Everyday Physics in Korean Contexts: A Culturally Grounded Physical Reasoning Benchmark

Jihae Jeong[1,*] **DaeYeop Lee**[1,2,*] **DongGeon Lee**[1] **Hwanjo Yu**[1,†]
[1]POSTECH [2]MODULABS
{wisdomjeong, daeyeoplee, donggeonlee, hwanjoyu}@postech.ac.kr

Abstract

Existing physical commonsense reasoning benchmarks predominantly focus on Western contexts, overlooking cultural variations in physical problem-solving. To address this gap, we introduce EPiK (Everyday Physics in Korean Contexts), a novel benchmark comprising 181 binary-choice questions that test physical reasoning within Korean cultural contexts—from kimchi (Korean food) to traditional fermentation. EPiK was constructed using a two-stage generation pipeline and verification pipeline to create culturally-authentic questions across 9 reasoning subtasks and 84 scenarios. Unlike approaches based on simple translation, our method generates questions organically from Korean contexts while upholding rigorous physical reasoning standards. Our evaluations reveal that Korean-specialized models consistently outperform general-purpose models of comparable size. This performance gap highlights the limitations of culturally-agnostic models and demonstrates the critical need for culturally-aware benchmarks to truly measure language understanding. We release EPiK to catalyze the development of AI capable of reasoning across a diverse spectrum of cultural contexts.

1 Introduction

Physical commonsense reasoning, the ability to understand and predict everyday physical phenomena, is fundamental to human intelligence and a critical capability for AI systems that interact with the real world (Ismayilzada et al., 2023; Pensa et al., 2024). While recent benchmarks (Bisk et al., 2020; Wang et al., 2023) have achieved substantial progress in evaluating this capability, they predominantly focus on Western, English-speaking contexts, leaving a significant gap in our understanding of how AI systems perform across diverse cultural and linguistic environments (Ponti et al., 2020; Shi et al., 2024).

This gap is particularly problematic because physical reasoning, despite appearing universal, is deeply intertwined with cultural context (Acharya et al., 2020; Shi et al., 2024; Acquaye et al., 2024; Shen et al., 2024). Consider the task of "preventing food spoilage during summer." While Western-centric benchmarks might emphasize refrigeration, Korean commonsense includes understanding fermentation dynamics in *ondol* (온돌, Korean traditional heating system) and the role of *a-gung-i* (아궁이, Korean traditional fireplace). Such culturally-embedded physical knowledge (Acharya et al., 2020; Liu et al., 2025) is essential for AI systems deployed in Korean contexts, yet remains unmeasured by existing benchmarks.

Current evaluation practices for Korean language models typically rely on translated versions of English benchmarks, which introduce multiple limitations (Sakai et al., 2024). First, translation often distorts the physical scenarios being described, as certain concepts lack direct equivalents across languages (Artetxe et al., 2020, 2023). Second, translated benchmarks fail to capture Korea-specific physical phenomena that arise from unique environmental conditions (e.g., distinct monsoon patterns), living arrangements (e.g., floor heating systems), and material culture (e.g., traditional cooking implements). Third, this approach perpetuates a Western perspective of *universal* physical reasoning, potentially overlooking diverse problem-solving strategies that emerge from different cultural contexts (Hershcovich et al., 2022; Koto et al., 2024; Myung et al., 2024; Liu et al., 2025).

To address these limitations, we introduce the EPiK (Everyday Physics in Korean Contexts) benchmark, a comprehensive dataset designed to evaluate physical commonsense reasoning within authentic Korean contexts. EPiK consists of 181 carefully curated binary-choice questions that require an understanding of both universal physical

[*]Both authors contributed equally to this work.
[†]Corresponding author.

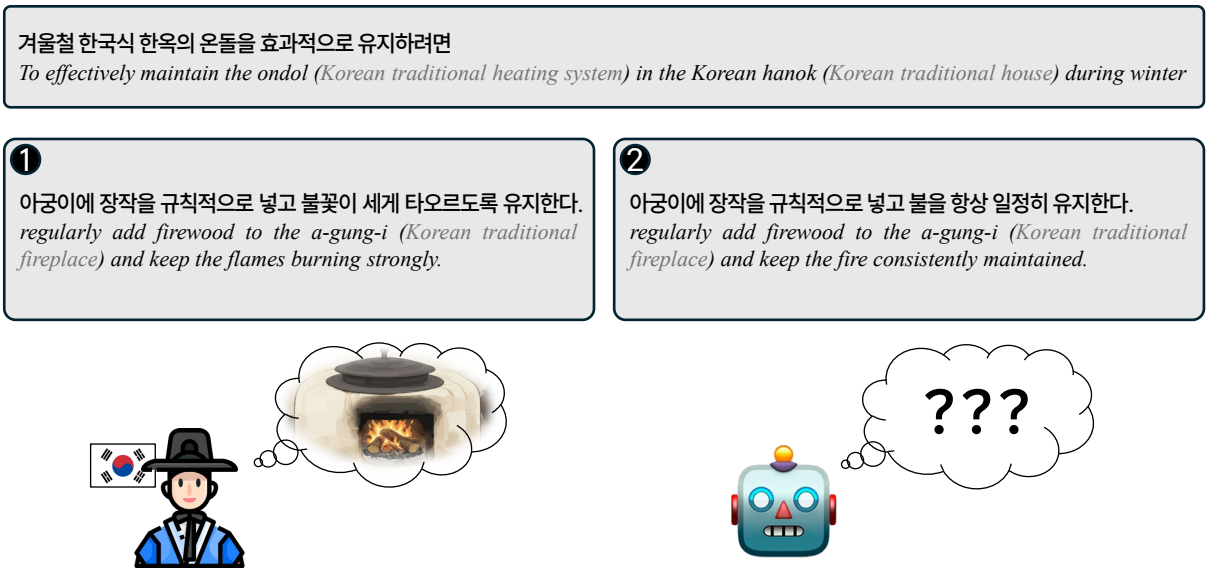

Figure 1: An illustrative example from our proposed EPiK benchmark. The problem requires understanding the traditional Korean *ondol* heating system, demonstrating the need to integrate physical reasoning with specific cultural contexts, a task that can be especially challenging for multi-lingual language models.

principles and Korea-specific applications. Unlike simple translations, our questions are generated from the ground up to reflect genuine Korean daily life scenarios while maintaining rigorous standards for physical reasoning complexity.

Our contributions are summarized as follows:

- We propose a rigorous methodology for building a culturally-grounded benchmark, featuring a systematic taxonomy that bridges universal physical principles with Korean-specific contexts and a two-stage pipeline with interactive verification and bias filtering to ensure question validity.
- Through extensive experiments on variety of models, we reveal a significant performance gap between general-purpose and Korean-specialized models, highlighting the critical need for culturally-aware evaluation.
- We publicly release the EPiK dataset as a novel resource to facilitate research on culturally-aware reasoning and to promote the development of more inclusive and globally competent models.

Our results demonstrate that Korean-specialized models consistently outperform general models on tasks requiring Korean-contextualized physical reasoning. This gap underscores the need for evaluation frameworks that respect linguistic and cultural diversity in assessing AI capabilities. Beyond Korean contexts, EPiK serves as a model for developing culturally-grounded benchmarks in other languages, contributing to a more comprehensive understanding of physical commonsense reasoning across human cultures.

2 Related Work

2.1 Physical Commonsense Reasoning Benchmark

Physical commonsense reasoning has emerged as a critical challenge in natural language understanding (NLU) (Davis, 2024). PIQA (Bisk et al., 2020) introduced physical commonsense as a textual reasoning task where, given a goal from everyday life, one must choose the more plausible method among two candidates. Despite being easy for humans, PIQA exposed a large gap for pretrained models, attributing difficulty to reporting bias in text-only pretraining (Paik et al., 2021).

Our benchmark follows the PIQA two-choice format and the same core objective (plausible physical action selection) but grounds both questions and answers in Korean daily life, thus serving as a 'Korean Cultural PIQA.'

2.2 Korean Commonsense Benchmark

Most commonsense evaluations have been English-centric or translation-driven (Ponti et al., 2020; Lin et al., 2021). Multilingual commonsense benchmarks often arise via translation or projection from English resources, which can import artifacts and dilute culturally prototypical solutions (Nie et al., 2024).

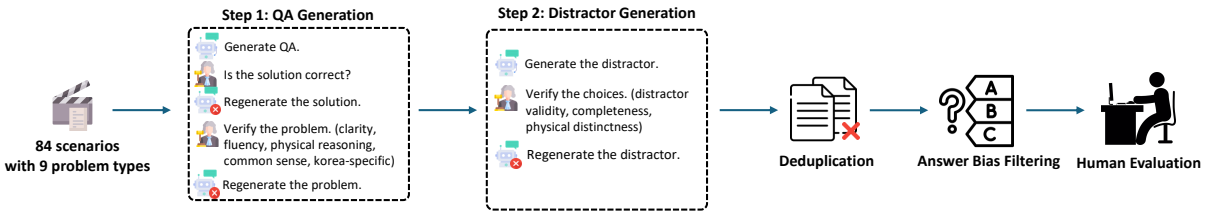

Figure 2: An overview of EPiK construction process. **1) Foundational Framework**: Define 84 scenarios and 9 problem types. **2) Two-stage Problem Generation and Verification**: Generate QA samples and verify the problems. **3) Deduplication 4) Answer Bias Filtering 5) Human Evaluation**

Recent efforts emphasize culturally grounded assessments, particularly for Korean language resources. CLIcK (Kim et al., 2024) collects language-and-culture QA from official exams and textbooks. KoCommonGEN v2(Seo et al., 2024) takes a different approach by reconstructing a commonsense generation dataset from scratch.

Complementary to these, EPiK is built from the ground up around Korean daily-life physical reasoning rather than translation, targeting safety-, tool-, and procedure-sensitive decisions that users routinely face. General NLU suites like KLUE (Park et al., 2021) provide broad coverage but do not specifically isolate physical commonsense in Korean.

3 Dataset Construction

EPiK is designed to evaluate the physical commonsense reasoning within the unique context of Korean culture. The problems are formulated to be solvable by native Korean speakers with general everyday knowledge of traditional and contemporary life. The dataset follows a two-alternative multiple-choice format, requiring a model to select the correct solution from two candidates. The entire construction pipeline is illustrated in Figure 2.

3.1 Foundational Framework: Scenarios and Task Taxonomy

To ensure comprehensive coverage and diversity, we first established a foundational framework consisting of background scenarios and a taxonomy of physical reasoning tasks.

Task Taxonomy To systematically guide the question generation process, we established a task taxonomy inspired by the PIQA dataset (Bisk et al., 2020). It comprises four high-level categories and nine fine-grained sub-categories capturing a wide range of physical reasoning skills. This taxonomy, detailed in Table 1, serves as a structured framework to ensure a balanced distribution of reasoning types. However, it is important to note that the generated questions often integrate concepts from multiple categories.

Scenario Sourcing To anchor our questions in diverse and realistic contexts, we source background scenarios from the TRACIE dataset (Zhou et al., 2021). We extracted and deduplicated the premises from its training set, yielding in a pool of 84 unique scenarios. For each scenario, we systematically generate questions corresponding to our nine task types, ensuring broad coverage of both situational contexts and reasoning challenges.

3.2 Two-Stage Generation and Verification Pipeline

We implemented two-stage pipeline to generate high-quality questions with unambiguous answers and plausible distractors. This approach decouples the creation of the correct solution from the distractor, allowing for rigorous quality control at each step. Implementation Details are in Appendix A, and the full prompts can be found in Appendix B.

3.2.1 Stage 1: Question and Correct Answer Generation

Generation The generation process is guided by a background scenario and a target question type from our taxonomy. The goal is to produce question-answer pairs where the solution requires an understanding of Korean cultural context for physical reasoning. To ensure stylistic diversity, we randomized parameters such as the maximum length of both the question and the answer (ranging from 10 to 24 characters, with a minimum combined length of 25 characters), and the final punctuation style of the question. Detailed prompts are provided in Appendix B.1.

Verification Each generated question-answer pair underwent a rigorous two-part verification pro-

Problem Type	Definition
기초 지식 및 속성 이해 *(Basic Knowledge and Attribute Comprehension)*	
물리 개념 및 원리 *Physical Concepts and Principles*	기본적인 물리 법칙과 원리에 대한 이해를 평가합니다. *Assesses understanding of fundamental physical laws and principles.*
객체 속성 및 기능 *Object Attributes and Functions*	사물의 재질, 무게, 형태, 질감 등 고유 속성과 가능한 상호작용을 이해하고 평가합니다. *Assesses understanding and evaluation of inherent properties like material, weight, shape, and texture, and potential interactions.*
물질의 상태와 변화 *States of Matter and Changes*	고체, 액체, 기체 상태의 특징과 녹는 것, 어는 것, 끓는 것, 녹이는 것 등 물질의 상태 변화의 원리를 이해하고 평가합니다. *Assesses understanding and evaluation of the characteristics of solid, liquid, and gas states, and the principles of state changes like melting, freezing, boiling, and dissolving.*
인과관계 및 동적 추론 *(Causality and Dynamic Reasoning)*	
물리적 결과 예측 *Predicting Physical Outcomes*	특정 행동이나 사건이 가져올 물리적 결과를 예측하는 능력을 평가합니다. *Assesses the ability to predict the physical outcomes of specific actions or events.*
물리적 원인 분석 *Analyzing Physical Causes*	이미 발생한 현상의 원인을 가장 타당한 물리적 관점에서 설명하는 능력을 평가합니다. *Assesses the ability to explain the cause of an already-occurred phenomenon from the most plausible physical perspective.*
목표 지향 및 응용 추론 *(Goal-Oriented and Applied Reasoning)*	
도구 및 절차 활용 *Utilizing Tools and Procedures*	목표 달성을 위해 적절한 도구를 선택하고 올바른 절차에 따라 사용하는 능력을 평가합니다. *Assesses the ability to select appropriate tools and follow correct procedures to achieve a goal.*
문제 해결 및 계획 *Problem Solving and Planning*	주어진 문제를 해결하기 위해 여러 단계의 물리적 행동을 계획하고 실행하는 능력을 평가합니다. *Assesses the ability to plan and execute multi-step physical actions to solve a given problem.*
물리적 타당성 판단 *Judging Physical Plausibility*	제시된 해결책이 물리적으로 가능한지, 현실 세계의 상식적인 도구와 개념에 기반하는지를 평가합니다. *Assesses whether a proposed solution is physically possible and based on common-sense tools and concepts in the real world.*
상황 및 제약 조건 추론 *(Situational and Constraint Reasoning)*	
위험성 및 안전성 평가 *Risk and Safety Assessment*	잠재적 위험을 예측하고 안전을 확보하기 위한 방법을 추론하는 능력을 평가합니다. *Assesses the ability to predict potential risks and infer methods to ensure safety.*

Table 1: The detailed taxonomy of the nine problem types designed for the EPiK. Our classification scheme is organized into four coarse-grained reasoning categories, each containing several fine-grained sub-categories.

cess:

1) Answer-Centric Verification This stage validates the correctness of the answer. If an answer is incorrect, the verifier provides a detailed rationale. This rationale serves as a corrective signal for the generator, which then produces a revised answer, creating an effective feedback loop (Appendix B.2).

2) Question-Centric Verification The stage assesses the quality of the question itself based on five criteria: (a) clarity, (b) presence of a physical reasoning component, (c) solvability with general Korean commonsense, (d) integration of Korean cultural elements, and (e) linguistic fluency. If a question-answer pair was discarded if it failed any criterion after three regeneration attempts (Appendix B.4).

3.2.2 Stage 2: Distractor Generation

Generation To create a compelling distractor, the generator introduces minimal but critical changes to the correct answer, applying slight lexical or phrasal perturbations while preserving sentence structure and style. This approach produces challenging negative examples that require a nuanced understanding of the underlying physical principle. Prompts are detailed in Appendix B.3

Verification Each generated distractor undergoes a dedicated verification procedure to ensure it meets a set of strict standards. The criteria for a valid distractor (Appendix B.5) are as follows:

1. It must be a demonstrably incorrect answer to the question.
2. It must be free of internal contradictions or logical fallacies.
3. The distinction between the correct answer and the distractor must hinge on a clear and identifiable physical principle, ensuring the choice is neither arbitrary nor ambiguous.

If a distractor fails to meet any of these criteria, it is regenerated. This iterative process is also repeated for a maximum of three attempts to secure a high-quality, challenging distractor for each question.

3.3 Deduplication

To ensure the novelty of our benchmark, we perform a deduplication step to filter out near-identical

instances. We employ the MinHash algorithm (Broder, 1997) to efficiently approximate the Jaccard similarity between all generated question pairs. After setting a similarity threshold of 0.6, we identify and remove 7 examples that are deemed near-duplicates. This filtering process enhances the diversity of the dataset and mitigates redundancy.

3.4 Answer Bias Filtering

To mitigate potential biases where models might identify the correct answer based on superficial cues rather than genuine reasoning, we implement a rigorous filtering stage. Artifacts such as disparities in answer length, specific word patterns, or inherent contradictions within the distractor can enable models to guess the correct option without true comprehension of the question.

To identify such flawed instances, we conducted an ablation experiment. We present only the two answer choices—the correct solution and the distractor—to a panel of Large Language Models (LLMs) without their corresponding questions. The models are tasked with predicting the correct answer from the choices alone. We include a third option, "Cannot be determined," for cases where the distinction is not inferable from the choices themselves, detailed in Appendix B.6.

For this task, we employ three distinct models: `gpt-4o-mini-2024-07-18`, `gpt-4.1-mini-2025-04-14`, and `gpt-4.1-nano-2025-04-14`. Any data instance where at least one of these models correctly identifies the answer is flagged and subsequently removed from the dataset. This rigorous filtering process ensures that only instances requiring genuine reasoning, rather than exploitation of statistical artifacts, remain in our benchmark.

3.5 Human Evaluation

To ensure the highest quality and validate our semi-automated generation pipeline, we conduct a comprehensive human evaluation phase. We employ a human expert with a bachelor's-level background in computer science and native proficiency in Korean. The evaluator is tasked with meticulously reviewing the dataset according to a detailed set of guidelines, which are provided in Appendix C. This manual verification aims to identify and filter out any remaining instances with subtle logical inconsistencies, cultural inaccuracies, ambiguous phrasing, or other quality issues that might have persisted through automated checks. Following this rigorous inspection of an initial pool of 335 candidate examples, 181 instances are certified as meeting all quality criteria, forming the final benchmark.

4 EPiK Benchmark

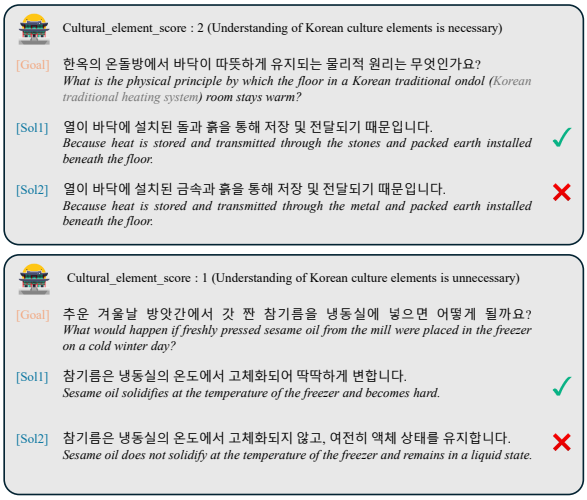

Figure 3: Illustration of the two levels of cultural dependency in our dataset. The top problem (score 2) is deeply rooted in specific Korean cultural knowledge, while the bottom problem (score 1) is culturally situated but solvable through universal scientific principles.

4.1 Dataset Description

EPiK benchmark is a collection of 181 binary-choice questions designed to evaluate physical reasoning in daily-life scenarios with a Korean cultural context. Each question consists of two candidate solutions, with exactly one correct answer. To prevent potential answer bias, the position of the correct answer is evenly distributed between the two options.

Each data sample contains the following fields:

- **problem_type_category**: The coarse-grained reasoning category of the question (e.g., basic knowledge and attribute comprehension, causality and dynamic reasoning, goal-oriented and applied reasoning, situational and constraint reasoning).
- **problem_type_subcategory**: A fine-grained description of the 9 reasoning types within the coarse category.
- **prompt**: The question statement or scenario to be solved.
- **solution0** and **solution1**: The two candidate solutions provided for the question.
- **label**: The index of the correct solution (0 or 1).

- **explanation**: A detailed explanation of why the correct answer is valid and why the alternative is incorrect.
- **korea_relevance**: Korean cultural elements utilized in the question and the answer.
- **physical_relevance**: Physical reasoning principles or elements used in the question and the answer.
- **cultural_element_score**: The degree to which the question leverages Korean-specific cultural elements. A score of 1 indicates that a Korean cultural element is mentioned, but the problem could be solved using alternative, non-Korean elements. A score of 2 indicates that Korean cultural elements are central to the question and answer, and understanding these elements is essential for correctly solving the problem.

Examples of EPiK can be found in Figure 3. The figure showcases two examples from our dataset, each representing a different level of cultural dependency as indicated by *cultural_element_score*.

For instance, the top example (score of 2) presents a scenario involving *ondol*, the traditional Korean underfloor heating system. Correctly answering this question requires an essential understanding of *ondol*'s installation and its principle of heating a floor from below. In contrast, the bottom example (score 1) features a question about storing *chamgireum* (sesame oil). Although it incorporates a Korean element, the problem can be solved with general scientific knowledge that most oils tend to solidify in a cold environment like a refrigerator. Thus, specific knowledge of *chamgireum* is not a prerequisite for solving the problem. This scoring system allows us to distinguish between problems that are merely set in a Korean context and those that are fundamentally rooted in Korean cultural knowledge.

4.2 Dataset Statistics

We analyze the statistical properties of our EPiK benchmark, focusing on two key aspects: the diversity of solution lengths and the distribution across problem categories.

Figure 4 illustrates the distribution of solution lengths within the EPiK benchmark. To encourage diversity in the length of generated text, we specified a random maximum word count for both the prompts and solutions during the data generation process. This approach resulted in the varied distribution, preventing the model from being biased towards solutions of a specific length.

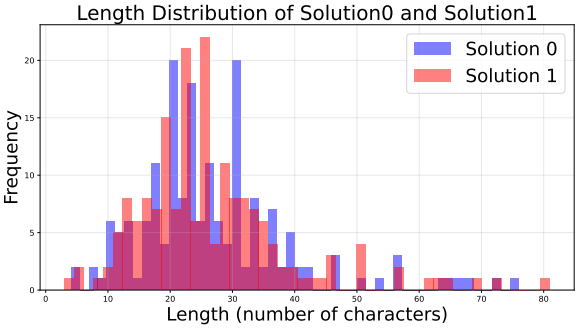

Figure 4: Distribution of solution lengths in the EPiK dataset. The varied distribution is a deliberate result of our data generation process, designed to mitigate potential length bias in model evaluation.

Category	Sub-category	
기초 지식 및 속성 이해 (Basic Knowledge and Attribute Comprehension)	물리 개념 및 원리 (Physical Concepts and Principles)	23 (12.71%)
	객체 속성 및 기능 (Object Attributes and Functions)	15 (8.29%)
	물질의 상태와 변화 (States of Matter and Changes)	22 (12.15%)
인과관계 및 동적 추론 (Causality and Dynamic Reasoning)	물리적 결과 예측 (Predicting Physical Outcomes)	31 (17.13%)
	물리적 원인 분석 (Analyzing Physical Causes)	18 (9.94%)
목표 지향 및 응용 추론 (Goal-Oriented and Applied Reasoning)	도구 및 절차 활용 (Utilizing Tools and Procedures)	19 (10.50%)
	문제 해결 및 계획 (Problem Solving and Planning)	12 (6.63%)
	물리적 타당성 판단 (Judging Physical Plausibility)	21 (11.60%)
상황 및 제약 조건 추론 (Situational and Constraint Reasoning)	위험성 및 안전성 평가 (Risk and Safety Assessment)	20 (11.05%)

Table 2: Distribution of the EPiK benchmark, broken down by our nine reasoning sub-categories.

Table 2 presents the distribution of samples across the nine pre-defined problem sub-categories, which are grouped under four main reasoning domains. The dataset is designed to have a relatively balanced representation of each category, ensuring a comprehensive evaluation of different physical reasoning abilities. However, it also includes many complex problems that integrate multiple sub-categories.

4.3 Two Solutions Analysis

4.3.1 Levenshtein Distance Distribution

Our benchmark was intentionally crafted to present a meaningful challenge by creating pairs of solu-

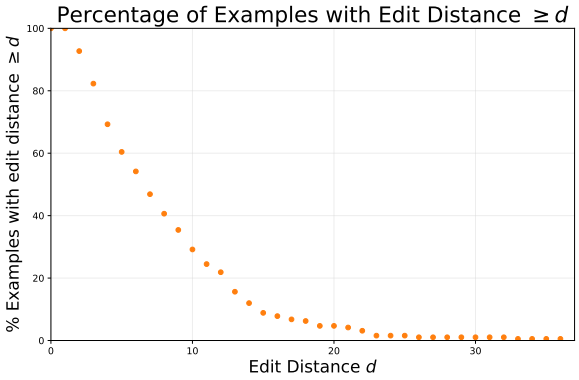

Figure 5: Distribution of the Levenshtein edit distance between the two solution candidates in the EPiK dataset. The strong skew towards lower distances confirms that most solution pairs have minimal lexical differences, requiring models to perform fine-grained semantic reasoning.

tion candidates with subtle lexical perturbations, while keeping their sentence structure, format, and length nearly identical. To quantitatively validate this design objective, we used the standard Levenshtein distance (Levenshtein, 1966), a metric that counts the minimum number of single-character edits—insertions, deletions, or substitutions—required to transform one solution into the other. As illustrated in Figure 5, the distribution of Levenshtein distances is heavily skewed toward smaller values, with the cumulative percentage of examples dropping off sharply as the edit distance increases. This confirms that the vast majority of our examples possess minimal lexical differences, thereby forcing models to rely on a deep semantic understanding to resolve fine-grained distinctions rather than superficial cues.

4.3.2 Analysis of Bias in Solutions

To ensure the quality of our dataset and verify that it primarily tests reasoning rather than reliance on superficial cues, we conducted an answer-only bias analysis. This experiment is designed to detect the presence of potential annotation artifacts—subtle, learnable patterns in the way correct and incorrect answers are phrased.

In this setup, models were presented with only the two solution candidates for each question, stripped of all contextual information from the question itself. The task was to predict the correct answer from these two options alone. For a two-alternative format, a random guess would yield an accuracy of 50%, which serves as our chance baseline. We evaluated a suite of contemporary models to probe for such biases.

Model	Accuracy (%)
`gpt-4o-mini-2024-07-18`	71.82
`gpt-4.1-mini-2025-04-14`	70.17
`gpt-4.1-nano-2025-04-14`	67.40
Baseline (random)	50.00

Table 3: Bias evaluation with two solutions. Models are tasked with selecting the correct solution from two options without seeing the corresponding question.

The performance of the models is detailed in Table 3. All models achieved scores significantly above the 50% chance baseline, with accuracies ranging from 67.40% to 71.82%. This indicates the presence of subtle yet detectable bias in solutions. However, these scores are also far from perfect, indicating that exploiting these artifacts is not a reliable strategy for solving the benchmark. Therefore, while some answer-only cues exist, we conclude that achieving high performance on the EPiK dataset still requires models to engage in genuine reasoning with the full question context.

5 Evaluation

5.1 Experiment Settings

All experiments were conducted on a single server equipped with an NVIDIA A100 GPU (80GB). To facilitate fast and memory-efficient inference, all response generation was handled using the vLLM (Kwon et al., 2023) library.

For decoding, we set *temperature*=0.7 across all experiments. We set *max_tokens*=4096 for reasoning models and *max_tokens*=3 for non-reasoning models.

We conducted a comprehensive evaluation on a wide range of open-source language models to assess performance on our benchmark. Our selection includes several models specialized for the Korean language: Kanana-1.5-2.1B-instruct, Midm-2.0-Mini-Instruct, A.X-3.1-Light, A.X-4.0-Light, Kanana-1.5-8B-Instruct, and EXAONE-4.0-32B (Bae et al., 2025). For comparison, we also evaluated leading general-purpose, instruction-tuned models across various sizes: Gemma-3-1B-it (Kamath et al., 2025), Qwen2.5-1.5B-Instruct (Yang et al., 2024), Gemma-2-2B-it (Rivière et al., 2024), Llama-3.2-3B-Instruct, Gemma-3-4B-it, Llama-3-8B-Instruct (Llama Team, AI @ Meta, 2024), Llama-3.1-8B-Instruct, Qwen3-8B, and Qwen3-

	Model	Accuracy (%)
	without Reasoning	
Korean-specialized Models	Kanana-1.5-2.1B-Instruct	78.45
	Midm-2.0-Mini-Instruct (2.3B)	77.90
	A.X-3.1-Light (7.2B)	86.19
	A.X-4.0-Light (7.2B)	**92.27**
	Kanana-1.5-8B-Instruct	89.50
General Models	Gemma-3-1B-it	55.80
	Qwen2.5-1.5B-Instruct	71.82
	Gemma-2-2B-it	66.85
	Llama-3.2-3B-Instruct	62.98
	Gemma-3-4B-it	**82.87**
	Llama-3-8B-Instruct	76.24
	Llama-3.1-8B-Instruct	81.22
	with Reasoning	
Korean-specialized Models	EXAONE-4.0-32B	82.32
General Models	Qwen3-8B	88.95
	Qwen3-32B	**91.71**
	Human Evaluation	
Native Korean		**96.69**

Table 4: Main evaluation results for Korean-specialized and General models on EPiK. Accuracy (%) is used as the metric. The best-performing model for each category is indicated in bold.

32B (Yang et al., 2025).

Evaluation of non-reasoning models was conducted on models with 8B parameters or fewer. All reported results are based on a single inference run for each model.

5.2 Results and Analysis

The main evaluation results for both Korean-specialized and general-purpose models on the EPiK benchmark are presented in Table 4.

Our primary finding is that Korean-specialized models consistently outperform general models of a comparable size. For instance, in the non-reasoning category, the A.X-4.0-Light (7.2B) model achieved the highest accuracy of all models at 92.27%, significantly surpassing the similarly sized Llama-3.1-8B-Instruct (81.22%). This trend demonstrates that specialized training on Korean language, data, and cultural contexts provides a distinct advantage for solving the problems presented in EPiK.

An interesting observation is the strong performance of the Chinese-specialized Qwen model family. We hypothesize that this is due to the cultural proximity between China and Korea, which may allow these models to better understand contextual nuances compared to models trained predominantly on Western data. Despite this, the top-performing models in their respective size classes are still the Korean-specialized ones, reinforcing the value of culturally specific training.

Furthermore, the results highlight a fascinating trade-off between model scale, reasoning abilities, and specialization. The A.X-4.0-Light model (7.2B), without any explicit reasoning prompt, outperformed even the much larger Qwen3-32B model (91.71%) which is a reasoning model. This suggests that for a culturally grounded benchmark like EPiK, a model's inherent understanding of specific cultural and linguistic nuances can be more critical than scale or generic reasoning capabilities alone. These findings strongly indicate the need for further research into the developing and evaluating language models on data that is not just multilingual, but deeply multicultural.

6 Conclusion

In this paper, we introduced EPiK, a new benchmark designed to evaluate physical commonsense reasoning within the rich context of Korean culture. Our work addresses a critical gap in existing benchmarks, which predominantly focus on Western-centric scenarios. Through a meticulous two-stage generation and verification pipeline, we constructed a dataset of 181 culturally authentic and physically grounded questions that move beyond simple translation, ensuring that problems are naturally situated in Korean daily life and traditions.

Our extensive evaluations provide concrete evidence for the importance of cultural-specific training. We found that Korean-specialized models demonstrate a clear performance advantage, with the A.X-4.0-Light model achieving the highest accuracy of 92.27%. More critically, our analysis revealed that deep cultural specialization can be more impactful than sheer model scale or generic reasoning frameworks.

These findings underscore the limitations of culturally-agnostic AI systems and prove that true language understanding requires more than scaled-up general knowledge. By releasing EPiK, we provide a valuable and challenging resource for the community, encouraging the development of more culturally aware and globally competent reasoning models.

Acknowledgments

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

References

Anurag Acharya, Kartik Talamadupula, and Mark A. Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning. *arXiv preprint arXiv 2009.05664*.

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between Ghana and the US. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9483–9502. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6489–6499. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7674–7684. Association for Computational Linguistics.

Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Seokhee Hong, Junwon Hwang, Taewan Hwang, Joonwon Jang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Euisoon Kim, Hyosang Kim, Jihoon Kim, Joonkee Kim, and 21 others. 2025. EXAONE 4.0: Unified large language models integrating non-reasoning and reasoning modes. *arXiv preprint arXiv 2507.11407*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Compression and Complexity of SEQUENCES 1997, Positano, Amalfitan Coast, Salerno, Italy, June 11-13, 1997, Proceedings*, pages 21–29. IEEE.

Ernest Davis. 2024. Benchmarks for automated commonsense reasoning: A survey. *ACM Comput. Surv.*, 56(4):81:1–81:41.

Daniel Hershcovich, Stella Frank, Heather C. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6997–7013. Association for Computational Linguistics.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9785–9821. Association for Computational Linguistics.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 191 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv 2503.19786*.

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, 20-25 May, 2024, Torino, Italy*, pages 3335–3346. ELRA and ICCL.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Trans. Assoc. Comput. Linguistics*, 12:1703–1719.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu Ma, Haotian Geng, David Katz, Ion Stoica, and Matei Zaharia. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*. ACM.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 1274–1287. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Trans. Assoc. Comput. Linguistics*, 13:652–689.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv 2407.21783*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 65–83, Bangkok, Thailand. Association for Computational Linguistics.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 823–835. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won-Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Tae Hwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, and 12 others. 2021. KLUE: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. A multi-layered approach to physical commonsense understanding: Creation and evaluation of an Italian dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 819–831. ELRA and ICCL.

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2362–2376. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv 2408.00118*.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14182–14214. Association for Computational Linguistics.

Jaehyung Seo, Jaewook Lee, Chanjun Park, Seongtae Hong, Seungjun Lee, and Heuiseok Lim. 2024. Ko-CommonGEN v2: A benchmark for navigating Korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2390–2415. Association for Computational Linguistics.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5668–5680. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério de Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4996–5025. Association for Computational Linguistics.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha S. Srinivasa. 2023. NEWTON: are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP*

2023, Singapore, December 6-10, 2023, pages 9743–9758. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *arXiv preprint arXiv 2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv 2412.15115*.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1361–1371. Association for Computational Linguistics.

A Implementation Details

The entire pipeline for dataset generation and verification was implemented using the OpenAI API. We strategically selected different models for each phase to best suit the task requirements. For the initial synthesis stage (Stage 1), the generation of questions and their correct answers was performed by `gpt-4o-2024-11-20`. In the subsequent distractor generation phase (Stage 2), we used `gpt-4o-2024-08-06`. For these generative steps, the decoding parameters were consistently set to a *temperature*=1.0 and *top_p*=1.0 to encourage lexical and structural diversity in the outputs.

For all verification processes, which demanded a high degree of logical scrutiny, we leveraged `gpt-5-2025-08-07`, a model specifically optimized for reasoning tasks. To ensure a thorough analytical process for these validation steps, we configured *reasoning_effort*=medium. This setting activates the model's enhanced reasoning capabilities, making it ideal for the critical task of validating the correctness and logical consistency of the generated data.

B Prompt Details

This section provides detailed prompt information.

B.1 Prompts for Generating Questions and Answers (Synthesis Stage 1)

System Prompt for Synthesis Stage 1

당신은 AI 모델의 물리적 추론 능력을 평가하기 위한 고품질 한국어 데이터셋을 제작하는 '데이터 생성 전문가' 입니다.
You are a "data generation expert" creating high-quality Korean datasets to evaluate AI models' physical reasoning abilities.
당신의 핵심 임무는 한국의 문화적 배경과 물리적 상식이 결합된 문제를 만드는 것입니다.
Your core mission is to create problems that combine Korean cultural context with physical commonsense knowledge.
이 문제들은 둘 중 하나라도 정확히 이해하지 못하면 풀 수 없는 고차원적인 추론을 요구해야 합니다.
These problems require high-level reasoning that cannot be solved unless at least one of these aspects is precisely understood.

Input Prompt Template 1 for Synthesis Stage 1

다음 지침에 따라, 한국 문화적 배경과 물리적 상식이 결합된 추론 문제를 하나의 JSON 객체 형식으로 생성해 주십시오.
Following the instructions below, generate a reasoning problem in JSON format that combines Korean cultural context with physical commonsense knowledge.

배경 시나리오 (한국적 맥락이 드러나도록 변형하여 반영할 것)
Background Scenario (Adapted to reflect Korean cultural context)
{Scenario}

문제 유형
Problem Type
{Type}: {Type Definition}

데이터 생성 지침
Data Generation Instructions
- 'problem':
- 배경 시나리오를 바탕으로 물리적 세계에 대한 상식과 한국 문화적 맥락이 동시에 드러나는 질문을 한국어로 생성합니다.
- Based on the background scenario, generate a question in Korean that simultaneously reflects commonsense knowledge about the physical world and Korean cultural context.
- 한국적인 문화 요소란 보편적인 문화 요소를 제외한 다른 국가와는 차별적인 한국만의 특색있는 문화 요소를 의미합니다. 한국의 전통적인 요소가 될 수도 있으며, 현대적인 한국의 요소가 될 수 있습니다.
- Korean cultural elements refer to distinctive features unique to Korea, excluding common global cultural elements. They can be traditional or modern Korean elements.
- 한국적인 문화 요소가 물리적 추론에 반드시 핵심적으로 영향을 미치도록 문제를 구성하세요.
- Ensure that the Korean cultural elements play a crucial role in the physical reasoning required to solve the problem.
- 배경 시나리오와 아주 작은 연관성만 있으면 됩니다. 큰 주제나 맥락 혹은 분위기가 비슷하거나, 사소한 사물, 소재, 동작 등이 겹쳐도 연관성으로 인정됩니다.
- Only a small connection to the background scenario is necessary. Similarity in main topic, context, atmosphere, or overlapping minor objects, materials, or actions is sufficient.

- 물리적 추론 능력을 평가할 수 있는 문제를 생성해야 합니다.
- Generate problems that can evaluate physical reasoning abilities.
- 평균적인 한국인의 상식 수준으로 풀 수 있는 문제를 생성해야 합니다.
- Problems should be solvable with the commonsense knowledge of an average Korean.
- 배경 설명이나 불필요한 서술을 포함하지 말고, 질문 문장만 간단하게 생성하세요.
- Do not include background explanations or unnecessary descriptions; generate only a concise question sentence.
- 최대 {prompt_max_words} 단어 이하로 생성하세요.
- Generate the question using no more than {prompt_max_words} words.
- ‘answer‘:
- ‘problem‘에 대한 올바른 정답을 한국어로 생성하세요. 단, 물리적으로 타당하고 현실적인 행동이어야 합니다.
- Generate the correct answer to the ‘problem‘ in Korean. It must be physically plausible and realistic.
- ‘problem‘의 문장을 반복하지 말고, 핵심적인 답만 간결하게 작성하세요.
- Do not repeat the problem sentence; provide a concise, core answer.
- 최대 {solution_max_words} 단어 이하로 생성하세요.
- Provide the answer using no more than {solution_max_words} words.
- ‘korea_relevance‘: ‘problem‘과 ‘answer‘에 명시되어 있는 한국적 문화 및 관습과 밀접한 관련이 있는 소재를 추출하여 명사(구)로 나열합니다. (keyword/keyphrase extraction)
- ‘korea_relevance‘: Extract keywords/keyphrases from the ‘problem‘ and ‘answer‘ that are closely related to Korean culture and customs.
- ‘physical_relevance‘: ‘problem‘과 ‘answer‘에 명시되어 있는 물리적 요소, 속성, 원리와 밀접한 관련이 있는 소재를 추출하여 명사(구)로 나열합니다. (keyword/keyphrase extraction)
- ‘physical_relevance‘: Extract keywords/keyphrases from the ‘problem‘ and ‘answer‘ that are closely related to physical elements, properties, or principles.
- ‘rationale‘: 질문에 대한 정답의 근거를 작성하세요.
- ‘rationale‘: Provide the rationale for the answer.

CAUTION
- 데이터 생성 지침을 반드시 준수해야 합니다.
- Strictly follow the data generation instructions.
- 생성된 문제와 풀이과정은 한국어 문맥에서 자연스럽게 읽혀야 합니다.
- The generated problem and solution should read naturally in Korean context.
- 생성된 문제와 풀이과정은 1문장이거나 2문장으로 구성되어야 합니다.
- The generated problem and solution should consist of one or two sentences.
- 생성된 문제는 반드시 한국의 문화, 관습, 생활환경에 대한 이해와 물리적 속성에 대한 지식이 결합되어야만 해결할 수 있어야 합니다.
- The generated problem must require a combination of understanding Korean culture, customs, living environment, and knowledge of physical properties to solve.
- 한국인이라면 누구나 정답을 알 수 있어야 합니다.
- The answer must be knowable by any Korean.

Input Prompt Template 2 for Synthesis Stage 1

다음 지침에 따라, 한국 문화적 배경과 물리적 상식이 결합된 추론 문제를 하나의 JSON 객체 형식으로 생성해 주십시오.
Following the instructions below, generate a reasoning problem in JSON format that combines Korean cultural context with physical commonsense knowledge.

배경 시나리오 (한국적 맥락이 드러나도록 변형하여 반영할 것)
Background Scenario (Adapted to reflect Korean cultural context)
{Scenario}

문제 유형
Problem Type
{Type}: {Type Definition}

데이터 생성 지침
Data Generation Instructions

- `problem`:
- 배경 시나리오를 바탕으로 물리적 세계에 대한 상식과 한국 문화적 맥락이 동시에 드러나는 질문을 한국어로 생성합니다.
- Based on the background scenario, generate a question in Korean that simultaneously reflects commonsense knowledge about the physical world and Korean cultural context.
- 한국적인 문화 요소란 보편적인 문화 요소를 제외한 다른 국가와는 차별적인 한국만의 특색있는 문화 요소를 의미합니다. 한국의 전통적인 요소가 될 수도 있으며, 현대적인 한국의 요소가 될 수 있습니다.
- Korean cultural elements refer to distinctive features unique to Korea, excluding common global cultural elements. They can be traditional or modern Korean elements.
- 한국적인 문화 요소가 물리적 추론에 반드시 핵심적으로 영향을 미치도록 문제를 구성하세요.

- Ensure that the Korean cultural elements play a crucial role in the physical reasoning required to solve the problem.
- 배경 시나리오와 아주 작은 연관성만 있으면 됩니다. 큰 주제나 맥락 혹은 분위기가 비슷하거나, 사소한 사물, 소재, 동작 등이 겹쳐도 연관성으로 인정됩니다.
- Only a small connection to the background scenario is necessary. Similarity in main topic, context, atmosphere, or overlapping minor objects, materials, or actions is sufficient.
- 문장을 끝맺지 마십시오. (ex. " 면", " 하기 위해서는")
- Do not end the sentence completely. (e.g., " if", " in order to")
- 물리적 추론 능력을 평가할 수 있는 문제를 생성해야 합니다.
- Generate problems that can evaluate physical reasoning abilities.
- 평균적인 한국인의 상식 수준으로 풀 수 있는 문제를 생성해야 합니다.
- Problems should be solvable with the commonsense knowledge of an average Korean.
- 배경 설명이나 불필요한 서술을 포함하지 말고, 질문 문장만 간단하게 생성하세요.
- Do not include background explanations or unnecessary descriptions; generate only a concise question sentence.
- 최대 {prompt_max_words} 단어 이하로 생성하세요.
- Generate the question using no more than {prompt_max_words} words.
- `answer`:
- `problem`에 대한 올바른 정답을 한국어로 생성하세요. 단, 물리적으로 타당하고 현실적인 행동이어야 합니다.
- Generate the correct answer to the `problem` in Korean. It must be physically plausible and realistic.
- `problem`의 문장을 반복하지 말고, 핵심적인 답만 간결하게 작성하세요.
- Do not repeat the `problem` sentence; provide a concise, core answer.
- 최대 {solution_max_words} 단어 이하로 생성하세요.
- Provide the answer using no more than {solution_max_words} words.
- `korea_relevance`: `problem`과 `answer`에 명시되어 있는 한국적 문화 및 관습과 밀접한 관련이 있는 소재를 추출하여 명사(구)로 나열합니다. (keyword/keyphrase extraction)
- `korea_relevance`: Extract keywords/keyphrases from the `problem` and `answer` that are closely related to Korean culture and customs.
- `physical_relevance`: `problem`과 `answer`에 명시되어 있는 물리적 요소, 속성, 원리와 밀접한 관련이 있는 소재를 추출하여 명사(구)로 나열합니다. (keyword/keyphrase extraction)
- `physical_relevance`: Extract keywords/keyphrases from the `problem` and `answer` that are closely related to physical elements, properties, or principles.
- `rationale`: 질문에 대한 정답의 근거를 작성하세요.
- `rationale`: Provide the rationale for the answer.

CAUTION
- 데이터 생성 지침을 반드시 준수해야 합니다.
- Strictly follow the data generation instructions.
- 생성된 문제와 풀이과정은 한국어 문맥에서 자연스럽게 읽혀야 합니다.
- The generated problem and solution should read naturally in Korean context.
- 생성된 문제와 풀이과정은 1문장이거나 2문장으로 구성되어야 합니다.
- The generated problem and solution should consist of one or two sentences.
- 생성된 문제는 반드시 한국의 문화, 관습, 생활환경에 대한 이해와 물리적 속성에 대한 지식이 결합되어야만 해결할 수 있어야 합니다.
- The generated problem must require a combination of understanding Korean culture, customs, living environment, and knowledge of physical properties to solve.
- 한국인이라면 누구나 정답을 알 수 있어야 합니다.
- The answer must be knowable by any Korean.

Figure 6: Prompt template for generating a question and an answer. To diversify the linguistic structure of the generated data, the input prompts were constructed by randomly selecting between two predefined templates, Template 1 and Template 2.

B.2 Regeneration Prompts for Incorrect Answers (Synthesis Stage 1-1)

System Prompt for Synthesis Step 1-1

당신은 주어진 문제와 오답, 그리고 오답에 대한 설명을 바탕으로 정답을 수정하는 '데이터 수정 전문가'입니다.
You are a "data correction expert" who revises the correct answer based on a given problem, an incorrect answer, and the explanation of the incorrect answer.
당신의 임무는 제공된 피드백을 정확히 반영하여 올바른 답변을 생성하는 것입니다.
Your task is to generate the correct answer by accurately reflecting the provided feedback.

Input Prompt Template for Synthesis Step 1-1

다음은 이전에 생성된 문제와 답변, 그리고 답변이 틀린 이유에 대한 설명입니다. 설명을 바탕으로 답변을 수정해 주십시오.
Below is a previously generated problem, its answer, and an explanation of why the answer is incorrect. Revise the answer based on this explanation.

문제
Problem
{problem}

기존 답변 (오답)
Original Answer (Incorrect)
{answer}

틀린 이유
Reason for Incorrectness
{correctness_rationale}

수정 지침
Correction Instructions
- `answer`:
- 틀린 이유를 참고하여 기존 답변을 물리적으로 타당하고, 한국 문화적 맥락에 맞는 정확한 답변으로 수정하세요.
- Revise the original answer to be physically plausible and consistent with the Korean cultural context, referring to the `reason for incorrectness`.
- `problem`의 문장을 반복하지 말고, 핵심적인 답만 간결하게 작성하세요.
- Do not repeat the `problem` sentence; provide a concise, core answer.
- `korea_relevance`: 수정된 `answer`와 기존 `problem`에 명시되어 있는 한국적 문화 및 관습과 밀접한 관련이 있는 소재를 추출하여 명사(구)로 나열합니다.
- `korea_relevance`: Extract keywords/keyphrases from the revised `answer` and the original `problem` that are closely related to Korean culture and customs.
- `physical_relevance`: 수정된 `answer`와 기존 `problem`에 명시되어 있는 물리적 요소, 속성, 원리와 밀접한 관련이 있는 소재를 추출하여 명사(구)로 나열합니다.
- `physical_relevance`: Extract keywords/keyphrases from the revised `answer` and the original `problem` that are closely related to physical elements, properties, or principles.
- `rationale`: 수정된 답변에 대한 정답의 근거를 새로 작성하세요.
- `rationale`: Provide the rationale for the revised answer.

CAUTION
- 데이터 생성 지침을 반드시 준수해야 합니다.
- Strictly follow the data generation instructions.
- 생성된 결과물은 하나의 JSON 객체 형식이어야 합니다.
- The generated output must be in a single JSON object format.

Figure 7: Prompt template for regenerating an answer.

B.3 Prompts for Generating Incorrect Answers (Synthesis Stage 2)

System Prompt for Synthesis Stage 2

당신은 AI 모델의 물리적 추론 능력을 평가하기 위한 고품질 한국어 데이터셋을 제작하는 '데이터 생성 전문가' 입니다.
You are a "data generation expert" creating high-quality Korean datasets to evaluate AI models' physical reasoning abilities.
당신의 핵심 임무는 제시된 질문과 정답을 바탕으로, 겉보기에 타당해 보이지만 실제로는 잘못된 그럴듯한 오답을 생성하는 것입니다.
Your core mission is to generate plausible incorrect answers that seem reasonable at first glance but are actually incorrect, based on the provided question and correct answer.

Input Prompt Template for Synthesis Stage 2

질문
Question
{problem}

정답
Correct Answer
{answer}

데이터 생성 지침
Data Generation Instructions
- `incorrect_answer`:
- 정답에서 특정 단어나 어구만 미묘하게 바꿔서 생성하세요.
- Generate by subtly modifying specific words or phrases from the correct answer.

- 나머지 문장 구조, 형식, 길이는 거의 동일하게 유지해 혼동을 유도하세요.
- Keep the remaining sentence structure, format, and length almost the same to induce confusion.
- 오답은 사람들이 흔히 가질 법하거나 과학적으로 들리는 그럴듯한 오개념을 기반으로 작성해야 합니다.
- The incorrect answer should be based on common misconceptions or scientifically plausible-sounding errors.
- 단순 부정만으로 만들지 마세요.
- Do not generate the incorrect answer by simply negating the correct answer.
- 문법적으로 완전하고 자연스럽게 읽혀야 합니다.
- Ensure the sentence is grammatically complete and reads naturally.
- 오답은 명백히 터무니없는 주장이어서는 안 됩니다. 그럴듯하지만 틀린 설명이어야 합니다.
- The incorrect answer should not be obviously absurd; it must be plausible but incorrect.
- 문제는 반드시 질문을 읽어야 풀 수 있으며, 질문이 제시되지 않으면 정답과 오답만으로는 구분이 불가능해야 합니다.
- The problem must be read to solve it; without the question, the correct and incorrect answers alone should not be distinguishable.
- 정답과 쉽게 구분되지 않도록 너무 단순하거나 쉽게 판별되는 오답은 피해야 합니다.
- Avoid overly simple or easily identifiable incorrect answers that can be easily distinguished from the correct one.
- 정답과 오답의 우열이 명백한 물리적 원리에 기반해야 합니다.
- The distinction between correct and incorrect answers must be based on clear physical principles.
- `comparison`: 정답이 왜 물리적으로 옳고, 오답은 왜 틀렸는지 명확하게 비교 설명합니다.
- `comparison`: Clearly explain why the correct answer is physically valid and why the incorrect answer is wrong.

Figure 8: Prompt template for generating an incorrect answer. (Synthesis Stage 2)

B.4 Prompts for Verifying Questions and Answers (Verification Stage 1)

System Prompt for Verification Stage 1

당신은 AI 모델의 물리적 추론 능력을 평가하기 위한 고품질 한국어 데이터셋을 평가하는 '데이터 평가 전문가'입니다.
You are a "data evaluation expert" tasked with evaluating high-quality Korean datasets designed to assess AI models' physical reasoning abilities.
당신의 핵심 임무는 주어지는 평가 기준과 요청에 따라 데이터를 평가하는 것입니다.
Your core mission is to evaluate the data according to the given evaluation criteria and instructions.

Input Prompt Template for verifying the correctness of an answer in Verification Stage 1

주어진 평가 기준에 따라 다음 데이터셋을 평가하십시오.
Evaluate the following dataset according to the given evaluation criteria.

평가 기준
Evaluation Criteria
- 주어진 문제에 대한 정답이 실제로 올바른 정답인지 평가하십시오. 틀린 답이라면 false, 실제로 올바른 정답이라면 true의 값으로 평가하십시오.
- Assess whether the given answer to the problem is actually correct. Use `false` if it is incorrect and `true` if it is correct.
- 평가에 대한 이유를 함께 서술하십시오. (`correctness_rationale`)
- Provide a rationale for your evaluation. (`correctness_rationale`)

평가 요청
Evaluation Request
- JSON 객체 형식으로 평가 결과를 생성하십시오.
- Generate the evaluation result in a JSON object format.
- 평가 기준에 따라 평가한 값을 'correctness: {{value}}'로 생성하십시오.
- Provide the evaluated value as 'correctness: {{value}}' according to the evaluation criteria.

데이터셋
Dataset

문제
Problem
{problem}

정답
Answer
{answer}

Input Prompt Template for verifying a problem in Verification Step 1

주어진 평가 기준에 따라 다음 데이터셋을 평가하십시오.
Evaluate the following dataset according to the given evaluation criteria.

평가 기준
Evaluation Criteria

명확성 평가 기준
Clarity Evaluation Criteria
- 주어진 문제와 정답이 명확한지 평가하십시오. (`is_problem_clear`)
- Evaluate whether the given problem and answer are clear. (`is_problem_clear`)
- 문제와 정답이 모호하거나, 해석에 혼동을 주어 의도를 파악하기 어렵다면 false로 평가하십시오.
- If the problem and answer are ambiguous or confusing to interpret, evaluate as false.
- 문제와 정답이 명확하고 혼동의 여지가 없다면 true의 값으로 평가하십시오.
- If the problem and answer are clear and unambiguous, evaluate as true.

물리적 추론 평가 기준
Physical Reasoning Evaluation Criteria
- 주어진 물리 요소가 문제와 정답에 포함되는지 여부를 평가하십시오.
- Evaluate whether the given physical elements are included in the problem and answer.
- 만약 주어진 물리 요소가 문제와 정답에 포함된다면 true, 포함되지 않는다면 false의 값으로 평가하십시오. (`physical_relevance_included`)
- If the physical elements are included, evaluate as true; if not, evaluate as false. (`physical_relevance_included`)
- 평가에 대한 이유를 함께 서술하십시오. (`physical_relevance_included_rationale`)
- Provide the rationale for this evaluation. (`physical_relevance_included_rationale`)

상식 평가 기준
Commonsense Evaluation Criteria
- 주어진 문제와 정답이 평균적인 한국인의 상식 수준으로 풀 수 있는지 true 또는 false의 값으로 평가하십시오. (`commonsense_feasible`)
- Evaluate whether the problem and answer can be solved using the commonsense level of an average Korean. (`commonsense_feasible`)
- 평가에 대한 이유를 함께 서술하십시오. (`commonsense_feasible_rationale`)
- Provide the rationale for this evaluation. (`commonsense_feasible_rationale`)

문화 집중 평가 기준
Cultural Focus Evaluation Criteria
- 한국적인 문화 요소란 보편적인 문화 요소를 제외한 다른 국가와는 차별적인 한국만의 특색있는 문화 요소를 의미합니다. 한국의 전통적인 요소가 될 수도 있으며, 현대적인 한국의 요소가 될 수 있습니다.
- Korean cultural elements refer to uniquely Korean elements distinct from universal or common global cultural elements. They may be traditional or modern Korean elements.
- 주어지는 문화 요소 중에서 위 한국적인 문화 요소 기준에 따라 한국적인 문화 요소에 포함되지 않는 요소를 제외하고 한국적인 문화 요소를 구성하십시오. (`korea_relevance_list`)
- From the given cultural elements, exclude those that do not meet the uniquely Korean criteria, and construct a list of Korean cultural elements. (`korea_relevance_list`)
- 문화 요소의 제외 또는 포함에 대한 이유를 함께 서술하십시오. (`korea_relevance_rationale`)
- Provide the rationale for inclusion or exclusion. (`korea_relevance_rationale`)

- 구성한 한국적인 문화 요소가 문제와 정답에 포함되지 않는다면 0, 한국적 문화 요소가 문제와 정답에 단순히 언급되며 다른 요소로 대체해도 문제가 성립한다면 1, 한국적 문화 요소가 문제와 정답에 핵심적으로 포함되며 한국적 문화 요소에 대한 이해가 문제 해결에 필수적이라면 2의 값을 평가하십시오. (`cultural_element_score`)
- If the constructed Korean cultural elements are not included in the problem and answer, evaluate as 0. If the Korean cultural elements are merely mentioned in the problem and answer but can be replaced with other elements without affecting the problem, evaluate as 1. If the Korean cultural elements are essential in the problem and answer, and understanding them is crucial for solving the problem, evaluate as 2. (`cultural_element_score`)
- 평가에 대한 이유를 함께 서술하십시오. (`cultural_element_score_rationale`)
- Provide the rationale for this evaluation. (`cultural_element_score_rationale`)

유창성 평가 기준
Fluency Evaluation Criteria
- 주어진 문제와 정답을 연결하였을 때, 한국어 문장이 자연스러운지 평가하십시오. (`fluency`)
- Evaluate whether the combined problem and answer form a natural Korean sentence. (`fluency`)
- 한국어 원어민이 썼다고 느껴질 만큼 자연스럽고 유창하면 true의 값으로 평가하십시오.
- If it sounds fluent and natural as if written by a native Korean, evaluate as true.
- 번역투의 부자연스러운 표현이 포함되어 있다면 false의 값으로 평가하십시오.
- If awkward, translation-like expressions are present, evaluate as false.

평가 요청
Evaluation Request
- JSON 객체 형식으로 평가 결과를 생성하십시오.
- Generate the evaluation result in a JSON object format.
- 명확성 평가 기준에 따라 평가한 값을 'is_problem_clear: {{value}}'로 생성하십시오.
- According to the clarity evaluation criterion, generate the value as 'is_problem_clear: {{value}}'.
- 물리적 추론 평가 기준에 따라 평가한 값을 'physical_relevance_included: {{value}}'로 생성하십시오.
- According to the physical reasoning evaluation criterion, generate the value as 'physical_relevance_included: {{value}}'.
- 물리적 추론 평가 기준에 따라 평가한 이유를 'physical_relevance_included_rationale: {{value}}'로 생성하십시오.
- Provide the rationale for the evaluation according to the physical reasoning criterion as 'physical_relevance_included_rationale: {{value}}'.
- 상식 평가 기준에 따라 평가한 값을 'commonsense_feasible: {{value}}'로 생성하십시오.
- According to the commonsense evaluation criterion, generate the value as 'commonsense_feasible: {{value}}'.
- 상식 평가 기준에 따라 평가한 이유를 'commonsense_feasible_rationale: {{value}}'로 생성하십시오.
- Provide the rationale for the evaluation according to the commonsense criterion as 'commonsense_feasible_rationale: {{value}}'.
- 문화 집중 평가 기준에 따라 주어진 문제와 정답을 평가하여 'cultural_element_score: {{value}}'로 생성하십시오.
- Evaluate the given problem and answer according to the cultural focus criterion and generate 'cultural_element_score: {{value}}'.
- 문화 집중 평가 기준에 따라 평가한 이유를 'cultural_element_score_rationale: {{value}}'로 생성하십시오.
- Provide the rationale for the evaluation according to the cultural focus criterion as 'cultural_element_score_rationale: {{value}}'.
- 구성한 한국적인 전통요소를 'korea_relevance_list: {{list}}'로 생성하십시오.
- Generate the composed Korean traditional elements as 'korea_relevance_list: {{list}}'.
- 문화 요소의 제외 또는 포함에 대한 이유를 'korea_relevance_rationale: {{value}}'로 생성하십시오.
- Provide the rationale for inclusion or exclusion of cultural elements as 'korea_relevance_rationale: {{value}}'.
- 유창성 평가 기준에 따라 평가한 값을 'fluency: {{value}}'로 생성하십시오.
- According to the fluency evaluation criterion, generate the value as 'fluency: {{value}}'.

데이터셋
Dataset

문제
Problem
{problem}

정답
Answer
{answer}

문화 요소
Cultural Elements
{korea_relevance}

물리 요소
Physical Elements
{physical_relevance}

Figure 9: Prompt template for verifying a question and an answer. (Verification Stage 1)

B.5 Prompts for Verifying Incorrect Answers (Verification Stage 2)

System Prompt for Verification Stage 2

당신은 AI 모델의 물리적 추론 능력을 평가하기 위한 고품질 한국어 데이터셋을 평가하는 '데이터 평가 전문가'입니다.
You are a "data evaluation expert" tasked with evaluating high-quality Korean datasets designed to assess AI models' physical reasoning abilities.
당신의 핵심 임무는 주어지는 평가 기준과 요청에 따라 데이터를 평가하는 것입니다.
Your core mission is to evaluate the data according to the given evaluation criteria and instructions.

Input Prompt Template for Verification Stage 2

주어진 평가 기준에 따라 다음 데이터셋을 평가하십시오.
Evaluate the following dataset according to the given evaluation criteria.

```
# 평가 기준
# Evaluation Criteria

## 오답 평가 기준
## Incorrect Answer Evaluation Criteria
- 주어진 오답이 주어진 문제에 대해 실제로 틀린 답인지 평가하십시오. (is_true_distractor)
- Evaluate whether the given incorrect answer is truly wrong for the given problem. (is_true_distractor)
- 만약 주어진 오답이 실제로 문제에 대한 정답이 아니라면 (즉 오답이라면) 오답에 대한 기능을 제대로 수행하고 있으므로 true의 값으로 평가하십시오.
- If the given incorrect answer is not actually the correct answer to the problem (i.e., it is indeed incorrect), then it serves its function properly and should be evaluated as true.
- 만약 주어진 오답이 실제로 문제에 대한 정답으로 볼 수 있거나, 다소 모호한 정답이라면 오답에 대한 기능을 제대로 수행하지 못하고 있으므로 false의 값으로 평가하십시오.
- If the given incorrect answer can actually be seen as a correct or somewhat ambiguous answer, then it does not serve its function as an incorrect answer and should be evaluated as false.

## 오답 완전성 평가 기준
## Completeness Evaluation Criteria for Incorrect Answers
- 오답 선택지 자체에 모순이나 불완전한 부분이 없는지 평가하십시오. (is_distractor_complete)
- Evaluate whether the incorrect answer option itself has no contradictions or incompleteness. (is_distractor_complete)
- 만약 오답 선택지 자체에 모순이나 불완전한 부분이 없다면 true의 값으로 평가하십시오.
- If the incorrect answer option has no contradictions or incompleteness, evaluate it as true.
- 만약 오답 선택지 자체에 모순이나 불완전한 부분이 있다면 false의 값으로 평가하십시오.
- If the incorrect answer option has contradictions or incompleteness, evaluate it as false.

## 오답 유형 평가 기준
## Type Evaluation Criteria for Incorrect Answers
- 정답과 오답의 우열이 명백한 물리적 원리에 기반하는지 평가하십시오. (is_phyiscally_distinct)
- Evaluate whether the superiority of the correct answer over the incorrect answer is clearly based on physical principles. (is_phyiscally_distinct)
- 만약 정답과 오답의 우열이 명백한 물리적 원리에 기반한다면 true의 값으로 평가하십시오.
- If the superiority of the correct answer over the incorrect answer is clearly based on physical principles, evaluate it as true.
- 만약 정답과 오답의 우열이 명백한 물리적 원리에 기반하지 않고 주관적인 선호, 비물리적인 요인등에 의한 것이라면 false의 값으로 평가하십시오.
- If the superiority is not based on physical principles but rather on subjective preferences or non-physical factors, evaluate it as false.

# 평가 요청
# Evaluation Request
- JSON 객체 형식으로 평가 결과를 생성하십시오.
- Generate the evaluation result in JSON object format.
- 오답 평가 기준에 따라 평가한 값을 'is_true_distractor: {{value}}'로 생성하십시오.
- According to the incorrect answer evaluation criterion, generate the value as 'is_true_distractor: {{value}}'.
- 오답 완전성 평가 기준에 따라 평가한 값을 'is_distractor_complete: {{value}}'로 생성하십시오.
- According to the completeness evaluation criterion for incorrect answers, generate the value as 'is_distractor_complete: {{value}}'.
- 오답 유형 평가 기준에 따라 평가한 값을 'is_phyiscally_distinct: {{value}}'로 생성하십시오.
- According to the type evaluation criterion for incorrect answers, generate the value as 'is_phyiscally_distinct: {{value}}'.

# 데이터셋
# Dataset

## 문제
## Problem
{problem}

## 정답
## Answer
{answer}

## 오답
## Incorrect Answer
{incorrect_answer}
```

Figure 10: Prompt template for verifying an incorrect answer. (Verification Stage 2)

B.6 Prompts for Filtering Answer Bias

System Prompt for Filtering Answer Bias

당신은 문제를 보지 않고, 주어진 선지만으로 정답을 찾을 수 있는지 판단해야 합니다.
You need to determine whether you can find the correct answer based solely on the given choices without seeing the question.
주의사항:
Instructions:
- 질문은 제공되지 않았습니다. 단지 두 선지만 보고 정답을 고르세요.
- The question is not provided. Choose the correct answer by looking only at the two options.
- 선지의 길이나 특정 단어 패턴 등 표면적 특징만 보고 판단하지 마세요.
- Do not make your judgment based only on superficial features such as the length of the choices or specific word patterns.
- 정답 번호만 출력하세요.
- Output only the number of the correct answer.
- 두 선지만으로 정답을 판단할 수 없다면 3을 출력하세요.
- If you cannot determine the correct answer from the two choices alone, output 3.
- 이유나 설명은 출력하지 마세요.
- Do not output any reasons or explanations.

Input Prompt Template for Filtering Answer Bias

다음 중 정답을 고르세요.
Select the correct answer from the following options.
1. solution0
2. solution1
3. 두 선지만으로 정답을 판단할 수 없다.
3. You cannot determine the correct answer from the two choices alone.
정답 번호만 출력하세요. 1 또는 2 또는 3 중 하나만 선택합니다.
Output only the number of the correct answer. Choose only one among 1, 2, or 3.

Figure 11: Prompt template for filtering answer bias.

C Human Evaluation Guide

To ensure a standardized and rigorous final review, we provided the human expert with a comprehensive evaluation guide. This guide outlines 5 key assessment criteria: accuracy, naturalness and logical consistency, cultural appropriateness, common sense, and plausibility. The complete set of instructions is available in Figure 12, which details the standards for accepting or rejecting a data instance.

Human Evaluation Guide

1. 평가 개요
1. Evaluation Overview
이 문서는 `dataset.jsonl` 파일에 포함된 한국어 질의응답 데이터의 품질을 평가하기 위한 가이드입니다. 각 항목은 다음 요소들을 포함하고 있습니다:
This document is a guide for evaluating the quality of Korean Q&A data included in the `dataset.jsonl` file. Each item includes the following:
- 문제 유형 카테고리
- Question Type Category
- 문제 유형 서브카테고리
- Question Type Subcategory
- 프롬프트 (질문)
- Prompt (Question)
- 두 개의 답변 (solution0, solution1)
- Two answers (solution0, solution1)
- 정답 레이블
- Correct answer label
- 메타데이터 (문화적 관련성, 물리적 관련성 점수 등)
- Metadata (cultural relevance, physical relevance scores, etc.)

2. 평가 항목
2. Evaluation Items
2.1 정확성 (Accuracy)
2.1 Accuracy
- 제시된 질문에 따라 두 답변 중 정답을 구하세요.
- Find the correct answer out of the two solutions provided based on the question.

2.2 자연스러움과 논리성 (Naturalness & Logical Consistency)
2.2 Naturalness & Logical Consistency
- 질문과 답변이 자연스럽고 올바른 한국어로 작성되었는지 확인하세요.
- Check if the question and answer are written in natural and correct Korean.
- 어색한 표현이나 문법 오류가 없는지 점검하세요.
- Check for awkward expressions or grammatical errors.
- 일상생활에서 실제로 사용할 법한 표현인지 고려하세요.
- Consider whether the expressions are commonly used in daily life.
- 질문과 답변 사이에 논리적 일관성이 유지되는지 확인하세요.
- Check if logical consistency is maintained between the question and answer.

2.3 문화적 적절성 (Cultural Appropriateness)
2.3 Cultural Appropriateness
- 한국 문화와 관련된 내용이 정확하게 반영되었는지 평가하세요.
- Evaluate whether content related to Korean culture is accurately reflected.
- 전통 문화에 대한 잘못된 정보가 없는지 확인하세요.
- Check for any misinformation about traditional culture.
- 문화적 민감성을 고려하여 부적절한 표현이 없는지 검토하세요.
- Review for inappropriate expressions, considering cultural sensitivity.

2.4 상식 (Common Sense)
2.4 Common Sense
- 문제가 한국인의 상식 수준에서 풀 수 있는지 확인하세요.
- Check if the problem can be solved at the level of a Korean person's common sense.

2.5 현실성 (Practicality / Plausibility)
2.5 Practicality / Plausibility
- 문제와 답변이 실제 상황에서 합리적이고 실행 가능해야 합니다.
- The question and answer must be reasonable and plausible in a real-world situation.
- 물리적, 사회적, 상식적으로 불가능하거나 극도로 비현실적인 선택지는 배제해야 합니다.
- Exclude options that are physically, socially, or common-sensically impossible or extremely unrealistic.

3. 평가 방법
3. Evaluation Method
1. 각 항목을 주의 깊게 읽고 이해하세요.
1. Read and understand each item carefully.
2. 제공된 정답 레이블이 올바른지 확인하세요.
2. Check if the provided correct answer label is correct.
3. 각 평가 항목(정확성, 자연스러움과 논리성, 문화적 적절성, 상식, 현실성)에 대해 0,1점으로 평가하세요.
3. Evaluate each item (Accuracy, Naturalness & Logical Consistency, Cultural Appropriateness, Common Sense, Practicality) with a score of 0 or 1.

4. 주의사항
4. Precautions
- 모든 평가는 객관적으로 진행해 주세요.
- Please conduct all evaluations objectively.

Figure 12: Guide for human evaluation.