HAE-RAE Bench: Evaluation of Korean Knowledge in Language Models

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee
Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim
EleutherAI, OnelineAI, MODULABS
spthsrbwls123@yonsei.ac.kr

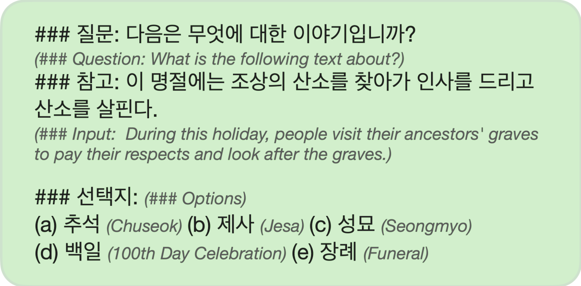
Abstract

Large language models (LLMs) trained on massive corpora demonstrate impressive capabilities in a wide range of tasks. While there are ongoing efforts to adapt these models to languages beyond English, the attention given to their evaluation methodologies remains limited. Current multilingual benchmarks often rely on back translations or re-implementations of English tests, limiting their capacity to capture unique cultural and linguistic nuances. To bridge this gap for the Korean language, we introduce the **HAE-RAE Bench**, a dataset curated to challenge models lacking Korean cultural and contextual depth. The dataset encompasses six downstream tasks across four domains: vocabulary, history, general knowledge, and reading comprehension. Unlike traditional evaluation suites focused on token and sequence classification or mathematical and logical reasoning, the HAE-RAE Bench emphasizes a model's aptitude for recalling Korean-specific knowledge and cultural contexts. Comparative analysis with prior Korean benchmarks indicates that the HAE-RAE Bench presents a greater challenge to non-Korean models by disturbing abilities and knowledge learned from English being transferred.

Keywords: Multilingual Evaluation, Cultural Bias

1. Introduction

Over time, LLMs and benchmark datasets have evolved in tandem, continually becoming more sophisticated and challenging, recognizing their reciprocal relationship. Despite the pivotal role played by benchmark datasets in advancing the capabilities of LLMs, multilingual evaluation tools remain primarily limited. Existing evaluation efforts often rely on translated versions of English datasets (Shi et al., 2022) or translation-specific benchmarks such as WMT 21 (Akhbardeh et al., 2021). While providing some insights into the models' performance across languages, this approach fails to fully capture the intricacies, nuances, and knowledge specific to each linguistic context.



질문: 다음은 무엇에 대한 이야기입니까?
(### Question: What is the following text about?)
참고: 이 명절에는 조상의 산소를 찾아가 인사를 드리고 산소를 살핀다.
(### Input: During this holiday, people visit their ancestors' graves to pay their respects and look after the graves.)

선택지: *(### Options)*
(a) 추석 *(Chuseok)* (b) 제사 *(Jesa)* (c) 성묘 *(Seongmyo)*
(d) 백일 *(100th Day Celebration)* (e) 장례 *(Funeral)*

Figure 1: Example instance from the HAE-RAE Bench. English translations are added for broader accessibility.

Some of the existing efforts to evaluate language models in Korean include Korean-NLI & STS (Ham et al., 2020), KLUE (Park et al., 2021), and KoBEST (Kim et al., 2022). Korean-NLI & STS is derived from machine and human translations of English datasets for natural language inference (NLI) and semantic textual similarity (STS). Accordingly, they hardly capture the unique nuances of the Korean language. KLUE is a Korean version of the GLUE benchmark (Wang et al., 2018), which supports a variety of tasks, including NLI, STS, and topic classification. Unfortunately, its adoption was limited due to its relatively simple tasks. The latest benchmark, KoBEST, is designed to assess a language model's ability to address questions that require advanced reasoning, like understanding passages of time or causality. However, with the advent of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023) and conversational agents built upon them, there is an increasing need to evaluate the cultural knowledge of language models to ensure they converse with native speakers without sounding incoherent. To address this issue, we introduce the **HAE-RAE Bench**, a Korean benchmark dataset originally crafted to capture culture-specific nuances inherent to the Korean language.

We evaluate ten language models including, Polyglot-Ko (Ko et al., 2023), UMT5 (Chung et al., 2023), Llama-2 (Touvron et al., 2023), GPT-3.5-Turbo and GPT-4 (OpenAI, 2023). Evaluation results reveal that multilingual LLMs suffer in solving the HAE-RAE Bench compared to Polyglot-Ko, a native language model trained on Korean from scratch. Furthermore, we our results hint that In-Context Learning (ICL) may be insufficient in steering a LLM to align with a specific culture. HAE-RAE Bench is publicly available for future research.[1]

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

LREC-COLING 2024, pages 7993–8007

2. Related Work

2.1. Language Model

Since the introduction of the Transformer architecture (Vaswani et al., 2017) and early derivatives like BERT (Devlin et al., 2018) and GPT (Radford et al.), research in English language models has expanded rapidly. With their instruction-following capabilities, InstructGPT (Ouyang et al., 2022) and Flan-T5 (Chung et al., 2022) further invigorated this interest. While most of these models primarily focus on English, there are notable exceptions for Chinese, with Qwen (QwenLM, 2023), Baichuan (Yang et al., 2023), and GLM (Zeng et al., 2022). Efforts aimed at narrowing the disparity in progress between English and other languages include:

1. Building language-specific models from scratch, such as Polyglot-Ko(Korean) (Ko et al., 2023), HyperCLOVA(Korean) (Kim et al., 2021), Japanese StableLM(Japanese) (StabilityAI, 2023), and ruGPT(Russian) (ai forever, 2023);

2. Developing multilingual models like BLOOM (Scao et al., 2022), MT5 (Xue et al., 2020), and UMT5 (Chung et al., 2023);

3. Adapting English models for other languages, as seen with Sabiá (Pires et al., 2023) and Chinese-LLaMA (Cui et al., 2023).

Following the advancement of multilingual language models, a critical research question arises: "How should the language-specific capabilities of these models be evaluated?" This underscores the necessity for benchmarks specifically curated to assess the multilingual ability of LLMs.

2.2. Multilingual Evaluation

Multitask benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) were introduced along with the English language models. Once these were saturated, they were followed by even bigger benchmarks such as MMLU (Hendrycks et al., 2020) and Big BENCH (Srivastava et al., 2022). Non-English evaluation research has mirrored this trend, predominantly through translation or re-implementation of existing English benchmarks. Examples include JGLUE (Kurihara et al., 2022), KLUE (Park et al., 2021), and CMMLU (Li et al., 2023), which are Japanese and Korean adaptations of GLUE and Chinese re-implementation of MMLU, respectively.

However, these benchmarks fall short of capturing the native knowledge encoded in the parameter of LLMs. This highlights the need for evaluation suites curated to assess the cultural context of a model. Recent research in this direction is BHASA (Leong et al., 2023), which aims at gauging the cultural depth of language models in Southeast Asian languages. Nonetheless, limitations are apparent: only 34 questions for Indonesian and 28 for Tamil in the entire dataset specifically address cultural representation tasks. In this paper, we introduce the **HAE-RAE Bench**, an evaluation set of 1.5K questions curated to assess Korean-specific knowledge in language models.

2.3. Korean Evaluation

Korean language model evaluation is also a field of interest, with resources emerging after English and Chinese. Examples include Korean-NLI & STS (Ham et al., 2020), KorFin-ASC (Son et al., 2023), KLUE (Park et al., 2021), and KoBEST (Kim et al., 2022). Korean-NLI & STS are based on translations of English datasets for natural language inference (NLI) and semantic textual similarity (STS), potentially missing Korean nuances. KorFin-ASC is derived from Korean news but concentrates on sentiment classification, specifically in the financial domain. KLUE mirrors the GLUE benchmark with Korean, covering tasks like Topic Classification, Semantic Textual Similarity, Natural Language Inference, and more. However, either translated or task-oriented, these benchmarks fail to fully assess language-specific models. Recent research include KoBEST (Kim et al., 2022), which features Korean re-implementations of HellaSwag (Zellers et al., 2019), COPA (Gordon et al., 2012), BOOLQ (Clark et al., 2019), SentiNeg (Savanur and Sumathi, 2023), and WiC (Pilehvar and Camacho-Collados, 2018). However, LLMs trained in massive amounts of English corpora may excel in these evaluation suites by leveraging their general problem-solving capabilities derived from the scale. The HAE-RAE Bench distinguishes itself from the above-mentioned Korean benchmarks by evaluating the depth of knowledge encoded in language models instead of their natural language understanding or reasoning abilities.

3. HAE-RAE Bench

The design principle behind the HAE-RAE Bench significantly differs from earlier Korean benchmark suites like KLUE (Park et al., 2021) or KoBEST (Kim et al., 2022). While previous benchmarks focused on evaluating natural language understanding or reasoning abilities, HAE-RAE emphasizes the depth of knowledge itself. This change is driven by the emergence of LLMs and conversational agents or search engines built on them. We posit that knowledge of Korean vocabulary, culture, geography, and history might be as crucial, if not more so, than traditional NLU tasks such as token

Category	Type	Total # of Question	Total # of Unique Morpheme	Avg. # of Words (std) per question	Avg. # of Words (std) per passage	Fertility Rate (std) Polyglot-Ko	Fertility Rate (std) Llama-2
Loan Words	{Q}	169	960	5.1 (0.3)	-	3.9 (0.3)	6.7 (0.6)
Rare Words	{Q}	405	2721	13.0 (3.4)	-	3.1 (0.3)	6.1 (0.4)
Standard Nomenclature	{Q}	153	1018	8.3 (0.5)	-	3.2 (0.4)	6.4 (0.6)
Reading Comprehension	{Q, P}	447	5825	7.1 (1.8)	69.6 (44.6)	2.5 (0.4)	6.0 (0.5)
General Knowledge	{Q, P}	176	2099	7.0 (3.0)	9.1 (13.6)	3.4 (0.6)	6.4 (0.9)
History	{Q}	188	1595	12.8 (3.5)	-	3.3 (0.4)	6.3 (0.6)

Table 1: HAE-RAE Bench Statistics.

or sequence classification in conversational situations. Accordingly, the resulting benchmark encompasses six downstream tasks: Loan Words(LW), Standard Nomenclature(SN), Rare Words(RW), General Knowledge(GK), History(HI) and Reading Comprehension(RC).

Statistics for the HAE-RAE Bench dataset are provided in Table 1. “Type" indicates the structure of the question. “Q" denotes that the instance comprises a question with multiple choices, while “Q, P" indicates the inclusion of an associated passage. We also present the fertility rate of the dataset, tokenized using different models: Polyglot-Ko (Ko et al., 2023), UMT5 (Chung et al., 2023), and Llama-2 (Touvron et al., 2023). The fertility rate (Ács, 2019) calculates the average number of sub-tokens generated per word. A fertility rate of 1 implies that the tokenizer’s vocabulary encompasses every word in the text. A higher fertility rate may suggest potential challenges for the tokenizer in grasping context. In our observation, the fertility rate increases for models with less emphasis on Koreans. To assess the relative complexity of the vocabularies in the HAE-RAE Bench, we compared its fertility rate with that of KoBEST, as shown in Table 2. Using the polyglot-ko tokenizer, the fertility rates for HAE-RAE Bench and KoBEST are 3.0 and 2.7, respectively. This suggests that the HAE-RAE Bench comprises less common words. Examples for each subset of the datasets are presented in section 10.3.

Dataset	Polyglot-Ko	UMT5	Llama-2
HAE-RAE Bench	3.00 (0.38)	3.56 (0.37)	6.33 (0.59)
KoBEST	2.70 (0.34)	3.39 (0.45)	6.44 (0.75)

Table 2: Fertility rate (std) of HAE-RAE Bench and KoBEST.

3.1. Loan Words

Task Description Loan words refer to vocabularies directly adopted from foreign languages. In South Korea, the National Institute of Korean Language (NIKL) [2] formulates corresponding Korean terms for such words. In this task, language models are given a foreign word along with five choices and are tasked to identify the correct Korean equivalent.

Creation The pairs of foreign words and their Korean equivalents are sourced from NIKL. Some Korean terms are infrequently used, either because the foreign word has been entrenched in society for a long time or because it’s a recent addition and not yet widely recognized. To ensure we focus on reasonably common terms, we filter the list to only include words present in both “Naver Knowledge Encyclopedia" [3] and “Daum Encyclopedia" [4], the two most widely used online encyclopedias in Korea. From the refined list, we randomly sampled 200 vocabularies. Incorrect options were selected from the remaining terms based on their Levenshtein distance (Levenshtein et al., 1966) to the correct answer. While Levenshtein distance may initially seem to prioritize syntax over semantics, it effectively captures both in Korean. “Han" (Chinese logograms) constitute about 55% of the Korean vocabulary. Accordingly, words with the same Korean letter have related meanings. Moreover, the structure of the Korean language involves compounding, where multiple “roots" (fundamental word units) merge to form new words. Consequently, words sharing similar meanings often include the same “root", making the syntactic and semantic distances in Korean words largely aligned. Finally, we applied a Levenshtein distance threshold of 3, omitting samples with fewer than four incorrect options meeting this criterion.

3.2. Standard Nomenclature

Task Description Standard Nomenclatures, published by NIKL, are unified terminology for domain-specific words. In this task, language models are presented with a specialized term along with five

[2] https://www.korean.go.kr/front/main.do

[3] https://terms.naver.com

[4] https://100.daum.net

options, with the objective of identifying the official term endorsed by NIKL.

Creation Pairs of domain-specific words and their official terms are collected from NIKL. We follow the approach in 3.1 to create questions.

3.3. Rare Words

Task Description The Rare Words task aims to probe language models' understanding of challenging vocabulary. Given a definition and five words, models are tasked with selecting the word that best suits the provided definition.

Creation We sourced pairs of definitions and challenging words from past episodes of the TV program "Woorimal Battle," [5] known for its challenging Korean vocabulary quizzes. We follow the approach in 3.1 to create questions.

3.4. General Knowledge

Task Description General Knowledge evaluates the model's familiarity with various aspects of the Korean cultural, using five-option multiple-choice questions.

Category	# of instances	Average Length
Tradition	17	35.2
Law	10	32.2
Geography	49	46
Korean Pop	50	42.3
Korean Drama	50	36.7

Table 3: The number of data instances for each category.

Creation We first identify five primary categories for general knowledge: tradition, law, geography, Korean pop, and Korean drama. We then crowd-sourced questions to fit these subcategories. Then, we remove overlapping, factually incorrect, and questions that fail to align with the defined category. We also conduct additional investigations to ensure that no superficial artifacts are not inadvertently introduced. Basic statistics for each subcategory are illustrated in Table 3.

Investigation Following Kaushik and Lipton (2018), we examined the performance of Polyglot-Ko-12.8B using question-only (Q-only) and context-only (C-only) settings. Polyglot-Ko-12.8B achieved scores of 25.57% and 23.86% for Q-only and C-only, respectively, while the complete setting outperformed both with a score of 32.95%. Although the Q-only and C-only settings are within 10% accuracy of the complete setting, it is worth noting that the model's lower bound is set at 20%. Therefore, we conclude that the dataset was crafted correctly to require both question and context to answer.

Metric	Full	Q-Only	C-Only	△ (*min*)
Acc	**32.95**	25.57	23.86	-7.38
Macro F1	**32.01**	24.35	23.64	-7.56

Table 4: Performance of Polyglot-Ko-12.8B on General Knowledge with truncated inputs.

3.5. History

Task Description The history task assesses the model's understanding of historical events. Presented with a question and five options, the model must identify the correct answer.

Creation We first sourced web pages tagged "Korean history" from Namuwiki, Korea's equivalent to Wikipedia, and randomly selected 40 pages. From each page, authors manually crafted five questions. We refer Malaviya et al. (2022) and filtered out 12 questions with overlapping tokens between questions and answers. Moreover, to investigate potential biases introduced while creating the wrong options, we analyzed two simple linguistic indicators: the probability of the longest option being correct was 21.53%, and for the shortest option, it was 17.01%. Through this process, we remove overly simplistic questions and investigate for potential biases.

3.6. Reading Comprehension

Task Description Reading comprehension tasks involve providing paired questions and passages along with four options. The materials for our Reading Comprehension (RC) tests were sourced from the Korean Language Ability Test (KLAT), an exam designed to evaluate proficiency in Korean as a second language.

Creation The tests were gathered from sample materials publicly released by the Korea Educational Testing Service (KETS). We omitted questions that required interpreting images. The sourced KLAT is divided into four proficiency tiers: three that correspond to the Common European Framework of Reference (CEFR) levels—A (beginner), B (intermediate), and C (advanced)—plus an introductory level below A for absolute beginners.

[5] https://program.kbs.co.kr/1tv/culture/woorimal/pc/index.html

3.7. Quality Check

To further filter the collected questions, we reviewed the entire dataset and conducted factual verification using online resources. In this process, we manually corrected 23 questions with labeling or crawling errors.

4. Evaluation Settings

4.1. Language Models

We evaluated ten models across varying sizes from four model families. From openly available models we selected (1) Korean-focused models: Polyglot-ko-1.3B/3.8B/5.8B/12.8B (Ko et al., 2023), (2) Multilingual models: UMT5-XL/XXL (Chung et al., 2023), and (3) English-centric models: Llama-2-7B/13B (Touvron et al., 2023). For analysis, we excluded models that do not disclose statistics on the number of pretrained Korean tokens. This leaves out Falcon (Penedo et al., 2023) and BLOOM (Scao et al., 2022) from our experiments. Additionally, we included GPT-3.5-Turbo and GPT-4 in our evaluation to gauge the efficacy of the HAE-RAE Bench in assessing state-of-the-art proprietary LLMs.

Polyglot-Ko Ko et al. (2023) is available in four sizes: 1.3B, 3.8B, 5.8B, and 12.8B, all built using the GPT-NeoX codebase. It was pretrained on a Korean-only corpus, with sizes ranging between 167B to 212B tokens. Despite its smaller pretraining budget compared to similar-sized English models, Polyglot-Ko achieved state-of-the-art results on KoBEST, a benchmark comprising five Korean language understanding and reasoning tasks (Kim et al., 2022).

UMT5 Chung et al. (2023) was originally trained in five sizes: small (77M), base (250M), large (800M), xlarge (3B), and xxlarge (13B), closely following the mT5 architecture(Xue et al., 2020). However, the large variant was not released publicly due to pretraining instability. The models are trained on a corpus of 1T tokens, which includes 14.8 billion Korean tokens. UMT5 surpasses mT5 in benchmarks such as XNLI (Conneau et al., 2018) and TyDi QA (Clark et al., 2020). As the small and base models do not have counterparts in the Polyglot-Ko suite, our experiments focus on the xlarge and xxlarge models.

Llama-2 Touvron et al. (2023) is available in three sizes: 7B, 13B, and 70B. It is trained on a corpus of 2T tokens, predominantly in English (89.7%), with Korean comprising a mere 0.06% or about 0.6B tokens. We utilize the version without finetuning. The Llama-2-70B model is excluded from our study due to the absence of a corresponding Korean model.

We employ the “log-likelihood" method implemented via LM-Eval-Harness (Gao et al., 2021) to evaluate the models. We compute the log-likelihood with each option concatenated to the question and select the one with the highest likelihood as the answer. All evaluations are implemented using bfloat16 precision in 0-shot, 5-shot, and 10-shot settings. We use accuracy as our primary metric.

HAE-RAE Bench aims to curate a dataset that challenges models lacking depth in Korean culture and knowledge, thereby guiding researchers in creating better Korean language models. To compare the ability of this benchmark to differentiate less native language models against prior benchmarks, we use KoBEST (Kim et al., 2022) as our baseline. We selected KoBEST as it offers a broad range of language understanding and reasoning tasks. KoBEST comprises five tasks: BoolQ, COPA, HellaSwag, WiC, and SentiNeg. However, given the findings of (Ko et al., 2023), that both monolingual and multilingual language models exhibit inconsistent performance on WiC, we omit this task from our assessment. While other available datasets may be adopted as baselines, they come with limitations. For instance, Korean-NLI&STS (Ham et al., 2020), being translated from English, is inherently more accessible for English models. KLUE (Park et al., 2021), despite being handcrafted, primarily focuses on basic NLU tasks like topic classification and NER. This makes it incapable of evaluating complex reasoning capabilities. Additionally, its test set is not publicly available.

5. Evaluation Results

Is HAE-RAE bench harder for foreign models? In Tables 5 and 6, we observe that the performance of LLMs scales with model size and the number of exemplars within the same suite. Nevertheless, despite their extensive training budgets, UMT5 and Llama-2 consistently fall short of their Polyglot-Ko counterparts. Furthermore, they rarely surpass the results of Polyglot-Ko-1.3B (0-shot). These results reaffirm the importance of language-specific corpora in learning cultural context and knowledge. It also highlights the effectiveness of HAE-RAE Bench, in assessing the language model's proficiency in Korean.

Our results illustrated in Tables 7, 8, and 9 suggest that the HAE-RAE Bench is particularly challenging for non-Korean models compared to the KoBEST benchmark. The performance gap between Polyglot-Ko and its counterparts is

		Loan Words			Standard Nomenclature			Rare Words		
Model	Params	n=0	n=5	n=10	n=0	n=5	n=10	n=0	n=5	n=10
Polyglot-Ko	1.3B	76.92	88.76	91.72	60.13	69.93	71.24	47.41	61.48	61.23
	3.8B	78.70	88.76	91.72	63.40	79.74	77.78	47.16	70.62	72.10
	5.8B	82.84	93.49	94.08	**66.67**	82.35	83.66	**56.79**	73.09	74.57
	12.8B	**87.57**	**94.67**	**94.67**	61.44	**84.97**	**86.93**	53.09	**75.31**	**76.05**
UMT5	3B	58.58	61.54	59.76	41.83	37.25	33.33	25.68	25.43	24.44
	13B	58.58	59.76	60.36	41.83	43.79	44.44	33.09	30.37	28.64
LLaMA-2	7B	66.86	73.96	75.15	39.22	49.02	50.98	29.38	39.26	39.01
	13B	66.86	77.51	78.11	49.02	57.52	64.05	32.35	42.47	43.95

Table 5: Evaluation results of the performance on Loan Words, Standard Nomenclature, and Rare Word tasks.

		History			General Knowledge			Reading Comprehension		
Model	Params	n=0	n=5	n=10	n=0	n=5	n=10	n=0	n=5	n=10
Polyglot-Ko	1.3B	60.11	78.19	77.13	26.70	30.68	28.98	34.45	37.81	37.14
	3.8B	69.15	86.17	85.11	28.41	**33.52**	33.52	40.49	42.06	40.04
	5.8B	79.79	85.11	81.91	29.55	27.84	28.41	40.72	42.73	41.39
	12.8B	**80.32**	**88.30**	**90.43**	**32.95**	**33.52**	**34.66**	**41.61**	**45.41**	**46.76**
UMT5	3B	14.36	12.77	14.36	22.73	19.32	19.32	25.28	24.83	25.28
	13B	21.59	18.09	19.15	21.81	25.00	19.32	29.75	25.28	27.74
LLaMA-2	7B	28.72	35.64	35.64	21.02	24.43	25.00	29.98	32.89	31.32
	13B	35.11	38.83	40.96	28.41	31.82	28.98	31.99	36.47	34.00

Table 6: Evaluation results of the performance on History, General Knowledge, and Reading Comprehension tasks.

	Polyglot-Ko		Δ	
Dataset	Params	Average	UMT5	Llama-2
HAE-RAE Bench	1.3B	51.0	-16.5	-15.1
	3.8B	54.6	-20.1	-18.7
	5.8B	59.4	-25.0	-23.5
	12.8B	59.5	-25.1	-23.6
KoBEST	1.3B	56.3	-6.3	-6.1
	3.8B	55.7	-5.7	-5.5
	5.8B	56.0	-6.0	-5.8
	12.8B	65.2	-15.2	-15.0

Table 7: Average Performance of Polyglot-Ko vs. UMT5-XXL and Llama-2-13B on HAE-RAE and KoBEST (0-shot).

	Polyglot-Ko		Δ	
Dataset	Params	Average	UMT5	Llama-2
HAE-RAE Bench	1.3B	61.1	-27.4	-13.7
	3.8B	66.8	-33.1	-19.4
	5.8B	67.4	-33.7	-20.0
	12.8B	70.4	-36.6	-22.9
KoBEST	1.3B	56.4	-8.4	7.8
	3.8B	64.7	-16.6	-0.5
	5.8B	68.0	-19.9	-3.8
	12.8B	71.4	-23.3	-7.2

Table 8: Average Performance of Polyglot-Ko vs. UMT5-XXL and Llama-2-13B on HAE-RAE and KoBEST (5-shot).

	Polyglot-Ko		Δ	
Dataset	Params	Average	UMT5	Llama-2
HAE-RAE Bench	1.3B	61.2	-28.0	-12.9
	3.8B	66.7	-33.4	-18.4
	5.8B	67.3	-34.1	-19.0
	12.8B	71.6	-38.3	-23.2
KoBEST	1.3B	55.0	-8.7	12.1
	3.8B	63.3	-16.9	3.9
	5.8B	68.0	-21.6	-0.8
	12.8B	71.8	-25.4	-4.6

Table 9: Average Performance of Polyglot-Ko vs. UMT5-XXL and Llama-2-13B on HAE-RAE and KoBEST (10-shot).

more pronounced on the HAE-RAE Bench than KoBEST across all exemplar counts. Notably, for Llama-2-13B, the margin narrows considerably on KoBEST with an increase in exemplars. This discrepancy highlights that the proposed benchmark is especially challenging for models not tailored in Korean and difficult to mitigate by in-context learning. The entire result for KoBEST is illustrated in section 10.1.

Does language frequency in the training corpora matter? Despite UMT5 being trained on a larger volume of Korean tokens, it underperforms Llama-2 on the HAE-RAE Bench. Moreover, the advantage of in-context learning is relatively minimal for UMT5. Our findings support previous claims that language-specific reasoning capabilities of language models are not solely tied to the number of dedicated tokens in the pretraining corpus (Shi et al., 2022). These results indicate that language models under the size of 20B parameters also transfer their in-context learning abilities to low-resource languages.

How important is the model size for the HAE-RAE Bench? In Table 10, we employ regression and Analysis of Variance(ANOVA) to examine the correlation between the parameter count of Polyglot-Ko models and their performance. To narrow the focus solely on the impact of model size, the analysis is limited to the Polyglot-Ko family, thus setting aside variables like corpus quality or model architecture. For the KoBEST benchmark, the results demonstrate a marked relationship between performance and model size, as indicated by the high R^2 value of 0.71 and the significant F-statistic and p-value. In contrast, for the HAE-RAE Bench, the model size explains only about a quarter of the performance variability. Additionally, the absence of statistical significance in both the regression and ANOVA for the HAE-RAE Bench implies that its evaluation is influenced by a broader spectrum of factors, pointing to challenges beyond just model size.

	Regression			**ANOVA**
Benchmark	β_0	β_1	R^2	F-statistic
HAE-RAE Bench	58.79	0.73	0.26	1.42
KoBEST	56.49	1.17	0.71*	8.23*

Table 10: Results from regression and ANOVA for the HAE-RAE and KoBEST benchmarks. An asterisk (*) denotes outcomes with a p-value less than 0.01, indicating statistical significance.

Can GPT-3.5/4 ace HAE-RAE Bench? In Table 11, the performance of GPT-3.5 and GPT-4 on the HAE-RAE Bench and KoBEST is presented. Unlike openly available models for which we leveraged a log probability method to gauge accuracy, these models do not provide log probabilities for individual tokens. Accordingly, we prompted the models to generate the number of the options they deemed correct. Direct comparison between these evaluation methods is not feasible. However, the method used for proprietary models is more challenging than the log-likelihood method applied to open models. The former entails generating answers from the entire vocabulary, whereas the latter restricts choices to five options. Notably, GPT-3.5 and GPT-4 achieved scores of 51.2% and 67.8% on the HAE-RAE Bench, respectively, indicating potential for further improvements. Conversely, their performances on KoBEST were 68.0% and 81.1%, suggesting narrower margins for improvement. In summary, state-of-the-art language models such as GPT-3.5 and GPT-4 have yet to master either the HAE-RAE Bench or KoBEST, though more room is left for the HAE-RAE Bench. The entire evaluation results for GPT-3.5 and GPT-4 models is available at section 10.2.

Can knowledge be transferred from English? Past research indicates that LLMs can internally transfer knowledge acquired in English to low-resource languages (Huang et al., 2023; Zhou et al., 2023). We employ Cross-lingual thought prompting (XLT) with GPT-3.5-Turbo and GPT-4 to investigate whether LLMs leverage abilities derived from English corpora to solve HAE-RA Bench. XLT (Shi et al., 2022) is a technique that aids the transfer of abilities learned in English to other languages. As illustrated in Table 11, English prompting enhances the performance of LLMs on both the HAE-RAE Bench and KoBEST. However, the gains for the HAE-RAE Bench are modest: 4.2 for GPT-3.5-Turbo and 0.4 for GPT-4. In contrast, the improvements on KoBEST are more substantial, with margins of 9.9 and 11.1, respectively. Given KoBEST's focus on language understanding and reasoning, we suspect that such skills are more seamlessly transferable across languages within models. On the other hand, the HAE-RAE Bench probes the nuances of cultural context and knowledge, which are challenging to learn from English tokens, thereby undermining the benefits of extensive training across various languages.

6. Error Analysis

Error analysis is essential to understand the common errors or likely biases of language model mistakes and identify areas of future research. Accordingly, we compare the results of Polyglot-Ko-12.8B (0-shot) and GPT-4 (Korean Prompting) for possible errors.

We first examine the answer distribution to see if either model has a bias toward selecting certain numbers. This is shown in Figure 2. We find that both models are less likely to guess " 5" compared to other numbers. This pattern can be traced back to the dataset composition: while most questions offer five multiple-choice options, the reading comprehension subset provides only four. Beyond this, neither model displays any notable trends.

In the **Rare Words**, **Loan Words**, and **Standard Nomenclature** subsets of the HAE-RAE Bench, incorrect options were generated using a sorting

	GPT-3.5-Turbo			**GPT-4**		
Dataset	Ko	En	Δ	Ko	En	Δ
HAE-RAE Bench	51.2	55.4	4.2	67.8	68.2	0.4
KoBEST	68.0	79.3	11.4	81.1	91.0	9.9

Table 11: Evaluation result of GPT-3.5-Turbo and GPT-4 on HAE-RAE Bench and KoBEST with zero shot setting. We use the snapshot from June 13th 2023 for both models. Ko and En denote the language of the prompt used.

	GPT-4		Polyglot-Ko-12.8B	
Dataset	Correct	Incorrect	Correct	Incorrect
Rare Words	1.58 (0.81)	1.58 (0.80)	1.58 (0.81)	1.58 (0.80)
Loan Words	1.58 (0.80)	1.56 (0.80)	1.58 (0.80)	1.57 (0.80)
Standard Nomenclature	1.58 (0.80)	1.55 (0.80)	1.57 (0.81)	1.56 (0.80)

Table 12: Average Levenshtein distance of options for correct and incorrect questions.

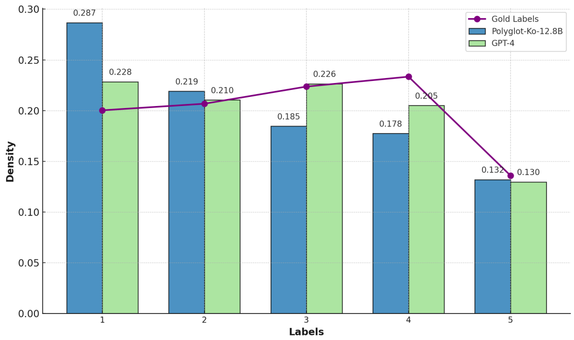


Figure 2: Density distribution of answer choices by Polyglot-Ko-12.8B, GPT-4, and Gold Labels.

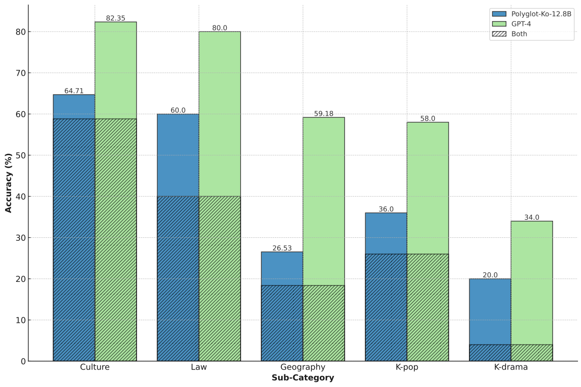


Figure 3: Accuracy of Polyglot-Ko-12.8B and GPT-4 on sub-categories of General Knowledge. The striped areas within each bar represent questions that both models answered correctly.

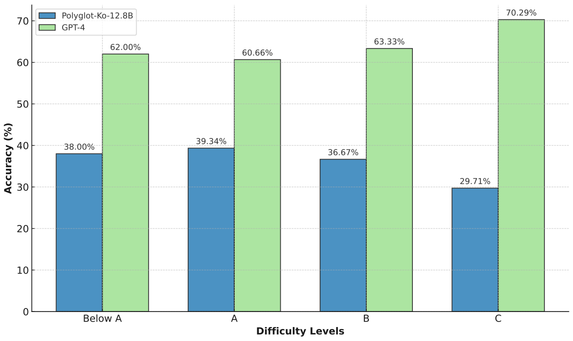


Figure 4: Accuracy of Polyglot-Ko-12.8B and GPT-4 on different levels of Reading Comprehension.

method based on Levenshtein distance. To investigate the impact of Levenshtein distance on model performance, we compare the average distances for options based on whether the model answered the question correctly. As shown in Table 12, no discernible difference in Levenshtein distance is observed for either model between correct and incorrect answers. We suspect the set Levenshtein distance threshold of 3 may not lead to meaningful variations in question difficulty. We review all incorrect questions for Polyglot-Ko-12.8B and GPT-4 to delve deeper. However, given the questions' simple structures, such as *"What is the {official loan word / correct standard nomenclature} for {word}?"* or *"Which word is suitable for the definition {def}?"*, we did not identify any syntactic characteristics that might explain the incorrect questions.

For the **General Knowledge** subset, we assess the performance of Polyglot-Ko-12.8B and GPT-4 across the subcategories, as shown in Figure 3. GPT-4 consistently outperforms Polyglot-Ko-12.8B. Polyglot-Ko-12.8B fares better in law and culture but lags in geography, K-pop, and especially K-drama. Given the need for up-to-date information in these areas, the model's weaker performance in K-pop and K-drama may stem from its knowledge cutoff. GPT-4 excels across all categories, likely benefiting from a diverse training set. Both models have the lowest scores in the K-drama category, suggesting either model limitations or ambiguous questions.

The **Reading Comprehension** subset of HAE-RAE Bench comprises four difficulty levels: Introductory (for absolute beginners), A (beginner), B (intermediate), and C (advanced), based on the Common European Framework of Reference (CEFR). In figure 4, we examine the performance of each model across these levels. Our findings indicate that GPT-4 consistently outperforms Polyglot-Ko-12.8B across all difficulty tiers, with the performance gap becoming more pronounced at higher levels (B and C). The performance of Polyglot-Ko-12.8B peaks at difficulty level A and declines, suggesting a limitation in handling more challenging questions.

7. License

HAE-RAE Bench is released under a CC BY-NC-ND license. This license prohibits remixing, redistribution, and commercial use of the dataset. This constraint is due to the reading comprehension subset, for which the copyright holder of KLAT has restricted commercial alterations. However, we do not anticipate this as a significant issue since benchmark datasets are rarely used for commercial purposes. Researchers can still freely download and evaluate their models using this dataset.

8. Conclusion

This paper introduces the HAE-RAE Bench, a dataset curated to evaluate the cultural knowledge encoded in language models. Unlike previous Korean language model evaluation suites, the HAE-RAE Bench is crafted to present a greater challenge to non-Korean models, disrupting their ability to guess answers based on in-context learning or scale-derived multilingualism. Our work is among the first to propose a non-task-oriented dataset aimed at assessing whether a language model's knowledge is adequate for roles like a domestic conversational agent or search engine. This research suggests a pathway for advancing non-English NLP, emphasizing the need for language models that are as proficient in language-specific knowledge as they are in language understanding and reasoning tasks.

Acknowledgments

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

9. Bibliographical References

ai forever. 2023. rugpt-3.5 13b.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli:

Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

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*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dohyeong Kim, Myeongjun Jang, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks. *arXiv preprint arXiv:2204.04541*.

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*.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian

Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6540, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Ramon Pires, Hugo Abonizio, Thales Rogério, and Rodrigo Nogueira. 2023. Sabi\'a: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

QwenLM. 2023. Introducing qwen-7b: Open foundation and human-aligned models (of the state-of-the-arts).

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

Sandhya Savanur and R. Sumathi. 2023. Sentineg: Algorithm to process negations at sentence level in sentiment analysis. *International Journal of Software Innovation*, 11:1–27.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Guijin Son, Hanwool Lee, Nahyeon Kang, and Moonjeong Hahm. 2023. Removing non-stationary knowledge from pre-trained language models for entity-level sentiment classification in finance. *arXiv preprint arXiv:2301.03136*.

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*.

StabilityAI. 2023. Japanese-stablelm-base-alpha-7b.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vicuna. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and

Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

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*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2023. Enhancing cross-lingual prompting with dual prompt augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11008–11020.

Judit Ács. 2019. Exploring bert's vocabulary.

10. Appendix

10.1. KoBEST Evaluation Results

The evaluation results for KoBEST are presented in table 15. Polyglot-Ko achieves the top scores in most settings, with exceptions in the 5-shot and 10-shot configurations for BoolQ.

10.2. GPT-3.5/4 Evaluation

In this section, we detail the prompts employed for the XLT evaluation of GPT-3.5 and GPT-4 models. We also provide their performance metrics for each downstream task in the HAE-RAE Bench and KoBEST datasets. The specific instructions used for XLT with both models are presented below. Table 14 and Table 13 contain the evaluation results for HAE-RAE Bench and KoBEST, respectively.

```
Read the given question, and choose the most suitable
answer. Answer your answer with the number of the
answer you think to be correct.
### question: {question}
### options:
(1) {option#1}
(2) {option#3}
(3) {option#3}
(4) {option#4}
(5) {option#5}
### answer:
```

Figure 5: Prompt used in our Direct Evaluation.

10.3. HAE-RAE Bench Examples

Starting from Figure 6, we present examples for each task alongside their translated versions.

Model	Lang	BoolQ	COPA	HellaSwag	SentiNeg
GPT-3.5-Turbo	Ko	82.34	57.88	40.00	91.69
	En	86.40	79.20	55.00	96.73
GPT-4	Ko	96.58	52.40	76.60	98.74
	En	96.65	95.70	73.00	98.49

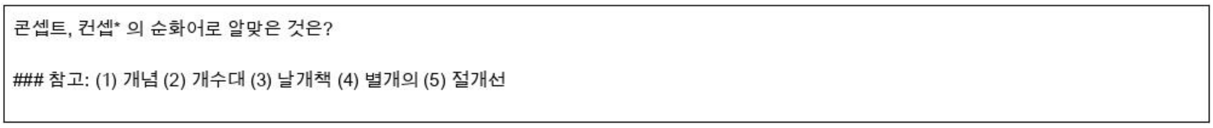
Table 13: Evaluation results for GPT-3.5 and GPT-4 on KoBEST

Model	Lang	Loan Words	Standard Nomenclature	Rare Words	History	General Knowledge	Reading Comprehension
GPT-3.5-Turbo	Ko	62.13	55.56	63.46	30.32	35.80	60.18
	En	**72.19**	67.32	61.73	30.85	42.61	57.72
GPT-4	Ko	70.41	67.32	**74.32**	**60.64**	54.55	**79.64**
	En	66.86	**79.08**	73.83	54.79	**55.68**	79.19

Table 14: Evaluation results for GPT-3.5 and GPT-4 on HAE-RAE Bench

		BoolQ			COPA			HellaSwag			SentiNeg		
Model	Params	n=0	n=5	n=10	n=0	n=5	n=10	n=0	n=5	n=10	n=0	n=5	n=10
Polyglot-Ko	1.3B	49.9	51.2	50.2	72.1	72.0	71.8	41.0	40.6	42.0	69.8	62.0	56.2
	3.8B	50.6	53.9	52.9	75.7	76.2	76.3	44.6	47.8	48.0	58.7	81.1	76.1
	5.8B	53.7	58.6	55.7	77.9	76.9	77.5	49.0	48.6	50.2	50.4	87.9	88.4
	12.8B	**56.7**	62.8	63.8	**79.6**	**81.0**	**80.2**	**49.0**	**51.0**	**49.8**	**91.7**	**90.7**	**93.5**
UMT5	3B	50.2	50.2	50.2	52.4	53.4	51.1	32.0	31.0	28.2	52.4	49.9	49.6
	13B	50.2	50.3	50.3	57.9	58.2	57.3	36.2	32.0	31.6	56.9	51.9	46.4
Llama-2	7B	50.9	58.8	55.8	56.1	58.1	58.2	41.8	43.2	43.2	48.9	58.2	57.2
	13B	50.6	**74.2**	**77.1**	59.1	64.1	63.1	41.6	44.8	42.6	50.6	73.8	85.9

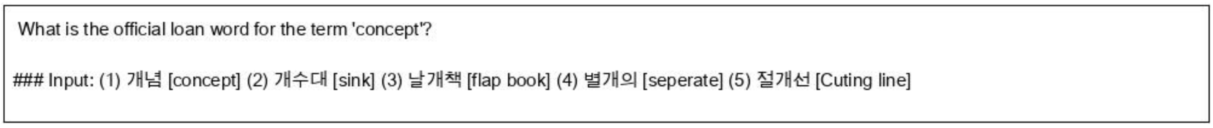
Table 15: Evaluation results for KoBEST.



```
콘셉트, 컨셉* 의 순화어로 알맞은 것은?

### 참고: (1) 개념 (2) 개수대 (3) 날개책 (4) 별개의 (5) 절개선
```

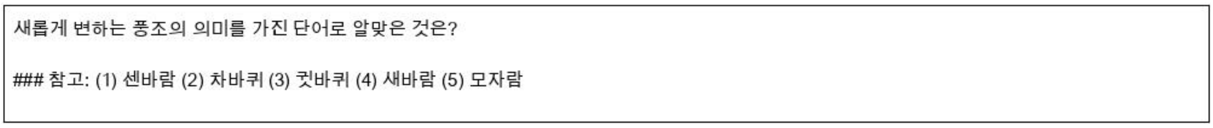
Figure 6: An example from the Loan Words subset.



```
What is the official loan word for the term 'concept'?

### Input: (1) 개념 [concept] (2) 개수대 [sink] (3) 날개책 [flap book] (4) 별개의 [seperate] (5) 절개선 [Cuting line]
```

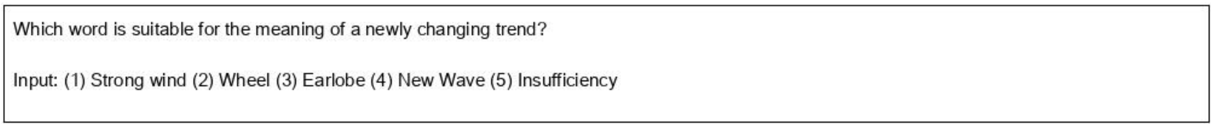
Figure 7: A translated example from the Loan Words subset.



```
새롭게 변하는 풍조의 의미를 가진 단어로 알맞은 것은?

### 참고: (1) 센바람 (2) 차바퀴 (3) 귓바퀴 (4) 새바람 (5) 모자람
```

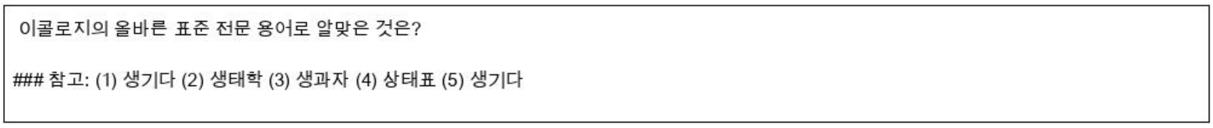
Figure 8: An example from the Rare Words subset.



```
Which word is suitable for the meaning of a newly changing trend?

Input: (1) Strong wind (2) Wheel (3) Earlobe (4) New Wave (5) Insufficiency
```

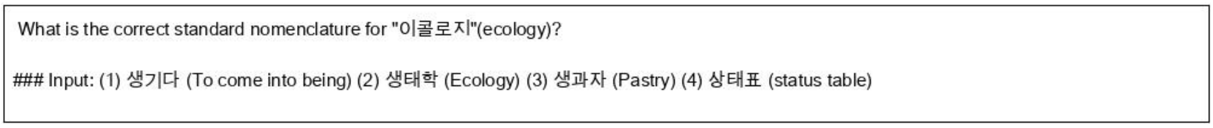
Figure 9: A translated example from the Rare Words subset.



```
이콜로지의 올바른 표준 전문 용어로 알맞은 것은?

### 참고: (1) 생기다 (2) 생태학 (3) 생과자 (4) 상태표 (5) 생기다
```

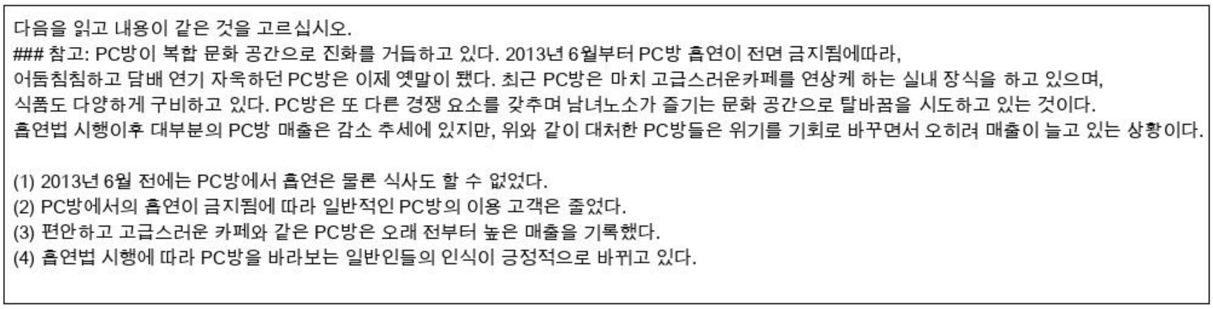
Figure 10: An example from the Standard Nomenclature subset.



```
What is the correct standard nomenclature for "이콜로지"(ecology)?

### Input: (1) 생기다 (To come into being) (2) 생태학 (Ecology) (3) 생과자 (Pastry) (4) 상태표 (status table)
```

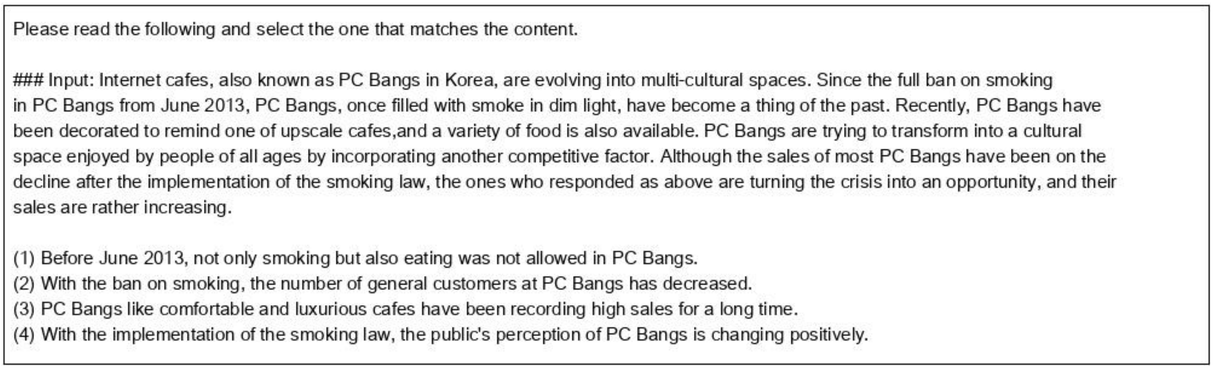
Figure 11: A translated example from the Standard Nomenclature subset.



다음을 읽고 내용이 같은 것을 고르십시오.
참고: PC방이 복합 문화 공간으로 진화를 거듭하고 있다. 2013년 6월부터 PC방 흡연이 전면 금지됨에따라,
어둠침침하고 담배 연기 자욱하던 PC방은 이제 옛말이 됐다. 최근 PC방은 마치 고급스러운카페를 연상케 하는 실내 장식을 하고 있으며,
식품도 다양하게 구비하고 있다. PC방은 또 다른 경쟁 요소를 갖추며 남녀노소가 즐기는 문화 공간으로 탈바꿈을 시도하고 있는 것이다.
흡연법 시행이후 대부분의 PC방 매출은 감소 추세에 있지만, 위와 같이 대처한 PC방들은 위기를 기회로 바꾸면서 오히려 매출이 늘고 있는 상황이다.

(1) 2013년 6월 전에는 PC방에서 흡연은 물론 식사도 할 수 없었다.
(2) PC방에서의 흡연이 금지됨에 따라 일반적인 PC방의 이용 고객은 줄었다.
(3) 편안하고 고급스러운 카페와 같은 PC방은 오래 전부터 높은 매출을 기록했다.
(4) 흡연법 시행에 따라 PC방을 바라보는 일반인들의 인식이 긍정적으로 바뀌고 있다.

Figure 12: An example from the Reading Comprehension subset.

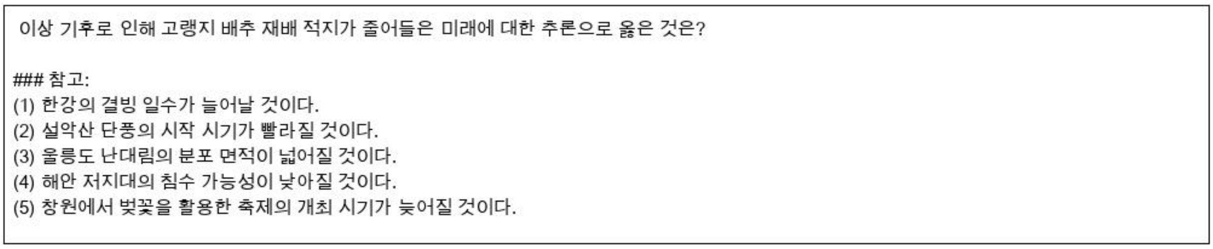


Please read the following and select the one that matches the content.

Input: Internet cafes, also known as PC Bangs in Korea, are evolving into multi-cultural spaces. Since the full ban on smoking
in PC Bangs from June 2013, PC Bangs, once filled with smoke in dim light, have become a thing of the past. Recently, PC Bangs have
been decorated to remind one of upscale cafes,and a variety of food is also available. PC Bangs are trying to transform into a cultural
space enjoyed by people of all ages by incorporating another competitive factor. Although the sales of most PC Bangs have been on the
decline after the implementation of the smoking law, the ones who responded as above are turning the crisis into an opportunity, and their
sales are rather increasing.

(1) Before June 2013, not only smoking but also eating was not allowed in PC Bangs.
(2) With the ban on smoking, the number of general customers at PC Bangs has decreased.
(3) PC Bangs like comfortable and luxurious cafes have been recording high sales for a long time.
(4) With the implementation of the smoking law, the public's perception of PC Bangs is changing positively.

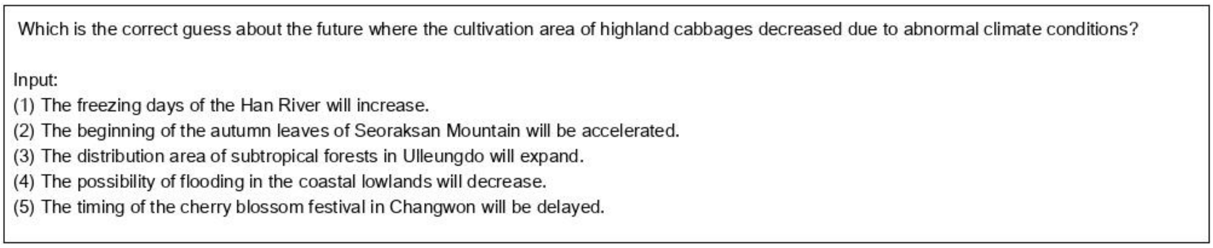
Figure 13: A translated example from the Reading Comprehension subset.



이상 기후로 인해 고랭지 배추 재배 적지가 줄어들은 미래에 대한 추론으로 옳은 것은?

참고:
(1) 한강의 결빙 일수가 늘어날 것이다.
(2) 설악산 단풍의 시작 시기가 빨라질 것이다.
(3) 울릉도 난대림의 분포 면적이 넓어질 것이다.
(4) 해안 저지대의 침수 가능성이 낮아질 것이다.
(5) 창원에서 벚꽃을 활용한 축제의 개최 시기가 늦어질 것이다.

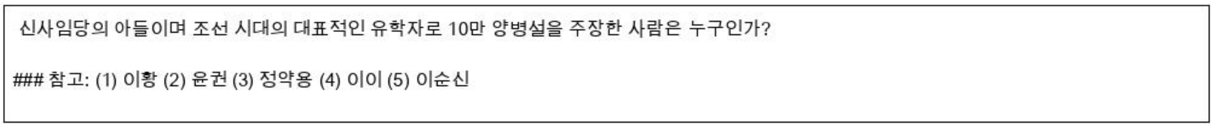
Figure 14: An example from the General Knowledge subset.



Which is the correct guess about the future where the cultivation area of highland cabbages decreased due to abnormal climate conditions?

Input:
(1) The freezing days of the Han River will increase.
(2) The beginning of the autumn leaves of Seoraksan Mountain will be accelerated.
(3) The distribution area of subtropical forests in Ulleungdo will expand.
(4) The possibility of flooding in the coastal lowlands will decrease.
(5) The timing of the cherry blossom festival in Changwon will be delayed.

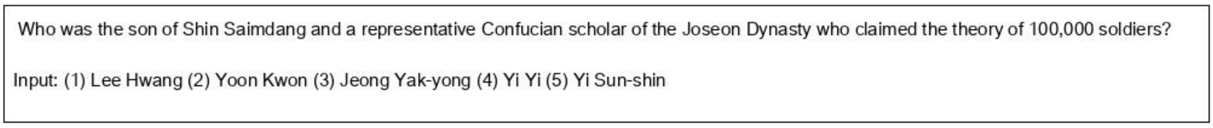
Figure 15: A translated example from the General Knowledge subset.



신사임당의 아들이며 조선 시대의 대표적인 유학자로 10만 양병설을 주장한 사람은 누구인가?

참고: (1) 이황 (2) 윤권 (3) 정약용 (4) 이이 (5) 이순신

Figure 16: An example from the History subset.



Who was the son of Shin Saimdang and a representative Confucian scholar of the Joseon Dynasty who claimed the theory of 100,000 soldiers?

Input: (1) Lee Hwang (2) Yoon Kwon (3) Jeong Yak-yong (4) Yi Yi (5) Yi Sun-shin

Figure 17: A translated example from the History subset.