KMMLU: Measuring Massive Multitask Language Understanding in Korean

Guijin Son[1,2,6], Hanwool Lee[2,3], Sungdong Kim[4,5], Seungone Kim[5], Niklas Muennighoff[7], Taekyoon Choi[4], Cheonbok Park[4], Kang Min Yoo[4], and Stella Biderman[2]

[1]Yonsei University, [2]EleutherAI, [3]NCSOFT AI, [4]NAVER Cloud, [5]KAIST AI, [6]OnelineAI, [7]Contextual AI

Abstract

We propose KMMLU, a Korean benchmark with 35,030 expert-level multiple-choice questions across 45 subjects ranging from humanities to STEM. While prior Korean evaluation tools heavily rely on translated versions of existing English benchmarks, KMMLU is collected from original Korean exams, thereby capturing linguistic and cultural aspects of the Korean language. Recent models struggle to show performance over 60%, significantly below the pass mark of the source exams (80%), highlighting the room for improvement. Notably, one-fifth of the questions in KMMLU require knowledge of Korean culture for accurate resolution. KMMLU thus provides a more accurate reflection of human preferences compared to translated versions of MMLU and offers deeper insights into LLMs' shortcomings in Korean knowledge. The dataset and codes are made publicly available for future research.

1 Introduction

Recent works often leverage translated versions of MMLU (Hendrycks et al., 2020) to evaluate the multilingual capabilities of large language models (LLMs) (OpenAI, 2023; Qwen, 2024; Chen et al., 2023; Zhao et al., 2024). However, naively translating English benchmarks into a language of interest faces critical limitations. First, machine translation can lead to a compromised dataset with issues like unnatural language, typos, and grammatical mistakes (Xia et al., 2019; Riley et al., 2023; Yao et al., 2023). Second, MMLU, designed primarily for English speakers, includes content that assumes knowledge of the American legal system or requires familiarity with English slang and culture (Lee et al., 2023; Jin et al., 2023; Son et al., 2023; Li et al., 2023a; ZaloAI-JAIST, 2023). Thus, while translated versions hint at multilingual proficiency, they fail to capture the cultural aspects that native speakers might consider to be crucial.

To address this issue for the Korean NLP community, we introduce KMMLU, a comprehensive benchmark consisting of 35,030 questions spanning 45 subjects. Unique to KMMLU is its sourcing: *all* questions are derived from Korean exams, ensuring authentic Korean language without any translated material. Additionally, our questions are *localized* to Korea: they reflect the cultural attitudes of Koreans, rather than Westerners (see Figure 1). Our comparative analysis, depicted in Figure 3, shows that KMMLU surpasses previously translated benchmarks by offering questions that are linguistically natural and steeped in the Korean cultural context, providing a deeper insight into LLM performance in Korean (Section 6). To address the shortcomings of Korean benchmarking, which relies heavily on translated (Park et al., 2024; Ham et al., 2020; Jin et al., 2023) or private datasets (Park et al., 2024, 2021; Lee et al., 2024), we make KMMLU and its corresponding evaluation codes publicly available [1].

We evaluate 24 different LLMs across 4 categories: (1) Multilingual Pretrained Models (Touvron et al., 2023; Young et al., 2024; Bai et al., 2023); (2) Multilingual Chat Models; (3) Korean Pretrained Models (Ko et al., 2023); and (4) Proprietary Models including those serviced in Korea (OpenAI, 2023; Team et al., 2023; Yoo et al., 2024). Our results show significant room for improvement, with GPT-4 scoring the highest at 59.95%. Surprisingly, we see little evidence of a "curse of multilinguality" (Conneau et al., 2019; Pfeiffer et al., 2022) discussed in previous work comparing BLOOM (Workshop et al., 2022) to monolingual English models (Biderman et al., 2023; Peng et al., 2023).

We conduct a detailed analysis to better understand how LLMs utilize Korean knowledge in

[1] https://huggingface.co/datasets/HAERAE-HUB/KMMLU

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

	Required Type of Korean Knowledge			
Category	**Cultural**	**Regional**	**Legal**	**Other**
STEM	Civil-Engineering	Ecology	Civil-Engineering	Biology
	What is not considered a major problem in urban areas of our country?	What does not belong to the ecology of the Korean Peninsula?	According to regulations, what is the minimum distance required between the outer wall of an apartment building and the boundary of roads?	What is not included in the search results when searching for microbial strains on the website of the Korean Collection for Type Cultures (KCTC)?
	우리나라 대도시지역의 주요 문제라고 볼 수 없는 것은?	한반도의 생태축에 속하지 않는 것은?	주택건설기준 등에 관한 규정상 도로 및 주차장의 경계선으로부터 공동주택의 외벽까지는 최소 얼마 이상을 띄워야 하는가?	국내 대표적인 미생물 균주 분양 기관인 생물자원센터(KCTC) 홈페이지에서 균주를 검색할 때 나타나는 검색 결과에 포함되지 않는 것은?
Applied Science	Geomatics	Maritime-Engineering	Energy-Management	Gas-Technology-and-Engineering
	During which period in the history of our country's cadastral system was the land register called "양전도장" (Yangjeon Dojang)?	What phenomenon would occur if the southwest wind blows for a long time on the east coast of our country?	What is the maximum area limit for constructing a solar power plant in a "management area" with only a "development activity permit"?	What was the main cause of the Daegu city gas explosion, one of the major urban gas accidents in South Korea?
	우리나라 지적제도사(地籍制度史) 중 토지대장을 양전도장(量田都帳)이라 명칭하였던 시대는?	우리나라 동해안에서 남서풍이 오래 불면 어떤 현상이 일어나겠는가?	'개발행위허가' 만으로 태양광 발전소를 건설할 수 있는 '관리지역'의 면적제한 기준은 최대 몇 m2 미만인가?	국내에서 발생한 대형 도시가스 사고 중 대구 도시가스 폭발사고의 주원인은?
HUMSS	Korean-History	-	Accounting	Management
	Which of the following descriptions is not correct about "대한국" (a nation that existed in the Korean Peninsula) ?	-	Under the Korean International Financial Reporting Standards (K-IFRS), which is not classified as a financial asset?	Which of the following is wrong regarding the recent changes in the retail management environment in our country?
	'대한국'에 대한 설명으로 옳지 않은 것은?	-	한국채택국제회계기준(K-IFRS)하에서 금융자산으로 분류되지 않는 것은?	최근 우리나라에서 찾아볼 수 있는 소매 경영 환경의 변화로 가장 옳지 않은 것은?
Other	Food-Processing	Agricultural-Sciences	Agricultural-Sciences	Health
	Which of the following is not a method for brewing traditional Korean alcoholic beverages such as Yakju or Takju?	Which of the following is incorrect for why the production of the F1 breed of cabbage is concentrated along the southern coast?	What is the correct registration procedure when applying for listing in the National Variety List?	Which of the following descriptions is correct regarding the items in the Korean Nurses' Code of Ethics?
	전통주인 약주나 탁주를 제조하는 제국방법이 아닌 것은?	우리나라에서 배추의 F1품종의 종자생산이 남해안과 그 인근 도서 지방에 집중되어 있는 이유를 설명한 것 중 옳지 않은 것은?	국가품종목록 등재신청시 등재 절차로 옳은 것은?	한국간호사 윤리강령의 항목에 대한 설명으로 옳은 것은?

Figure 1: Examples of questions from KMMLU categorized by the type of Korean knowledge required. English translations are added for broader accessibility.

question-answering. Initially, we observe that, despite GPT-4's overall excellence, it displays gaps in areas demanding *localized knowledge*, demonstrating the importance of localizing benchmarks. For example, in Korean History, GPT-4 (OpenAI, 2023) achieves a 35% success rate compared to HYPERCLOVA X (Yoo et al., 2024), a Korean-specific LLM, which scores 44%. Notably, HYPERCLOVA X is unique in its consistent improvement with the use of Chain-of-Thought (CoT) prompting, indicating the challenge non-Korean LLMs face in producing accurate Korean explanations.

2 Related Work

2.1 Benchmarks for Large Language Models

Benchmarks are essential for accurately understanding and tracking the evolving capabilities of LLMs. Traditionally, benchmarks focused on linguistic tasks (Rajpurkar et al., 2016; Wang et al., 2019b,a), but with the recent surge of more capable LLMs, such approaches have become obsolete. To address this gap, new benchmarks have emerged, focusing on higher-level abilities such as commonsense reasoning (Clark et al., 2018; Sakaguchi et al., 2021; Zellers et al., 2019), mathematical reasoning (Hendrycks et al., 2021; Cobbe et al., 2021), code generation (Chen et al., 2021; Li et al., 2023b), and multi-turn conversations (Zheng et al., 2023). Notably, some efforts have concentrated on evaluating the capabilities via expansive datasets covering a wide range of knowledge-based topics (Hendrycks et al., 2020; Srivastava et al., 2022; Sawada et al., 2023). Most famously, MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2020) spans 57 subjects, evaluating LLMs across various disciplines. While many of these efforts have primarily focused on the English language, there has been progress in adapting and creating similar benchmarks for other languages (Li et al., 2023a; Huang et al., 2023; Zeng, 2023; Ghahroodi et al., 2024; Koto et al., 2024, 2023; Tam et al., 2024). Although KMMLU may be similar to the mentioned works in its motivation, it is the first of its kind in Korea, providing

Korean Benchmarks	Translated	Public?	# Category	# Instances	Type
KorNLI & KorSTS	O	O	2	6,389	Understanding
KLUE	X	X	8	38,500	Understanding
Ko-H5	O	X	5	25,700	Reasoning
KoBBQ	△	O	268	76,048	Bias
HAE-RAE Bench.	X	O	6	1,538	Cultural Knowledge
CLiCK	X	O	11	1,995	Cultural Knowledge
KorNAT	X	△	16	10,000	Cultural Alignment
KMMLU (Ours)	X	O	45	35,030	Expert Know. & Reasoning

Table 1: Overview of Korean benchmarks. The "# Instances" denotes the size of the test set. The KLUE dataset has not released its test set, thus we do not consider it public. The complete release of the KorNAT dataset is scheduled for the future and is currently unavailable as indicated by △. Parts of KoBBQ are translated.

a valuable foundation for Korean NLP research.

2.2 Korean Benchmarks

Prior benchmarks for the Korean language specialized on natural language understanding tasks, including natural language inference (Ham et al., 2020), machine reading comprehension (Lim et al., 2019), and hate speech detection (Moon et al., 2020). The Korean Language Understanding Evaluation (KLUE) benchmark (Park et al., 2021), analogous to the General Language Understanding Evaluation (GLUE) (Wang et al., 2019b), compiled eight downstream tasks aimed at gauging the comprehension of the Korean language. Nonetheless, these benchmarks exhibited limitations in evaluating reasoning abilities, rendering them insufficient for assessing LLMs. The Ko-H5 (Park et al., 2024) aims to overcome these shortcomings by offering an extensive array of reasoning benchmarks, such as HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), and TruthfulQA (Lin et al., 2021). Nevertheless, they also rely on machine/human translations that are susceptible to errors. Moreover, their datasets are private and they do not provide evaluation for models larger than 14 billion parameters, thus limiting its transparency and usefulness.

Recent benchmarks developed for Korean have shifted their focus toward preserving the linguistic and cultural nuances, going beyond mere translation of original English benchmarks. In this context, Jin et al. (2023) developed the Korean Bias Benchmark for Question Answering, which is derived from the original BBQ dataset (Parrish et al., 2022). This benchmark is specifically adjusted to align with Korean cultural contexts. Additionally, efforts have been made to create authentic Korean datasets from scratch to capture subtle cultural nuances. For instance, the HAE-RAE Benchmark (Son et al., 2023) and CLIcK (Kim et al., 2024) assess cultural and linguistic understanding specific to Korea. Furthermore, KorNAT (Lee et al., 2024) was introduced to specifically evaluate alignment with national characteristics of South Korea, focusing on two key aspects: social values and common knowledge.

The KMMLU benchmark extends beyond previous efforts as the first extensive knowledge benchmark covering a broad and deep array of fields, including Biology, Chemistry, Criminal Law, Taxation, Electrical Engineering, Aviation Engineering, and others. Additionally, it is unique in that the dataset, prompts, and evaluation codes are publicly available, ensuring transparent and reproducible assessment, Importantly, it is built entirely from Korean exams, providing a representation of Korean culture and linguistic nuances.

3 KMMLU

3.1 Task Overview

KMMLU is a collection of 35,030 multiple-choice questions spanning 45 categories, including HUMSS (Humanities and Social Science), STEM (science, technology, engineering, and mathematics), Applied Science, and other professional-level knowledge. Within STEM, the focus is on topics emphasizing scientific principles, from the natural and physical sciences to technological and engineering disciplines. Meanwhile, Applied Science encompasses industry-specific subjects such as Aviation Engineering and Maintenance, Gas Technology and Engineering, and Nondestructive Testing. HUMSS covers an extensive range of subjects, including history and psychology, offering in-depth insights into the diverse facets of human society and culture. The remaining subjects that do not fit into any of the three categories are put into Other.

We predominantly source the questions from Ko-

rean License Tests, notably, some of the license tests KMMLU draws from exams that requires at least 9 years of industry experience. In addition, KMMLU includes questions that require an understanding of cultural, regional, and legal knowledge to solve, as shown in Figure 1. For further details, see Table 2.

Category	# Questions
Prerequisites	
None	59,909
1 Prerequisite Test	12,316
2 Prerequisite Tests	776
2+ Years of Experience	65,135
4+ Years of Experience	98,678
9+ Years of Experience	6,963
Question Type	
Positive	207,030
Negation	36,747
Split	
Train	208,522
Validation	225
Test	35,030
Total	243,777

Table 2: Overview of questions in KMMLU: This table summarizes questions by number of prerequisites for human examinees, whether the question contains negation, and train/validation/test splits.

3.2 Dataset Creation

Our dataset is a compilation of questions from 533 diverse sources, spanning the Public Service Aptitude Test (PSAT), Korean License Tests, and the College Scholastic Ability Test (CSAT). This collection includes a broad academic spectrum, from high school to professional levels.

Initially, we collected 371,002 questions using automatic crawling. We then implement heuristic filters to erase duplicated samples or parsing errors, including stopwords, regex patterns, and model-based classifiers. Additionally, the format is standardized by excluding questions with fewer than four options and adjusting those with more than four. This filtering reduces the dataset by 34% to 243,777 questions. The significant reduction in dataset size stems from two main factors: firstly, we prioritize quality over quantity, employing broad filters to eliminate any questionable content, even at the cost of removing some valid samples. Secondly, we observe a high degree of similarity among questions, especially those that are seasonally repeated, resulting in extensive deduplication.

We gather human accuracy data from actual test-takers where available. Approximately 90% of our dataset's exams include human performance data, with an average accuracy of 62.6%. Most of the license exams in the dataset require an 80% score to pass. For the PSAT, the average passing score of the last 5 years has been 83.7%. Thus, achieving over 80% on KMMLU can be considered the equivalent of the minimum performance of a human expert, while the best experts are likely to score close to 100%. The dataset is structured into three components: a training set, a few-shot development set, and a test set. The few-shot development set features five questions per subject to support in-context learning (Brown et al., 2020). The training set includes 208,522 questions, suitable for both hyperparameter tuning and model training. The test set is a collection of questions with the lowest human accuracies, each subject consisting of 100 instances at a minimum and 35,050 questions in total. However, it is important to note that comparing human accuracy directly may not be desirable due to variations in test origins and test-taker populations, with some groups being more professional than others. While collecting the human data ourselves could potentially address this issue, we could not do so due to budget constraints.

Before finalization, we released the dataset to the public for six months. During this period, we received five issues reported by the community, and 741 instances were modified accordingly. Additionally, the 35,030 questions in the test set underwent manual review to remove copyrighted materials. We replaced 147 instances, including copyrighted materials. No additional errors were identified during this process. Finally, we conduct an analysis based on Xu et al. (2024)'s method to look for potential data leakages. We observe both open and proprietary LLMs fail to recall the KMMLU benchmark implying low likelihood of benchmark contamination. For details, see Section C.

3.3 CoT Exemplar Creation

Chung et al. (2022) devise 5-shot of exemplars to test CoT reasoning over MMLU (Hendrycks et al., 2020) [2]. Similarly, we create 5-shot of CoT exemplars for each subject to test models' reasoning capabilities on our benchmark. However, writing an accurate rationale for expert-level tests with various ranges is a difficult problem. Although the ideal solution might be to invite experts for each test, we decide to leverage assistance from various LLMs,

[2] github.com/jasonwei20/flan-2

considering resource constraints. Specifically, we employ two LLMs, GPT-4 and HyperCLOVA X, with diverse prompt techniques, zero-shot CoT (Kojima et al., 2022) and browsing-augmented CoT [3].

First, we elicit rationale and corresponding answers from the LLMs using both prompt techniques. Besides, we utilize a majority voting method, self-consistency (Wang et al., 2022), over ten reasoning paths obtained by oversampling. As a result, this step produces 4×10 rationales for each input, i.e., $4 = 2$ LLMs and 2 prompt types. Then, we choose the top-4 rationales ordering by longer and less repetitive output. Finally, authors manually select the most appropriate rationale among the top-4 and revise it with thorough inspections if necessary. For quality control, we ensure two workers for each question. We find about 87% of agreement between two workers at the first iteration. We iteratively validate the remaining conflicted examples. In total, we create $45 \times 5 = 225$ exemplars for the CoT inference within our benchmark. Please see Appendix H for more details.

3.4 KMMLU-HARD

KMMLU comprises 35,030 questions, outnumbering its predecessors, MMLU (Hendrycks et al., 2020) and CMMLU (Li et al., 2023a). Thus, in addition to KMMLU, we create KMMLU-HARD for more targeted and efficient evaluation. The KMMLU-HARD subset includes 4,104 questions that at least one of the following models—GPT-3.5 TURBO, GEMINI PRO, HYPERCLOVA X, and GPT-4—fails to answer correctly. These questions are equally distributed across all categories, each containing 23 to 100 questions.

4 Experimental Setup

4.1 Evaluation Methodology

In our evaluations of LLMs on KMMLU, we employ two distinct settings for a comprehensive comparison. First, the Direct method prompts the model to generate the most plausible option via greedy decoding. In this process, each model generates a response from its entire vocabulary, which makes $\frac{1}{vocab\_size}$ its random baseline. Second, CoT allows the model to generate text freely and leverages RegEx to parse the results. By generating a sequence of reasoning before the final answer, CoT has succeeded in aiding LLMs to solve reasoning-heavy tasks. Models are set to use greedy decoding for the CoT generation. All evaluations in this paper, regardless of the method, are done in a few-shot setting with five exemplars. Due to hardware constraints, we run our experiments with open models using 8-bit quantization.

[3] It is similar to ReAct prompting (Yao et al., 2022).

4.2 Models

In our study, to provide a comprehensive overview of existing LLMs in answering expert-level Korean questions, we evaluate 24 models varying in size, language, and training phase.

The 24 models include:

1. Multilingual Pretrained Models: LLAMA-2 (7B, 13B, 70B) (Touvron et al., 2023), QWEN (7B, 14B, 72B) (Bai et al., 2023), and Yi (6B, 34B) (Young et al., 2024);
2. Multilingual Chat Models: Chat versions of LLAMA-2, QWEN, and YI;
3. Korean Pretrained Models: POLYGLOT-KO (1.3B, 3.8B, 5.8B, 12.8B) (Ko et al., 2023);
4. Proprietary Models: GPT-3.5/4 (OpenAI, 2023)[4], GEMINI PRO (Team et al., 2023) and HYPERCLOVA X (Yoo et al., 2024)[5].

The inclusion of English & Chinese bilingual models aims to explore potential spillover effects, given the historical influence of Chinese Hanja on the Korean language. Further details on the models are provided in Appendix E and Table 18. Evaluation results for newer models are updated in Tables 20 and 21. It should be noted that the results in no way change the conclusions of our paper.

5 Evaluation Results

Pretraining Compute We compare the performance of 27 LLMs using the Direct method in Table 3. We observe a clear trend across pretrained and finetuned models, where those with a larger computing budget exhibit superior performance.[6] This scaling behavior indicates that increased computing resources - reflected in the number of parameters and the size of the training corpus - enhance a model's capacity to handle complex language tasks more accurately. Notably, despite being trained exclusively in Korean, POLYGLOT-KO-12.8B's

[4] We use the 0613 version for both GPT models.

[5] We use the HCX-L version.

[6] Unlike other models studied in this paper, the larger POLYGLOT-KO models were trained for *fewer* tokens than the smaller ones, explaining the non-monotone performance.

Model	STEM	Applied Science	HUMSS	Other	Average
Multilingual Pretrained Models					
LLAMA-2-7B	24.68	25.90	25.06	24.30	25.00
LLAMA-2-13B	33.81	33.86	26.26	30.86	31.26
LLAMA-2-70B	41.16	38.82	41.20	40.06	40.28
YI-6B	35.47	34.23	33.46	35.70	34.70
YI-34B	44.31	40.59	47.03	43.96	43.90
QWEN-7B	22.74	23.83	9.44	17.59	18.52
QWEN-14B	36.68	35.85	21.44	29.26	30.92
QWEN-72B	50.69	47.75	54.39	50.77	50.83
Multilingual Chat Models					
LLAMA-2-7B-CHAT	28.60	29.03	26.01	27.10	27.71
LLAMA-2-13B-CHAT	30.36	29.09	26.40	29.05	28.73
LLAMA-2-70B-CHAT	35.98	34.36	32.19	35.35	34.47
YI-6B-CHAT	35.58	34.55	34.39	35.95	35.11
YI-34B-CHAT	41.83	38.05	46.94	42.05	42.13
QWEN-7B-CHAT	20.26	22.16	8.67	15.70	16.82
QWEN-14B-CHAT	32.78	33.94	19.31	26.75	28.33
QWEN-72B-CHAT	47.57	46.26	49.05	46.33	47.28
Korean Pretrained Models					
POLYGLOT-KO-1.3B	28.77	28.02	26.99	28.11	27.97
POLYGLOT-KO-3.8B	29.68	31.07	26.59	29.54	29.26
POLYGLOT-KO-5.8B	29.18	30.17	26.73	29.12	28.83
POLYGLOT-KO-12.8B	29.27	30.08	27.08	30.55	29.26
Proprietary Models					
GPT-3.5-TURBO	44.64	42.11	40.54	42.61	42.47
GEMINI-PRO	51.30	49.06	49.87	50.61	50.18
HYPERCLOVA X	50.82	48.71	59.71	54.39	53.40
GPT-4	**59.95**	**57.69**	**63.69**	**58.65**	**59.95**

Table 3: Average accuracy(%) calculated using the Direct method in a 5-shot setting across the entire test set. The highest-scoring model across the entire table is highlighted in **bold**, and the best model within each category is underlined. Random guessing has an accuracy of 25% on all subjects. Please see Tables 22-26 for detailed results.

performance only marginally exceeds the random baseline of 25%, is on par with that of the English-centric LLAMA-2-13B, and lags behind YI and QWEN models of similar size. This emphasizes the importance of long training runs in achieving high performance: while POLYGLOT-KO-12.8B is approximately compute-optimally trained (Hoffmann et al., 2022), the order of magnitude increase in the training data size brings substantial increases in the performance of these non-optimally trained models. This disparity in training resources is further illustrated in Figure 2, where POLYGLOT-KO's significantly lower training budget compared to its counterparts is evident.

Fine-Tuning In Table 3, we also observe that fine-tuning Pretrained Models do not necessarily lead to better performance. In our experiments, models often exhibit minor performance differences between their base and chat versions. This aligns with past studies that suggested fine-tuning methods such as supervised fine-tuning, direct preference optimization, or reinforcement learning to have minor improvements in the knowledge of language models (Bi et al., 2024). Interestingly, QWEN-72B and LLAMA-2-70B experience -3.55% and -5.81% of performance drop respectively. We suspect that the ability to solve Korean questions in pretrained models of different languages originally stems from their failure to filter out Korean text from their pretraining corpora perfectly. However, datasets used during the post-training process are often curated with greater precision, possibly excluding all non-target languages. Therefore, such might harm the Korean language proficiency of such models.

Multilinguality at Scale The "curse of multilinguality" (Conneau et al., 2019; Pfeiffer et al., 2022) refers to the apparent decrease in model capabilities

Model	STEM		Applied Science		HUMSS		Other		Total	
	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT
QWEN-72B-CHAT	24.36	19.00	24.25	18.67	18.52	16.50	23.09	18.38	22.59	18.18
HYPERCLOVA X	14.36	28.00	14.58	24.83	20.62	**30.21**	18.90	**25.59**	17.06	**27.11**
GPT-3.5-TURBO	22.36	23.27	21.00	23.67	19.74	15.35	21.30	20.25	21.10	20.70
GPT-4-TURBO	**28.64**	**30.91**	**28.25**	**34.84**	**33.37**	19.68	**30.55**	20.10	**30.52**	25.28

Table 4: 5-shot accuracy on KMMLU-Hard subset (Section 3.4) according to prompting method, Direct and CoT (Wei et al., 2022). Please see Table 27 for detailed results.

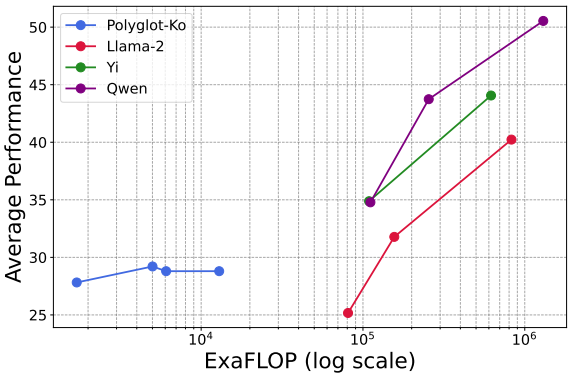

Figure 2: Average performance of POLYGLOT-KO, LLAMA-2, YI, and QWEN models. ExaFLOP on the xaxis represents the scale of computational operations, where each unit corresponds to 10^{18} floatingpoint operations (FLOPs). The total FLOPs are estimated as $6 \times \#param \times \#tokens$ (Kaplan et al., 2020).

when models are trained on multilingual corpora. While the curse can be severe for small models, it has been observed for masked language models that it weakens with scale (Goyal et al., 2021; Pfeiffer et al., 2022). Empirically, this seems not to be the case for the decoder-only BLOOM model (Workshop et al., 2022) as several papers have found that monolingual English models substantially outperform BLOOM on English tasks (Biderman et al., 2023; Peng et al., 2023). In contrast, we find evidence of positive transfer between languages, with large multilingual models like LLAMA-2, YI, and QWEN substantially outperforming the monolingual POLYGLOT-KO. Though the multilingual models are trained on an order of magnitude more tokens than POLYGLOT-KO, they encounter much less Korean text during their pretraining phases. For instance, LLAMA-2 is trained on 2 trillion tokens, with only 0.06% in Korean, amounting to 1.2 billion tokens. YI employs a language filter to exclude languages other than Chinese and English, and QWEN mentions that a significant portion of its data is in English and Chinese. In comparison, POLYGLOT-KO models are trained on 167 billion to 219 billion tokens depending on the model size. Our results show that scaled decoder-only models acquire capabilities in languages they are severely undertrained in, a finding that aligns with prior work (Muennighoff et al., 2023).

Chain-of-Though Prompting We employ a few-shot CoT prompting (Wei et al., 2022), leveraging 5-shot exemplars (Section 3.3) to examine whether advanced prompting method could improve performance. Since the CoT prompting requires much longer sequence generation than the Direct method, we compare four LLMs based on the KMMLU-Hard subset, considering resource constraints [7]. In Table 4, we find that only HYPERCLOVA X reliably improves the performances across categories with the CoT prompting, while other LLMs often show degradation with the CoT. In particular, GPT-3.5-TURBO and GPT-4-TURBO show better performances with CoT on STEM and Applied Science, but drastic performance drops on HUMSS. We presume the Korean-specific context in the HUMSS category is relatively hard to generalize by learning other languages, resulting in unfaithful explanations (Turpin et al., 2023).

6 Are Translated Benchmarks Enough?

The simplest approach to scale multilingual benchmarks involves translation. Efforts to translate MMLU have been implemented repeatedly (Park et al., 2023; Lai et al., 2023), some including professional human translators (OpenAI, 2024; Institute, 2024), offering clear benefits by enabling parallel performance comparisons and insights. However, KMMLU, sourced directly from texts originally written in Korean, provides irreplaceable advantages and evaluations. This section compares KMMLU with translated benchmarks (Section 6.1) and further analyzes performances related to Korean contexts. Further details are in Appendix B.5.

[7] We utilize GPT-4-Turbo (gpt-4-0125-preview) instead of GPT-4 for the same reason.

6.1 Analysis of Korea-Specific Instances

To provide a deeper insight into how KMMLU differs from past efforts that translate MMLU (Park et al., 2023; Chen et al., 2023), we compare the two on two fronts: the naturalness of phrasing and the necessity for specialized Korean knowledge. For the analysis, we randomly selected ten questions from each category within both datasets, resulting in 570 questions from a Korean-translated MMLU and 450 questions from KMMLU—two authors evaluated each question.

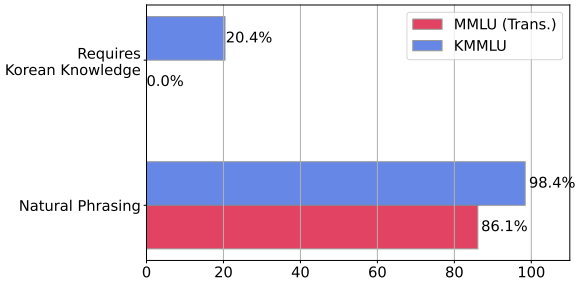

Figure 3: Comparison of MMLU (translated via GPT-4) and KMMLU (ours).

Figure 3 reveals a difference in how the two subsets appear to native Korean speakers. KMMLU questions are significantly more natural and culturally relevant, highlighting the limitations of MMLU in reflecting the nuances of the Korean language and cultural specifics. MMLU, derived from American tests, is inherently biased towards western culture. We provide examples in Table 11. Conversely, 20.4% of KMMLU requires understanding Korean cultural practices, societal norms, and legal frameworks. Leveraging the data from instructkr (2024), a Korean equivalent to Chiang et al. (2024), we measure the impact of such characteristics. Analyzing the ranking correlations of KMMLU and a translated MMLU against Elo rankings from 3253 human votes, Table 6 confirms that KMMLU more accurately mirrors human preferences, as shown by closer Levenshtein distances and stronger ranking correlations.

Metric	KMMLU	MMLU-Tran.
Levenshtein Distance	3	8
Spearman Ranking Corr.	0.94	0.86

Table 5: Correlation with the scores of KMMLU and translated MMLU against human-rated elo rankings.

6.2 KMMLU-KOR

To specifically assess models on questions that necessitate Korean knowledge, we introduce the KMMLU-KOR subset, a collection of 1,305 hand-picked questions (Further details in Appendix I). Table 6 presents a performance analysis, focusing on the discrepancy between general performance and Korea-specific questions. HyperCLOVA X, shows the smallest performance decline at 3.4%, whereas GPT-3.5-Turbo and GPT-4 experience drops of 7.11% and 8.74%, respectively. This underscores the significance of incorporating Korean linguistic and cultural contexts in model training to enhance performance on KMMLU.

Model	Original	KOR Sub.	Δ
QWEN-72B	50.83	42.56	−8.27
QWEN-72B-CHAT	47.28	35.13	−12.15
GPT-3.5-TURBO	42.47	35.36	−7.11
GEMINI-PRO	50.18	40.83	−9.35
HYPERCLOVA X	53.40	50.00	−3.40
GPT-4	59.95	51.21	−8.74

Table 6: Performance comparison on the original KMMLU and the KMMLU-KOR subset.

Error Analysis We observe HyperCLOVA X to correctly answer 47 questions that GPT-4 misses. Of these, 18 pertain to history, 11 to accounting and taxation, and 7 to geology. We pinpoint 73 Korea-specific questions that all models fail, including 31 concerning the Security Services Industry Act, a law first implemented in 2017 and frequently amended since. This highlights gaps in models' understanding of recent legislative changes. Additionally, we also find 23 questions that use the phrase "our country," presuming knowledge that this refers to Korea, further complicating comprehension for non-specialized models.

7 Conclusion

In this work, we introduce the **KMMLU** Benchmark—a comprehensive compilation of 35,030 expert-level multiple-choice questions spanning 45 subjects, all sourced from original Korean exams without any translated content. Our findings highlight significant room for improvement in the Korean proficiency of state-of-the-art LLMs. We discover that the improvements in the performance of non-Korean LLMs stem from capabilities unrelated to Korean, underscoring the importance of Korean pre-training for better performance in Korea-specific contexts. We expect the KMMLU benchmark to aid researchers in identifying the shortcomings of current models, enabling them to assess and develop better Korean LLMs effectively.

8 Limitations

While we put our greatest effort into creating a benchmark with extensive coverage, there are some limitations that future research will need to address. First, due to concerns over copyright issues, we removed a substantial number of questions from the Korean language, medical, and financial domains, thereby creating coverage gaps. Secondly, the recent surge of chat-aligned LLMs has cast doubt on the effectiveness of traditional benchmarks for assessing generative abilities and instruction-following skills. While MMLU continues to be a de facto standard for evaluating a broad range of knowledge, there is a shifting trend towards using dedicated LLM Judges and crowdsourced human preferences, such as the LMSys Chatbot Arena (Zheng et al., 2023), for assessing generative capabilities. Future efforts should aim to expand Korean benchmarking tools to include assessments of generative abilities. Moreover, the potential misuse of benchmarks may pose societal risks. Optimizing solely for benchmarks may create models that perform poorly in real-world applications and should be avoided.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS:

New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Alexandra Institute. 2024. Multilingual mmlu. https://huggingface.co/datasets/alexandrainst/m_mmlu.

instructkr. 2024. ko-chatbot-arena-leaderboard. https://huggingface.co/spaces/instructkr/ko-chatbot-arena-leaderboard.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2023. Kobbq: Korean bias benchmark for question answering. *arXiv preprint arXiv:2307.16778*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Sungho Park, et al. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. *arXiv preprint arXiv:2310.04928*.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Taekyoon Choi L. Junbum. 2023. llama-2-koen-13b.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoung Pil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, et al. 2023. Square: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. *arXiv preprint arXiv:2305.17696*.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. Kornat: Llm alignment benchmark for korean social values and common knowledge. *arXiv preprint arXiv:2402.13605*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. Beep! korean corpus of online news comments for toxic speech detection. *arXiv preprint arXiv:2005.12503*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

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. *arXiv preprint arXiv:2405.20574*.

Chanjun Park, Hwalsuk Lee, Hyunbyung Park, Hyeonwoo Kim, Sanghoon Kim, Seonghwan Cho, Sunghun Kim, and Sukyung Lee. 2023. Open ko-llm leaderboard. https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*.

Qwen. 2024. Qwen 1.5.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Parker Riley, Timothy Dozat, Jan A Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. Frmt: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2023. Hae-rae bench: Evaluation of korean knowledge in language models. *Preprint*, arXiv:2309.02706.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Zhi-Rui Tam, Ya-Ting Pai, Yen-Wei Lee, Sega Cheng, and Hong-Han Shuai. 2024. An improved traditional chinese evaluation suite for foundation model. *arXiv preprint arXiv:2403.01858*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

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*.

ZaloAI-JAIST. 2023. Vmlu.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Dataset Details

In this section, we provide additional details on the KMMLU benchmark.

A.1 Category Distribution

The KMMLU dataset consists of three subsets Train, Validation and Test. In Figures 4 and 5 we visualize the distribution of each sub-category for the train and test set respectively. The validation set contains 5 instances for all sub-categories.

A.2 Source Distribution

Source	# Instances
License Tests	235,976
PSAT	7373
CSAT	428

Table 7: Source distribution

KMMLU sources questions from 533 diverse exams, including the PSAT, Korean License Tests, and the CSAT. Table 7 presents the distribution of the sources, with License Tests comprising the majority at 235,976 samples.

Figure 6 provides an overview of the years each question was sourced from. The questions spans 25 years starting from 1999 to 2023. The average number of instances for each year is 9030.52, with the most at 2015 by 13583 samples and the least at 1999 with 170 samples.

A.3 Copyrights and License

To prevent copyright issues, we conduct a manual review of the *Test* and *Validation* sets of the KMMLU benchmark. We remove instances from exams with restrictive licenses, primarily in the medical and financial domains. We also filter questions that include segments of Korean literature where the respective authors hold copyright and are challenging to manage individually. However, the *Train* set consists of 208,522 instances, making

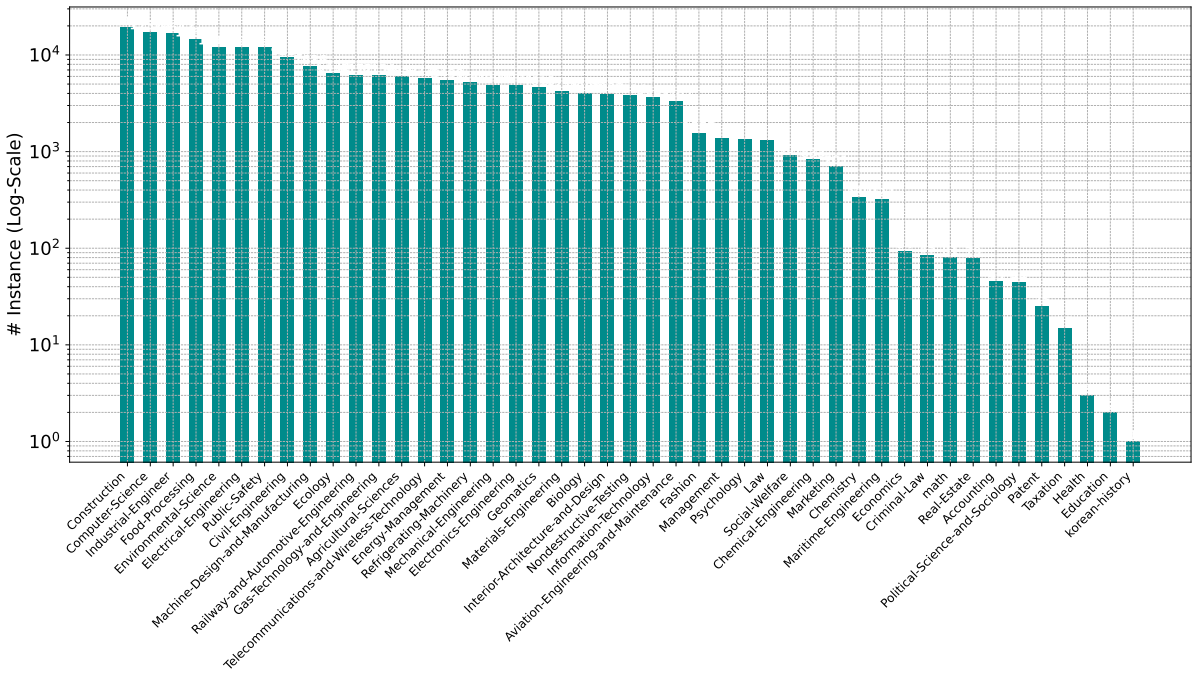

Figure 4: Sub-Category distribution of the train set in log-scale.

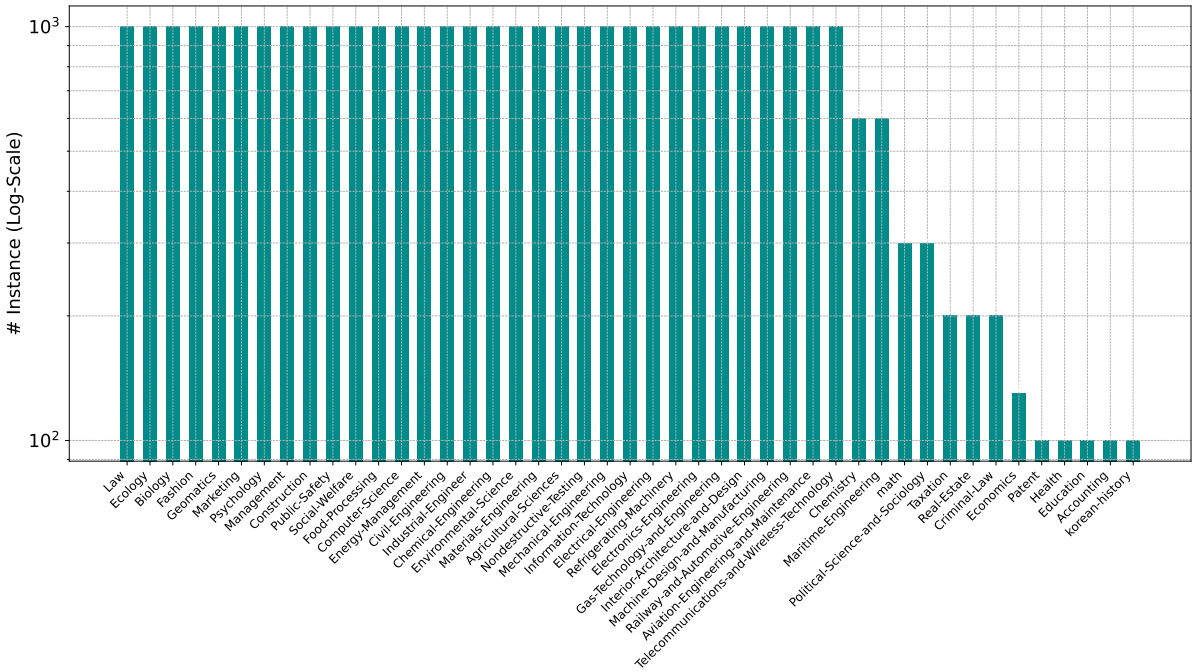

Figure 5: Sub-Category distribution of the test set in log-scale.

manual review impractical. Instead, we use insights from our manual review of the test and validation subsets to identify and remove questions in the train set that originate from sources identified as private. Additionally, we collect headers and footers from each webpage during the crawling process to remove those containing copyright information. The final dataset is published under a CC-BY-ND license and is freely available at HuggingFace[8]. Additionally during our manual review we do not identify any personally identifiable information or offensive content.

For better reproducibility, our evaluation codebase[9] is built using Eleuther AI's LM-Eval-Harness

[8] https://huggingface.co/datasets/HAERAE-HUB/KMMLU

[9] https://github.com/EleutherAI/

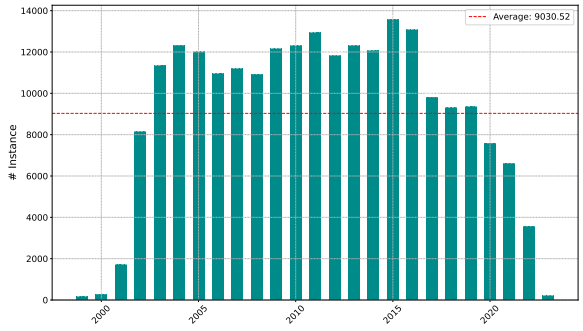

Figure 6: Overview of the years each question was sourced from.

(MIT License) (Gao et al., 2023). This includes the evaluation code, generation configuration for all settings (open and proprietary models, direct and CoT settings), and our prompts and CoT exemplars. All components are available under the MIT license.

B Additional Analysis

B.1 How Does Continual Pretraining Affect Korean Proficiency?

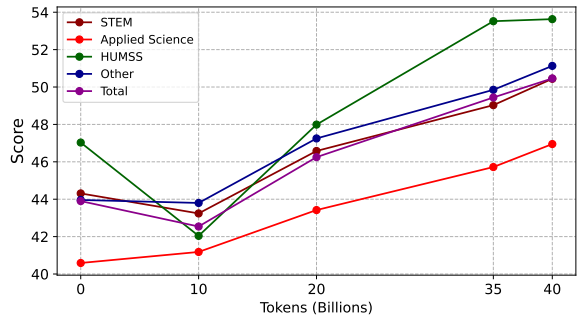

Figure 7: Performance of YI-KO-34B on the KMMLU benchmark on each checkpoint.

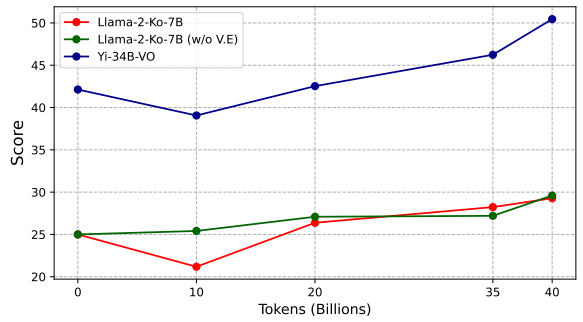

Figure 8: Performance of LLAMA-2-KO-7B (both with and without vocabulary expansion) and YI-KO-34B on the KMMLU benchmark for each checkpoint.

`lm-evaluation-harness`

Models from the YI series, initially trained for bilingual (English and Chinese) usage, significantly outperform the POLYGLOT-KO models. This performance gap further expands with the continual pretraining of these models. In Table 8, we assess the YI-KO 6B and 34B models, each continually trained for an additional 60 billion and 40 billion tokens, respectively, after expanding their vocabulary to include Korean. Additionally, we analyze the available checkpoints of the YI-KO-34B model in Figure 7, noting a consistent performance increase following an initial decline at the first checkpoint. This early drop is likely due to the expanded vocabulary, which can disrupt training initially (Zhao et al., 2024).

Additionally, in Figure 8, we evaluate the performance of three models—LLAMA-2-KO-7B (both with and without vocabulary expansion) and YI-KO-34B—across 40 billion tokens of continual pretraining.[10] Our analysis reveals two key insights: First, while vocabulary expansion initially leads to a drop in performance, the models subsequently recover and stabilize after surpassing 10 billion tokens of training. This challenges previous findings that suggested vocabulary expansion is unfavorable at such training scales. Interestingly, Table 12 shows that the perplexity of YI-KO-34B is higher than its original model YI-34B, likely due to the undertraining of newly added tokens, indicating room for improvement. Second, learning outcomes vary despite being trained on the same dataset, suggesting that differences in inherent model capabilities significantly influence performance.

B.2 Do Machines Handle Problems with Negation Effectively?

Table 9 demonstrates a notable trend in language model performance on the KMMLU test set: models perform better on questions that include negations. This finding contrasts with previous studies (Hosseini et al., 2021; Li et al., 2023a) that identified LLMs to suffer when dealing with negated questions. However, this does not suggest that negation in Korean presents a lower difficulty level than in other languages. Instead, the improved performance may be attributed to the nature of the questions in KMMLU, where negation is more common in declarative knowledge questions, which are generally easier for models to handle compared to procedural knowledge questions (Hendrycks et al.,

[10]The YI-KO-6B model is excluded as intermediary checkpoints are unavailable.

Model	STEM	Applied Science	HUMSS	Other	Average
LLAMA-2-KO-7B	31.60	32.50	26.33	30.00	30.10
YI-KO-6B	40.69	39.52	40.50	41.60	40.55
YI-KO-34B	50.44	46.95	53.63	51.13	50.46

Table 8: 5-shot accuracy on KMMLU with models Continual Pretrained on Korean.

2020). For example, the math subset, which is the most challenging subset for most LLMs, does not include any negated questions. Furthermore, Table 10 illustrates that only 20% of STEM and 19% of Applied Science questions include negation, in contrast to 45% in the HUMSS subset.

Models	W Negation	W/O Negation
LLAMA-2-70B	40.2	40.08
YI-34B	47.26	42.43
QWEN-72B	53.57	48.82
GEMINI-PRO	55.05	48.63
GPT-3.5-TURBO	45.61	40.39
GPT-4	65.53	57.88

Table 9: Comparison of accuracy between questions with and without negation. Evaluation is done in 5-shot setting using the Direct Method.

Category	% of Negated Q.
STEM	20.54%
Applied Science	19.16%
HUMSS	45.76%
Other	34.83%
Math	0.00%
Electrical Eng.	9.70%
Aviation Eng. & Maint.	14.40%

Table 10: Ratio of Negated Questions in each category.

B.3 When Do Korean Proprietary Models Outperform GPT-4?

Figure 9 provides a comparative performance analysis between the top-performing Korean model, HYPERCLOVA X, and GPT-4 across various disciplines, with detailed numerical results available in Appendix 9. The comparison shows that GPT-4 generally outperforms HYPERCLOVA X in most subjects, with performance differentials ranging from a significant 22.0% in Accounting to a marginal 0.5% in Taxation. Although specific details on the pretraining of both models are kept private HYPERCLOVA X is designed for bilingual (English and Korean) use, while GPT-4 supports many more languages. This observation corroborates our earlier findings from Section 5 that the curse of multilinguality diminish as models scale. Notably, HYPERCLOVA X demonstrates superior performance over GPT-4 in Korean History and Criminal Law. This is likely attributable to HYPERCLOVA X's specialized focus on the Korean language, which presumably enhances its proficiency in topics requiring regional-specific knowledge and understanding.

These characteristics are evident in Table 19, which compares the performance of models on the KMMLU-KOR subset. HYPERCLOVA X and GPT-4 scores 50.00 and 51.21, respectively. The performance gap narrows as GPT-4 experiences a decrease of 8.74 points, while HYPERCLOVA X sees a smaller decline of 3.4 points. This indicates that HYPERCLOVA X is more resilient to questions on Korean knowledge, maintaining closer to its original performance.

B.4 Do Machines Also Err Where Humans Often Do?

In Figure 10, we compare the performance of LLMs against human accuracy. The findings indicate that LLMs do not exhibit a performance trend that correlates with human performance. Instead, the models display similar performance levels irrespective of the variance in human accuracy. This observation aligns with insights from the (Hendrycks et al., 2020), which reported that GPT-3 achieved a higher score in College Mathematics at 35.0%, compared to 29.9% in Elementary Mathematics, suggesting that the model's performance does not necessarily scale with the complexity of the task as judged by human standards. Interestingly, the models demonstrate a strong correlation with each other, implying that despite being trained on distinct datasets, they possess similar capabilities. This phenomenon indicates that there may be underlying commonalities in how these models process and generate responses, leading to a similar performance trend.

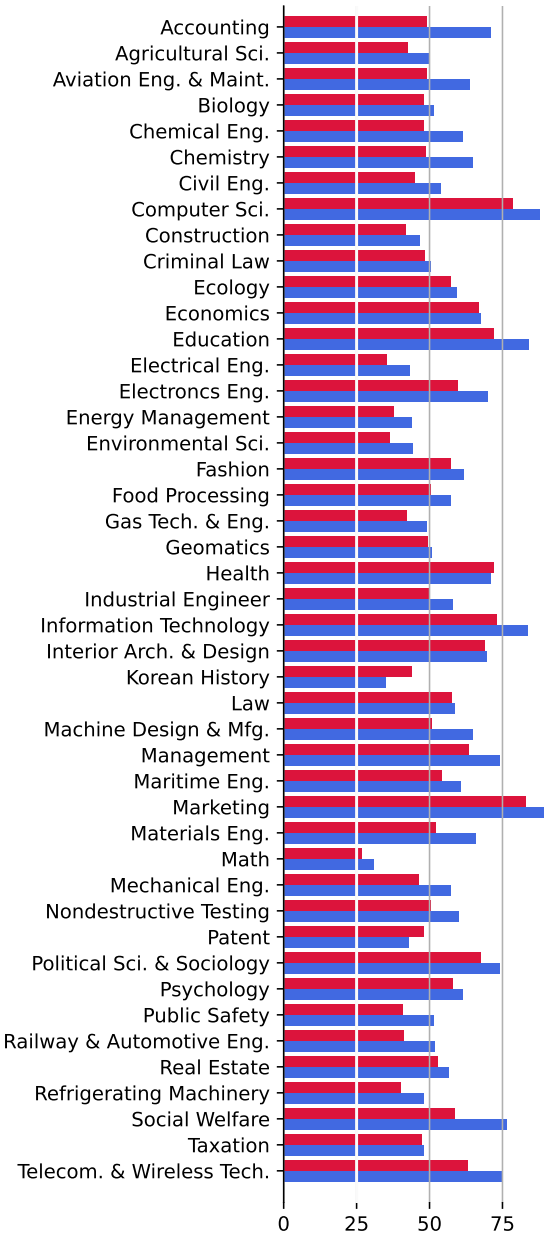

Figure 9: Comparison of GPT-4(Blue) and HYPERCLOVA X(Red) using the Direct method in a 5-shot setting.

B.5 Why are native benchmarks important?

Continuing from Section 6, we highlight the importance of original benchmarks over translated versions. Figure 3 illustrates that MMLU questions often reflect U.S.-centric content, such as categories like *"high_school_government_and_politics,"* which require understanding of the American governmental system, and *"miscellaneous,"* which assumes knowledge of American slang. These examples, shown in Table 11, reveal the cultural biases inherent in the dataset. Regardless of translation quality, using such content as a measure

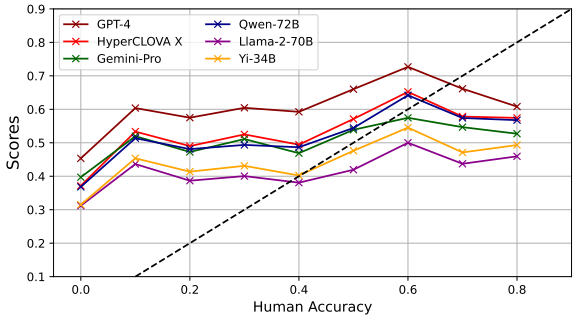

Figure 10: Comparison of model performance and human accuracy. Model performance is calculated using the Direct method in a 5-shot setting.

of Korean language competency is fundamentally flawed. For instance, proficiency in questions about the U.S. Sixth Amendment does not translate into effectiveness of a legal LLM for Korean contexts.

C Contamination Check

The questions for the KMMLU benchmark are sourced from publicly available Korean tests. Accordingly, there may be concerns about whether the sources have been included in the pretraining corpora of the tested language models. Unfortunately, none of the examined models disclose their training data, which precludes a direct assessment of potential knowledge leakage. As an alternative, we refer to the methods of Xu et al. (2024) to estimate potential leakage using perplexity and n-gram accuracy. We sample 2,000 instances from the KMMLU test set and paraphrase them using GPT-4-TURBO. The paraphrasing is conducted twice to reduce potential bias, with temperatures of 0.7 and 0.8. See Figure 14 for the prompts used.

In Table 12, we observe that all models demonstrate lower perplexity on the paraphrased versions of the benchmark. While this does not entirely rule out the possibility of whether the models have encountered the benchmark during pretraining, the preference for a rewritten version—which does not exist online and thus could not have been included in the pretraining—implies that the models are not familiar with the original material.

For proprietary models where perplexity examinations are unavailable, we use n-gram accuracy. First, we combine the question and answer parts with a single space for each sample, creating a combined text X. Second, we uniformly sample K (i.e., 5) starting points within the interval from 2 to $|X|$. The text from the beginning to each starting point

Subset	HIGH SCHOOL GOVERNMENT AND POLITICS
Question	Which of the following statements does NOT accurately describe voting behavior in the United States?
Subset	HIGH SCHOOL US HISTORY
Question	This question refers to the following information. "I was once a tool of oppression And as green as a sucker could be And monopolies banded together To beat a poor hayseed like me. "The railroads and old party bosses Together did sweetly agree; And they thought there would be little trouble In working a hayseed like me. . . ." —"The Hayseed" The song, and the movement that it was connected to, highlight which of the following developments in the broader society in the late 1800s?
Subset	PROFESSIONAL ACCOUNTING
Question	The FASB makes changes to the Accounting Standards Codification by issuing?
Subset	PROFESSIONAL LAW
Question	In which of the following stages does an indigent person not have the Sixth Amendment right to counsel?
Subset	US FOREIGN POLICY
Question	Who was the first American president to visit communist China?
Subset	MISCELLANEOUS
Question	What US president is mentioned by name in the opening theme song of TV's 'All in the Family'?

Table 11: Example questions in MMLU targetted towards American history or culture.

Model	Original	Paraphrased
POLYGLOT-KO-12.8B	9.961	8.996
YI-34B	3.006	2.533
YI-KO-34B	12.810	9.538
LLAMA-2-70B	3.067	2.589
QWEN-72B	10.750	7.234
YI-34B-CHAT	3.354	2.783
LLAMA-2-70B-CHAT	5.431	3.846
QWEN-72B-CHAT	15.625	9.875

Table 12: Comparision of perplexity calculated from the KMMLU Benchmark. The Paraphrased column denotes the average perplexity calculated from the two versions each generated with temperatures 0.7 and 0.8.

serves as the prompt, with the subsequent n-gram used as the prediction target. The prediction results are shown in Figure 11.

GPT-4 and HYPERCLOVA X demonstrate low average scores for predicting target n-grams, with 0.004 and 0.001 for N=3 and N=5 in GPT-4, and 0.014 and 0.005 for N=3 and N=5 in HYPERCLOVA X, respectively. However, GPT-4 scores above 0.4 in 21 instances for N=3 and 9 instances for N=5. Examining the samples in Tables 13 and 14 reveals that the target is a repetition of the prompt, suggesting the model's guesswork rather than data contamination. HYPERCLOVA X exhibits a marginally higher average score than GPT-4, scoring above 0.4 in 69 instances for N=3 and 37 instances for N=5. Tables 15, 16, and 17 present similar findings, with the model occasionally repeating the target verbatim despite its absence from the prompt, potentially indicating leakage. Nonetheless, given the low average scores, we hypothesize that the model might have encountered internet sources reproducing exam segments, rather than contamination of the entire source.

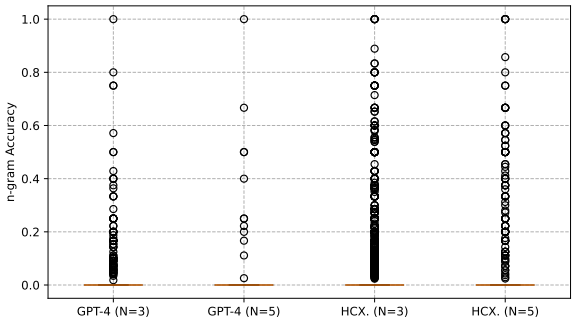

Figure 11: Comparison of model performance and human accuracy. Model performance is calculated using the Direct method in a 5-shot setting.

Model	GPT-4
Prompt	큰날개와 꼬리날개에 의한 무게중심 주위의 키놀이모멘트 관계식은? (단, MC.G : 무게중심 주위의 키놀이 모멘트, MC.G WING : 큰날개에 의한 키놀이 모멘트, MC.G TAIL : 꼬리날개에 의한 키놀이
Target	모멘트) MC.G = MC.G WING + MC.G TAIL
Generation	큰날개와 꼬리날개에 의한 무게중심 주위의 키놀이모멘트 관계식은 MC.G = MC.G WING + MC.G TAIL 입니다.
Accuracy (N=3, N=5)	(1.0, 0.8)

Table 13: An example where GPT-4 scores high n-gram accuracy. fbox highlights the overlapping n-grams.

Model	GPT-4
Prompt	(주)한공은 종업원 홍길동 씨가 퇴사하여 퇴직금 6,000,000원을 보통예금계좌에서 지급하였다. 홍길동 씨의 퇴사직전 (주)한공의 퇴직급여충당부채는 10,000,000원이다. 홍길동 씨 퇴직금지급에 대한 분개로 옳은 것은?(차) 퇴직급여충당부채
Target	: 6,000,000원, (대) 보통예금 : 6,000,000원
Generation	(주)한공은 종업원 홍길동 씨가 퇴사하여 퇴직금 6,000,000원을 보통예금계좌에서 지급하였다. 홍길동 씨의 퇴사직전 (주)한공의 퇴직급여충당부채는 10,000,000원이다. 홍길동 씨 퇴직금지급에 대한 분개로 옳은 것은?(차) 퇴직급여충당부채(차) 퇴직급여충당부채 6,000,000원 (대) 보통예금 6,000,000원
Accuracy (N=3, N=5)	(1.0, 0.8)

Table 14: An example where GPT-4 scores high n-gram accuracy. fbox highlights the overlapping n-grams.

Model	HYPERCLOVA X
Prompt	큰날개와 꼬리날개에 의한 무게중심 주위의 키놀이모멘트 관계식은? (단, MC.G : 무게중심 주위의 키놀이 모멘트, MC.G WING : 큰날개에 의한 키놀이 모멘트, MC.G TAIL : 꼬리날개에 의한 키놀이
Target	모멘트) MC.G = MC.G WING + MC.G TAIL
Generation	무게중심 주위의 키놀이 모멘트(MC.G)는 큰날개에 의한 키놀이 모멘트(MC.G WING)와 꼬리날개에 의한 키놀이 모멘트(MC.G TAIL)의 합으로 구할 수 있습니다. - MC.G = MC.G WING + MC.G TAIL
Accuracy (N=3, N=5)	(1.0, 0.8)

Table 15: An example where HYPERCLOVA X scores high n-gram accuracy

Model	HYPERCLOVA X
Prompt	기계식주차장의 기준에 관한 기술이 잘못 된 것은?대형기계식주차장은 직경 4m 이상의 방향전환장치와 그 방향전환장치에
Target	접한 너비 1m 이상의 여유공지가 있어야 한다.
Generation	접한 너비 1m 이상의 여유공지가 있어야 합니다.
Accuracy (N=3, N=5)	(1.0, 1.0)

Table 16: An example where HYPERCLOVA X scores high n-gram accuracy

D License

The KMMLU benchmark is released under a CC-BY-ND license. This license prohibits the distribution of the remixed or transformed version of the dataset. The code and prompts for evaluation are released by an MIT license via LM-Eval-Harness.

E Evaluated Models

Polyglot-Ko (Ko et al., 2023). Introduced by the Polyglot Team of EleutherAI POLYGLOT-KO is a comprehensive suite of Korean-centric autoregressive language models featuring models with 1.3, 3.8, 5.8, and 12.8 billion parameters. The models

Model	HYPERCLOVA X
Prompt	철골콘크리트조의 기둥에서 내화구조의 기준으로 옳은 것은?작은지름 25cm 이상으로서 철골을
Target	두께 5cm 이상의 콘크리트로 덮은 것
Generation	철골콘크리트조의 기둥에서 내화구조의 기준으로 옳은 것은?작은지름 25cm 이상으로서 철골을 철골철근콘크리트조의 기둥에서 내화구조의 기준은 다음과 같습니다... 4. 철골보를 두께 5cm 이상의 콘크리트로 덮은 것...
Accuracy (N=3, N=5)	(1.0, 1.0)

Table 17: An example where HYPERCLOVA X scores high n-gram accuracy

are pretrained on Korean corpus ranging from 167 to 219 billion tokens.

Llama-2 (Touvron et al., 2023). LLAMA-2 is a suite of large language models ranging from 7 to 70 billion parameters developed by Meta. The models are pretrained on 2 trillion tokens, and whether Korean is included is not reported. The suite also provides LLAMA-2-CHAT an aligned version for instruction-following and ssafety.

Yi (Young et al., 2024). The YI model, developed by 01.AI, is a series of bilingual language models available in two variants: 6B and 34B. It employs an architecture similar to LLAMA-2 and is pretrained on a multilingual corpus of 3 trillion tokens. Additionally, the model features chat versions tailored for instruction-following.

Qwen (Bai et al., 2023). QWEN is a suite of bilingual language models developed by Alibaba Cloud, with variants spanning from 1.8 billion to 72 billion parameters. Each model within the series is pretrained on a dataset of 3 trillion tokens. The QWEN also includes specialized chat models designed for following instructions.

GPT-3.5 & GPT-4 (OpenAI, 2023). Developed by OpenAI, the GPT series is renowned for exhibiting state-of-the-art performance across various benchmarks and tasks, including exceptional instruction-following capabilities. Specific details regarding the parameter count and the scope of the training data are not open to the public.

Gemini (Team et al., 2023). GEMINI is a series of models developed by Google, encompassing four variants: Nano-1, Nano-2, Pro, and Ultra. In our experiments, we utilize GEMINI-PRO. Details regarding the parameter count and the dataset used for training are not disclosed.

HyperCLOVA X (Yoo et al., 2024). HYPERCLOVA X, developed by NAVER, is a bilingual language model proficient in both English and Korean.

F Compute Resources

Models with openly available weights were evaluated on an internal cluster comprising 4 NVIDIA A100 HBM2e 80GB PCIe GPUs. We performed 44 direct evaluations on the original KMMLU set and 2 CoT evaluations on the KMMLU-HARD set, totaling approximately 300 A100 GPU hours.

G Prompting Format

For evaluation, we use the following prompting formats. For n-shot evaluation, the identical format is repeated.

Direct Evaluation Prompt

{Question}
A. {A}
B. {B}
C. {C}
D. {D}
정답:

Figure 12: Prompt used in our Direct Evaluation.

CoT Evaluation Prompt

질문: {Question}
A. {A}
B. {B}
C. {C}
D. {D}
정답: 차근 차근 생각해봅시다. 회계학 관련 정보를 위해 위키피디아를 참조하겠습니다.

Figure 13: Prompt used in our CoT Evaluation.

For the paraphrasing in Section C, we use the following prompting format.

Model	# Params	Access	Language
English-Centric / Bilingual Pretrained Models			
LLAMA-2 (Touvron et al., 2023)	7B, 13B, 70B	Weights Available	En
YI (Young et al., 2024)	6B, 34B	Weights Available	En / Zh
QWEN (Bai et al., 2023)	7B, 14B, 72B	Weights Available	En / Zh
English-Centric / Bilingual Finetuned Models			
LLAMA-2-CHAT (Touvron et al., 2023)	7B, 13B, 70B	Weights Available	En
YI-CHAT (Young et al., 2024)	6B, 34B	Weights Available	En / Zh
QWEN-CHAT (Bai et al., 2023)	7B, 14B,72B	Weights Available	En / Zh
Korean Pretrained Models			
POLYGLOT-KO (Ko et al., 2023)	1.3B, 3.8B, 5.8B, 12.8B	Open Source	Ko
Korean Continual Pretrained Models			
LLAMA-2-KO (L. Junbum, 2023)	7B	Weights Available	En / Ko
YI-KO (L. Junbum, 2023)	6B, 34B	Weights Available	En / Zh / Ko
Proprietary Models			
GPT-3.5-TURBO	undisclosed	API	-
GPT-4 (OpenAI, 2023)	undisclosed	API	-
GEMINI-PRO (Team et al., 2023)	undisclosed	API	-
HYPERCLOVA X (Yoo et al., 2024)	undisclosed	API	-

Table 18: Overview of the 31 LLMs evaluated in this paper.

H More details for CoT Exemplar Creation

We use the zero-shot CoT prompt of Figure 15 to collect the exemplar CoTs for our dataset. We request to use browsing for more accurate explanations if it is available. For GPT-4, we manually input the prompt to the ChatGPT Web interface (chat.openai.com). For HyperCLOVA X, we devise 3-shot demonstrations to generate relevant queries to the NAVER search engine (www.naver.com). Then, we concatenate top-3 search results to generate explanations.

I KMMLU-KOR Subset

For more targeted evaluation, alongside KMMLU-HARD, we introduce a second subset, KMMLU-KR. KMMLU-KR is a collection of questions that specifically require Korean knowledge to solve so that users can assess a language model's proficiency in Korea-specific contexts. We initially adopted a keyword filter to collect 2,149 candidate questions. Then, two authors independently reviewed each question, eliminating irrelevant entries and categorizing the pertinent ones into four distinct categories: *Cultural*, which delves into Korean history and societal norms; *Regional*, focusing on the geographical details of Korea; *Legal*, concerning Korea's legal and governmental structures; *Others*, comprising questions that demand knowledge of the Korean context but do not fit into the previous categories. Following the filtering, 1,305 questions remained, constituting the KMMLU-KOR subset. Please refer to Figure 1 for detailed examples of each category. For evaluation results, see Table 19.

J Evaluation Results

In this section, we present the results of our evaluation, broken down by category for each model assessed. Tables 22-26 include results using the Direct method. Table 27 presents the results evaluated using the CoT method. Figure 9 presents a comparative performance analysis between the most capable Korean model, HYPERCLOVA X, and GPT-4.

Model	Cultural	Legal	Regional	Other	Total
Multilingual Pretrained Models					
LLAMA-2-7B	21.51	22.77	38.46	25.53	23.26
LLAMA-2-13B	29.34	27.16	50.00	28.87	28.13
LLAMA-2-70B	35.33	34.05	30.77	40.14	34.88
YI-6B	31.33	33.01	50.00	33.58	33.13
YI-34B	37.35	41.02	38.46	47.01	41.18
QWEN-7B	10.47	10.08	26.92	21.28	11.63
QWEN-14B	22.16	21.00	34.62	26.76	22.04
QWEN-72B	37.95	43.20	34.62	46.27	42.56
Multilingual Chat Models					
LLAMA-2-7B-CHAT	23.84	26.72	42.31	28.37	26.78
LLAMA-2-13B-CHAT	27.11	24.76	46.15	29.10	25.95
LLAMA-2-70B-CHAT	27.71	30.83	53.85	29.85	30.62
YI-6B-CHAT	25.58	32.85	42.31	32.62	31.98
YI-34B-CHAT	37.95	41.02	42.31	45.52	41.09
QWEN-7B-CHAT	6.36	8.97	26.92	16.78	9.81
QWEN-14B-CHAT	20.36	17.56	34.62	26.76	19.34
QWEN-72B-CHAT	34.10	35.27	30.77	37.58	35.13
Korean Pretrained Models					
POLYGLOT-KO-1.3B	30.81	29.00	34.62	26.24	29.00
POLYGLOT-KO-3.8B	25.58	25.88	26.92	21.99	25.48
POLYGLOT-KO-3.8B	24.10	27.79	19.23	26.12	26.90
POLYGLOT-KO-12.8B	30.12	28.03	26.92	24.63	27.85
Korean Continual Pretrained Models					
YI-KO-6B	34.94	36.65	30.77	44.78	37.11
YI-KO-34B	44.19	44.70	42.31	51.77	45.14
Proprietary Models					
GPT-3.5-TURBO	38.73	34.15	26.92	41.61	35.36
GEMINI-PRO	45.18	39.68	42.31	43.28	40.83
HYPERCLOVA X	**47.90**	50.65	42.31	51.41	50.00
GPT-4	41.57	**52.31**	**46.15**	**58.21**	**51.21**

Table 19: Average accuracy(%) calculated using the Direct method in a 5-shot setting on the KMMLU-KOR subset. We report the micro-average accuracy due to the imbalance on each category. The highest-scoring model across the entire table is highlighted in **bold**, and the best model within each category is underlined.

Models	HUMSS	Other	Applied Science	STEM
Mistral-7B-v0.3	35.48	36.24	35.88	39.40
Mistral-8x7B-v0.1	42.01	39.14	37.48	40.62
Mistral-8x22B-v0.1	23.35	22.14	21.15	23.14
Gemma-2-2B	29.16	30.67	31.39	32.76
Gemma-2-9B	43.85	47.67	46.99	48.96
Gemma-2-27B	49.25	49.74	49.29	51.97
Qwen1.5-0.5B	23.75	17.55	15.59	17.38
Qwen1.5-1.8B	26.45	23.78	20.88	22.82
Qwen1.5-4B	32.32	34.66	35.12	35.36
Qwen1.5-7B	37.97	39.30	36.51	39.38
Qwen1.5-14B	45.08	44.85	42.96	46.29
Qwen1.5-32B	47.98	48.90	45.34	48.73
Qwen1.5-72B	54.38	52.85	50.13	51.65
Qwen1.5-110B	60.24	58.23	54.55	57.16
Qwen2-0.5B	28.51	29.63	28.70	28.77
Qwen2-1.5B	38.04	37.98	35.55	37.62
Qwen2-7B	51.44	49.29	46.46	49.08
Qwen2-72B	67.68	63.84	63.72	65.88
Meta-Llama-3.1-8B	40.53	42.18	39.38	42.85
Meta-Llama-3.1-70B	53.81	52.43	49.31	53.84

Table 20: 5-shot accuracy using the Direct method for 20 recent LLMs broken down by category.

Models	HUMSS	Other	Applied Science	STEM
Mistral-7B-Instruct-v0.3	33.52	30.74	28.85	33.52
Mistral-8x7B-Instruct-v0.1	41.90	39.10	38.49	42.10
Mistral-8x22B-Instruct-v0.1	47.58	47.63	45.19	48.67
Mistral-Large-Instruct-2407	63.08	59.79	60.32	61.68
Gemma-2-2B-it	33.52	34.77	31.58	34.56
Gemma-2-9B-it	47.12	48.00	43.50	47.08
Gemma-2-27B-it	53.07	50.89	48.43	52.45
Qwen1.5-0.5B-Chat	25.86	24.07	26.18	25.00
Qwen1.5-1.8B-Chat	27.59	24.52	21.67	23.26
Qwen1.5-4B-Chat	32.28	33.00	32.77	33.36
Qwen1.5-7B-Chat	37.37	39.65	37.60	40.48
Qwen1.5-14B-Chat	43.90	44.05	40.84	43.92
Qwen1.5-32B-Chat	47.28	48.20	43.66	47.00
Qwen1.5-72B-Chat	52.76	51.16	49.51	50.84
Qwen1.5-110B-Chat	53.25	50.49	49.74	52.55
Qwen2-0.5B-Instruct	29.37	30.98	31.03	30.78
Qwen2-1.5B-Instruct	37.74	38.55	34.51	37.46
Qwen2-7B-Instruct	48.52	45.79	43.49	46.94
Qwen2-72B-Instruct	65.40	62.88	62.21	64.25
Meta-Llama-3.1-8B-Instruct	42.02	41.86	39.57	42.98
Meta-Llama-3.1-70B-Instruct	53.60	49.50	50.47	54.01
Command-R-v01	43.20	41.27	36.47	40.08
Command-R-Plus	48.99	48.61	45.81	47.84

Table 21: 5-shot accuracy using the Direct method for 23 recent LLMs, broken down by category.

Category	Polyglot-Ko				Yi-Ko	
	1.3B	3.8B	5.8B	12.8B	6B	34B
accounting	30.0	32.0	32.0	30.0	38.0	46.0
agricultural_sciences	27.0	30.3	30.1	32.0	32.7	39.6
aviation_engineering_and_maintenance	30.2	29.7	29.9	30.7	36.1	45.2
biology	24.0	26.7	28.6	25.3	32.1	41.7
chemical_engineering	25.3	27.9	24.7	24.7	36.2	49.7
chemistry	30.3	25.2	26.0	29.2	40.8	52.33
civil_engineering	27.4	31.8	31.9	34.3	38.0	45.3
computer_science	32.1	35.9	34.8	33.9	61.5	73.9
construction	33.6	31.0	31.7	32.0	34.7	38.3
criminal_law	26.0	29.0	29.5	28.5	31.5	37.0
ecology	28.7	29.4	31.8	32.7	45.2	57.4
economics	23.8	26.2	24.6	24.6	41.5	58.5
education	23.0	20.0	24.0	25.0	53.0	71.0
electrical_engineering	29.3	32.5	32.0	32.6	34.9	36.4
electronics_engineering	30.5	30.0	35.2	33.3	47.1	56.6
energy_management	28.8	26.5	24.5	26.9	30.0	38.4
environmental_science	26.1	32.9	27.3	30.9	33.9	40.8
fashion	27.0	29.5	29.2	29.8	46.1	50.9
food_processing	27.3	31.8	33.5	29.4	36.1	45.8
gas_technology_and_engineering	31.9	30.9	30.2	30.9	32.5	38.5
geomatics	29.2	30.0	31.1	31.0	41.6	46.9
health	26.0	32.0	27.0	25.0	52.0	73.0
industrial_engineer	27.4	32.3	33.1	31.2	43.0	48.7
information_technology	34.2	34.1	34.0	30.8	57.1	70.9
interior_architecture_and_design	32.4	29.6	29.7	31.8	47.3	61.1
korean_history	34.0	26.0	25.0	31.0	33.0	42.0
law	26.0	24.2	24.4	23.9	41.8	53.6
machine_design_and_manufacturing	28.7	34.0	26.9	30.3	39.9	45.9
management	27.6	27.7	27.7	28.0	43.7	61.9
maritime_engineering	24.8	31.5	26.7	26.5	44.0	51.8
marketing	24.4	30.6	26.4	33.5	69.6	81.2
materials_engineering	30.9	30.2	30.1	26.9	39.8	51.8
math	30.0	21.3	20.0	24.7	24.0	31.3
mechanical_engineering	24.2	31.5	27.1	26.9	38.0	44.1
nondestructive_testing	26.4	32.1	34.2	30.3	39.0	50.6
patent	29.0	23.0	22.0	31.0	32.0	34.0
political_science_and_sociology	25.7	25.7	25.7	25.7	41.7	64.7
psychology	26.5	25.9	27.7	25.9	40.1	53.8
public_safety	28.5	30.7	31.5	31.3	32.1	40.6
railway_and_automotive_engineering	23.6	29.0	28.9	26.8	34.7	39.6
real_estate	27.0	27.5	29.5	32.0	45.0	60.0
refrigerating_machinery	27.0	28.9	29.7	28.3	30.0	37.9
social_welfare	25.3	28.9	30.0	28.8	44.7	58.5
taxation	29.0	27.0	23.5	26.5	36.5	43.0
telecommunications_and_wireless_technology	28.6	33.9	34.1	32.2	52.4	60.4

Table 22: 5-shot accuracy using the Direct method for Polyglot-Ko, and Yi-Ko broken down by category.

Category	LLama-2-7B		LLama-2-13B		LLama-2-70B	
	Org.	Chat	Org.	Chat	Org.	Chat
accounting	25.0	22.0	20.0	16.0	34.0	26.0
agricultural_sciences	23.7	31.0	29.6	27.4	33.6	32.7
aviation_engineering_and_maintenance	23.7	26.8	30.3	26.8	35.9	33.0
biology	23.6	26.4	28.8	25.2	33.0	28.1
chemical_engineering	27.0	28.5	32.7	31.3	38.5	33.1
chemistry	26.8	26.7	30.3	27.7	41.8	32.3
civil_engineering	26.9	32.1	33.8	31.1	36.4	35.4
computer_science	24.1	28.0	47.4	41.5	67.3	58.9
construction	22.9	31.3	30.1	28.2	31.8	33.6
criminal_law	26.5	26.5	30.0	22.0	30.0	25.0
ecology	16.8	28.0	32.5	31.0	43.7	38.7
economics	27.7	30.8	27.7	38.5	45.4	40.0
education	24.0	29.0	26.0	28.0	56.0	38.0
electrical_engineering	27.4	29.4	34.0	28.0	30.8	32.3
electronics_engineering	33.0	32.2	38.8	31.5	47.1	39.9
energy_management	23.5	25.4	26.6	24.8	30.8	28.9
environmental_science	27.5	30.4	32.9	29.0	28.3	29.6
fashion	27.8	30.0	32.2	32.4	41.8	36.2
food_processing	17.4	24.3	31.1	26.6	33.9	29.9
gas_technology_and_engineering	22.3	28.0	29.1	26.4	31.4	29.6
geomatics	26.9	31.0	35.4	30.5	40.2	36.9
health	22.0	21.0	30.0	25.0	53.0	42.0
industrial_engineer	24.5	28.9	36.5	34.3	41.9	38.6
information_technology	27.3	29.3	44.4	37.3	62.8	52.0
interior_architecture_and_design	28.3	30.2	36.0	33.0	47.8	40.8
korean_history	26.0	21.0	25.0	25.0	32.0	23.0
law	24.4	25.5	26.5	27.6	40.8	34.9
machine_design_and_manufacturing	24.0	29.2	34.1	27.9	41.8	35.0
management	24.0	25.5	29.7	27.1	47.8	37.2
maritime_engineering	30.0	30.3	32.8	29.7	40.3	34.7
marketing	24.0	25.1	38.7	37.2	70.7	57.4
materials_engineering	21.2	28.5	29.0	26.2	40.4	30.8
math	25.0	28.3	24.3	26.7	27.0	23.7
mechanical_engineering	25.3	29.4	34.6	28.0	31.0	30.5
nondestructive_testing	24.8	29.9	34.2	25.8	41.5	32.1
patent	25.0	24.0	26.0	26.0	33.0	25.0
political_science_and_sociology	23.7	27.7	25.3	30.7	47.3	36.0
psychology	24.7	24.9	25.2	23.5	39.1	28.0
public_safety	28.5	30.4	32.6	31.0	33.0	34.0
railway_and_automotive_engineering	22.7	26.7	31.2	27.3	32.4	30.0
real_estate	23.5	24.5	24.5	25.0	32.0	26.5
refrigerating_machinery	24.2	26.3	28.7	27.8	30.1	30.8
social_welfare	26.7	28.8	31.9	27.6	47.8	35.0
taxation	23.0	24.5	21.5	24.5	33.0	31.0
telecommunications_and_wireless_technology	27.9	29.5	44.4	35.1	54.2	44.0

Table 23: 5-shot accuracy using the Direct method for LLAMA-2 (original and chat versions) broken down by category.

Category	Yi-6B		Yi-34B	
	Org.	Chat	Org.	Chat
accounting	29.0	30.0	46.0	45.0
agricultural_sciences	32.7	29.5	36.0	34.7
aviation_engineering_and_maintenance	31.9	30.9	36.9	34.8
biology	28.9	29.4	32.5	30.9
chemical_engineering	31.8	31.5	40.8	40.7
chemistry	36.7	35.0	47.5	42.3
civil_engineering	32.8	33.1	40.9	36.9
computer_science	54.0	56.8	72.1	72.0
construction	30.9	30.9	34.7	30.4
criminal_law	34.5	36.5	39.0	37.5
ecology	34.3	35.1	46.7	44.5
economics	36.9	36.9	43.1	48.5
education	40.0	44.0	58.0	62.0
electrical_engineering	33.0	31.5	33.3	28.4
electronics_engineering	41.9	43.2	50.4	50.1
energy_management	28.8	30.5	33.8	32.7
environmental_science	31.5	29.5	34.1	29.5
fashion	33.8	35.0	43.3	40.9
food_processing	29.6	31.6	38.1	36.6
gas_technology_and_engineering	27.7	27.5	30.8	28.5
geomatics	34.9	36.6	41.6	38.9
health	40.0	44.0	59.0	52.0
industrial_engineer	36.3	35.9	43.1	41.1
information_technology	51.9	51.3	69.0	66.5
interior_architecture_and_design	38.5	39.5	48.3	49.0
korean_history	30.0	24.0	34.0	36.0
law	30.3	31.3	42.9	42.0
machine_design_and_manufacturing	33.2	33.6	40.6	37.9
management	35.5	38.0	57.7	54.4
maritime_engineering	36.7	39.2	44.2	43.0
marketing	57.4	57.7	74.6	74.9
materials_engineering	30.2	30.1	39.4	36.9
math	26.7	29.0	29.7	31.0
mechanical_engineering	29.9	28.6	35.5	30.0
nondestructive_testing	33.0	34.2	42.6	39.0
patent	33.0	31.0	38.0	40.0
political_science_and_sociology	36.0	37.0	55.0	51.7
psychology	28.3	29.9	44.1	41.4
public_safety	30.8	29.4	34.1	30.2
railway_and_automotive_engineering	33.0	32.0	33.7	29.4
real_estate	37.0	37.5	44.5	44.0
refrigerating_machinery	29.0	29.4	33.0	29.9
social_welfare	37.0	37.2	55.1	53.9
taxation	30.5	33.5	42.5	44.0
telecommunications_and_wireless_technology	41.9	41.5	55.3	51.7

Table 24: 5-shot accuracy using the Direct method for Yi (original and chat versions) broken down by category.

Category	QWEN-7B		QWEN-14B		QWEN-72B	
	Org.	Chat	Org.	Chat	Org.	Chat
accounting	9.0	9.0	25.0	15.0	15.0	46.0
agricultural_sciences	28.8	34.3	38.5	24.7	34.1	40.4
aviation_engineering_and_maintenance	23.3	33.6	49.2	19.9	31.9	48.7
biology	15.2	29.0	40.5	15.4	26.5	39.7
chemical_engineering	19.0	32.7	50.8	17.9	28.3	45.2
chemistry	24.3	44.2	54.3	21.2	37.7	50.7
civil_engineering	17.6	31.3	46.5	17.5	31.7	46.7
computer_science	32.0	54.2	75.7	30.6	52.8	76.4
construction	21.7	32.1	38.0	12.4	20.0	26.0
criminal_law	4.5	12.5	40.0	4.0	9.0	36.5
ecology	34.2	46.2	52.4	35.3	45.7	53.1
economics	5.4	10.8	60.0	3.8	9.2	54.6
education	10.0	29.0	71.0	7.0	29.0	74.0
electrical_engineering	22.3	27.7	34.8	18.8	26.9	35.0
electronics_engineering	14.2	30.3	59.3	16.1	32.1	62.9
energy_management	26.4	32.3	40.3	22.1	29.8	38.2
environmental_science	26.4	32.6	38.0	28.0	34.4	41.4
fashion	32.6	42.8	49.6	29.6	41.6	48.7
food_processing	8.3	19.0	45.8	5.7	12.9	36.8
gas_technology_and_engineering	16.7	26.0	39.9	12.0	21.4	31.2
geomatics	22.8	31.8	43.8	19.5	29.4	41.8
health	9.0	30.0	71.0	13.0	28.0	61.0
industrial_engineer	23.6	42.3	49.0	22.1	41.7	47.1
information_technology	38.9	56.5	74.2	24.2	42.1	63.5
interior_architecture_and_design	19.5	37.3	58.8	17.2	34.3	58.6
korean_history	2.0	9.0	37.0	2.0	10.0	30.0
law	6.0	15.6	50.2	6.9	14.0	45.6
machine_design_and_manufacturing	24.4	37.9	51.0	23.0	33.8	48.4
management	8.9	23.7	64.4	8.1	23.1	58.7
maritime_engineering	21.3	40.8	49.8	18.0	31.8	43.3
marketing	37.8	59.7	85.1	37.1	60.2	81.9
materials_engineering	15.1	29.0	50.2	7.2	20.9	37.9
math	18.7	26.7	36.7	20.3	22.7	28.7
mechanical_engineering	12.8	26.1	41.5	14.5	25.4	46.4
nondestructive_testing	27.2	40.9	48.4	26.4	38.7	48.5
patent	7.0	16.0	39.0	4.0	11.0	33.0
political_science_and_sociology	11.7	30.7	62.0	13.3	27.3	56.7
psychology	18.4	31.1	51.5	15.2	30.1	45.4
public_safety	7.9	14.1	40.3	7.5	16.0	41.0
railway_and_automotive_engineering	20.5	31.7	40.1	22.2	31.6	39.2
real_estate	2.0	7.5	53.0	3.5	8.5	45.0
refrigerating_machinery	18.9	29.1	39.4	18.0	27.6	37.2
social_welfare	25.0	41.0	64.7	22.0	38.2	60.1
taxation	3.0	7.5	42.5	4.0	7.5	32.0
telecommunications_and_wireless_technology	39.1	50.0	64.2	36.6	50.7	64.4

Table 25: 5-shot accuracy using the Direct method for QWEN (original and chat versions) broken down by category.

Category	GEMINI-PRO	HCX.	GPT-3.5-TURBO	GPT-4
accounting	44.0	46.0	42.0	71.0
agricultural_sciences	42.5	42.7	34.1	50.2
aviation_engineering_and_maintenance	53.0	49.0	43.5	63.9
biology	46.5	47.9	35.0	51.3
chemical_engineering	51.9	47.9	42.9	61.3
chemistry	50.2	48.8	45.0	64.8
civil_engineering	47.6	45.1	41.2	53.9
computer_science	75.0	78.5	66.1	87.7
construction	37.6	41.9	34.9	46.7
criminal_law	39.0	48.5	32.5	50.5
ecology	52.6	57.3	47.0	59.2
economics	53.1	65.4	40.8	67.7
education	58.0	72.0	40.0	84.0
electrical_engineering	39.1	35.3	34.8	43.2
electronics_engineering	60.2	59.8	52.1	69.9
energy_management	38.1	37.6	33.9	43.9
environmental_science	38.0	36.3	34.8	44.4
fashion	53.0	57.2	46.6	61.7
food_processing	50.1	50.3	39.6	57.4
gas_technology_and_engineering	42.0	42.3	34.5	49.0
geomatics	41.7	49.4	41.8	50.9
health	65.0	72.0	50.0	71.0
industrial_engineer	50.7	50.2	43.3	58.1
information_technology	72.3	73.1	66.3	83.7
interior_architecture_and_design	63.5	69.1	51.0	69.8
korean_history	41.0	42.0	32.0	35.0
law	48.5	58.7	40.2	58.6
machine_design_and_manufacturing	54.4	50.8	43.9	64.9
management	59.7	64.3	51.2	74.1
maritime_engineering	51.2	54.3	45.2	60.8
marketing	81.0	83.1	71.1	89.3
materials_engineering	53.8	52.1	43.5	66.0
math	26.7	26.7	30.3	31.0
mechanical_engineering	48.7	46.3	38.9	57.3
nondestructive_testing	52.9	50.6	42.8	59.9
patent	37.0	52.0	34.0	43.0
political_science_and_sociology	57.7	66.7	47.7	74.0
psychology	47.0	58.7	37.0	61.3
public_safety	41.3	41.0	36.5	51.5
railway_and_automotive_engineering	42.8	41.2	34.7	51.7
real_estate	45.0	53.0	37.0	56.5
refrigerating_machinery	40.7	40.0	33.9	48.1
social_welfare	60.6	61.6	49.6	76.4
taxation	40.0	48.0	33.0	48.0
telecommunications_and_wireless_technology	63.7	63.0	54.8	74.9

Table 26: 5-shot accuracy using the Direct method for GEMINI-PRO, GPT-3.5-TURBO, GPT-4 and HYPERCLOVA X broken down by category.

Category	QWEN-72B-CHAT	HCX.	GPT-3.5-TURBO	GPT-4
accounting	21.7	17.4	19.6	26.1
agricultural_sciences	13.0	14.0	15.0	13.0
aviation_engineering_and_maintenance	21.0	24.0	26.0	38.0
biology	21.0	24.0	15.0	14.0
chemical_engineering	17.0	31.0	26.0	43.0
chemistry	22.0	30.0	29.0	44.0
civil_engineering	17.0	25.0	20.0	16.0
computer_science	25.0	36.0	18.0	25.0
construction	26.0	28.0	18.0	24.0
criminal_law	9.0	24.0	9.0	8.0
ecology	12.0	24.0	16.0	11.0
economics	23.8	33.3	26.2	28.6
education	17.4	26.1	0.0	26.1
electrical_engineering	11.0	24.0	20.0	30.0
electronics_engineering	23.0	20.0	34.0	48.0
energy_management	18.0	15.0	25.0	26.0
environmental_science	16.0	22.0	17.0	27.0
fashion	20.0	29.0	24.0	16.0
food_processing	17.0	24.0	21.0	28.0
gas_technology_and_engineering	19.0	29.0	25.0	31.0
geomatics	18.0	24.0	20.0	24.0
health	8.7	26.1	26.1	21.7
industrial_engineer	13.0	27.0	19.0	22.0
information_technology	28.0	33.0	41.0	46.0
interior_architecture_and_design	21.0	37.0	29.0	24.0
korean_history	11.4	47.7	18.2	9.1
law	13.0	35.0	11.0	17.0
machine_design_and_manufacturing	19.0	32.0	23.0	32.0
management	26.0	24.0	20.0	23.0
maritime_engineering	21.0	27.0	19.0	21.0
marketing	29.0	18.0	17.0	18.0
materials_engineering	21.0	24.0	20.0	24.0
math	18.0	32.0	31.0	51.0
mechanical_engineering	17.0	25.0	20.0	36.0
nondestructive_testing	19.0	23.0	27.0	24.0
patent	18.0	23.5	23.5	11.8
political_science_and_sociology	24.4	27.8	4.4	14.4
psychology	16.0	36.0	14.0	9.0
public_safety	21.0	30.0	13.0	12.0
railway_and_automotive_engineering	12.0	25.0	19.0	29.0
real_estate	10.1	25.8	10.1	14.6
refrigerating_machinery	18.0	26.0	26.0	38.0
social_welfare	13.0	35.0	36.0	51.0
taxation	5.2	26.0	10.4	4.2
telecommunications_and_wireless_technology	25.0	30.0	30.0	38.0

Table 27: 5-shot accuracy using the CoT method for QWEN-72B-CHAT, GPT-3.5-TURBO, GPT-4 and HYPER-CLOVA X broken down by category.

You are a problem rewriter. Your task is to paraphrase the problem and the answer presented below.

Please follow the instructions below:

1. Please paraphrase the problem by rewording it with new expressions and sentence structures.
2. Please do not change the essence of the problem and the answer.
3. Please make sure not to deviate too much from the original content, and maintain the style as much as possible.
4. Please write in coherent Korean. It should sound natural to a native Korean speaker.

Please write "Rewritten Question: <question>" to output your rewritten question without any additional information, and write "Rewritten Answer: <answer>" to output your rewritten answer without any additional information.

There is an example for your reference:

Original Question: (주)한국은 20× 1년 1월 초 A사 지분 상품을 ₩ 10,000 에 매입하면서 매입 수수료 ₩ 500을 현금으로 지급하고, 기타 포괄손익 - 공정가치 측정 금융자산으로 분류하였다. 20× 1년 12월 말 A사 지분 상품의 공정가치가 ₩8,000이라면, 20× 1년 말 (주)한국이 인식할 A사 지분 상품 관련 평가손익은?
Original Answer: 금융자산 평가손실 (기타 포괄손익) ₩2,500

Rewritten Question: 한국 회사가 20× 1년 1월에 A사 지분 상품을 ₩ 10,000에 구매하여 현금으로 ₩ 500의 매입 수수료를 내고, 이를 기타 포괄손익 - 공정가치 측정 금융자산으로 분류했습니다. 20× 1년 12월 말에 A사 지분 상품의 공정가치가 ₩8,000이라면, 20× 1년 말에 한국 회사가 A사 지분 상품과 관련된 평가손익은 얼마입니까?
Rewritten Answer: 한국 회사의 금융자산 평가손실은 ₩2,500입니다.

Original Question: {Question}
Original Answer: {Answer}

Figure 14: Prompt used in paraphrasing for contamination check.

CoT Elicitation Prompt

다음은 {category}에 대한 객관식 질문입니다.
정확한 답을 하기 위해 반드시 웹 브라우징을 활용하시오. 먼저 자세한 정답 추론/해설 과정을 한글로 생성하세요.
그리고나서, 최종 답변은 반드시 다음과 같은 포맷으로 답해야 합니다. '따라서, 정답은 (A|B|C|D)입니다.'

질문: {Question}
선택지:
(A). {A}
(B). {B}
(C). {C}
(D). {D}
정답 해설: 차근 차근 생각해보겠습니다.

Figure 15: Zero-shot CoT prompt used in our CoT exemplar creation.