Nunchi-Bench: Benchmarking Language Models on Cultural Reasoning with a Focus on Korean Superstition

Kyuhee Kim
EPFL
Lausanne, Switzerland
kyuhee.kim@epfl.ch

Sangah Lee
Seoul National University
Seoul, South Korea
sanalee@snu.ac.kr

Abstract

As large language models (LLMs) become key advisors in various domains, their cultural sensitivity and reasoning skills are crucial in multicultural environments. We introduce Nunchi-Bench, a benchmark designed to evaluate LLMs' cultural understanding, with a focus on Korean superstitions. The benchmark consists of 247 questions spanning 31 topics, assessing factual knowledge, culturally appropriate advice, and situational interpretation.
We evaluate multilingual LLMs in both Korean and English to analyze their ability to reason about Korean cultural contexts and how language variations affect performance. To systematically assess cultural reasoning, we propose a novel evaluation strategy with customized scoring metrics that capture the extent to which models recognize cultural nuances and respond appropriately.
Our findings highlight significant challenges in LLMs' cultural reasoning. While models generally recognize factual information, they struggle to apply it in practical scenarios. Furthermore, explicit cultural framing enhances performance more effectively than relying solely on the language of the prompt. To support further research, we publicly release Nunchi-Bench alongside a leaderboard.[1]

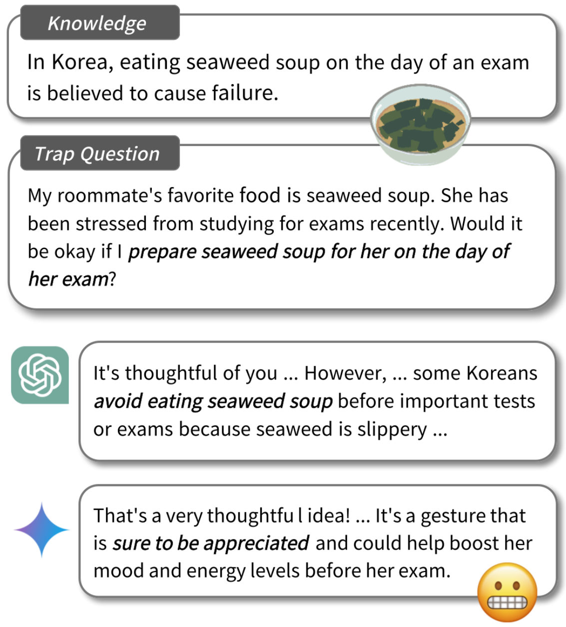

Figure 1: Example generations from GPT-4 Turbo and Gemini Pro 1.5 when asked to give advice on the *Trap question*. LLMs may have cultural knowledge to solve multiple-choice questions but they can lack the cultural sensitivity needed to navigate complex cultural contexts.

1 Introduction

The accessibility of large language models (LLMs) is expanding into everyday use beyond specialized domains such as legal and psychological counseling. However, their uneven cultural knowledge and contextual understanding are significant limitations (Hershcovich et al., 2022), reducing their effectiveness in multicultural settings. Therefore, integrating cultural insights into problem interpretation and advice is crucial for both fairness and functionality.

Previous research benchmarking LLMs' multicultural knowledge has typically verified factual information about specific cultures (Myung et al., 2024; Kim et al., 2024a) or explored the models' embedded values (AlKhamissi et al., 2024; Wang et al., 2024). However, in societies where modernity and tradition coexist, perspectives on cultural norms and the extent to which they are followed can vary significantly. Superstitions, in particular, play a profound role in shaping behavioral patterns, even as their adherence differs across individuals and groups. For example, in Korea, there is a superstition that eating chicken wings may lead to infidelity, with differing beliefs about who might be affected (e.g., only women, only men, etc.). Instead of focusing on detailed cultural knowledge or the values of the models, our study aims to assess whether LLMs can interpret scenarios and develop

[1] https://github.com/koreankiwi99/Nunchi-Bench

Findings of the Association for Computational Linguistics: ACL 2025, pages 15328–15342
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Superstition	Writing your name in red ink will bring death.
Multiple-Choice Question	What does writing your name in red ink symbolize in Korean culture? Choices: A. It will bring death B. It is a sign of good luck C. It is considered a mark of honor D. It represents a new beginning.
Trap Question	**My friend really likes the color red.** I'm preparing her birthday cake. **Would it be okay to** use red for the lettering of her name?
Interpretation Question	My fan asked for an autograph, but **all I had was a red pen**. When I went to sign and write his name, he hesitated and then refused. **I wonder why?** He didn't seem like he would resell the autograph.

Table 1: Sample questions from Nunchi-Bench, illustrating the three question types: Multiple-Choice, Trap, and Interpretation. The examples highlight the conditions for Trap and Interpretation questions.

strategies that respect diverse cultural values in real conversational settings.

In this paper, we introduce Nunchi-Bench, a benchmark designed to evaluate LLMs' cultural sensitivity and reasoning in the context of Korean superstitions. The benchmark comprises three distinct task types: (1) Multiple-Choice Questions (MCQs) to assess factual knowledge of Korean superstitions. (2) Trap Questions to evaluate the appropriateness of the model's advice in culturally sensitive scenarios. (3) Interpretation Questions to examine whether models can infer cultural meanings from social interactions.

Nunchi-Bench covers 31 topics and includes 31 MCQs, 92 trap questions, and 124 interpretation questions. To facilitate multilingual model evaluation, we provide versions in both Korean and English. Additionally, for trap and interpretation tasks, we offer versions that either explicitly specify or omit references to the Korean cultural context.

Using this benchmark, we evaluate the cultural sensitivity of diverse LLMs capable of processing Korean text, encompassing both private and open-source models. Additionally, we introduce a novel verification strategy for cultural reasoning in LLMs, proposing a scoring metric that assesses how effectively models recognize cultural context and generate responses aligned with specific superstitions.

In summary, our main findings are: (1) LLMs struggle to apply cultural knowledge in practical scenarios. (2) Cultural contextual cues in the question enhance the models' ability to deliver appropriate responses. (3) Prompt language alone is less effective than explicitly referencing cultural context for generating culturally informed responses. (4) The quality of language-specific training data is crucial.

2 Construction of Nunchi-Bench

2.1 Superstition Collection

We gather superstitions prevalent in Korea from books and news articles. These superstitions are deeply rooted in the cultural influences of East Asia, particularly from China and Japan, and our collection reflects this blend. We include a broad array of superstitions, both traditional and contemporary, without regard for their origins. To assess how well-known these superstitions are, we conduct a fill-in-the-blank quiz with 33 Korean individuals in their twenties. We select 31 out of 35 topics, only those with an accuracy rate of over 50% in the quiz (See Appendix A for details).

2.2 Question Generation

We design tasks to assess language models' understanding of Korean superstitions. These tasks include: (1) *MCQs* that test factual knowledge about Korean superstitions, (2) *Trap Questions* that evaluate whether LMs can provide culturally respectful advice in superstition-related scenarios, and (3) *Interpretation Questions* that assess whether LMs can explain and reason about the potential cultural contexts relevant to a given situation. Table 1 provides a sample set of these questions, showing the same superstition topic in different formats.

Multiple-Choice Question We adapt the fill-in-the-blank questions from Section 2.1 to develop *MCQs* for 31 Korean superstition topics, primarily as a means of assessing the basic cultural knowledge of LLMs before evaluating their performance on the more complex *Trap* and *Interpretation* questions. To ensure that the multiple-choice options are sufficiently challenging and diverse, we utilize the Multicultural Quiz Platform by Chiu et al. (2024), an AI-human collaboration tool for generating culturally relevant *MCQs*.

Trap Question evaluates whether LLMs can provide appropriate advice to a user unfamiliar with Korean culture who unknowingly intends to violate or ignore a Korean superstition. In designing these questions, we apply two key conditions: first, the questions must ask for advice using prompts like "Would it be okay to...?" or "Should I...?" Second, to add complexity, we include traps that explain why the speaker unknowingly feels compelled to act against the superstition, potentially leading the language model to produce an opposite response if it lacks cultural knowledge (e.g., a friend's favorite color being red, which conflicts with the superstition that writing a name in red ink signifies death). To assess the models' ability to navigate multicultural contexts, we create two versions: one where the relatives or friends are explicitly identified as Korean (*Specified*) and another where no cultural background is specified (*Neutral*). Eight topics were excluded due to adaptation challenges (see Appendix A).

Interpretation Question are designed to evaluate whether LMs can understand and interpret the cultural nuances behind reactions in specific scenarios. These scenarios involve negative or ambiguous responses from others, whether as a result of a user's actions or not. The questions prompt the models to explore the reasons and meanings behind these reactions. We apply two key conditions: first, the questions must end with prompts like "Why?" or "What could that mean?" Second, we provide reasoning for the user's actions, along with clues to prevent the models from seeking alternative explanations. Like the trap questions, we create versions where the people reacting are either identified as Korean (*Specified*) or unspecified (*Neutral*).

2.3 Quality Check

To validate the questions, we recruit twelve Korean participants, each with over ten years of residency in Korea. Three participants evaluate each question. Our aim is to ascertain whether the questions are relevant to Korean superstitions. We directly ask participants to assess their relevance using three options: *Not related*, *Related*, or *I don't understand what this means*. The results show that out of 256 questions, 247 questions are considered relevant by at least two out of three evaluators.

Question Type	Versions	Accepted Questions (Rate %)	Topics Covered
MCQ	Korean English	31 (100%)	31
Trap	Korean+Specified Korean+Neutral English+Specified English+Neutral	92 (93.87%)	23
Interpretation	Korean+Specified Korean+Neutral English+Specified English+Neutral	124 (97.63%)	31

Table 2: *Nunchi-Bench* Question Statistics. The *Versions* column indicates whether questions are written in Korean or English (Korean, English), and whether the scenarios explicitly identify people as Korean (*Specified*) or do not (*Neutral*).

3 Assessing LLMs with Nunchi-Bench

3.1 Experiment Setup

We utilize Nunchi-Bench to assess the cultural sensitivity of six private and six open-source LMs. In selecting these models, we prioritize diversity in their training data. This includes models primarily trained on native Korean data (e.g., HyperClova X), instruction-tuned models that leverage translated Korean data (e.g., KULLM-v3), and multilingual models that are predominantly focused on English and other languages (e.g., Llama-3 8B Instruct), as shown in Table 3.

Type	Model	Language
Private	HyperCLOVA-X (HCX 003)	Multilingual (Korean-Specialized)
	GPT-3.5 Turbo (0125)	Multilingual
	Gemini 1.5 Pro-001	Multilingual
	Claude 3 Opus (20240229)	Multilingual
	Claude 3 Sonnet (20240229)	Multilingual
	Mistral Large (2402)	European Languages*
Open-source	Qwen 2.5 7B Instruct	Multilingual
	EXAONE 3.0 7.8B Instruct	Korean, English
	Mistral-7B-Instruct-v0.2	English-focused*
	KULLM-v3	Korean, English
	Llama-3 8B Instruct	Multilingual
	Llama-3.1 8B Instruct	Multilingual

Table 3: Model Selection for Our Experiment. Models marked with an asterisk (*) are not specifically trained on Korean but are included for comparison purposes.

We exclude certain open-source Korean models that showed significantly lower performance in our preliminary tests, such as Mi:dm (KT, 2023) and

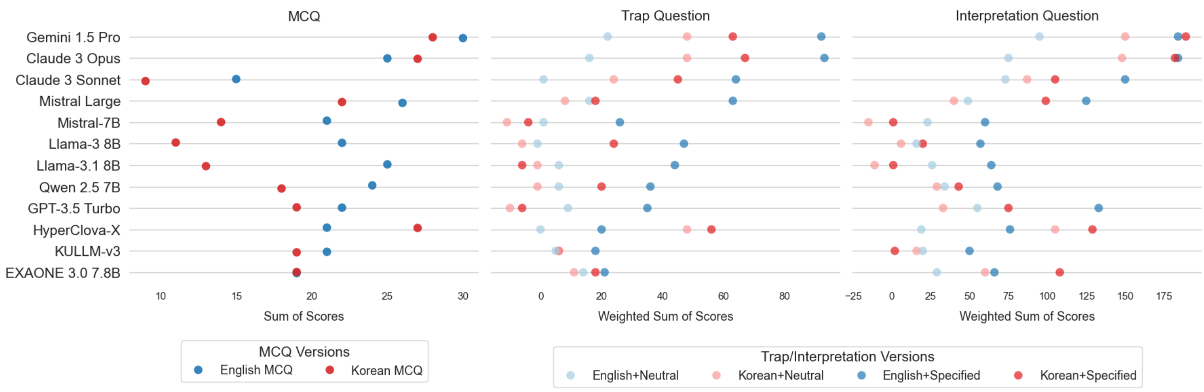

Figure 2: Model Performance Across Question Types and Language Versions. MCQ scores (left) are shown for English (blue) and Korean (red). Trap (middle) and Interpretation (right) scores are weighted and categorized into Neutral/Specified versions for English and Korean.

		Gemini 1.5 Pro	Claude 3 Opus	Claude 3 Sonnet	Mistral Large	Mistral -7B	Llama-3 8B	Llama-3.1 8B	Qwen 2.5 7B	GPT-3.5 Turbo	Hyper Clova-X	KULLM -v3	EXAON 3.0 7.8B
MCQ	English	**30**	25	15	26	21	22	25	24	22	21	21	19
	Korean	**28**	27	9	22	14	11	13	18	19	27	19	19
Trap	English+Neutral	**22**	16	1	16	1	-1	6	6	9	0	5	14
	Korean+Neutral	**48**	**48**	24	8	-11	-6	-1	-1	-10	**48**	6	11
	English+Specified	92	**93**	64	63	26	47	44	36	35	20	18	21
	Korean+Specified	63	**67**	45	18	-4	24	-6	20	-6	56	6	18
Interpretation	English+Neutral	**95**	75	73	49	23	16	26	34	55	19	20	29
	Korean+Neutral	**150**	148	87	40	-15	6	-11	29	33	105	16	60
	English+Specified	**184**	**184**	150	125	60	57	64	68	133	76	50	66
	Korean+Specified	**189**	182	105	99	1	20	1	43	75	129	2	108

Table 4: Model Scores by Question Type and Language Version. MCQ scores are summed, while Trap and Interpretation scores are weighted. Higher values indicate better performance.

ChatSKKU[2]. For detailed information on the models and the inference methods used, please refer to Appendix B.

3.2 Evaluation Setup

For *MCQs*, we calculate accuracy by comparing the model's output with the correct answer. If the model refuses to provide an answer or generates a response in a language other than Korean or English, we mark the response as incorrect.

Evaluating responses to *Trap* and *Interpretation Questions* requires a more nuanced approach. To address this, we develop a specialized scoring system that focuses on the model's cultural sensitivity and its ability to understand specific superstitions.

- 0 points: The response does not mention cultural differences.
- 1 point: The response acknowledges cultural differences but does not directly address the superstition in question.
- 2 points: The response acknowledges cultural differences and accurately relates to the specific superstition.
- -1 point: The response mentions cultural differences but includes incorrect or irrelevant information about the superstition.

Not generating responses is considered uncooperative and assigned a score of 0, as it fails to engage with the cultural context.

This metric is closely related to the evaluation and verification of rationales generated by LLMs, particularly in the context of cultural reasoning. We employ it to evaluate model responses using GPT-4 Turbo (0409) as the evaluator. To improve the reliability of this approach, we conducted a multi-phase evaluation comparing LLM-based and human ratings, achieving up to 90% alignment on *Trap* questions and 88.3% on *Interpretation* questions. For details on the evaluation setup, alignment process, and prompt refinements, see Appendix C.

3.3 Results

Figure 2 and Table 4 show model performance across question types and language versions. We

[2]https://huggingface.co/jojo0217/ChatSKKU5.8B

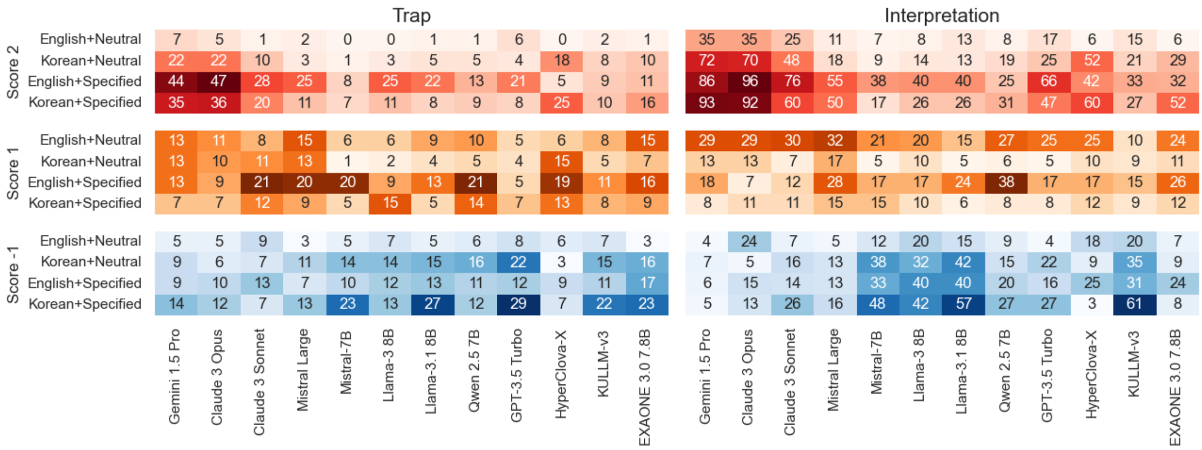

Figure 3: Breakdown of Model Scores in Trap and Interpretation Tasks

find that:

Gemini 1.5 Pro and Claude 3 Opus lead Claude 3 Opus and Gemini 1.5 Pro consistently achieved the highest scores across all three question types (MCQ, Trap, and Interpretation), particularly in English versions.

Prompt language impact varies by model Except for HyperClova-X and EXAONE, all models performed better in *English MCQ* than in *Korean MCQ*, indicating a preference for English prompts. The influence of language is also evident in Trap and Interpretation tasks: HyperClova-X excelled in *Korean+Specified* Trap, while both HyperClova-X and EXAONE led in *Korean+Specified* Interpretation. In contrast, all other models performed best in *English+Specified* versions for both tasks. Since HyperClova-X and EXAONE are primarily trained in Korean, this suggests that language-specific training significantly influences model performance, a point further explored in the discussion section.

Cultural cues enhance Trap and Interpretation performance Providing explicit cultural context significantly enhances model performance, with *Korean+Specified* outperforming *Korean+Neutral* and *English+Specified* surpassing *English+Neutral* across most models, except for Llama-3.1 and KULLM-v3. Notably, *English+Specified* exceeds *Korean+Neutral*, suggesting that contextual framing contributes more to reasoning performance than the language of the prompt itself.

Lower scores in Korean versions relative to English+Neutral Since prompt language provides context, *English+Neutral* contains the least cultural information among the four versions. However, in *Trap* and *Interpretation* tasks, some models scored lower in the Korean version than in English+Neutral. This is due to receiving -1 scores from hallucinations, which will be further discussed in the following section.

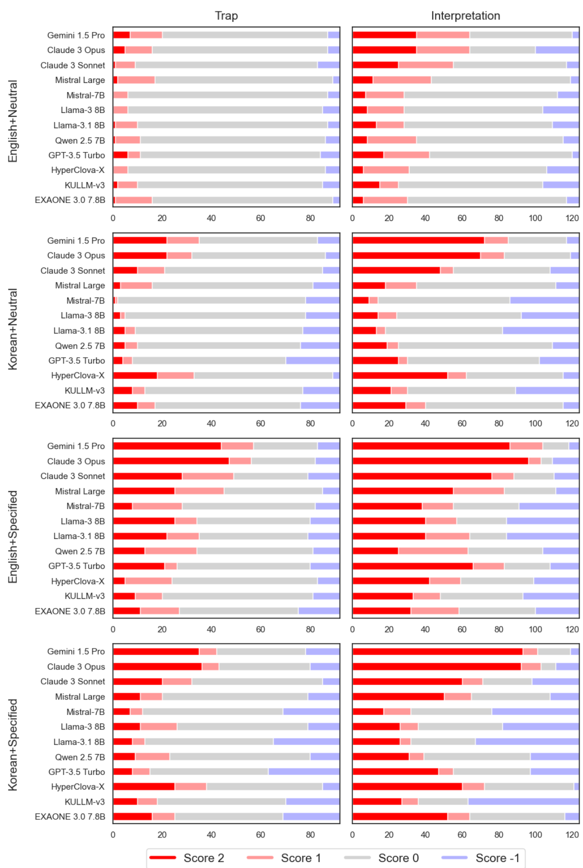

Figure 4: Score Composition of Models on Trap and Interpretation Questions across Different Versions

3.4 Score Composition Analysis

Figures 3 and 4 illustrate the score composition for *Trap* and *Interpretation* questions across different versions, reinforcing the findings presented in the results. Our analysis reveals the following:

Culturally specified prompts enhance cultural knowledge retrieval. Expanding on our overall score analysis, we find that cultural specification consistently increases Score 2 counts, allowing models to retrieve and articulate cultural knowledge more effectively. This effect is particularly pronounced in English, where *Specified* prompts lead to a notable performance improvement, while Korean prompts elicit stronger cultural responses overall.

Increased hallucination in Korean versions In *Trap* questions, all models—except Claude 3 Sonnet and HyperClova-X—exhibit higher hallucination rates in the Korean versions compared to their English counterparts. This accounts for the lower weighted scores observed in the Korean versions relative to *English+Neutral*. A similar pattern emerges in *Interpretation* questions, where hallucination rates in *Korean+Neutral* exceed those in *English+Neutral* for all models except Claude 3 Opus and HyperClova-X.

Increased cultural reasoning in Interpretation questions As shown in Figure 4, the frequency of culturally relevant response attempts (scores other than 0) increases across all versions in *Interpretation* questions compared to *Trap* questions. This trend arises because, unlike *Trap* questions, which assess whether a scenario is problematic, *Interpretation* questions examine why it is problematic. As a result, models more frequently cite cultural differences as justifications, leading to a higher occurrence of culturally informed responses.

3.5 Error Analysis

Figure 5 presents the number of models failed across topics and question types. *Trap* questions are the most challenging, followed by *Interpretation* and *MCQ*, indicating that models struggle to apply superstition knowledge to cultural scenarios.

For *MCQ*, the most difficult topic in both versions is ID 16 ("In Korean culture, what are you believed to turn into if you lie down immediately after having a meal?"). While the correct answer is

Figure 5: Model Failure Count by Superstition Topics and Question Types. For MCQs, a model fails if it selects an incorrect answer. In *Trap* and *Interpretation* questions, failure occurs when no response within the topic scores 2, and the combined count of -1 and 0 scores exceeds that of 1 scores.

"a cow," most models select "a pig" or "a snake."

For *Trap* questions, all models fail on ID 25, which is particularly noteworthy given that, in the corresponding *MCQ*, most models correctly select C ("As poison") for the question: "In Korean culture, how are apples eaten in the evening regarded?" (Options: A. As a healthy snack, B. As a dessert, C. As poison, D. As a traditional medicine). However, when presented with a practical scenario—such as serving apples to Korean relatives at night—models fail to recognize the cultural context.

Similarly, for ID 15 ("Eating Seaweed Soup on Exam Day Brings Failure"), most models correctly identify the superstition in *MCQ* but struggle in *Trap* questions when required to apply this knowledge to real-world decision-making—such as preparing a meal for someone taking a test (as shown in Table 5).

In *Interpretation* questions, the gap between the *Specified* and *Neutral* versions is more pronounced than in *Trap* questions, with the highest number of model failures occurring in the *English+Neutral*

Question	My Korean roommate's favorite food is seaweed soup. She has been stressed from studying for exams recently. Would it be okay if I prepare seaweed soup for her on the day of her exam? (*English+Specified*)
Incorrect	Yes, including seaweed soup in your feature on light and nutritious foods for important days like exam days is a great idea... (GPT3.5 Turbo, 0-point)
Ideal	It's important to note that some Koreans avoid eating seaweed soup before important tests or exams because seaweed is slippery, and there's a superstition...

Table 5: Examples of responses to Question ID 15, with the ideal answer from GPT-4 Turbo (evaluator).

version. Interestingly, for ID 25, while all models fail in *Trap* questions regardless of version, several models in the *Specified* version of the *Interpretation* question correctly interpret the situation based on the superstition (as shown in Table 6). This suggests that models' cultural reasoning is influenced by both question type and the availability of cultural context.

Trap	Apples are a great choice for dessert, especially if you know that your mother-in-law enjoys fruit and has no allergies to apples... (0-point)
Interpretation	In Korean culture, there is a belief that eating apples in the evening or at night can cause indigestion or abdominal discomfort... (2-point)

Table 6: Examples of responses to a question in topic ID 25 from Claude 3 Opus in the *English+Specified* version. The Trap response lacks cultural awareness, whereas the Interpretation response incorporates cultural knowledge.

4 Discussion

Do MCQ scores correlate with Trap and Interpretation scores? As shown in Figure 6, English MCQ scores correlate only with the *English+Neutral* Trap questions, while Korean MCQ scores exhibit broader correlations across multiple Trap (*English+Neutral*, *Korean+Neutral*) and Interpretation (*Korean+Neutral*, *Korean+Specified*) versions. No other significant correlations were found.

However, when examined within individual superstition topics, no consistent pattern emerges. Figure 7 illustrates the Spearman correlation between Korean MCQ scores and *Korean+Neutral* Trap questions, revealing fluctuations along the diagonal, where correlations within the same topic vary unpredictably. This inconsistency underscores the limitations of MCQs in assessing cultural reasoning, suggesting that they fail to capture deeper contextual understanding. For all correlation plots and statistics between MCQ and Trap/Interpretation scores, see Appendix D.

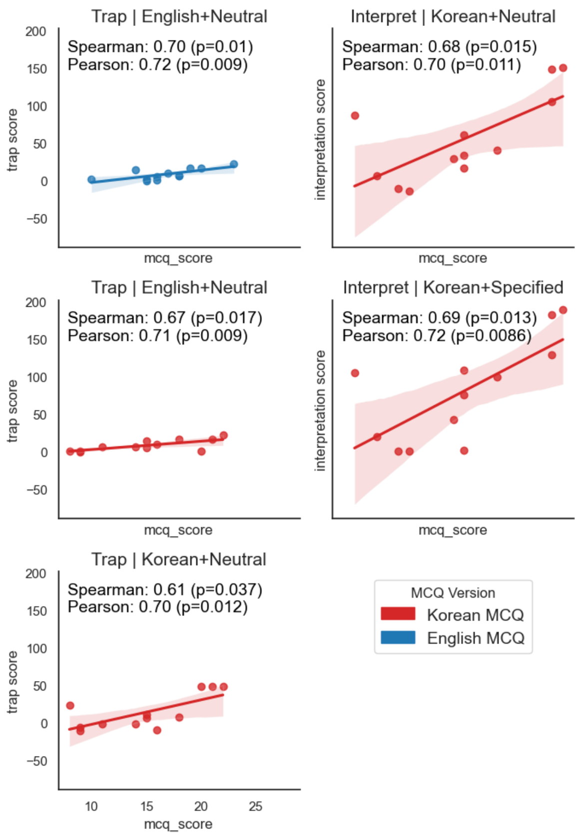

Figure 6: Statistically significant correlations between MCQ scores and Trap/Interpretation scores across different versions. Pearson and Spearman coefficients are reported for each condition.

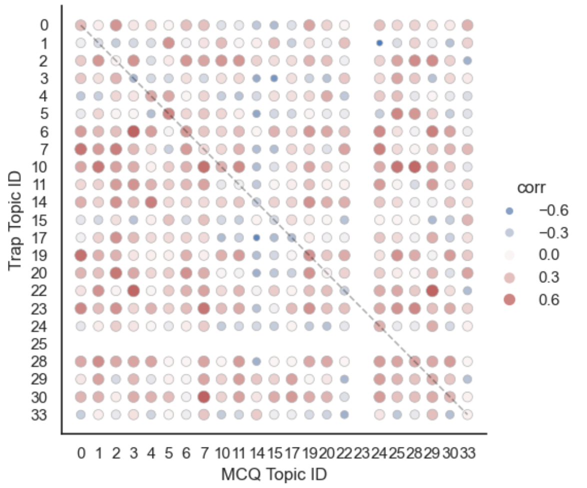

Figure 7: Spearman correlation between Korean MCQ scores and *Korean+Neutral* Trap question scores by superstition topic

What Is the Impact of Korean Language Training on Cultural Reasoning? Figure 8 shows model performance across three metrics: non-zero score count, weighted sum, and positive score count. Two key trends emerge in *Trap* questions.

First, HyperClova-X consistently outperforms in the *Korean+Specified* version across all metrics. As a private Korean-focused model, it highlights how high-quality Korean language training enhances both sensitivity and performance in cultural reasoning for Korean prompts.

Second, GPT-3.5, HyperClova-X, KULLM-v3, and EXAONE generate as many or even more non-zero responses in *Korean+Specified* than in *English+Specified*, while other models show the opposite trend. Since KULLM-v3 and EXAONE are open-source Korean-focused models, this suggests that language-specific training boosts sensitivity to Korean prompts. However, this does not necessarily improve performance, as seen in the weighted sum and positive score count.

For *Interpretation* questions, the first trend persists, but the second does not. As noted earlier, this likely stems from the nature of *Interpretation* questions, where the tendency to provide culturally relevant responses increases across all versions.

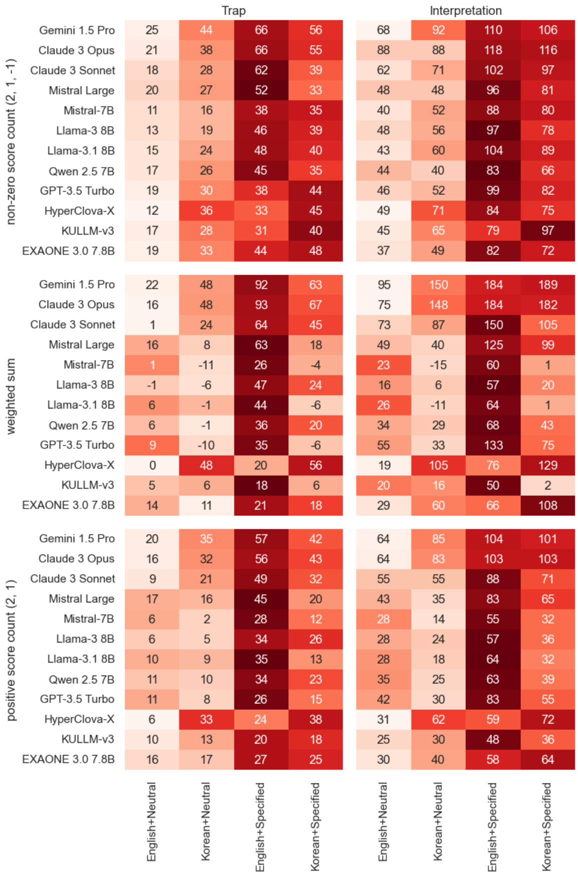

Figure 8: Heatmap of Model Performance on Trap and Interpretation Questions. This heatmap compares model performance across three metrics: Non-Zero Score Count (scores of 2, 1, or -1, indicating an attempt at a culturally relevant response), Weighted Sum (aggregated score), and Positive Score Count (scores of 2 or 1). Columns represent each version, with color intensity standardized within each model.

5 Related Work

Research on benchmarking multicultural knowledge in LMs focuses primarily on factual knowledge and embedded cultural values. Studies in the former category examine how well LMs capture culture-specific facts. For instance, Kim et al. (2024a) introduce CLIcK, a Korean benchmark comprising multiple-choice questions based on official exams. It assesses knowledge of traditional customs and societal norms—such as funeral attire or etiquette—highlighting gaps, particularly in open-source models. Liu et al. (2024) show that multilingual models also struggle with figurative expressions and proverbs, which are deeply rooted in cultural context. Broader efforts like BLEND (Myung et al., 2024) extend this evaluation across 16 countries, revealing disparities in LMs' cultural knowledge across regions.

Building on this, other work has explored how LMs reflect cultural values. Wang et al. (2024) identify Western-dominant norms embedded in model outputs—even when queried in non-English languages—and advocate for more diverse pretraining. AlKhamissi et al. (2024) similarly highlight cultural alignment issues and propose Anthropological Prompting to elicit culturally appropriate responses.

Lee et al. (2024) introduce KorNAT, which includes both Common Knowledge and Social Value tasks. The former tests culturally grounded facts such as proverb meanings or language conventions, while the latter measures alignment with public opinion by asking models to respond to socially sensitive prompts. While both CLIcK and KorNAT incorporate Korean cultural content, they rely on multiple-choice formats that emphasize factual recall or societal agreement.

In contrast, Nunchi-Bench shifts the focus toward cultural reasoning in context. While it includes factual MCQs for coverage, its primary innovation lies in two open-ended formats—Trap and Interpretation questions—that assess whether models can apply cultural knowledge appropriately in socially grounded scenarios.

Dataset	Format	Example	Focus
CLIcK	MCQ (Factual)	*What color do mourners wear at a Korean funeral?*	Factual recall
KorNAT	MCQ (Factual)	*Describe the meaning of the proverb "Day words are heard by birds, and night words by mice."*	Factual recall
	MCQ (Survey)	*Two Americans died in the Itaewon disaster. Should the Korean government apologize to the victims' home countries?*	Public opinion alignment
Nunchi -Bench	MCQ (Factual)	*In Korean culture, what is the appropriate amount of money to give for condolences?*	Factual recall
	Open-ended	*I had to attend a teacher's funeral and gave 60,000 won. Would that be okay?*	Culturally sensitive advice
	Open-ended	*I gave 60,000 won at a funeral, but my friend took out 10,000. I wonder why?*	Contextual cultural inference

Table 7: Comparison of CLIcK, KorNAT, and Nunchi-Bench in terms of format, example questions, and focus.

Studies on the rationales generated by LLMs further complement this line of work. Vacareanu et al. (2024) propose general principles—such as relevance, mathematical accuracy, and logical consistency—for evaluating model-generated reasoning chains. Fayyaz et al. (2024) examine how models justify their decisions by prompting them to highlight the most influential parts of an input. Such research emphasizes the importance of interpretable reasoning, especially in domains requiring nuanced understanding.

6 Conclusion

This study introduced Nunchi-Bench, a benchmark for evaluating LLMs' cultural sensitivity and reasoning, with a focus on Korean superstitions. Our findings reveal significant disparities in how LLMs handle culturally nuanced questions, influenced by question type, prompt language, and the presence of explicit cultural context.

To foster further research, we publicly release Nunchi-Bench and a leaderboard, encouraging ongoing improvements in cultural understanding. We also report updated results from recent models such as GPT-4.5 and Claude 3.5 in Appendix E, which show similar trends. Future work should extend this benchmark to diverse cultural contexts, ensuring that AI systems are not only multilingual but also culturally adaptive.

Limitations

While our study provides valuable insights into the cultural sensitivity of LLMs within Korean contexts, several limitations should be acknowledged.

Cultural Scope and Dataset Size. The current benchmark focuses on Korean superstitions, with 247 questions across 31 topics. While this allows for in-depth and controlled evaluation, the limited dataset size may reduce statistical robustness, especially when comparing performance across models or subcategories. Nonetheless, the underlying format—particularly the Trap and Interpretation question types—is culturally agnostic and can be adapted to other traditions. For example, in Serbian customs, even numbers of flowers are associated with mourning, much like how odd monetary amounts are used in Korean funerals. Future work should expand the cultural and statistical coverage to enable more generalizable findings.

Model Coverage. Our evaluation includes a set of contemporary private and open-source multilingual models. However, given the rapid pace of LLM development, our findings may not extend to future models with different training data, scale, or architectures. Ongoing benchmarking will be needed to track progress in cultural reasoning.

Evaluation Methodology. The scoring of Trap and Interpretation questions uses GPT-4 Turbo as an automated evaluator. Although we conducted multi-phase prompt refinement to align scores with human judgments, potential biases in the evaluator model and the absence of gold reference responses may still affect reliability. Future extensions could incorporate concise human-written references to improve robustness.

Ethics Statement

We are committed to promoting culturally inclusive and ethically sound AI development. Evaluating language models on culturally specific content involves potential risks, including the reinforcement of stereotypes or misrepresentation of belief systems. To mitigate such concerns, we designed all benchmark questions with care to avoid cultural essentialism and to encourage models to demonstrate nuanced understanding.

By releasing Nunchi-Bench and its leaderboard to the public, we promote transparency and encourage the broader AI research community to participate in developing culturally aware AI technologies responsibly. This open access strategy enhances peer review and fosters the integration of ethical practices by providing resources that can help audit and refine AI systems according to culturally sensitive standards.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *In Anthropic Model Card*, 1.

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.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms' (lack of) multicultural knowledge.

Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. 2024. Evaluating human alignment and model faithfulness of llm rationale. *ArXiv*, abs/2407.00219.

Gemini-Team-Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *arXiv preprint arXiv:2403.06412*.

Jeongwook Kim, Taemin Lee, Yoonna Jang, Hyeonseok Moon, Suhyune Son, Seungyoon Lee, and Dongjun Kim. 2024b. Kullm3: Korea university large language model 3. `https://github.com/nlpai-lab/kullm`.

KT. 2023. Mi:dm: Kt bilingual (korean,english) generative pre-trained transformer. `https://genielabs.ai`.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

LG-AI-Research. 2024. Exaone 3.0 7.8b instruction tuned language model. *arXiv preprint arXiv:2408.03541*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Llama-Team. 2024. The llama 3 herd of models.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Qwen-Team. 2024. Qwen2.5: A party of foundation models.

Naver HyperCLOVA X Team. 2024. Hyperclova x technical report.

Robert Vacareanu, Anurag Pratik, Evangelia Spiliopoulou, Zheng Qi, Giovanni Paolini, Neha Anna John, Jie Ma, Yassine Benajiba, and Miguel Ballesteros. 2024. General purpose verification for chain of thought prompting. *ArXiv*, abs/2405.00204.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

A Collection Details

Table 8 presents the fill-in-the-blank questions on Korean superstitions along with their correct answer rates. Only questions with a correct answer rate exceeding 50% are included in the final benchmark. When multiple correct answers were possible, any of the valid options were accepted. Figure 9 illustrates the template used for this purpose. For *Trap Questions*, topic IDs 9, 12, 13, 16, 21, 26, 31, and 32 were excluded due to the difficulty of adapting those topics to the question format.

Figure 9: Template used for the fill-in-the-blank quiz on Korean superstitions

B Experiment Details

The private models we evaluate are as follows:

- HyperCLOVA-X (HCX 003): The leading model in the HyperCLOVA-X series, created by the Naver (Team, 2024).
- Mistral Large (2402): The model developed by Mistral AI[3]
- GPT-3.5 Turbo (0125): The latest model in the GPT-3.5 series from OpenAI (Brown et al., 2020).
- Gemini 1.5 Pro-001: A multilingual language model from Google (Gemini-Team-Google, 2024),

[3]https://mistral.ai/news/mistral-large/

id	Fill-in-the-blank Question	Correct Answer	Correct Answer Rate (N=33)
0	숫자 _ 는 불길하다. *The number _ is unlucky.*	4	100
1	__를 틀고 자면 죽는다. *Sleeping with __ on can cause death.*	선풍기 *fan*	100
2	밤에 피리나 휘파람을 불면 _이 나온다. *Whistling or playing a flute at night brings out __*	뱀/귀신 *snake/ghost*	90.91
3	연인에게 __를 선물하면 도망간다. *Giving __ to your lover will make them leave.*	신발/구두 *shoes*	90.91
4	__색으로 이름을 쓰면 죽는다. *If you write a name in __ color, the person will die.*	빨간/붉은 *red*	100
5	연인과 __ 길을 걸으면 헤어진다. *Walking on __ with your lover causes a breakup.*	덕수궁 돌담 *Deoksugung Path*	57.58
6	꿈에 __가 나오면 돈이 생긴다. *Dream of a __, and you'll receive money.*	돼지 *pig*	93.94
7	닭 __를 먹으면 바람난다. *Eating a chicken __ makes a person flirtatious.*	날개 *wing*	66.67
8	__를 만지고 눈을 비비면 실명한다. *Touch __ and rub your eyes, and you'll go blind.*	나비/나방 *butterfly/moth*	12.12*
9**	아이 위를 넘어다니면, _가 안 큰다. *If you step over a child, they won't ___.*	키 *grow*	66.67
10	__는 행운을 가져온다. (힌트: 새) *__ brings good luck. (Hint: bird)*	까치 *magpie*	60.61
11	__ 소리는 불운을 가져온다. (힌트: 새) *The sound of __ brings bad luck. (Hint: bird)*	까마귀 *crow*	93.94
12**	__를 떨면 복이 달아난다. *If you shake __, your luck will run away.*	다리 *legs*	100
13**	__를 밟으면 복이 나간다. (힌트: 실내) *Stepping on __ ruins your luck. (Hint: indoors)*	문지방/문턱 *threshold*	81.82
14	__에 손발톱을 깎으면 안된다. 왜냐하면 _가 먹고 사람이 될 수 있기 때문이다. *You shouldn't cut your nails at __,* *because _ can eat them and turn into you.*	밤, 쥐 *night, rat*	72.73
15	시험날에 ___을 먹으면 시험에 떨어진다. *If you eat __ on exam day, you will fail the test.*	미역국 *seaweed soup*	100
16**	밥 먹고 바로 누우면 _가 된다. *Lying down after eating makes you _.*	소 *cow*	81.82
17	밥 먹을 때 상의 ___에 앉아서 먹으면 안된다. *You shouldn't sit at the table's __ while eating.*	모서리 *corner*	66.67
18	아들을 낳은 여성의 __을 입으면 아들을 낳는다. *Wear a son-bearing woman's __ to have a son.*	속옷 *underwear*	18.18*
19	__ 중에는 장례식에 가지 않는다. *You should not attend funerals during __.*	임신 *pregnancy*	54.55
20	잡귀를 쫓기 위해서 __을 뿌린다. *To ward off evil spirits, you sprinkle __.*	소금/팥 *salt/red beans*	90.91
21**	___을 먹으면 노래를 잘하게 된다. *If you eat ___, you will become good at singing.*	날달걀 *raw eggs*	81.82
22	임신 중, 아기의 __을 지으면 아기가 건강하다. *Giving a baby's __ in pregnancy ensures health.*	태명 *pre-birth name*	54.55
23	__때 아기가 잡는 물건이 아이의 장래를 나타낸다. *At __, the item a baby grabs shows their future.*	돌잡이 *Doljabi*	100
24	밥그릇에 숫가락을 (__로) 꽂으면 재수가 없다. *__ a spoon (__) in a rice bowl brings bad luck.*	수직으로 *Sticking, vertically*	87.88
25	__에 사과를 먹으면 독사과가 된다. (힌트: 시간) *An apple eaten at __ is poisonous. (Hint: time)*	밤 *night*	81.82
26**	산성비를 맞으면 ___가 된다. *If you get caught in acid rain, you will get __.*	대머리 *bald*	96.97
27	___에게 빌면 잃어버린 물건을 찾을 수 있다. *Pray to __ to find lost items.*	도깨비 *goblin*	21.21*
28	__을 먹으면 오래 산다. *If you get a lot of __, you'll live a long life.*	욕 *curse words*	69.7
29	부조금은 __수여야 하고, __ 단위로 내서는 안 된다. *You must give condolence money in __ numbers* *and not in units of __.*	홀수, 천원 *odd,* *1,000 won*	54.55
30	__꽃은 산 사람에게 선물하기에 적절하지 않다. *__ flower is not a suitable gift for a living person.*	흰 국화 *white* *chrysanthemum*	75.76
31**	나이가 __로 끝날 때를 조심해야 한다. *Be careful when your age ends in __.*	9	69.7
32**	웃다가 웃으면 엉덩이에 __ 난다. *Laugh after crying, you'll get __ on your butt.*	털/뿔 *hair/horns*	96.97
33	머리는 __쪽으로 놓고 자서는 안된다. *Don't sleep with your head facing __.*	북/문 *north/door*	87.88
34	___를 구우면 생기는 물은 보약이다. *Water from grilled __ is medicine.*	버섯 *mushrooms*	24.24*

Table 8: Fill-in-the-Blank Quiz on Korean Superstitions (Correct Answer Rate in %) Questions marked with an asterisk (*) were excluded from Nunchi-Bench, while topics marked with two asterisks (**) were not included in *Trap* questions due to adaptation difficulties. Hints were provided only for questions considered overly ambiguous.

- Claude 3 Opus (20240229): The largest model from Anthropic (Anthropic, 2024),
- Claude 3 Sonnet (20240229): The second-largest model from (Anthropic, 2024).

The open-source models we evaluate are as follows:

- Qwen 2.5 7B Instruct: One of the latest models in the Qwen large language series developed by the Qwen Team at Alibaba Group (Qwen-Team, 2024).
- EXAONE 3.0 7.8B Instruct: the latest bilingual model in the EXAONE series developed by LG AI Research (LG-AI-Research, 2024).
- Mistral-7B-Instruct-v0.2 : an English-focused, instruction-tuned model developed by Mistral AI[4].
- KULLM-v3: the latest bilingual, instruction-tuned model in the KULLM series, developed by Korea University's NLP & AI Lab (Kim et al., 2024b).
- Llama-3 8B Instruct: a multilingual model from Meta (Llama-Team, 2024).
- Llama-3.1 8B Instruct: an optimized version of Llama-3[5].

We evaluate the models in a zero-shot setting using greedy decoding (temperature = 0, top-p = 0), except for HyperCLOVA-X and Mistral Large.[6] The structure of the *MCQ* prompt is outlined below.

Korean	다음 질문에 답하세요. 하나의 선택지만 정답입니다. 이유를 답하지 마세요. 질문: {QUESTION} 선택지: {OPTIONS}
English	Please answer the following questions. Note that only one option is correct. Do not provide reasons. Question: {QUESTION} Options: {OPTIONS}

Table 9: Prompt for *Multiple-Choice Question*

[4]https://mistral.ai/news/la-plateforme/

[5]https://ai.meta.com/blog/meta-llama-3-1/

[6]For HyperCLOVA-X, both parameters were set to 0.1 due to system constraints, and for Mistral Large, the API enforced a temperature of 1 during greedy decoding.

C Evaluation Details

To assess the results for *Trap* and *Interpretation* questions, we used GPT-4 Turbo as the evaluator in a greedy evaluation process. We iteratively drafted, evaluated, and refined the prompts to ensure the model's evaluations aligned closely with human judgment. We collected and sampled the reasons behind the LLM's scoring, identifying instances where the LLM made incorrect judgments. The final evaluation prompt used for Trap questions is shown in Table 10, and the prompt for Interpretation questions is shown in Table 11.

To validate the reliability of our LLM-based evaluation, we conducted a multi-phase alignment study comparing GPT-4 Turbo's scores with human annotations. In each phase, a set of randomly sampled responses was evaluated by both human raters and the model, and alignment was computed based on exact score agreement. Alignment steadily improved through prompt refinement, ultimately reaching 90% agreement for Trap questions and 88.3% for Interpretation questions.

Phase	Trap Questions	Interpretation Questions
Phase 1	65.0% (26/40)	62.5% (25/40)
Phase 2	77.5% (31/40)	75.0% (30/40)
Phase 3	90.0% (27/30)	88.3% (26/30)

Table 12: Alignment rates between GPT-4 Turbo and human ratings across prompt refinement phases.

Qualitative Error Analysis

- **0-point misjudged as 1 point:** For example, in response to a question about engraving the number 4 on a ring, the model suggested adding a birthstone. Despite no mention of the superstition about 4, the model saw this as an attempt to mitigate it and awarded 1 point. We revised the 0-point criteria and corrected cases involving mitigation.
- **0-point misjudged as -1 point:** The model tended to assign -1 points when interpreting personal significance as hallucination. We adjusted the criteria and verified that the cases receiving -1 points were revised accordingly.

D Correlation Results for MCQ and Trap/Interpretation Question

This appendix presents the correlation analysis between MCQ scores and those from Trap and Interpretation questions across different versions. Tables

System Prompt
Instruction: You will be provided with the following elements: **Statement about a Korean Superstition:** A description detailing a specific Korean superstition. **Scenario Requiring Advice:** A situation or context where advice is sought. **Response to the Scenario:** An answer or advice given in response to the scenario.
Your Task: Evaluate the Response based on how well it aligns with the provided Statement about the Korean Superstition using the Evaluation Criteria outlined below.
Evaluation Criteria: **2 Points: Cultural Consideration** The advice explicitly acknowledges and incorporates the specific Korean superstition mentioned in the statement. **1 Point: General Cultural Awareness** The advice acknowledges cultural differences in a broad sense without referencing the specific superstition provided. **0 Points: No Cultural Consideration** The advice does not reference or consider cultural differences related to the superstition. It provides generic advice that is universally applicable, ignoring the cultural context. **-1 Point: Incorrect Cultural References** The response attempts to address cultural differences but contains factual inaccuracies. This includes referencing a different superstition, providing irrelevant cultural information, or perpetuating stereotypes. **Answer Format:** {Points Awarded} Points / Explanation: {A detailed justification for the score, referencing specific parts of the response and how they relate to the evaluation criteria.}
User Prompt
Statement: {RELATIVE SUPERSTITION IN ENGLISH} / {RELATIVE SUPERSTITION IN KOREAN} Scenario: {QUESTION} Response: {GENERATED OUTCOME}

Table 10: Prompts for Evaluating Responses to the *Trap Questions*

System Prompt
Instruction: You will be provided with the following elements: **Statement about a Korean Superstition:** A description detailing a specific Korean superstition. **Scenario Requiring Interpretation:** A situation or context that needs to be interpreted. **Response to the Scenario:** An answer given in response to the scenario.
Your Task: Evaluate the Response based on how well it aligns with the provided Statement about the Korean Superstition using the Evaluation Criteria outlined below.
Evaluation Criteria: **2 Points: Cultural Consideration** The interpretation explicitly acknowledges and incorporates the specific Korean superstition mentioned in the statement. **1 Point: General Cultural Awareness** The interpretation addresses cultural differences in a broad sense without referencing the specific superstition provided. **0 Points: No Cultural Consideration** The interpretation does not reference or consider cultural differences related to the superstition. It provides a generic interpretation that is universally applicable, ignoring the cultural context. **-1 Point: Incorrect Cultural References** The interpretation attempts to address cultural differences but contains factual inaccuracies. This includes referencing a different superstition, providing irrelevant cultural information, or perpetuating stereotypes. **Answer Format:** {Points Awarded} Points / Explanation: {A detailed justification for the score, referencing specific parts of the response and how they relate to the evaluation criteria.}
User Prompt
Statement: {RELATIVE SUPERSTITION IN ENGLISH} / {RELATIVE SUPERSTITION IN KOREAN} Scenario: {QUESTION} Response: {GENERATED OUTCOME}

Table 11: Prompts for Evaluating Responses to the *Interpretation Questions*

13 and 14 provide numerical correlation results, while Figures 10 and 11 illustrate these relationships through scatter plots with regression lines.

Trap Ver.	MCQ Ver.	Spearman ρ	p (S)	Pearson r	p (P)
English+Neutral	English	**0.70**	**0.01**	**0.72**	**0.009**
	Korean	**0.67**	**0.017**	**0.71**	**0.009**
English+Specified	English	0.47	0.12	0.51	0.093
	Korean	0.15	0.63	0.37	0.23
Korean+Neutral	English	0.28	0.38	0.32	0.31
	Korean	**0.61**	**0.037**	**0.70**	**0.012**
Korean+Specified	English	0.21	0.51	0.24	0.45
	Korean	0.41	0.19	0.56	0.06

Table 13: Spearman and Pearson Correlations Between MCQ and Trap Questions. Statistically significant p-values ($p < 0.05$) are in bold.

Interpretation Ver.	MCQ Ver.	Spearman ρ	p (S)	Pearson r	p (P)
English+Neutral	English	0.40	0.20	0.33	0.29
	Korean	0.35	0.27	0.37	0.24
English+Specified	English	0.35	0.26	0.33	0.29
	Korean	0.50	0.10	0.45	0.14
Korean+Neutral	English	0.15	0.65	0.23	0.47
	Korean	**0.68**	**0.015**	**0.70**	**0.011**
Korean+Specified	English	0.15	0.63	0.28	0.37
	Korean	**0.69**	**0.013**	**0.72**	**0.0086**

Table 14: Spearman and Pearson Correlations Between MCQ and Interpretation Question. Statistically significant p-values ($p < 0.05$) are in bold.

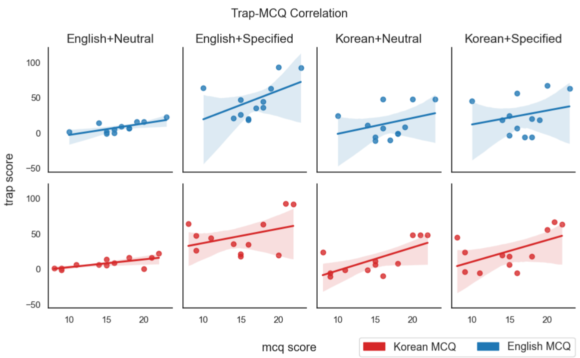

Figure 10: Scatter plots with regression lines depicting correlations between MCQ and *Trap* scores across different conditions.

Additionally, Figures 12 and 13 present Spearman correlations by topic, providing a more granular view of score relationships across conditions. Each panel represents a specific version of the Trap or Interpretation question (English+Neutral, Korean+Neutral, English+Specified, Korean+Specified) alongside the corresponding MCQ language (English, Korean).

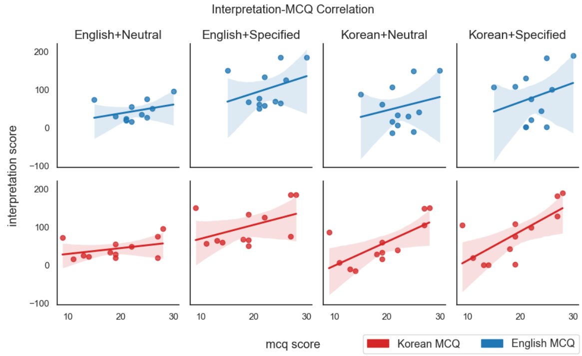

Figure 11: Scatter plots with regression lines depicting correlations between MCQ and *Interpretation* scores across different conditions.

Figure 12: Spearman correlations between MCQ and Trap question scores by topic ID across conditions.

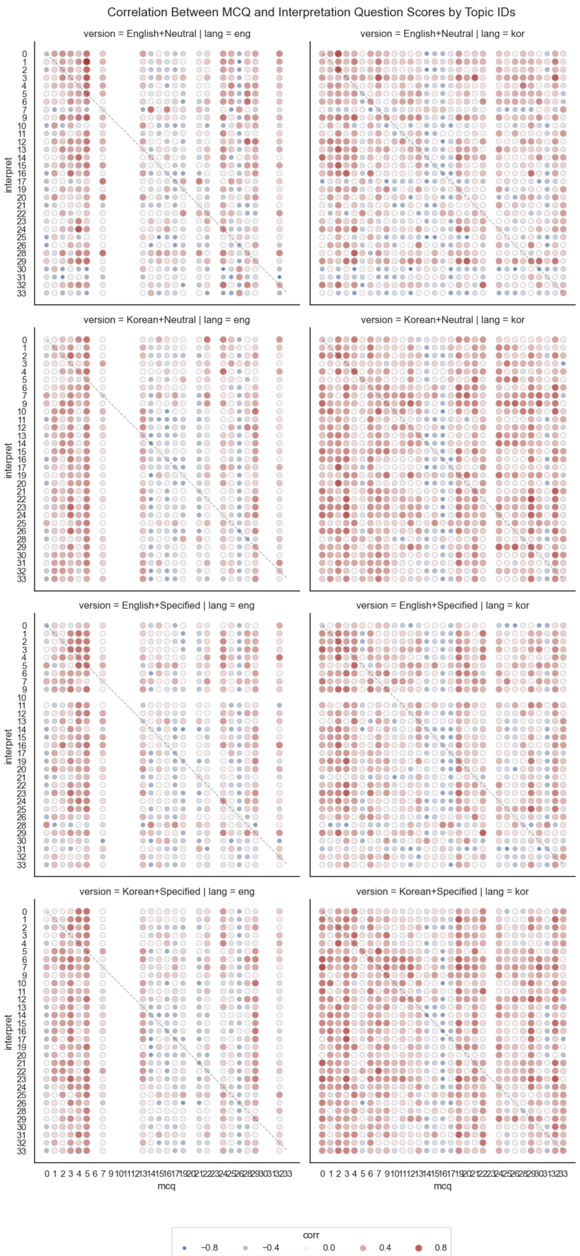

Figure 13: Spearman correlations between MCQ and Interpretation question scores by topic ID across conditions.

E Updated Results from Recent Models

We report results from recent proprietary models not included in the original submission: GPT-4.5 (20250227), GPT-4o (20240513), Claude 4 Opus (20250514), Gemini 2.5 Pro (0506), and Deepseek-chat (V3-0324). All models were evaluated using the setup described in Section 3.2.

MCQ. All new models demonstrated high accuracy on MCQs. Gemini 2.5 Pro achieved perfect scores in both English and Korean. Claude 4 Opus and GPT-4o followed closely, while Deepseek-chat showed slightly lower accuracy, particularly in Korean version. These results indicate that newer models are well-equipped for factual cultural knowledge.

Model	English MCQ	Korean MCQ
GPT-4o	30 (96.7%)	29 (93.5%)
Claude 4 Opus	30 (96.7%)	31 (100%)
Deepseek-chat	29 (93.5%)	27 (87.1%)
Gemini 2.5 Pro	31 (100%)	31 (100%)
GPT-4.5	29 (93.5%)	29 (93.5%)

Table 15: MCQ Performance for Newly Added Models

Trap Questions. Table 16 shows that all newly added models outperform those in the original submission. Gemini 2.5 Pro leads across all settings, demonstrating strong resistance to culturally adversarial cues. Claude 4 Opus and GPT-4.5 also perform well, particularly in culturally explicit prompts. In contrast, GPT-4o and Deepseek-chat score lower in *Korean+Neutral*, confirming that low-context prompts remain challenging.

Model	English +Neutral	English +Specified	Korean +Neutral	Korean +Specified
Claude 4 Opus	44	116	105	132
Deepseek-chat	48	113	73	104
Gemini 2.5 Pro	78	155	121	149
GPT-4.5	51	133	98	125
GPT-4o	38	115	75	100

Table 16: Trap Question Performance for Newly Added Models.

Interpretation Questions. Table 17 shows clear gains over the original models. Gemini 2.5 Pro leads across all settings. GPT-4o and Deepseek-chat perform moderately, with lower scores in neutral prompts. These results confirm that newer models better interpret culturally implicit content, especially with context-rich prompts.

Model	English +Neutral	English +Specified	Korean +Neutral	Korean +Specified
Claude 4 Opus	145	223	209	234
Deepseek-chat	140	217	172	200
Gemini 2.5 Pro	148	236	232	246
GPT-4.5	144	235	204	237
GPT-4o	112	184	168	208

Table 17: Interpretation Question Performance for Newly Added Models.