Benchmarking Korean Idiom Understanding: A Comparative Analysis of Local and Global Models

Xiaonan Wang **Seoyoon Park** **Hansaem Kim**
Yonsei University
nan@yonsei.ac.kr seoyoon.park@yonsei.ac.kr khss@yonsei.ac.kr

Abstract

Although an increasing number of multilingual LLMs (large language models) have begun to support Korean, there remains a notable lack of benchmark datasets specifically designed to evaluate their proficiency in Korean cultural and linguistic understanding. A major reason for this gap is that many available benchmarks in Korean are adapted from English originals via translation, which often fails to reflect the unique cultural context embedded in the Korean language. Even the few benchmark datasets based on native Korean data that involve cultural content typically focus on tasks such as bias or hate speech detection, where cultural knowledge serves merely as topical background rather than being integrated as a core component of semantic understanding. To address this gap, we introduce the Korean Idiom Matching Benchmark (KIM Bench), which consists of 1,175 instances. Idioms are culture-specific and often untranslatable, making them ideal for testing models' cross-cultural semantic understanding. Using KIM Bench, We evaluate global and Korean native models. Our analysis show that larger and locally trained models better capture idiom semantics and cultural nuances, while chain-of-thought prompting may reduce accuracy. Models still struggle with deep semantic and contextual understanding. KIM Bench offers a compact tool for cross-cultural evaluation and insights into improving performance on culturally grounded tasks.

1 Introduction

With the rapid advancement of large language models (LLMs), research on evaluating these models has increased dramatically in recent years (Chang et al., 2024). Early evaluations primarily focused on models' basic language understanding (Wang et al., 2018) and generation (Zhong et al., 2022) capabilities. However, as these models have demonstrated growing performance across multilingual and multidisciplinary tasks, assessing their cross-cultural understanding has become increasingly crucial. The complexity of culture and the diversity of languages make this task progressively more challenging.

Idiom	누워서 떡 먹기
Literal	Eating rice cake while lying down
Meaning	Describes something that is very easy to do

Table 1: An example of metaphor in idiom. The sense of "누워서 떡 먹기" should be inferred figuratively rather than interpreted literally using the meanings of the constituent characters.

Many existing benchmarks primarily target English, covering tasks ranging from basic language understanding to complex reasoning (Chang et al., 2024). However, in non-English-speaking regions, localized benchmarks are often created by translating English datasets into the target language (Shi et al., 2022). Although this approach is convenient, it introduces significant limitations: Western cultural biases embedded in the source data are often preserved in translation, leading to cultural distortions in evaluation. As a result, such benchmarks fail to reflect the linguistic specificity and cultural diversity of the target language.

To address this challenge, a growing number of culturally distinct communities have initiated efforts to develop benchmarks that are tailored to reflect their unique cultural characteristics and linguistic intricacies. For example, Koto et al. (2024) proposed IndoCulture, a benchmark designed to explore language models' ability to understand the diverse cultures and geographical factors across 11 Indonesian provinces. Inoue et al. (2024) introduced Heron-Bench, which evaluates language models' understanding within the Japanese con-

Proceedings of Recent Advances in Natural Language Processing,pages 1341–1351
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-156

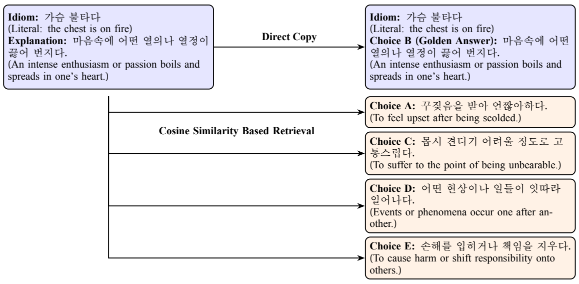
text through image-question pairs. Das et al. (2023) developed a Bengali dataset to assess language models' understanding of cultural biases related to gender, religion, and national identity. However, benchmarks targeting Korean culture remain insufficient. Although some efforts have been made to evaluate language models in the Korean context, most focus on tasks such as bias detection or hate speech classification, where cultural knowledge serves merely as topical background rather than being integrated into the core of semantic understanding. As a result, these benchmarks lack the incorporation of cultural richness and semantic nuance, limiting comprehensive evaluation of models' ability to understand Korean culturally grounded linguistic phenomena. To address the lack of benchmarks for evaluating cultural understanding in Korean, we introduce **KIM Bench** (Korean Idiom Matching Benchmark), a dataset centered on idiom matching. Idioms are highly condensed, metaphorical, and non-literal expressions deeply embedded in specific cultural contexts. Since each culture has its own idiomatic system that is often difficult to translate directly, understanding idioms requires both linguistic competence and cultural knowledge. In practical applications such as translation, education, and dialogue generation, failure to correctly interpret idioms can lead to serious semantic misunderstandings. KIM Bench captures diverse semantic patterns and rich cultural nuances, providing a focused testbed for evaluating language models' culturally grounded semantic reasoning and local adaptability.

Our main contributions are summarized as follows:

- We present **KIM Bench**[1], a novel benchmark consisting of 1,175 samples, designed to evaluate language models' comprehension of Korean idioms, which require both linguistic and cultural understanding.

- We perform a systematic comparison of multilingual GPT models and the Korean-native HyperCLOVA X, and provide detailed error analyses that reveal model-specific challenges in idiom interpretation.

- We empirically demonstrate that chain-of-thought prompting, while beneficial in logical reasoning tasks, can introduce semantic noise and impair model performance in culture-dependent scenarios.



Idiom: 가슴 불타다
(Literal: the chest is on fire)
Explanation: 마음속에 어떤 열의나 열정이 끓어 번지다.
(An intense enthusiasm or passion boils and spreads in one's heart.)

Idiom: 가슴 불타다
(Literal: the chest is on fire)
Choice B (Golden Answer): 마음속에 어떤 열의나 열정이 끓어 번지다.
(An intense enthusiasm or passion boils and spreads in one's heart.)

Choice A: 꾸짖음을 받아 언짢아하다.
(To feel upset after being scolded.)

Choice C: 몹시 견디기 어려울 정도로 고통스럽다.
(To suffer to the point of being unbearable.)

Choice D: 어떤 현상이나 일들이 잇따라 일어나다.
(Events or phenomena occur one after another.)

Choice E: 손해를 입히거나 책임을 지우다.
(To cause harm or shift responsibility onto others.)

Figure 1: An example from Korean Idiom Matching Benchmark

2 Related Work

2.1 Large Language Model

Large Language Model (LLM) typically refers to deep neural network models with hundreds of millions, billions, or even trillions of parameters. The core idea behind these models is that by training on massive amounts of textual data, the models can gradually ”learn” the vocabulary, grammar, and underlying semantic associations in natural language, enabling them to generate coherent and semantically logical text within a given context. In practice, autoregressive language models are trained by maximizing the conditional probability of each token given its preceding tokens (Naveed et al., 2024; Shanahan, 2024; Zhao et al., 2024; Kumar et al., 2025). The conditional probability is typically modeled using Transformer networks based on multi-head attention or other neural network architectures (Vaswani et al., 2017). The development of English models in LLMs has been particularly prominent, with the GPT (Generative Pre-trained Transformer) series being the most representative. GPT, developed by OpenAI, is a series of generative language models based on the Transformer architecture. By leveraging large-scale unsupervised pre-training and task-specific fine-tuning, these models have achieved remarkable results in natural language generation and understanding tasks (Brown et al., 2020). From GPT-1 to GPT-4, the scale and performance of these models have continuously improved, demonstrating strong capabilities in context understanding, complex reasoning, and multi-turn dialogue generation (OpenAI, 2024a). Meanwhile, models tailored to other languages have also been advanc-

[1] The dataset is available at KIM-Bench on GitHub.

ing. For instance, HyperCLOVA X, developed by Naver, is a LLM system specifically optimized for Korean and multilingual tasks related to Korean culture (Kim et al., 2021). HyperCLOVA X utilizes advanced multi-head attention mechanisms and extensive Korean datasets to accurately capture Korean syntactic and semantic characteristics.

2.2 Evaluating Large Language Models

LLMs have made remarkable progress in natural language understanding and generation (Annepaka and Pakray, 2025), which has spurred the development of benchmarks using standardized datasets and evaluation tasks to compare models quantitatively (Li et al., 2024). Early benchmarks, such as GLUE (Wang et al., 2018) and SuperGLUE (Sarlin et al., 2020), focused on English-centric tasks like sentiment analysis, natural language inference, and question answering, establishing performance baselines. As the need to assess broader cognitive abilities emerged, benchmarks like HellaSwag (Zellers et al., 2019) and CosmoQA (Huang et al., 2019) incorporated commonsense and causal reasoning, enhancing assessment robustness. With many tasks in early benchmarks reaching human-level performance, new benchmarks such as MMLU-Pro (Wang et al., 2024b) and KULTURE Bench (Wang et al., 2024a) were developed to better reflect modern models' capabilities across diverse subjects and complex tasks. In response to the global demand for LLM applications, researchers have translated English benchmarks into other languages (Shi et al., 2022); however, these multilingual benchmarks often fail to capture cultural context and expression habits, leading to biased evaluations (Choenni et al., 2024). Consequently, native benchmarks have become essential. For example, benchmarks such as CMoralEval (Yu et al., 2024) for Chinese culture, ITALIC (Seveso et al., 2025) for Italian culture, and ILMAAM (Nacar et al., 2025) for Arabic-speaking communities have been developed to evaluate cultural understanding in their respective contexts. In Korea, benchmarks such as the Korean Bias Benchmark (Jin et al., 2024), HAERAE Bench (Son et al., 2024), and CLIcK (Kim et al., 2024) primarily assess sociolinguistic bias and surface-level cultural references using question-and-answer formats. However, they rarely incorporate data that directly reflects culturally embedded expressions or traditional knowledge, limiting their capacity to evaluate deeper semantic and cultural understanding in Korean. To address this gap, we introduce KIM (Korean Idiom Matching) Bench, which incorporates traditional Korean idioms into a multiple-choice format and enable a deeper evaluation of cultural sensitivity.

3 Korean Idiom Matching Benchmark

3.1 Characteristics of Korean Idioms

Idioms are a common linguistic phenomenon and are referred as "관용구" (gwanyonggu) in Korean. Due to its compact structure and rich expressiveness, idioms are extensively used in everyday conversations and various forms of written text. The main challenge in machine reading comprehension involving idioms lies in effectively capturing the extended meanings of idioms. The non-compositional nature and metaphorical meanings of many idioms (see an example in Table 1) make their translation particularly challenging, drawing substantial interest from researchers (Shao et al., 2017). The meanings of such idioms often diverge from the literal interpretations of their constituent elements. Typically rooted in ancient cultural narratives, these idioms have preserved their figurative meanings throughout the evolution of language over time. For instance, "미역국을 먹다"(to eat seaweed soup) has a metaphorical meaning, which originates from a historical event: In 1907, the Japanese colonial authorities forcibly disbanded the Korean military, pushing the country into a deeper crisis. Many soldiers lost their jobs and fell into financial hardship. In Korean, the word "해산" (haesan) can mean both ”disbandment” and ”childbirth”, as they share the same pronunciation. Due to this phonetic similarity, people associated the disbandment of the military with the Korean tradition of eating seaweed soup after childbirth. As a result, in 1947, the term "미역국을 먹다"(to eat seaweed soup) was officially recorded in the ”Great Dictionary” as a slang expression referring to ”the dissolution of an organization or the dismissal of an individual from a position.” Over time, this phrase has further evolved to commonly symbolize failing an exam or losing a job, becoming widely used in everyday language. Thus, understanding and accurately interpreting expressions requires familiarity with Korean cultural and historical origins.

3.2 Construction of Korean Idiom Matching (KIM) Benchmark

To evaluate language models' ability of understanding Korean idioms, we constructed a matching dataset where the model must select the correct idiom explanation from five given options. Each sample consists of one Golden Answer and four distractor options, which are generated based on semantic similarity between idiom explanations. This approach ensures a more challenging task by increasing the difficulty of distinguishing between closely related meanings. Figure 1 presents an example in KIM Bench.

3.2.1 Raw Data Acquisition

This dataset's idiom data primarily comes from two authoritative Korean linguistic resources. First, "표준국어대사전"[2] (the Standard Korean Language Dictionary), compiled by the National Institute of the Korean Language (NIKL), is one of the most authoritative linguistic resources in Korea. It contains approximately 510,000 entries, including standard Korean, dialects, and North Korean vocabulary, providing a rich collection of idioms and proverbs that serve as a fundamental reference for Korean language research and learning. Second, "우리말샘"[3] (Urimalsam), an open Korean linguistic database launched by NIKL in 2016, includes around 1.5 million lexical entries, covering everyday expressions, regional dialects, and specialized terminology. This database adopts a participatory approach, allowing continuous expansion to reflect the evolving nature of the Korean language in contemporary society. These two resources not only ensure high authority and extensive coverage but also provide a rich cultural and linguistic foundation, making them essential sources for constructing this dataset. The data collection process is conducted manually, with all idioms and their corresponding explanations extracted one by one from the aforementioned resources by researchers. To ensure the representativeness and diversity of the dataset, we specifically focus on commonly used idioms in Korean society. All collected idioms are reviewed by experts in Korean Studies-related fields to verify their actual frequency and prevalence in modern Korean society. During the data collection process, all idioms with multiple interpretations are systematically recorded to account for their semantic variability. This approach is adopted to ensure comprehensive coverage, as many Korean idioms exhibit polysemy, with meanings that shift depending on the contextual framework in which they are used. As a result, the final dataset comprises 1,175 idioms, each accompanied by its respective interpretations.

Similarity Range	Cosine Similarity Interval
Range 1	0.50 – 0.60
Range 2	0.61 – 0.70
Range 3	0.71 – 0.80
Range 4	0.81 – 0.90

Table 2: Cosine similarity ranges for negative choices selection

3.2.2 Data Cleaning

After collecting the data, we conduct a thorough cleaning process to ensure consistency and accuracy. We remove unnecessary elements such as unrelated numbers, extraneous parentheses, and other irrelevant information that did not contribute to the understanding of the idioms. Additionally, we eliminate duplicate samples to maintain the uniqueness of the dataset. In instances where an idiom's explanation referenced another idiom instead of directly conveying its meaning, such as "'간 붓다'를 속되게 이르는 말" (a vulgar expression for '간 붓다'), we replace these explanations with the actual meaning of the idiom, "지나치게 대담해지다" (to become excessively bold). As a result, each idiom is now presented with a clear and self-contained definition.

3.2.3 Candidates Retrieval

The golden answer for each sample is based on the dictionary definition, serving as the authoritative interpretation of the idiom. To generate challenging distractor options, we utilize cosine similarity, a widely used metric in natural language processing that measures the semantic similarity between text representations by computing the cosine of the angle between two vectors. The cosine similarity formula is as follows (Manning et al., 2008):

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \qquad (1)$$

where A and B represent the embedding vectors of two idiom explanations, $A \cdot B$ denotes their dot product, and $\|A\|\|B\|$ is the product of their

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

[3] https://opendict.korean.go.kr/main

magnitudes. The similarity score ranges from -1 to 1, where 1 indicates identical meanings, 0 suggests no correlation, and -1 represents completely opposite meanings. In this research, we leverage the pre-trained multilingual BERT model (bert-base-multilingual-cased) (Devlin et al., 2019) to generate embeddings for each idiom explanation. Specifically, we extract the [CLS] token representation as the global semantic representation of each sentence. We then compute the cosine similarity between this representation and the embeddings of all other idiom explanations in the dataset to quantify their semantic proximity. To ensure a balanced level of difficulty for distractor options, we categorize cosine similarity values into four predefined ranges, and one incorrect option is randomly selected from each range (see Table 2). To prevent language models from inferring the correct answer based on the position of the options, the five candidate choices in each sample (including the golden answer and four distractor options) are randomly shuffled. This ensures that the model must analyze the meaning of each option comprehensively rather than relying solely on positional cues. Using the above method, we generate five candidate options for each idiom, with one being the golden answer and the remaining four randomly selected distractor options based on similarity intervals.

Direct Prompt

다음은 한국어 관용구와 이에 대한 다섯 가지 설명입니다.
가장 적합한 정답을 선택하고 **A, B, C, D, E** 중 하나의 글자만 출력하세요.
(*Below are a Korean idiom and five candidate meanings. Select the most appropriate answer and output only one letter from* ***A, B, C, D, or E****.*)

관용구(Idiom): `{idiom}`
선택지(Options): `{options_text}`

정답이라고 생각하는 알파벳 한 글자 (**A, B, C, D, E** 중 하나)만 출력하세요. 예: `A`
(*Output only the single letter* (*one of* ***A, B, C, D, or E***) *that you believe is correct. For example:* `A`)

Figure 2: Direct prompting template used for idiom comprehension evaluation

4 Experiments

4.1 Models

In this research, we evaluate several advanced language models developed by OpenAI, including GPT-4o, GPT-4o-mini, o1, o1-mini and o3-mini (OpenAI, 2024b,c). GPT-4o is an enhanced version of GPT-4 that supports multimodal inputs (text, audio, image, and video) and demonstrates significant improvements in processing non-English languages. GPT-4o-mini is a lightweight variant of GPT-4o, optimized for computational efficiency while maintaining strong text understanding and generation performance. o1 and o3-mini is a model specifically designed for complex reasoning tasks, capable of performing high-quality inference and enhancing depth of thought through inference-time scaling and reflection. o1-mini is a streamlined version of o1, balancing reasoning capabilities with improved efficiency. Additionally, to compare the performance with international models, we test HyperClovaX (Kim et al., 2021) a major Korean language model primarily trained on Korean-language data developed by Naver. By including HyperClovaX in our experiments, we aim to assess how well specialized Korean models perform compared to OpenAI's multilingual models.

To evaluate the performance of large language models in understanding Korean idioms, we employ two different prompting methods: Direct Prompting and Chain-of-Thought (CoT) Prompting. In the Direct Prompting round, we provide all the models with inputs containing idioms and five candidate explanations, asking the models to select the most appropriate meaning. The specific prompt used in this process is shown in Figure 2. Specifically, for GPT-4o, we further request the model to provide a rationale for its chosen answer to analyze its decision-making process. The primary goal of the Direct Prompting method is to assess the models' ability to comprehend and select answers based on semantics and context, without explicit reasoning cues. To further evaluate the models' reasoning capabilities, we apply the Chain-of-Thought Prompting method to different models. In this method, we explicitly guide the models through the problem-solving process by providing step-by-step reasoning instructions before selecting the final answer. The aim of using CoT prompting was to explore whether this technique could enhance the models' performance in understanding idioms (Kojima et al., 2022), particularly in tasks requiring deeper semantic comprehension and reasoning. The prompt used in this process is shown in Figure 3.

Chain-of-Thoughts Prompt

다음은 한국어 관용구와 이에 대한 다섯 가지 설명입니다.
당신은 이 문제를 단계적으로 생각해야 하며 답이라고 생각하는 알파벳 한 글자(**A, B, C, D, E** 중 하나)만 출력하세요.
(*Below is a Korean idiom and five descriptions related to it. You should analyze the problem step by step and select the option that best fits the explanation of the idiom. Output only a single letter* (*one of* ***A, B, C, D, or E***) *that you believe is the correct answer.*)

관용구(Idiom): `{idiom}`
선택지(Options): `{options_text}`

최종 선택한 답변을 반드시 대괄호로 묶어서 표시하세요. 예를 들어: `[A]`
(*Enclose your final chosen answer in square brackets. For example:* `[A]`)

Figure 3: Chain-of-Thought prompting template used for idiom comprehension evaluation

4.2 Metrics

This study employs accuracy as the evaluation metric to measure the models' performance in understanding Korean idioms. Accuracy is defined as the ratio of correctly predicted answers to the total number of test samples, calculated as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{2}$$

During the evaluation, the models are required to select one answer from five candidate options for each sample, where one is the golden answer and the other four are incorrect options. The accuracy score directly reflects the models' ability to understand the semantics of idioms and distinguish between similar options.

5 Results and Discussions

5.1 Overall Model Performance

The Direct Prompting results on the KIM benchmark reveal notable variations in model capabilities for cross-cultural language understanding as shown in Figure 4. The experimental data indicate that full-scale models outperform their mini counterparts; specifically, GPT4o and o1 achieved accuracies of 87.32% and 89.02%, respectively, while the mini versions GPT4o-mini and o1-mini recorded only 80.85% and 70.55%. This pronounced difference underscores the critical role of model scale in capturing the complex semantic nuances of idiomatic expressions, as larger models, with their richer representations and enhanced reasoning capabilities, are better equipped to handle ambiguous or abstract contexts. Moreover, the results highlight the significance of localized training data. HyperClovaX achieves the highest accuracy of 90.38%, demonstrating that datasets enriched with Korean linguistic and cultural contexts substantially improve the model' s understanding of Korean idioms.

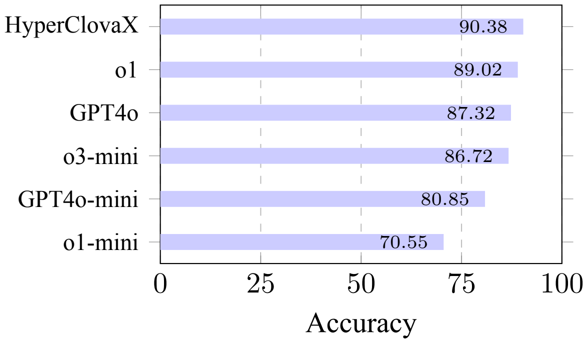


Figure 4: Accuracies of Different Models under Direct Prompting

5.2 Does chain-of-thought (CoT) help LLMs' Korean idiom understanding?

In recent years, Chain-of-Thought (CoT) prompting has been demonstrated to significantly enhance model performance in tasks requiring multi-step reasoning, such as mathematical problem solving, logical reasoning, and programming. For instance, Wei et al. (2022) proposed Chain-of-Thought Prompting and found that CoT substantially improved LLMs' performance in mathematical and commonsense reasoning tasks. However, as shown in Table 3, our experimental results indicate that after applying CoT, the accuracy of GPT4o-mini increased only slightly, from 80.85% to 81.45%, while that of o1-mini rose modestly from 70.55% to 73.11%. For these smaller-scale models, such limited improvement suggests that CoT may provide a clearer reasoning path, helping them better elucidate semantic clues within abstract idioms; nevertheless, the overall gains remain minimal. In contrast, both larger general models (HyperClovaX, GPT4o) and specialized inference-oriented models (o1, o3mini) showed performance degradation after employing CoT. Specifically, among the general models, HyperClovaX's accuracy dropped significantly

Model	Accuracy (w/o CoT)	Accuracy (w/ CoT)	Difference (Δ)
GPT4o-mini	80.85	81.45	+0.60
o1-mini	<u>70.55</u>	<u>73.11</u>	+2.56
GPT4o	87.32	**86.47**	-0.85
HyperClovaX	**90.38**	83.83	-6.55
o1	89.02	83.66	-5.36
o3-mini	86.72	76.68	-10.04

Table 3: Accuracy comparison with and without Chain-of-Thought (CoT) in Korean Idiom Matching. The highest value in each column is **bolded**, and the lowest is <u>underlined</u>.

from 90.38% to 83.83%, while GPT4o decreased slightly from 87.32% to 86.47%. The inference-oriented models experienced even greater declines: o1 fell from 89.02% to 83.66%, and o3mini exhibited a dramatic decrease from 86.72% to 76.68%. These results suggest that for models already possessing strong semantic representations or established logical reasoning capabilities, additional CoT processes might introduce redundant analysis or reasoning noise, disrupting their original efficient matching mechanisms or internal inference pathways. Particularly, inference-oriented models, which inherently include explicit reasoning mechanisms, could accumulate additional errors or irrelevant information due to the extra CoT steps, thereby significantly reducing accuracy. We hypothesize that the Korean idiom-matching task differs fundamentally from typical logical reasoning tasks, relying more heavily on cultural background knowledge, contextual appropriateness, and quick recognition of fixed expressions rather than explicit multi-step logical deductions. Thus, CoT did not produce the anticipated positive effects in this task; instead, it sometimes negatively impacted performance, especially for models already equipped with robust semantic or inference capabilities, and yielded only slight benefits for smaller-scale models.

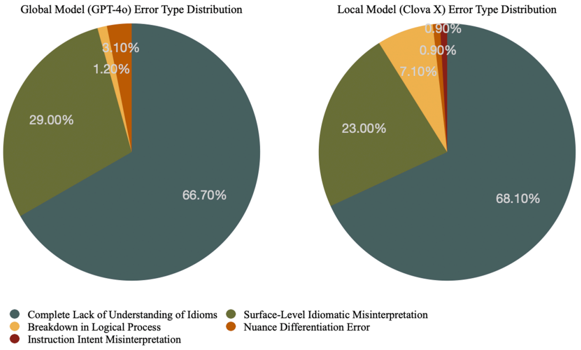


Figure 5: Error Type Distribution Comparison: Global Model (GPT-4o) vs. Local Model (HyperClovaX)

5.3 What types of errors occur in model responses to Korean idiom test?

We conduct a detailed error analysis of the models in the Korean idiom matching task, with a focus on comparing the performance of the local Korean model HyperClovaX and the global model GPT4o under Direct Prompting. Based on our observations, we categorize the errors into five types: Surface-Level Idiomatic Misinterpretation, Complete Lack of Understanding of Idioms, Breakdown in Logical Process, Nuance Differentiation Error, and Instruction Intent Misinterpretation. The first type, Surface-Level Idiomatic Misinterpretation, refers to instances where the model, despite understanding the basic meaning of a term, relies solely on its superficial features to make random guesses, resulting in semantic misinterpretations. The second type, Complete Lack of Understanding of Idioms, is characterized by the model' s failure to grasp the true meaning of idiomatic expressions, sometimes producing hallucinated explanations. The third type, Breakdown in Logical Process, indicates that even when the model captures the correct semantic clues, it may ultimately select an incorrect answer due to breaks in the reasoning chain or the accumulation of noise. In addition, Nuance Differentiation Error occurs when the model struggles to differentiate between highly similar options, while Instruction Intent Misinterpretation arises from a misinterpretation of the input prompt, causing a failure to accurately capture the intended task requirements. We also compared the proportion of errors made by the global model GPT-4o and the local model HyperClovaX, as illustrated in Figure 5. Overall, both models face significant challenges in understanding Korean idioms in the matching task. Among the error types, Complete Lack of Understanding of Idioms has the highest proportion, at 66.7% for GPT4o and 68.1% for HyperClovaX, indicating clear deficiencies in capturing the intrinsic meanings of idioms. Additionally, the Surface-Level Idiomatic Misinterpretation error is also quite common, accounting for 29% of errors in GPT4o and 23% in HyperClovaX. This suggests that when processing Korean idioms, both models tend to rely on superficial cues and make arbitrary inferences, failing to grasp the deeper cultural connotations. In contrast, the rates of Breakdown in Logical Process and Nuance Differentiation Error are relatively low, at 1.2% and 3.1% for GPT4o, and 7.1% and 0.9% for HyperClovaX, respectively; furthermore, in terms of Instruction Intent Misinterpretation, both models remain at extremely low levels (0% for GPT4o and 0.9% for HyperClovaX), indicating that they are generally capable of understanding the input instructions correctly. When comparing the two models, several notable differences emerge. Although both models display similar fundamental difficulties in idiom comprehension, GPT4o demonstrates greater stability in its reasoning process, with a Breakdown in Logical Process rate of only 1.2% compared to HyperClovaX' s 7.1%, reflecting the global model's advantage in handling complex reasoning tasks. At the same time, GPT4o exhibits a slightly higher error rate in Surface-Level Idiomatic Misinterpretation (29% versus HyperClovaX's 23%), indicating a greater tendency to rely on superficial cues when processing ambiguous information. Conversely, HyperClovaX performs better in differentiating subtle expressions, with a "Nuance Differentiation Error" rate of only 0.9%, far lower than GPT4o's 3.10%. In general, while GPT4o excels in internal reasoning stability, HyperClovaX shows an advantage in capturing subtle linguistic nuances.

5.4 To what extent do humans outperform language models in understanding Korean idioms?

To compare human and model performance, we randomly selected 50 Korean idiom comprehension questions from our constructed benchmark and invited 20 native Korean speakers to answer them. The results showed that the average accuracy of the native speakers was 91.3%. Meanwhile, among several language models we tested, the best performer is HyperClovaX with an accuracy of 90.38%. This result indicates that our benchmark is highly challenging. The questions widely involve semantic ambiguities, cultural allusions, and metaphorical expressions, which place high demands on the participants' language comprehension and cultural background knowledge. Therefore, even among native speakers, the average accuracy did not exceed 95%, demonstrating a certain degree of cognitive challenge. Moreover, despite the inherent difficulty of the questions, the best-performing model, HyperClovaX, only 0.92 percentage points lower than that of the native speakers. This data reflects that in handling Korean idiom comprehension tasks, which are both highly complex and context-dependent, the performance of current domestic models is very close to that of native speakers, and their understanding of idioms is steadily approaching human levels.

6 Conclusion

This paper presents KIM Bench, a novel benchmark specifically designed to evaluate LLMs' ability to understand Korean idioms. Our experiments encompass global models from the OpenAI series as well as the Korean native model HyperClovaX and reveal several key findings. First, larger-scale models perform better on this task, indicating that model scale plays a significant role in capturing the subtle nuances of complex semantics. Second, localized training data is crucial in enhancing the understanding of culturally characteristic expressions. Third, our analysis of CoT prompting shows that, in the idiom matching task, CoT may introduce unnecessary noise that, in some cases, reduces model performance, likely because this task relies more on understanding cultural background and sensitivity to contextual cues. Furthermore, detailed error analysis indicates that both global and native models exhibit issues, including superficial comprehension errors and a failure to accurately capture the deeper, context-dependent meanings of idioms. These findings not only highlight the inherent challenges of cross-cultural language evaluation but also underscore the importance of developing benchmarks that accurately reflect the subtle nuances of language and culture. In general, KIM Bench provides a challenging tool for assessing models' ability to understand Korean idioms and offers a new perspective for further research into culturally nuanced language processing.

Limitations

There is a limitation regarding the scope of model evaluation in this study. The experiments only included some global models and the Korean native model HyperClovaX, without covering other types or scales of language models. This limitation restricts a comprehensive assessment of LLMs' cross-cultural understanding capabilities and may not fully reflect the performance differences across various model architectures and sizes in handling culturally relevant tasks.

Acknowledgements

This work was supported by the Yonsei University Research Fund of 2024 (2024-22-0317).

References

Yadagiri Annepaka and Partha Pakray. 2025. Large language models: a survey of their development, capabilities, and applications: Large language models: a survey of their development, capabilities,... *Knowl. Inf. Syst.*, 67(3):2967–3022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Rochelle Choenni, Sara Rajaee, Christof Monz, and Ekaterina Shutova. 2024. On the evaluation practices in multilingual nlp: Can machine translation offer an alternative to human translations?

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. 2024. Heron-bench: A benchmark for evaluating vision language models in japanese.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean Bias Benchmark for Question Answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

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.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neu-*

ral Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. Llm post-training: A deep dive into reasoning large language models.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024. A survey on benchmarks of multimodal large language models.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader, and Anis Koubaa. 2025. Towards inclusive Arabic LLMs: A culturally aligned benchmark in Arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

OpenAI. 2024a. Gpt-4 technical report.

OpenAI. 2024b. Gpt-4o system card.

OpenAI. 2024c. Openai o1 system card.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Murray Shanahan. 2024. Talking about large language models. *Commun. ACM*, 67(2):68–79.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024a. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11817–11837, Bangkok, Thailand. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.