KMMLU:
Measuring Massive Multitask Language Understanding in Korean

Anonymous ACL submission

Abstract

We propose KMMLU, a new Korean benchmark with 35,030 expert-level multiple-choice questions across 45 subjects ranging from humanities to STEM. Unlike previous Korean benchmarks that are translated from existing English benchmarks, KMMLU is collected from original Korean exams, capturing linguistic and cultural aspects of the Korean language. We test 26 publically available and proprietary LLMs, identifying significant room for improvement. The best publicly available model achieves 50.54% on KMMLU, far below the average human performance of 62.6%. This model was primarily trained for English and Chinese, not Korean. Current LLMs tailored to Korean, such as POLYGLOT-KO, perform far worse. Surprisingly, even the most capable proprietary LLMs, e.g., GPT-4 and HYPERCLOVA X, achieve 59.95% and 53.40%, respectively. This suggests that further work is needed to improve Korean LLMs, and KMMLU offers the right tool to track this progress. We make our dataset publicly available on the Hugging Face Hub and integrate the benchmark into EleutherAI's Language Model Evaluation Harness.

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, as illustrated in Figure 1, naively translating English benchmarks into a target 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 and government or requires

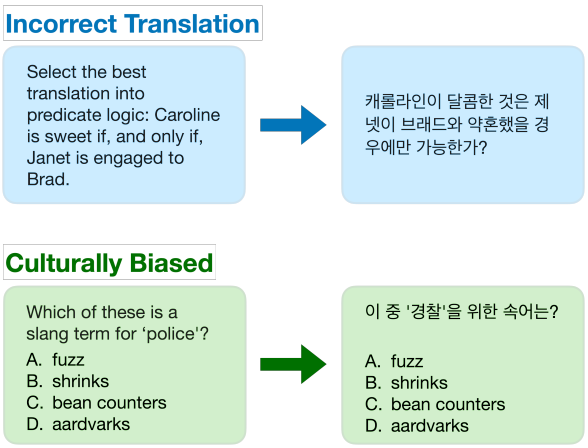

Figure 1: Undesirable questions in translated versions of MMLU.

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 might hint at multilingual proficiency, they often do not fully capture the linguistic or 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 an authentic Korean language without any translated material. Additionally, our questions are *localized* to Korea: they reflect the topics and cultural attitudes of Koreans, rather than Westerners (see Figure 2).

We evaluate 26 different LLMs across 5 categories: (1) Multilingual Pretrained Models (Touvron et al., 2023; 01.AI; Bai et al., 2023); (2) Multilingual Chat Models (Touvron et al., 2023; 01.AI; Bai et al., 2023); (3) Korean Pretrained Models (Ko et al., 2023); (4) Korean Continual Pretrained Models (L. Junbum, 2023b); and (5) Proprietary Models

including those serviced in Korea (OpenAI, 2023; Team et al., 2023; Kim et al., 2021). Our results show significant room for improvement, with GPT-4 scoring the highest at 59.95%, while the average accuracy of human test-takers stands at 62.6%. 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).

Finally, we conduct a detailed analysis to deepen our understanding of how large language models (LLMs) utilize Korean knowledge in question-solving. Initially, we observe that, despite GPT-4's overall excellence, it displays notable 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, a Korean-specific LLM, which scores 44%. Our further analysis reveals that the performance boosts seen in non-Korean LLMs are attributed more to their overall capabilities rather than a deep understanding of Korean, stressing the importance of targeted Korean pre-training to improve their effectiveness in Korea-specific tasks. 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 and reliable Korean explanations.

2 Related Work

2.1 Benchmarks for Large Language Models

Benchmarks are essential for accurately understanding and tracking the evolving capabilities of large language models (LLMs). Traditionally, benchmarks (Rajpurkar et al., 2016; Wang et al., 2019b,a) focused on primary linguistics tasks, 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 ranging from basic mathematics to complex areas like law and computer science, 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, especially Chinese (Li et al., 2023a; Huang et al., 2023; Zeng, 2023).

2.2 Korean benchmarks

Benchmarks for Korean language models have followed a similar path: starting with single-task assessments (Youngmin Kim, 2020), evolving to collections of these tasks (Park et al., 2021), and more recently expanding to include tests of reasoning abilities (Kim et al., 2022) and cultural knowledge (Son et al., 2023). Translated versions of English benchmarks are also widely adopted (Park et al., 2023). However, there is a lack of broad domain benchmarks that facilitate the evaluation of broader Korean proficiency.

3 KMMLU

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	36,777
Split	
Train	208,522
Validation	225
Test	35,030
Total	243,777

Table 1: 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.

	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 소매경영환경의 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 2: Examples of questions from KMMLU categorized by the type of Korean Knowledge required. English translations are added for broader accessibility.

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 with emphasis on 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 Korean License Tests, notably, some of the license tests KMMLU draws from require up to 9 years of industry experience to take the test. This underscores the depth and practicality of the knowledge KMMLU tests. In addition, KMMLU includes questions that require an understanding of cultural, regional and legal knowledge to solve, as shown in Figure 2. For further details, see Table 1.

3.2 Dataset Creation

The dataset comprises questions sourced from 533 different freely available tests, including the Public Service Aptitude Test (PSAT) and various Korean License Tests. These questions are of professional-level difficulty. We conduct several rounds of human validations, randomly sampling 100 instances and applying heuristic rules to eliminate crawling errors, ultimately reducing the benchmark by approximately 34%, from 371,002 to 243,777 questions. Questions with less than four options were excluded, and for those with more than four, excess incorrect choices were randomly removed to ensure exactly four answer choices per question. Before finalizing, a preview version of the benchmark was made available online for two months to allow open-source contributors to identify issues, which we subsequently addressed. Finally, we manually

reviewed questions to remove instances containing potentially copyrighted material to the best of our ability.

We also gathered human accuracy data from actual test-takers when available. We found data on human performance for approximately 90% of the exams in the dataset, with the average human performance being 62.6%. The dataset is divided into a training set, a few-shot development set, and a test set. The few-shot development set contains five questions per subject. The training set, which may be used for hyperparameter tuning or training, comprises 208,522 questions. The test set contains at least 100 examples per subject and includes questions with the lowest human accuracy rates for a total of 35,030 questions. In cases with insufficient human accuracy data, questions were chosen randomly.

3.3 CoT Exemplar Creation

Chung et al. (2022) devise 5-shot of exemplars to test Chain-of-Thought (CoT) reasoning over MMLU (Hendrycks et al., 2020) [1]. Inspired by them, we create 5-shot of CoT exemplars for each subject to test models' reasoning capabilities over our benchmark. However, writing an accurate rationale for expert-level tests with various ranges is a non-trivial problem. Although the ideal solution might be to invite experts for each test, we decide to leverage assistance from various LLMs, 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 [2].

First, we elicit rationale (reasoning) 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 according to heuristics, 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 D for more details.

3.4 KMMLU Hard

KMMLU stands out for its considerable size, comprising 35,030 questions across its test subsets, surpassing MMLU (Hendrycks et al., 2020) and CMMLU (Li et al., 2023a), which contain approximately 10,000 instances each. Given the substantial resources required to run the full subset, we introduce a more manageable subset termed KMMLU Hard. This subset selects questions that at least one of the proprietary models fails to answer correctly in the direct approach. It encompasses a total of 4,104 questions, with categories ranging from a minimum of 23 to a maximum of 100 questions.

4 Experimenal 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. This method is widely employed throughout this paper as it applies to proprietary and weight-available LLMs. Second, Chain-of-Thought (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. 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 all experiments with 8-bit quantization.

4.2 Models

In our study, to provide a comprehensive overview of existing LLMs in answering expert-level Korean questions, we evaluate 26 models varying in size, language, and training phase (pre-trained or supervised fine-tuned).

The 26 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) (01.AI);

2. Multilingual Chat Models: Chat versions of LLAMA-2, QWEN, and YI; 2023);

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

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

3. Korean Pretrained Models: POLYGLOT-KO (1.3B, 3.8B, 5.8B, 12.8B) (Ko et al., 2023);

4. Korean Continual Pretrained Models:LLAMA-2-KOEN (L. Junbum, 2023b) and YI-KOEN (L. Junbum, 2023a); and

5. Proprietary Models: GPT-3.5, GPT-4 (OpenAI, 2023), GEMINI PRO (Team et al., 2023) and HYPERCLOVA X (Kim et al., 2021).

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 B and Table 7.

5 Evaluation Results

Pretraining Compute We compare the performance of 24 LLMs using the Direct method in Table 2. We observe a clear trend across pretrained and fine-tuned models, where those with a larger computing budget exhibit superior performance[3]. This scaling behavior indicates that increased computing resources - reflected in the number of parameters and the size of 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 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 3, where Polyglot-Ko's significantly lower training budget compared to its counterparts is evident.

Fine-Tuning In Table 2, we also observe that 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.

[3]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.

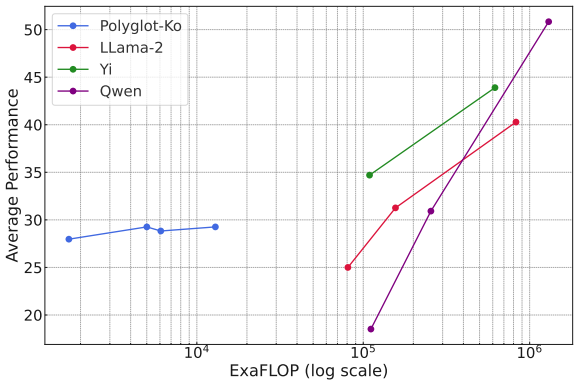

Figure 3: Average performance of POLYGLOT-KO, LLAMA-2, YI, and QWEN models versus pretraining budget.

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.

No Curse of Multilinguality at Scale The "curse of multilinguality" (Conneau et al., 2019; Pfeiffer et al., 2022) refers the apparent decrease in model capabilities when models are trained on multilingual corpora. While the curse can be severe for small models, it has been widely noted as generally ameliorated by scaling for masked language models (Goyal et al., 2021; Pfeiffer et al., 2022). Empirically, this seems to not be the case for the decoder-only BLOOM (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). By contrast, we see no evidence of a curse of multilinguality here, with large multilingual models like LLAMA-2, YI, and QWEN substantially outperforming the monolingual POLYGLOT-KO. While the multilingual models have been trained for an order of magnitude more tokens than POLYGLOT-KO, we view it as implausible to suppose that they've seen more text than POLYGLOT-KO, as none of

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**
Korean Continual Pretrained Models					
LLAMA-2-KOEN-13B	35.32	34.19	31.43	33.85	33.71
YI-KOEN-6B	**40.69**	**39.52**	**40.50**	**41.60**	**40.55**
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 2: Average accuracy(%) calculated using the Direct method in a 5-shot setting across the entire test set. We report the macro-average accuracy across subjects within each category. 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 8-12 for detailed results.

the papers is Korean even mentioned in the discussion of the training data. While many papers have observed that large models can be good at multiple languages, as far as we are aware this is the first work to explicitly document evidence that the curse of multilinguality goes away as decoder-only models are scaled.

Comparing Disciplines In evaluations on KMMLU, we observe LLMs demonstrate relatively balanced performance across the four categories: STEM, Applied Science, HUMSS, and Others. However, LLMs have considerable knowledge shortfalls subject-wise. Notably, GPT-4's performance ranges markedly, achieving its highest score in Marketing at 89.3% while performing as low as 31.0% in Math. Regarding average performance, LLMs excel in areas like Marketing, Computer Science, Information Technology, and Telecommunications and Wireless Technology. Conversely, they generally show weaker performance in subjects requiring specific cultural or regional knowledge. For instance, LLMs perform worst in Korean History, followed by Math, Patent, Taxation, and Criminal Law. Notably, subjects like Patent, Taxation, and Criminal Law demand an understanding of Korean legal systems and according statutory interpretation, suggesting a potential area for improvement in LLMs' handling of region-specific content. For a full breakdown by topic area, see Figure 10 in the Appendix F.

6 Analysis

We conduct detailed analyses to gain a thorough understanding of the KMMLU benchmark and the varying performance of LLMs across different contexts. Initially, we compare the cultural representation in translated versions of MMLU with KMMLU. Subsequently, we study the Korea-specific questions included in the KMMLU benchmark. We also investigate the reasons behind the underperformance of POLYGLOT-KO, a model specifically pre-trained on Korean corpora, compared to its multilingual counterparts. Lastly, we assess the effectiveness of chain-of-thought prompts in enhancing model performance.

6.1 Analysis of Korea-Specific Instances in KMMLU

In this work, we source the KMMLU benchmark exclusively from exams crafted in Korean. 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 MMLU and 450 questions from KMMLU—two authors evaluated each question.

In Figure 2, we categorized questions that necessitate Korean knowledge into four distinct types. The *Cultural* type includes questions that require an understanding of Korean history and societal norms. The *Regional* type involves questions that demand geographical knowledge of Korea. The *Legal* type involves questions concerning Korea's legal and governmental systems. The *Others* type includes questions that need Korean knowledge but do not fit into the previous categories. We manually annotate a subset from the KMMLU benchmark. However, it should be noted that this subset is not comprehensive of all culturally specific questions within the KMMLU. Instead, we employ a keyword-based filtering approach to collect 1,455 potential questions and manually classify them as per the designated categories.

The performance of selected models on this subset compared to questions that are identified as *not* requiring any Korea-specific knowledge is detailed in Table 3. Our analysis reveals three notable patterns. First, continual pretraining of LLAMA-2-7B on Korean text for 60 billion tokens enhances its performance on Korea-specific questions. Second, models that undergo training specifically on Korean text consistently perform better on Korean-specific questions. Finally, HYPERCLOVA X outperforms GPT-4 in addressing Korean-specific questions, highlighting the deficiencies in GPT-4's understanding of Korean-specific content.

Model	Korea-Specific	General
# of Instance	645	810
LLAMA-2-7B	20.66	**23.95**
LLAMA-2-KOEAN-7B	**24.13**	20.37
LLAMA-2-13B	28.55	**29.88**
POLYGLOT-12.8B	**28.39**	27.90
QWEN-72B	44.79	**52.34**
GEMINI-PRO	42.94	**48.64**
GPT-3.5-TURBO	39.59	**42.47**
GPT-4	54.89	**60.49**
HYPERCLOVA X	**55.21**	54.32

Table 3: Average accuracy of selected models on questions that require knowledge specific to Korea compared to questions that don't. For each model, the larger score is **bolded**.

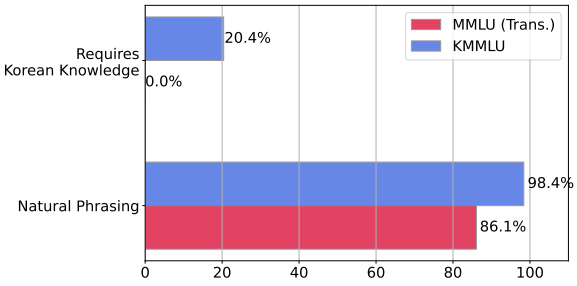

Figure 4: Comparison of MMLU (translated to Korean via GPT-4) and KMMLU (ours) in terms of naturalness and the necessity of Korean knowledge to solve.

Figure 4 reveals a difference in how the two datasets 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, inherently lacks questions about Korean culture. Conversely, 20.4% of KMMLU requires understanding Korean cultural practices, societal norms, and legal frameworks. This disparity is evident in categories like *"high_school_government_and_politics"* in MMLU, which lean heavily towards U.S.-centric content, assuming familiarity with the American governmental system, and the *"miscellaneous"* category, which presupposes knowledge of American slang,

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 Chain-of-Thought (CoT) (Wei et al., 2022). Please see Table 13 for detailed results.

underscoring the cultural bias embedded within the dataset. However, it is crucial to note that while 20.4% of KMMLU's content is culturally specific, the remainder broadly assesses a language model's general knowledge, including areas such as mathematics, which are universally applicable and not tied to any particular country's knowledge base.

6.2 Why Do Korean Models Show Lower Performance on KMMLU?

In Figure 5, we examine the performance of publicly accessible models on the MMLU (Hendrycks et al., 2020) and KMMLU benchmarks. For models predominantly trained on non-Korean texts, we observe that performance on KMMLU improves with model scaling; however, the performance margin between MMLU and KMMLU remains consistent, suggesting that the observed improvements are attributed to enhancements in capabilities other than Korean. Conversely, POLYGLOT-KO, despite its lower overall performance due to a smaller training budget, demonstrates superior proficiency in Korean compared to English.

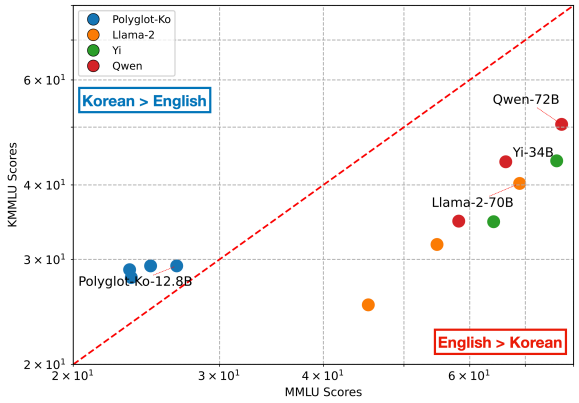

Figure 5: Comparing model performance on MMLU and KMMLU. Regression analysis on LLAMA-2, QWEN, and YI models exhibits a slope of 0.51.

6.3 Can Chain-of-Thought prompting improve performance on KMMLU?

We employ a few-shot Chain-of-Thought (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 [4]. 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).

7 Conclusion

In this study, 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.

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

8 Limitations

In this study, we put our greatest effort into creating a benchmark that offers extensive coverage and is suitable for evaluating proficiency in Korean. Nevertheless, there are some limitations that future research will need to address.

Firstly, although the KMMLU benchmark encompasses 45 categories, it does not include questions from the medical and financial domains. This is primarily because exams in these fields within Korea often do not release their past tests and questions. Additionally, our benchmark lacks Korean reading comprehension questions. Although we initially gathered a substantial number of these questions, most were excluded due to concerns over copyright issues, especially those involving segments of Korean literature.

Secondly, the recent surge of highly 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, for assessing generative capabilities. Future efforts should aim to expand Korean benchmarking tools to include assessments of generative abilities.

References

01.AI. Building the next generation of open-source and bilingual llms. https://huggingface.co/01-ai/Yi-34B. Accessed: 2024-01-03.

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.

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.

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*.

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*.

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*.

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*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424.

Dohyeong Kim, Myeongjun Jang, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks.

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.

Taekyoon Choi L. Junbum. 2023a. beomi/yi-ko-6b.

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

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*.

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.

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.

OpenAI. 2023. Gpt-4 technical report.

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*.

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.

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*.

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.

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*.

Seungyoung Lim;Hyunjeong Lee;Soyoon Park;Myungji Kim Youngmin Kim. 2020. KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension. *Journal of KIISE*, 47:577–586.

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 Additional Analysis

A.1 Do Machines Also Err Where Humans Often Do?

In Figure 6, 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.

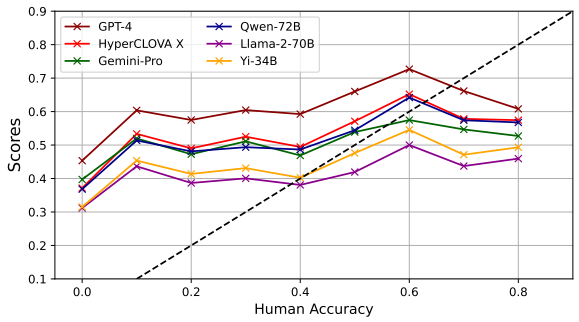

Figure 6: Comparison of model performance and human accuracy. Model performance is calculated using the Direct method in a 5-shot setting.

A.2 Do Machines Handle Problems with Negation Effectively?

Table 5 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., 2020). For example, the math subset, which is the most challenging subset for most LLMs, does not include any negated questions. Furthermore, Table 6 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 5: 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 6: Ratio of Negated Questions in each category.

B 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 are pre-trained 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 pre-trained 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 (01.AI). 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 pre-trained on a

Model	# Params	Access	Language
English-Centric / Bilingual Pretrained Models			
LLAMA-2 (Touvron et al., 2023)	7B, 13B, 70B	Weights Available	En
YI (01.AI)	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 (01.AI)	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-KOEN (L. Junbum, 2023a)	13B	Weights Available	En / Ko
YI-KOEN (L. Junbum, 2023b)	6B	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 (Kim et al., 2021)	undisclosed	API	-

Table 7: Overview of the 31 LLMs evaluated in this paper.

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 pre-trained 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 (Kim et al., 2021). HYPERCLOVA X, developed by NAVER, is a bilingual language model proficient in both English and Korean.

C Prompting Format

For evaluation, we use the following prompting formats.

Direct Evaluation Prompt

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

Figure 7: 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 8: Prompt used in our CoT Evaluation.

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). {option_A}
(B). {option_B}
(C). {option_C}
(D). {option_D}
정답 해설: 차근 차근 생각해보겠습니다.

Figure 9: Zero-shot CoT prompt used in our CoT exemplar creation.

D More details for CoT Exemplar Creation

We use the zero-shot CoT prompt of Figure 9 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.

E License

The KMMLU benchmark is released under a CC-BY-NC-ND license. This license prohibits remixing, redistribution, and commercial use of the dataset.

F Evaluation Results

In this section, we present the results of our evaluation, broken down by category for each model assessed. Tables 8-12 include results using the Direct method. Table 13 presents the results evaluated using the CoT method. Figure 10 presents a comparative performance analysis between the most capable Korean model, HYPERCLOVA X, and GPT-4.

Figure 10 presents a comparative performance analysis between the most capable Korean model, HYPERCLOVA X, and GPT-4 across each discipline. Detailed numerical results are provided in Appendix 10. The comparison reveals that, in most subjects, GPT-4 surpasses HYPERCLOVA X, with the performance margin varying significantly – from a high of 22.0% in Accounting to a narrow 0.5% in Taxation. 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.

Category	Polyglot-Ko				Yi-KoEN	Llama-2-KoEN
	1.3B	3.8B	5.8B	12.8B	6B	13B
accounting	30.0	32.0	32.0	30.0	38.0	23.0
agricultural_sciences	27.0	30.3	30.1	32.0	32.7	29.7
aviation_engineering_and_maintenance	30.2	29.7	29.9	30.7	36.1	31.1
biology	24.0	26.7	28.6	25.3	32.1	28.7
chemical_engineering	25.3	27.9	24.7	24.7	36.2	34.4
chemistry	30.3	25.2	26.0	29.2	40.8	30.3
civil_engineering	27.4	31.8	31.9	34.3	38.0	32.9
computer_science	32.1	35.9	34.8	33.9	61.5	50.6
construction	33.6	31.0	31.7	32.0	34.7	29.8
criminal_law	26.0	29.0	29.5	28.5	31.5	30.0
ecology	28.7	29.4	31.8	32.7	45.2	35.8
economics	23.8	26.2	24.6	24.6	41.5	40.8
education	23.0	20.0	24.0	25.0	53.0	41.0
electrical_engineering	29.3	32.5	32.0	32.6	34.9	33.8
electronics_engineering	30.5	30.0	35.2	33.3	47.1	38.5
energy_management	28.8	26.5	24.5	26.9	30.0	26.3
environmental_science	26.1	32.9	27.3	30.9	33.9	31.8
fashion	27.0	29.5	29.2	29.8	46.1	36.6
food_processing	27.3	31.8	33.5	29.4	36.1	28.8
gas_technology_and_engineering	31.9	30.9	30.2	30.9	32.5	27.3
geomatics	29.2	30.0	31.1	31.0	41.6	37.2
health	26.0	32.0	27.0	25.0	52.0	36.0
industrial_engineer	27.4	32.3	33.1	31.2	43.0	35.1
information_technology	34.2	34.1	34.0	30.8	57.1	47.2
interior_architecture_and_design	32.4	29.6	29.7	31.8	47.3	39.9
korean_history	34.0	26.0	25.0	31.0	33.0	24.0
law	26.0	24.2	24.4	23.9	41.8	32.2
machine_design_and_manufacturing	28.7	34.0	26.9	30.3	39.9	37.7
management	27.6	27.7	27.7	28.0	43.7	33.2
maritime_engineering	24.8	31.5	26.7	26.5	44.0	36.3
marketing	24.4	30.6	26.4	33.5	69.6	54.1
materials_engineering	30.9	30.2	30.1	26.9	39.8	33.3
math	30.0	21.3	20.0	24.7	24.0	27.7
mechanical_engineering	24.2	31.5	27.1	26.9	38.0	33.8
nondestructive_testing	26.4	32.1	34.2	30.3	39.0	34.2
patent	29.0	23.0	22.0	31.0	32.0	26.0
political_science_and_sociology	25.7	25.7	25.7	25.7	41.7	31.0
psychology	26.5	25.9	27.7	25.9	40.1	29.7
public_safety	28.5	30.7	31.5	31.3	32.1	32.6
railway_and_automotive_engineering	23.6	29.0	28.9	26.8	34.7	30.6
real_estate	27.0	27.5	29.5	32.0	45.0	30.0
refrigerating_machinery	27.0	28.9	29.7	28.3	30.0	28.8
social_welfare	25.3	28.9	30.0	28.8	44.7	33.9
taxation	29.0	27.0	23.5	26.5	36.5	27.0
telecommunications_and_wireless_technology	28.6	33.9	34.1	32.2	52.4	44.2

Table 8: 5-shot accuracy using the Direct method for Polyglot-Ko, Yi-KoEN-6B, and Llama-2-KoEN-6B 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 9: 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 10: 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 11: 5-shot accuracy using the Direct method for QWEN (original and chat versions) broken down by category.

Category	GEMINI-PRO	GPT-3.5-TURBO	GPT-4	HYPERCLOVA X
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 12: 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	HYPERCLOVA X	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 13: 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.

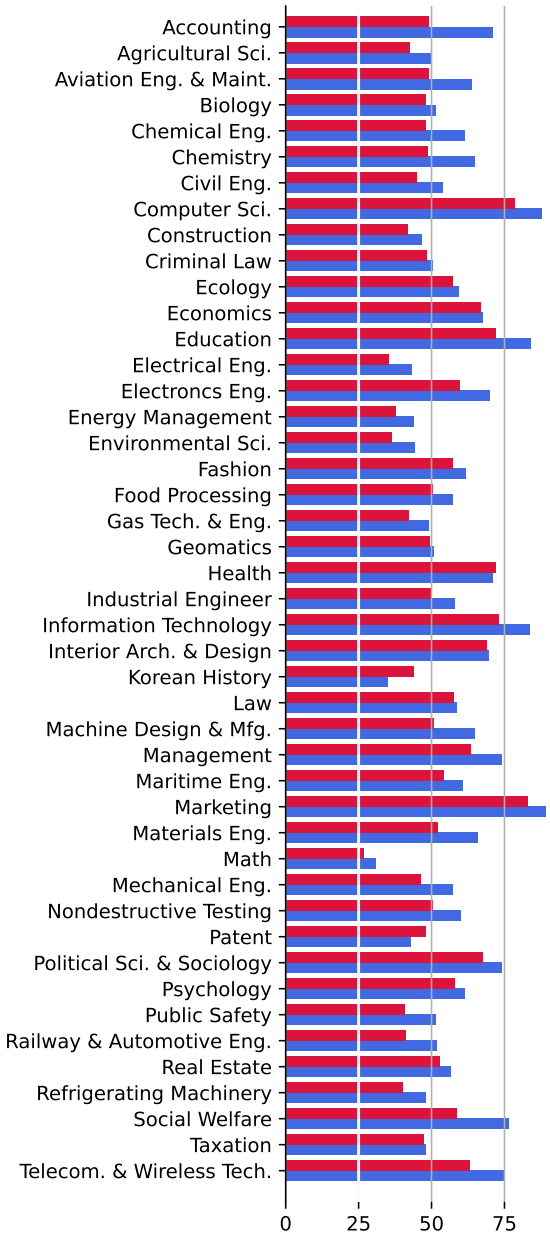

Figure 10: Comparison of GPT-4 and HYPERCLOVA X using the Direct method in a 5-shot setting. GPT-4 in Blue and HYPERCLOVA X in Red.