

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



Table of Contents

1. Main Contributions
2. Dataset Creation: HAE-RAE Bench
3. Evaluation Results

Main Contributions

The first dataset to evaluate cultural, local knowledge for Korea

Summary:

Multilingual evaluation typically focuses on assessing a language model's ability to perform specific tasks, such as summarization, extraction, or translation.

However, this approach does not effectively measure the model's grasp of the ethnolinguistic knowledge necessary to engage in meaningful conversations with native speakers of the language.

Accordingly, we make a new dataset to evaluate the “cultural” and “language-specific” knowledge of a model.

Dataset Creation: HAE-RAE Bench

The dataset is consisted of six subsets.

Category	Type	Total # of		Avg. # of Words (std)		Fertility Rate (std)	
		Question	Unique Morpheme	per question	per passage	Polyglot-Ko	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.

The design principle behind HAE-RAE Bench, is to collect general-level knowledge questions. A model that fails to answer might be inadequate the converse with korean native speakers.

Dataset Creation: HAE-RAE Bench

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:

1. Source pairs of foreign words and korean equivalents from NIKL
2. Filter based on whether they listed on Korean Encyclopedias
3. Convert to MCQA format by selecting wrong options using Levenshtein distance

Dataset Creation: HAE-RAE Bench

Standard Nomenclature

Task Description:

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

Creation:

We follow an identical process with the Loan Words subset.

Dataset Creation: HAE-RAE Bench

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:

1. Source pairs of foreign words and korean equivalents from "Woorimal Battle"
2. We follow an identical process with the Loan Words subset.

Dataset Creation: HAE-RAE Bench

General Knowledge

Task Description:

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

Creation:

1. Define five primary categories for general knowledge: tradition, law, geography, Korean pop, and Korean drama.
2. Crowd source questions to fit these subcategories.
3. Remove overlapping, factually incorrect, and questions that fail to align with the defined category

Dataset Creation: HAE-RAE Bench

General Knowledge

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.

Metric	Full	Q-Only	C-Only	Δ (<i>min</i>)
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.

Dataset Creation: HAE-RAE Bench

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:

1. Source web pages tagged "Korean history" from Namuwiki, Korea's equivalent to Wikipedia, and randomly selected 40 pages.
2. Manually craft 5 questions for each page, remove overlapping questions.

Dataset Creation: HAE-RAE Bench

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:

1. Collect test materials publicly released by the Korea Educational Testing Service (KETS)

Evaluation Results

Polyglot-Ko, UMT5, Llama-2

Model	Params	Loan Words			Standard Nomenclature			Rare Words		
		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.

Model	Params	History			General Knowledge			Reading Comprehension		
		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.

Is HAE-RAE bench harder for foreign models?

Yes, as shown table 5,6 despite UMT5 and Llama-2 trained on a substantially larger training budget, Polyglot-Ko outperforms Bothe models.

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. This hints that other factors, the quality of the corpora, or the entire training budget may affect the models' performance on a new language.

Evaluation Results

GPT-3.5 and GPT-4

Dataset	GPT-3.5-Turbo			GPT-4		
	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.

Can proprietary models ace HAE-RAE Bench?

No, top-performing models like GPT-4 also suffer in scoring high scores.

Interestingly, despite all the questions being written in Korean, using a English prompt improves performance. We suspect model is failing to retrieve knowledge despite having it when prompted in Korean.

HAE-RAE Bench

Now being widely used as a defacto-standard for evaluation of Korean language models


Model	HAE-RAE Bench (0-shot)						All
	<i>LW</i>	<i>RW</i>	<i>SN</i>	<i>RC</i>	<i>HI</i>	<i>GK</i>	
SOLAR API (solar-1-mini-chat)	61.54	77.53	62.09	73.83	81.38	50.57	70.55
GPT-3.5 (gpt-3.5-turbo-0125)	55.62	51.11	49.67	62.42	27.13	45.45	51.17
GPT-4 (gpt-4-0125-preview)	38.46	63.70	43.79	81.21	79.79	60.22	65.60
HGX-S	79.88	83.95	85.62	75.17	88.30	51.14	77.89
HGX-L	82.25	93.09	86.93	81.43	94.15	59.09	84.14



Table 12: Detailed results on HAE-RAE Bench (Son et al., 2023). We report the performance across six areas: loan words (LW), rare words (RW), standard nomenclature (SN), reading comprehension (RC), history (HI), and general knowledge (GK). HGX-L outperforms other closed-source models in all but one area.

HAE-RAE Bench

Available through HuggingFace and LM-Eval-Harness

Hugging Face

Models Datasets Spaces Posts Docs Pricing 




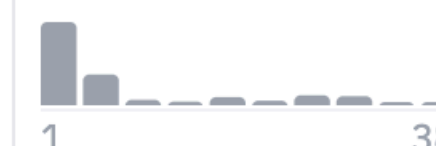


Datasets: HAERAE-HUB / **HAE_RAE_BENCH_1.0**   like 0

ArXiv: [arxiv:2309.02706](#) Tags: [Croissant](#)

Dataset card Viewer Files and versions Community **2** Settings


Dataset Viewer [Auto-converted to Parquet](#) [API](#) [View in Dataset Viewer](#)

Subset (6) **General Knowledge** · 176 rows Split (1) **test** · 176 rows

question	a	b	c	d	e
string · lengths	string · lengths	string · lengths	string · lengths	string · lengths	string · len
					
26 374	1 45	1 39	1 38	1 44	1
다음은 무엇에 대한 이야기입니까? ### 참고:...	추석	제사	성묘	백일	장례
다음 중 중 대화가 이루어지는 장소로 가장 적...	식당	병원	학교	도서관	백화점

Downloads last month **2333**

[Use in Datasets library](#) [Edit dataset card](#)



Size of downloaded dataset files: **600 kB**

Size of the auto-converted Parquet files: **353 kB** Number of rows: **1,538**