al., 2024)	7.0	6.4	35.4	33.2	23.3	54.0	53.1	50.7	20.3	39.6
EXAONE-3.0-7.8B-Instruct (Research, 2024)	7.8	7.1	39.3	34.1	21.9	64.5	59.1	56.9	24.2	45.5
Meta-Llama-3-8B-Instruct(Dubey et al., 2024)	8.0	6.0	37.3	39.2	22.3	46.5	46.9	43.3	20.5	35.5
Meta-Llama-3.1-8B-Instruct(Meta, 2024)	8.0	5.3	38.2	36.7	19.7	42.5	41.9	40.6	18.4	32.3
Yi-1.5-9B-Chat (Young et al., 2024)	9.0	8.2	37.5	38.6	20.7	47.0	43.7	45.0	22.5	36.0
gemma-2-9b-it (Team et al., 2024)	9.0	6.7	41.7	41.8	20.0	63.0	61.3	59.3	29.8	48.5
Phi-3-medium-4k-instruct (Abdin et al., 2024)	14.0	12.6	48.7	41.6	17.3	34.8	34.8	32.0	17.7	27.4
gemma-2-27b-it (Team et al., 2024)	27.0	18.8	49.6	47.3	24.6	74.5	69.6	68.5	33.9	55.9
Yi-1.5-34B-Chat (Young et al., 2024)	34.0	18.9	61.5	44.2	25.1	64.0	57.4	55.4	25.8	45.4
Meta-Llama-3.1-70B-Instruct(Meta, 2024)	70.0	20.3	67.5	49.5	31.5	63.2	65.6	62.6	31.2	50.8
Qwen2-72B-Instruct (Yang et al., 2024)	72.0	21.7	69.1	49.4	32.3	76.0	74.1	71.9	36.0	58.7
Open Source VLM										
InternVL2-1B (Chen et al., 2024b)	1.0	33.5	59.6	65.2	35.0	0.8	0.4	0.9	0.4	0.6
InternVL2-2B (Chen et al., 2024b)	2.0	35.4	62.0	74.0	35.7	2.2	2.0	3.3	1.7	2.2
Qwen2-VL-2B-Instruct (Wang et al., 2024a)	2.0	42.9	65.4	76.5	40.2	13.2	13.0	12.2	8.4	11.0
paligemma-3b-mix-448 (Beyer* et al., 2024)	3.0	29.1	65.3	69.8	33.4	8.2	8.7	8.7	4.9	7.1
InternVL2-4B (Chen et al., 2024b)	4.0	57.0	71.5	78.7	46.5	1.5	2.0	1.7	0.9	1.4
Phi-3.5-vision-instruct (Abdin et al., 2024)	4.2	44.8	68.6	77.8	39.3	15.0	17.0	13.1	4.6	10.9
Qwen2-VL-7B-Instruct (Wang et al., 2024a)	7.0	53.2	66.7	71.5	59.1	49.5	46.9	42.0	16.9	34.3
llava-1.5-7b-hf (Liu et al., 2024a)	7.0	30.9	67.3	53.0	30.8	3.2	4.6	4.8	3.2	3.9
llava-v1.6-vicuna-7b-hf (Liu et al., 2024b)	7.0	35.2	71.7	53.9	34.0	3.0	2.8	1.9	1.6	2.1
InternVL2-8B (Chen et al., 2024b)	8.0	58.2	61.9	65.9	53.3	12.2	11.7	8.0	4.0	7.9
llama3-llava-next-8b-hf (Liu et al., 2024b)	8.0	37.1	70.5	55.8	35.1	10.2	7.8	7.2	2.6	6.0
llava-1.5-13b-hf (Liu et al., 2024a)	13.0	26.6	49.3	57.6	37.5	11.8	8.1	7.4	4.6	7.1
llava-v1.6-vicuna-13b-hf (Liu et al., 2024b)	13.0	37.0	71.5	60.3	34.9	5.0	5.0	7.2	6.9	6.3
cogvlm2-llama3-chat-19B (Hong et al., 2024)	19.0	40.0	59.3	74.7	43.5	5.8	6.7	4.6	6.1	5.9
InternVL2-26B (Chen et al., 2024b)	26.0	59.5	60.3	84.4	46.6	8.8	6.5	7.2	1.3	5.0
llava-v1.6-34b-hf (Liu et al., 2024b)	34.0	44.6	63.6	83.6	50.7	25.0	0.0	50.0	0.0	15.0
InternVL2-40B (Chen et al., 2024b)	40.0	58.3	70.5	87.7	51.5	49.3	0.0	36.8	11.9	20.8
llava-next-72b-hf (Liu et al., 2024b)	72.0	51.9	79.4	77.1	44.9	49.0	45.0	39.4	10.6	30.7
InternVL2-Llama3-76B (Chen et al., 2024b)	76.0	64.1	81.7	87.0	55.1	10.9	7.3	11.1	4.3	7.5
llava-next-110b-hf (Liu et al., 2024b)	110.0	55.1	85.4	83.1	48.7	19.8	23.0	20.9	12.0	17.6
Closed Source LLM										
gemini-1.5-pro(2024.05)(Google, 2024)	N/A	19.1	68.3	53.9	32.7	80.0	81.7	81.9	44.0	66.4
HyperCLOVA-X(2024.09)(Yoo et al., 2024)	N/A	20.9	83.8	50.7	29.1	82.0	84.6	85.1	51.2	70.9
claude-3-5-sonnet-20240620(Anthropic, 2024)	N/A	27.6	80.0	61.5	54.2	86.5	86.3	86.1	60.5	76.0
gpt-4o-2024-05-13(OpenAI, 2024)	N/A	36.4	84.5	63.4	56.8	82.5	82.0	84.4	52.5	70.8
Closed Source MLLM										
gemini-1.5-pro(2024.05)(Google, 2024)	N/A	52.5	80.6	81.9	58.0	87.0	88.5	86.1	52.4	73.3
HyperCLOVA-X(2024.09)(NAVER Cloud, 2024)	N/A	57.0	**93.3**	79.1	44.8	83.5	88.1	86.1	55.7	74.0
claude-3-5-sonnet-20240620(Anthropic, 2024)	N/A	65.9	88.4	**93.3**	67.4	94.0	93.3	90.7	62.8	80.6
gpt-4o-2024-05-13(OpenAI, 2024)	N/A	**62.5**	89.2	**93.3**	**69.5**	**95.0**	**95.4**	**94.4**	**66.1**	**83.4**

Table 3: **Results on various conventional benchmarks and KoNET.** These are achieved under the condition with CoT prompting and an off-the-shelf OCR API.

tably, open-source models such as EXAONE and Qwen2 have shown strong performance in Korean and East Asian contexts, highlighting the need for greater focus on non-English languages in future research and open-source AI development.

4.3 Further Analyses

Q1: Do MLLMs perform better on KoNET due to their support for multimodal inputs?

Table 3 indicates unexpected results, with MLLMs sometimes lagging behind LLMs on KoNET, contrary to other benchmarks. We analyze model pairs sharing LLM backbones in Table 4. Without the off-the-shelf OCR assistance, closed-source MLLMs demonstrate competitive performance, comparable to LLMs with OCR support. However, many open-source MLLMs do not perform as effectively, revealing a specific challenge with text recognition in the Korean context.

Q2: Can CoT prompting improve performance on KoNET?

As shown in Table 4, CoT generally enhances performance across all models. Notably, this improvement is more pronounced in high-performing closed-source models compared to open-source models. This suggests that while CoT is beneficial, some open-source models are not yet fully optimized for reasoning in the Korean context, making CoT less effective.

Q3: Do AI models have similar error patterns to students?

We compare human error rates on 327 questions with AI error rates. The human error rates in KoCSAT are derived from the Korean College Scholastic Ability Test (KoCSAT), which plays a crucial role in university admissions in South Korea. This exam is a large-scale standardized assessment taken by hundreds of thousands of students each year, who systematically prepare and sit for the test under controlled conditions. In this study, human error rates are calculated based on data from approximately 505K students, using official statistics published by the Korea Institute for Curriculum and Evaluation (KICE[3]). KICE is the official national institution responsible for the development and evaluation of all exams included in KoNET.

To analyze error rates, we explore variability in model responses by assigning different personas (Safdari et al., 2023) and adjusting parameters like temperature. Using gpt-4o-2024-05-13, the strongest of our test models, we create 10 personas,[4] generating 10 responses per persona for a total of 120 responses. For gpt-4o-2024-05-13, gemini-1.5-pro, HyperCLOVA-X, and claude-3-5-sonnet-20240620, we use three personas (‘student,’ ‘teacher,’ and ‘professor’),[5] also generating 10 responses per persona for a total of 120 responses. This setup addresses the challenge of limited high-performing AI models by using personas to expand the response pool, thus enabling comprehensive trend comparisons between AI models and student groups.

Figure 2 indicates a weaker than expected positive correlation. Detailed analysis shows AI models excel in comprehension tasks, likely due to human attention lapses, while humans perform better in memorization tasks, especially in long-tail questions for exams like the CSAT. These outcomes align with expectations and underscore the benchmark’s value by integrating human error data, providing a rich resource for future studies.

Model	Size (B)	Mode	wo OCR		w OCR	
			Direct	CoT	Direct	CoT
Qwen2-1.5B-Instruct	1.5	Text			14.7	19.2
		Vision	9.8	11.2	10.8	11.0
Phi-3.5-mini-instruct	3.8	Text			27.1	21.8
		Vision	21.1	4.4	24.9	10.9
Qwen2-7B-Instruct	7.0	Text			33.1	39.6
		Vision	21.9	33.9	35.7	34.3
Meta-Llama-3.1-70B-Instruct	70.0	Text			53.7	50.8
		Vision	22.1	4.2	45.5	30.7
gemini-1.5-pro	N/A	Text			64.3	66.4
		Vision	32.7	47.8	71.1	73.3
HyperCLOVA-X	N/A	Text			67.2	70.9
		Vision	**69.5**	**75.2**	69.5	74.0
claude-3-5-sonnet-20240620	N/A	Text			70.4	76.0
		Vision	40.2	73.5	71.1	80.6
gpt-4o-2024-05-13	N/A	Text			70.1	70.8
		Vision	66.0	74.9	**74.8**	**83.4**

Table 4: **Comparison on common backbones.** This shows various LLMs with their corresponding MLLMs.

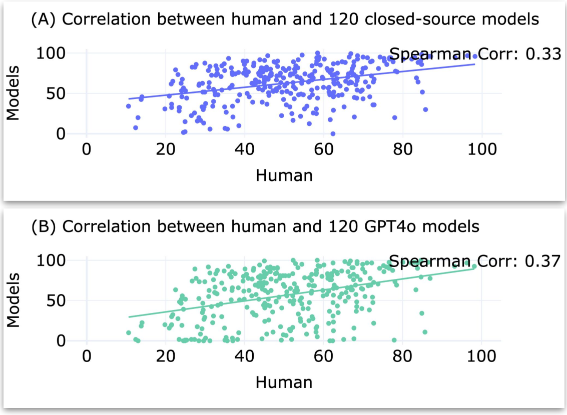


Figure 2: **Correlation analysis of error rates.** The x-axis shows human error rates, and the y-axis displays error rates from closed-source models. Appendix C.3 offers a detailed discussion on the methods used to calculate these error rates.

5 Conclusion

We present KoNET as a benchmark for evaluating multimodal generative AI models using Korean educational tests. Our findings reveal varying performance with multimodal inputs and highlight specific challenges. The disparity between open and closed-source models points to the need for advancements in open-source models within non-English contexts. Our analysis of human error rates offers valuable insights into AI and human performance comparisons. Through KoNET, we aim to encourage research in multimodal and multilingual AI, thereby promoting inclusivity and diversity.

Limitations

While KoNET serves as a valuable resource for assessing the intellectual capabilities of models through Korean educational tests, it does have certain limitations. Similar to many current benchmarks, KoNET primarily adheres to a multiple-choice QA format, which may not fully capture a model’s capacity to articulate problem-solving

[3] https://www.kice.re.kr

[4] Personas include ‘student,’ ‘teacher,’ ‘professor,’ ‘engineer,’ ‘scientist,’ ‘mathematician,’ ‘doctor,’ ‘lawyer,’ ‘master student,’ and ‘PhD student.’

[5] Each persona undergoes 10 repeated experiments.

processes. Although a small proportion of the questions are subjective (see Table 2), these generally involve short-response formats. To address this, future work could focus on evaluating models' reasoning abilities by incorporating rationales behind their answers. This advancement necessitates the development of comprehensive reference answers and a consideration of the increased computational costs involved.

Moreover, as is common with all benchmarks, periodic updates to the test set are necessary to mitigate potential biases and data contamination upon public release. Given that KoNET is based on annually updated national tests, it is inherently suited for regular renewal. We anticipate that our dataset construction methodology, along with the open-source dataset builder, will empower the research community to continuously update KoNET, ensuring its ongoing relevance and utility in advancing AI systems to better meet diverse needs.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anthropic. 2024. Claude 3.5 sonnet. Accessed: 9 Oct. 2024.

Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai*. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Korean Education Centre. 2020. Education in Korea — koreaneducentreinuk.org. http://koreaneducentreinuk.org/en/education-in-korea. [Accessed 14-10-2024].

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan

Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Google. 2024. Gemini 1.5 pro.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *Preprint*, arXiv:2408.16500.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396.

Jin-hwa Kim, Soohyun Lim, Jaesun Park, and Hansu Cho. 2019. Korean Localization of Visual Question Answering for Blind People. In *AI for Social Good workshop at NeurIPS*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Meta. 2024. LLaMA 3.1. Accessed: 23 Jul. 2024.

NAVER Cloud. 2024. HyperCLOVA X Vision. Accessed: 19 Aug. 2024.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Gemini 1.5 pro.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, SeongHwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024. Open Ko-LLM leaderboard: Evaluating large language models in Korean with Ko-h5 benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3220–3234, Bangkok, Thailand. Association for Computational Linguistics.

LG AI Research. 2024. Exaone 3.0 7.8b instruction tuned language model. *arXiv preprint arXiv:2408.03541*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *Preprint*, arXiv:2402.11548.

Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, et al. 2024. Cs-bench: A comprehensive benchmark for large language models towards computer science mastery. *arXiv preprint arXiv:2406.08587*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024b. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on Machine Learning*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Details on the KoNET Construction

KoNET encompasses a wide range of subjects across each exam, as detailed in Table 5. For K-GED (comprising KoEGED, KoMGED, KoHGED), core subjects are included as common components, while each exam features additional unique subjects. The KoCSAT comprises core subjects and optional subjects, with each optional subject further divided into specialized areas. Although students typically select specific subjects for their exams, this study includes questions from all subjects to ensure comprehensive coverage. All images within KoNET are presented in gray-scale, encapsulating the question, answer choices, and comprehension elements within a single image—a format that varies across problems. We adopt the simplest input method to evaluate both LLMs and MLLMs models. Each provided image is structured to contain both the question and all the information necessary to solve it. For text input, no additional text is provided beyond instruction-following prompts and OCR tokens (See Figure 4). This input format also allows us to indirectly assess the MLLMs models' overall understanding of the image and their ability to recognize Korean characters.

KoNET is constructed by parsing publicly available official PDFs from the Korea Institute of Curriculum and Evaluation[6]. We remain mindful of licensing issues, acknowledging the inherent copyright of these questions. However, details regarding specific licensing terms remain elusive; the only guidance available from the Korea Institute of Curriculum and Evaluation indicates permission for non-commercial use. We uphold the copyrights of the original owners with utmost respect. Rather than distributing the data directly, we provide dataset builder code that allows users to convert downloaded official PDFs into benchmark-ready formats. In this paper, we include images that mimic various question types rather than actual problem images. The rendered images in the form of test sheets, based on these mimicked images, are shown in Figure 3. Actual problem images can be generated and reviewed using the provided dataset builder.

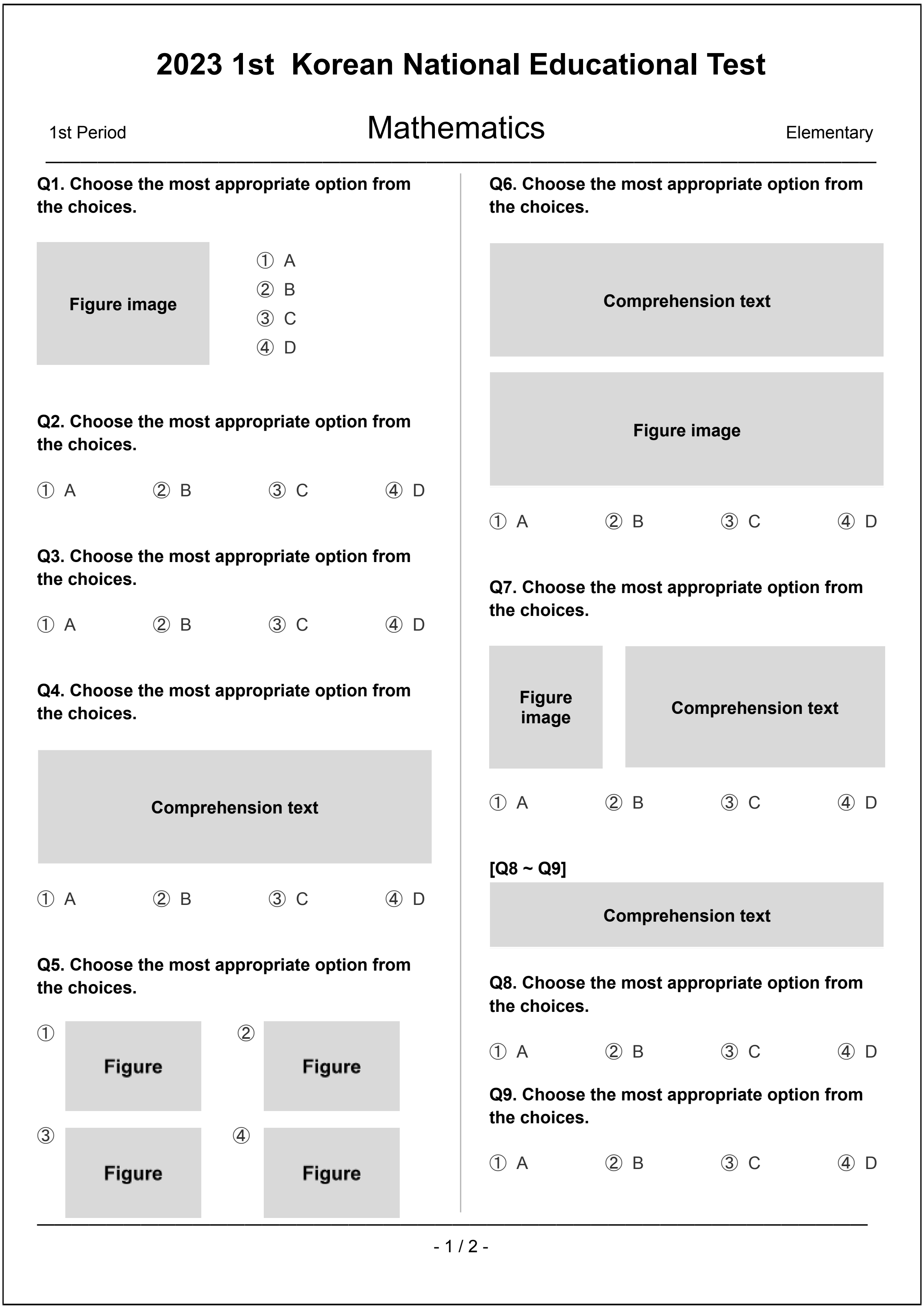
Test	Subjects
KoEGED	Korean, English, Mathematics, Social Studies, Science, Music, Physical Education, Ethics, Art, Practical
KoMGED	Korean, English, Mathematics, Social Studies, Science, Music, Physical Education, Ethics, Art, Information, Technology
KoHGED	Korean, English, Mathematics, Social Studies, Science, Music, Physical Education, Ethics, Art, Technology, Korean History
KoCSAT	Korean (Common), Korean (Speech Writing), Korean (Language and Media), Mathematics (Common), Mathematics (Statistics), Mathematics (Calculus), Mathematics (Geometry), English, Korean History, Social Studies (Every Ethics), Social Studies (Ethical Ideology), Social Studies (Korean Geography), Social Studies (International Geography), Social Studies (East Asia History), Social Studies (International History), Social Studies (Economics), Social Studies(Politics and Law), Social Studies(Social Culture), Science (Physics I), Science (Chemistry I), Science (Bio Science I), Science (Earth Science I), Science (Physics II), Science (Chemistry II), Science (Bio Science II), Science (Earth Science II), Job Studies (Successful Career Life), Job Studies (Agricultural Technology), Job Studies (General Industry), Job Studies (Commercial Economy), Job Studies (Fisheries Shipping Industry), Job Studies (Human Development), Second Language (German), Second Language (French), Second Language (Spanish), Second Language (Chinese), Second Language (Japanese), Second Language (Russian), Second Language (Arabic), Second Language (Vietnamese), Second Language (Chinese characters)

Table 5: **List of subjects categorized under various Korean educational tests.** KoEGED represents subjects for elementary-level general education (10 subjects), KoMGED covers middle-level general education (11 subjects), and KoHGED encompasses high school-level general education (11 subjects). KoCSAT includes the 41 subjects evaluated in the Korean College Scholastic Ability Test, spanning multiple disciplines, including languages, mathematics, sciences, social studies, and job studies.

B Details of the Used Prompts

In this study, we use Korean prompts to generate and assess the response generation capabilities of the models. Two types of prompts are employed: the Direct prompt and the Chain of Thought (CoT) prompt. The Direct prompt involves extracting an-

[6] https://www.kice.re.kr



2023 1st Korean National Educational Test

1st Period

Mathematics

Elementary

Q1. Choose the most appropriate option from the choices.

Figure image

① A
② B
③ C
④ D

Q2. Choose the most appropriate option from the choices.

① A ② B ③ C ④ D

Q3. Choose the most appropriate option from the choices.

① A ② B ③ C ④ D

Q4. Choose the most appropriate option from the choices.

Comprehension text

① A ② B ③ C ④ D

Q5. Choose the most appropriate option from the choices.

① Figure ② Figure
③ Figure ④ Figure

Q6. Choose the most appropriate option from the choices.

Comprehension text

Figure image

① A ② B ③ C ④ D

Q7. Choose the most appropriate option from the choices.

Figure image

Comprehension text

① A ② B ③ C ④ D

[Q8 ~ Q9]

Comprehension text

Q8. Choose the most appropriate option from the choices.

① A ② B ③ C ④ D

Q9. Choose the most appropriate option from the choices.

① A ② B ③ C ④ D

- 1 / 2 -

Figure 3: **Illustrative Representation of the KoNET.** The test includes various types of questions, such as those requiring comprehension of images and queries, reading and understanding of lengthy texts, and simple knowledge-based queries.

swers directly from the provided options for each question. Conversely, the CoT prompt allows the model to reason through the problem to infer the answer. Additionally, a Judge prompt is used within

	Test 1	Test 2	Test 3	Test 4	Test 5
Accuracy	96.9%	98.3%	98.2%	97.4%	98.2%

Table 6: **Agreement Rate Between Human Evaluation and Judge Model.** When using the LLM-as-a-Judge approach, results may vary slightly with each evaluation. To ensure consistency, we conduct evaluations five times to assess whether the LLM-as-a-Judge method aligns closely with answers annotated manually by the authors. When considering the authors' evaluation results as the ground truth, we find that the accuracy is consistently high. This suggests that LLMs can reliably substitute human evaluators with a high degree of confidence.

the CoT framework to evaluate the responses generated by comparing them with the correct answers. While the original prompts are in Korean, English translations are also provided for reference. The format of these prompts is exemplified in Figure 4.

C Additional Analysis

C.1 On the Performance Gap Between LLMs and MLLMs

Figure 5 illustrates the score distribution of LLMs and MLLMs on both conventional benchmarks and KoNET. As shown in our work, the KoNET reveals a distinct distribution pattern compared to traditional benchmarks. Notably, MLLMs underperform relative to LLMs. As analyzed in the paper, we suggest that public LLMs may actually achieve better performance when supported by Korean OCR and many commercially available MLLMs are less effective in processing non-English contexts. This finding provides a novel perspective for model analysis that diverges from traditional benchmarks.

C.2 Comparison of LLM-as-a-Judge with Manual Grading

To see whether LLM-as-a-Judge provide similar user experience or performance to manual grading, we conduct an additional analysis on this. Given the multiple-choice nature of the tests and the potential for varying text responses, we adopt the LLM-as-a-Judge strategy to ensure grading accuracy. Table 6 indicates that this approach closely mirrors manual grading results, demonstrating its reliability and potential as an efficient evaluation method.

C.3 Analysis of Human Error Rates

We employ the error rates from the KoCSAT to assess and compare the performance of models against human performance. Human error rates range from 10.6% to 98.2%, as illustrated in Figure 6.

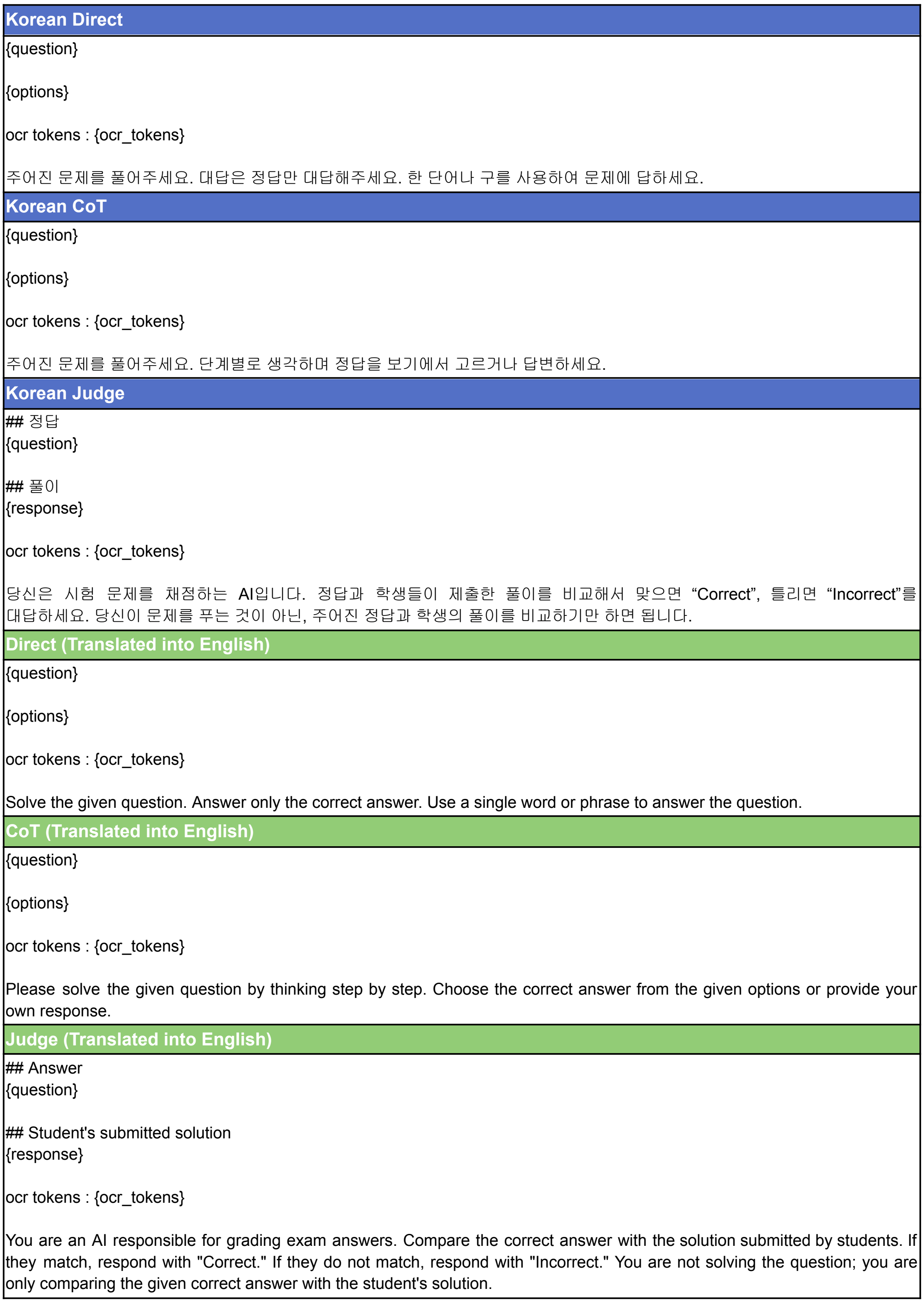
In the first analysis, we calculate model error rates using four closed-source MLLM APIs. For each model, we configure ten personas (i.e., different system messages), set the temperature to 1.0, and generate outputs three times.

In the second analysis, we utilize the GPT-4o model across ten personas, generating twelve distinct responses per persona. We then compute the model error rates and compare them with the human error rates. Figure 7 illustrates the distribution of error rates across subjects, while Figure 8 provides a point-by-point comparison of human and model error rates.

This rigorous analysis enhances our understanding of model performance relative to human benchmarks, offering valuable insights into the strengths and limitations of current MLLMs in processing complex educational content.

C.4 Multilingual Ability Assessment

We assess multilingual capabilities using specific subjects from KoNET. The KoCSAT includes subjects for nine different languages. Traditionally, multilingual capabilities are evaluated by translating English-based benchmarks into other languages or by making indirect comparisons using benchmarks crafted in different linguistic regions. However, the multilingual subjects in KoCSAT consist of independent questions with comparable difficulty levels, enabling a more equitable and valid comparison of multilingual abilities. Figure 9 illustrates the multilingual capabilities across different model types.



Korean Direct

{question}

{options}

ocr tokens : {ocr_tokens}

주어진 문제를 풀어주세요. 대답은 정답만 대답해주세요. 한 단어나 구를 사용하여 문제에 답하세요.

Korean CoT

{question}

{options}

ocr tokens : {ocr_tokens}

주어진 문제를 풀어주세요. 단계별로 생각하며 정답을 보기에서 고르거나 답변하세요.

Korean Judge

정답
{question}

풀이
{response}

ocr tokens : {ocr_tokens}

당신은 시험 문제를 채점하는 AI입니다. 정답과 학생들이 제출한 풀이를 비교해서 맞으면 “Correct”, 틀리면 “Incorrect”를 대답하세요. 당신이 문제를 푸는 것이 아닌, 주어진 정답과 학생의 풀이를 비교하기만 하면 됩니다.

Direct (Translated into English)

{question}

{options}

ocr tokens : {ocr_tokens}

Solve the given question. Answer only the correct answer. Use a single word or phrase to answer the question.

CoT (Translated into English)

{question}

{options}

ocr tokens : {ocr_tokens}

Please solve the given question by thinking step by step. Choose the correct answer from the given options or provide your own response.

Judge (Translated into English)

Answer
{question}

Student's submitted solution
{response}

ocr tokens : {ocr_tokens}

You are an AI responsible for grading exam answers. Compare the correct answer with the solution submitted by students. If they match, respond with "Correct." If they do not match, respond with "Incorrect." You are not solving the question; you are only comparing the given correct answer with the student's solution.

Figure 4: **Examples of prompt formats used in the study.** These include Direct prompts for answer extraction, CoT (Chain-of-Thought) prompts for reasoning-based inference, and Judge prompts for evaluating the accuracy of generated responses.

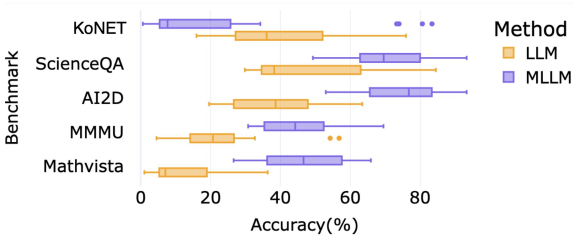


Figure 5: **Performance of LLMs and MLLMs across Previous benchmarks and KoNET.** These present a performance comparison between LLMs and MLLMs across various benchmarks, including KoNET. These illustrate the accuracy distribution for each model type, but KoNET shows a different distribution trend between LLMs and MLLMs compared to other benchmarks.

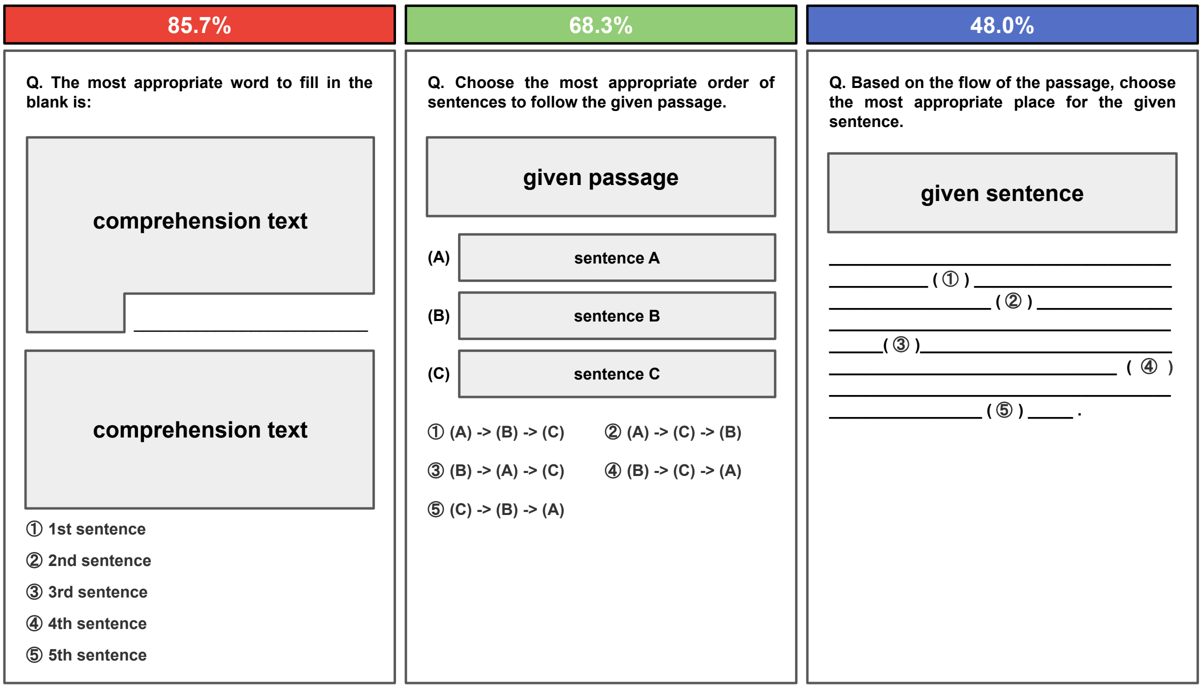


Figure 6: **Examples of human error rate.** These illustrates human error rates across three types of comprehension tasks: sentence selection (left), sentence ordering (middle), and sentence insertion (right). The percentages at the top represent the error rates calculated based on responses from students. Higher error rates indicate more challenging tasks requiring deeper comprehension. Notably, as the complexity of the comprehension text increases, the error rate also rises, suggesting a greater cognitive load in understanding and structuring the given information.

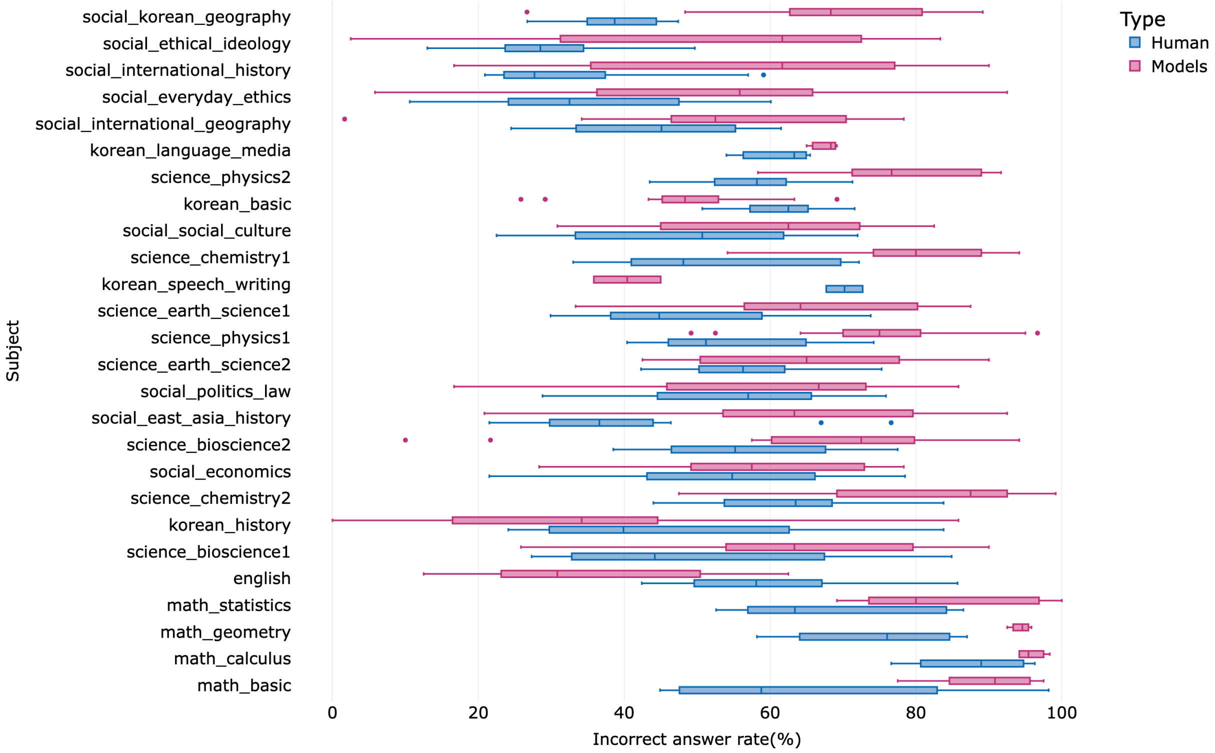


Figure 7: **Distribution of human and models error rate by subjects.** These compares the error rate distributions between humans (blue) and models (pink) across various academic subjects. The x-axis represents the error rate, while the y-axis lists different subjects, covering social sciences, natural sciences, Korean language, history, and mathematics. The varying distributions highlight the differences in performance between humans and models, with some subjects showing a greater disparity.

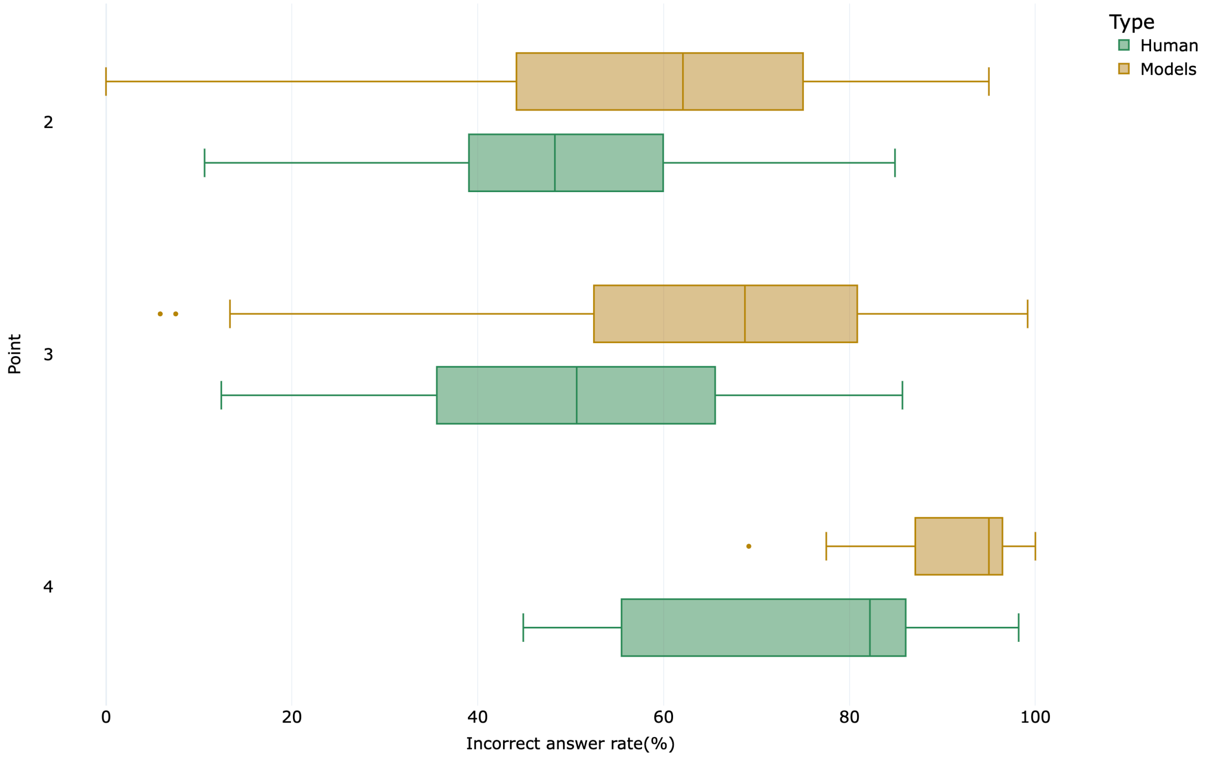


Figure 8: **Distribution of human and models error rate by points.** These presents the error rate distribution of humans (green) and models (brown) based on different point values assigned to questions. The x-axis represents the percentage of incorrect answers, while the y-axis categorizes questions by their point values. Higher-point questions generally require deeper reasoning and comprehension, which is reflected in the increasing error rates for both humans and models.

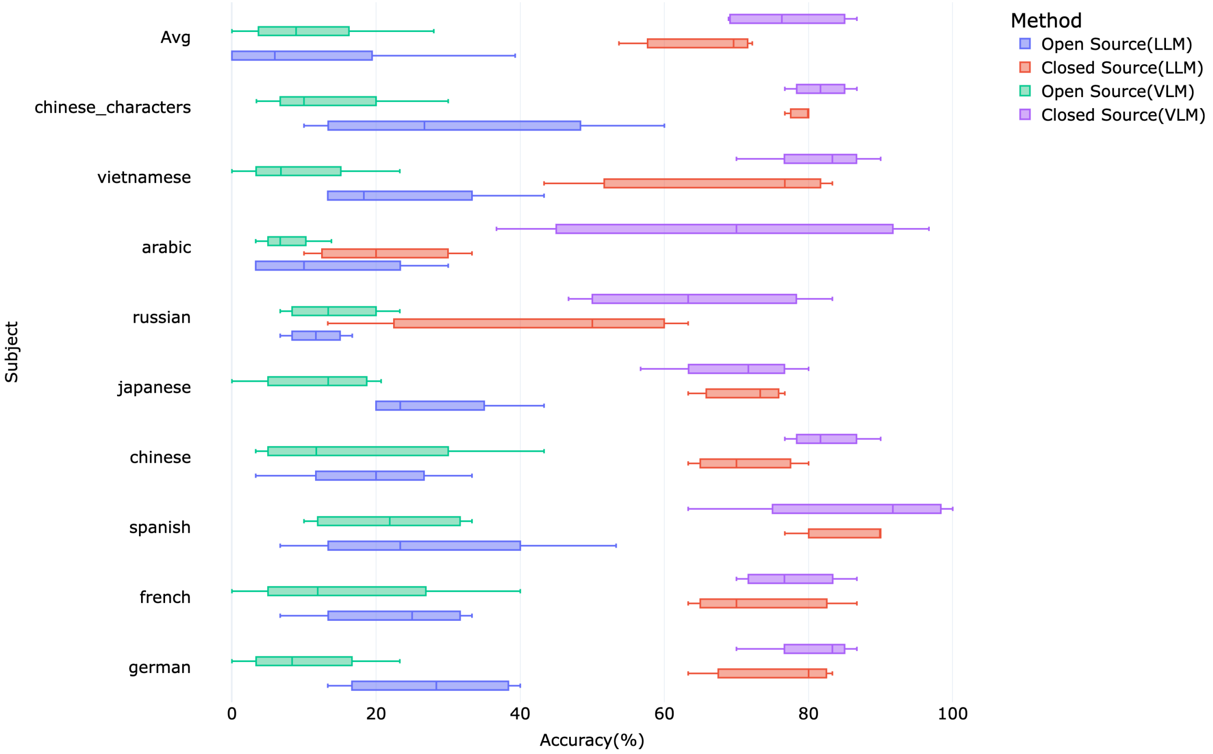


Figure 9: **Performance of multilingual ability.** These illustrations depict the accuracy distribution of various models across multiple languages, highlighting their multilingual capabilities. The x-axis represents accuracy percentages, while the y-axis lists different languages. In general, Open Source models tend to support a narrower range of languages fluently compared to Closed Source models. However, even among Closed Source LLMs, performance tends to decline for certain languages; for instance, Arabic differs from English in writing direction, which can impact model performance.