Culturally-Grounded Physical Commonsense Reasoning: A New Bilingual Dataset for Hindi and Telugu

Aditi Gupta , Chaithra Reddy Nerella
International Institute of Information Technology, Hyderabad
Hyderabad, India
{aditi.gu, chaithra.nerella}@research.iiit.ac.in

Abstract

For artificial intelligence to be truly useful globally, it must understand the world as different cultures do. Current AI models, trained predominantly on English, Western-centric data, often fail when faced with reasoning tasks grounded in other cultural contexts. To address this gap, we created a new bilingual benchmark for evaluating physical commonsense reasoning in Hindi and Telugu for the MRL 2025 Shared Task. Our dataset contains 280 original examples that test an AI's understanding of situations, objects, and traditions specific to India. Each question is presented in the PIQA (Bisk et al., 2020) format, offering two nearly identical solutions, one of which is physically correct and the other subtly wrong. Our methodology includes expert manual authoring with LLM-assisted generation, followed by a manual human validation by native speakers to ensure cultural authenticity and physical accuracy for every example in both languages. This work contributes a challenging new resource for developing and testing more culturally-aware AI models.

1 Introduction

To make an AI system understand the *unwritten rules of the physical world*, a skill that humans acquire naturally through everyday experience. This includes basic physical reasoning, such as knowing that a *glass pot will shatter if dropped*, while a *steel pot might only dent*, culturally specific sentences and other commonsense related tasks.

This is known as physical commonsense reasoning, a critical area of modern AI research. However, most existing tools for testing this ability are designed exclusively in English and implicitly assume a Western cultural context. This creates a significant blind spot because the way commonsense knowledge is learned, represented, and applied is deeply influenced by local culture, objects, and traditions.

This paper addresses the critical lack of culturally relevant benchmarks for Hindi and Telugu, two major world languages spoken by hundreds of millions of people. Simply translating existing English datasets is insufficient. For instance, a translated question about the physical behavior of a *baseball bat* does not test a model's understanding of how a *gilli-danda* (a traditional Indian street-game stick) behaves. While their physical properties may be similar, the cultural context, crucial for true understanding, is lost. As a result, we currently have limited insights into how well modern AI models can reason about the world outside of an English-speaking, Western context.

To address this gap, we introduce *Physical Reasoning and Multilingual Authenticated Annotations*, a new PIQA-style benchmark (Bisk et al., 2020) developed as part of the MRL 2025 Shared Task. Our contributions are threefold:

1. We provide a new benchmark of 280 examples for physical commonsense reasoning in Hindi and Telugu.

2. We focus specifically on culturally grounded scenarios, including *local foods, festivals, and tools*, making it a unique test of both physical and cultural commonsense.

3. We employ a challenging minimal-pair design, where correct and incorrect answers are nearly identical, forcing AI models to reason about subtle physical differences rather than rely on superficial cues.

2 Related Work

Our work is situated at the intersection of commonsense reasoning and multilingual natural language processing, addressing a specific gap in culturally-aware physical reasoning.

2.1 Foundational Commonsense Reasoning Benchmarks

The challenge of endowing machines with commonsense has a long history, with early efforts like the Winograd Schema Challenge (Levesque et al., 2012) require reasoning beyond simple statistical patterns. Over time, the field has specialized and the Datasets like Social IQa (Sap et al., 2019) were developed to test reasoning about social situations, while others like CommonsenseQA (Talmor et al., 2019) provided large-scale multiple-choice questions.

Our work is most directly inspired by benchmarks focused on physical reasoning. The primary model for our format is PIQA (Bisk et al., 2020), which introduced the effective two-solution, minimal-pair structure for testing knowledge of physical interactions. Similarly, HellaSwag (Zellers et al., 2019) demonstrated the power of *adversarial filtering*, using a model to help generate incorrect answers that are cleverly designed to trick other models but are obvious to humans. Our dataset adopts the robust PIQA format but fundamentally shifts the content from generic, globally-assumed scenarios to culturally-specific ones.

2.2 Multilingual and Cross-Lingual Evaluation

The challenge of commonsense reasoning has been extended to the multilingual domain, primarily through two approaches. The first involves creating broad, multi-task benchmarks like XGLUE (Liang et al., 2020) that includes tasks like question answering and natural language inference. The second approach focuses on specific reasoning types, such as the cross-lingual causal reasoning dataset XCOPA (Ponti et al., 2020).

While these resources are invaluable, they often use a *translate-test* methodology, where an English original is translated into other languages. This process can miss the rich, unique commonsense embedded in different cultures. For example, a translated question about the properties of “making a peanut butter sandwich” fails to test a model’s knowledge about the physical properties of making a local Indian dish like *dosa*. This Dataset avoids this issue by creating original, non-translated content for each language.

2.3 Culturally-Specific Datasets for Indian Languages

Within the Indian subcontinent, there has been a significant effort to develop benchmarks capturing the region’s vast linguistic diversity. Foundational efforts in Natural Language Understanding (NLU), such as IndicGLUE (Kakwani et al., 2020), have been vital for tasks like sentiment analysis and named entity recognition. This has been complemented by work in Natural Language Generation (NLG) with the IndicNLG Benchmark (Kumar et al., 2022), which provides datasets for tasks like summarization and question generation.

More recently, datasets like IndicSQuAD (Arora et al., 2025) have focused on culturally-aware question-answering. While these datasets are crucial for building models that can process and generate text in an Indian context, a dedicated resource for evaluating physical commonsense reasoning, deeply tied to local objects, materials, and traditional practices, has been a missing piece. We target this challenging reasoning domain in a structured, PIQA-style format for Hindi and Telugu.

3 Dataset Creation

The dataset was created following the guidelines of the MRL 2025 Shared Task. Our goal was to build a benchmark that is both physically grounded and culturally authentic for two Indian languages: Hindi and Telugu. We used a hybrid approach that combined the creative ability of a large language model with the judgment and intuition of human annotators.The process combined large language model (LLM) assistance with manual verification by native speakers. This balance helped us maintain both linguistic fluency and cultural authenticity.

3.1 Language Selection

We chose Hindi and Telugu as they are two of the most widely spoken Indian languages, yet both remain underrepresented in reasoning benchmarks. The Hindi portion of the dataset draws on Standard Hindi, which is widely understood across Northern India. Many prompts were inspired by cultural practices such as food preparation, household activities, and regional crafts. The Telugu portion is based on Standard Telugu, spoken in Telangana and Andhra Pradesh, and reflects daily life in those regions, from traditional agricultural practices to the handling of clay utensils. By covering these two languages, we aimed to capture both scale and diversity: Hindi provides broad coverage, while Telugu allows for representation of a Dravidian language with distinct cultural references.

3.2 Item Generation Strategy

The dataset was built using a three-stage strategy. First, native speakers wrote a small set of seed examples. These were carefully designed to highlight scenarios that rely on physical reasoning, such as preparing food in traditional utensils or handling clay pots. Second, we used a large language model, Gemini, to expand the dataset. The seed items served as few-shot prompts that guided the model toward generating further examples in the same style. However, not all outputs were directly usable. Each generated item was reviewed by native speakers, who edited or discarded those that were ambiguous, too generic, or physically inaccurate. This iterative process of LLM assistance followed by human refinement allowed us to build the dataset without sacrificing quality.

3.3 Prompt Generation

Prompts were intentionally varied in style to avoid monotony and to encourage deeper reasoning. Some were phrased as direct questions, others as incomplete sentences, and some as short scenarios. The topics covered a broad range of domains, including household chores, cooking, agricultural practices, and cultural rituals. For instance, a Hindi prompt might describe the preparation of matka kulfi, while a Telugu prompt could involve handling earthen pots during a festival. In all cases, the prompts were chosen to make physical reasoning central to solving the task, whether it was about fragility, weight, balance, or material behavior.

3.4 Solution Generation

For each prompt, we created two possible solutions. Following the PIQA-style format, the two options were written to be as close as possible in wording, often differing by only one or two words. This ensures that success on the task depends on reasoning about meaning, not surface-level differences. The correct solution was always physically and culturally valid, while the incorrect solution was written to be grammatically sound and seemingly plausible but ultimately wrong. For example, in a Hindi item about the deg-bhapka system of extracting perfume, one option involved steam passing through a bamboo pipe (correct), while the alternative described ash passing through the pipe (incorrect). Similarly, a Telugu prompt about a clay pot falling to the ground offered “breaking into pieces” as the correct answer and “bouncing back” as the distractor. These minimal yet meaningful contrasts made the dataset both realistic and challenging.

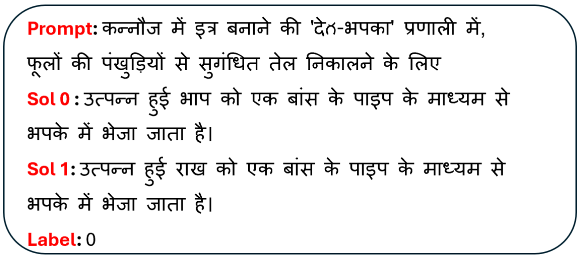
4 Dataset Analysis

The dataset is organized in a straightforward tab-separated format with five columns. The goal column contains the main prompt, which may appear as a short scenario, a guiding question, or an incomplete sentence in either Hindi or Telugu. The two candidate answers are placed in the solution0 and solution1 columns. These are written to be nearly identical in structure, differing only in one or two words, but with only one representing the physically correct choice. The label column marks the correct answer, using 0 or 1 to indicate whether solution0 or solution1 is valid. Finally, the language column specifies whether the example is in Hindi or Telugu. This simple yet precise structure makes the dataset easy to read, annotate, and apply in experiments on multilingual physical commonsense reasoning.

4.1 Quantitative Statistics.

Our dataset contains a total of 280 examples, with a split of 180 in Hindi and 100 in Telugu. Each entry has a prompt, two possible solutions, and a label identifying the correct choice.

For Telugu, the prompts are on average about ten words long, while the candidate solutions are slightly shorter, averaging nine words each. This balance ensures that the items are rich enough to capture reasoning without being unnecessarily long. The Hindi portion shows similar lengths, keeping the two languages consistent.



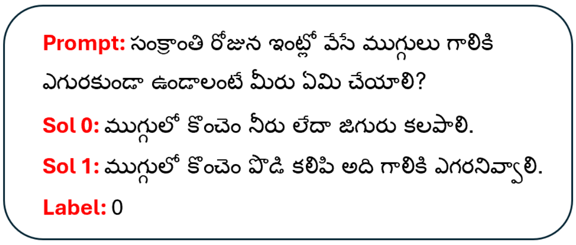
Prompt: कन्नौज में इत्र बनाने की 'देग-भपका' प्रणाली में, फूलों की पंखुड़ियों से सुगंधित तेल निकालने के लिए

Sol 0 : उत्पन्न हुई भाप को एक बांस के पाइप के माध्यम से भपके में भेजा जाता है।

Sol 1: उत्पन्न हुई राख को एक बांस के पाइप के माध्यम से भपके में भेजा जाता है।

Label: 0

Figure 1: Sample example of Hindi language



Prompt: సంక్రాంతి రోజున ఇంట్లో వేసే ముగ్గులు గాలికి ఎగురకుండా ఉండాలంటే మీరు ఏమి చేయాలి?

Sol 0: ముగ్గులో కొంచెం నీరు లేదా జిగురు కలపాలి.

Sol 1: ముగ్గులో కొంచెం పొడి కలిపి అది గాలికి ఎగరనివ్వాలి.

Label: 0

Figure 2: Sample example of Telugu language

4.2 Cultural Diversity

A key strength of the dataset lies in its cultural grounding. Many prompts draw from everyday life in India, as well as traditional practices and regional knowledge. These are situations that cannot simply be reproduced by translating English datasets.

A sample Hindi example from the dataset is shown in Figure 1. This example is about the process of oil extraction from flowers and herbs in a traditional distillation setup, commonly used in cultural practices like making perfumes or essential oils. The correct answer is that "steam" is used in distillation process while the incorrect answer suggests "ash" is used in the process.

Similarly, the example in Figure 2, is about a traditional practice for the Sankranti festival. The prompt asks how to keep a rangoli (muggu) from blowing away in the wind. The correct answer is that a binder like "water" or "glue" is used to make it stick, while the incorrect answer suggests "dry powder" is used in the process.

5 Conclusion

In this work, we presented a new dataset for physical commonsense reasoning in Hindi and Telugu, two languages that are widely spoken but remain under-represented in NLP resources. Our dataset goes beyond direct translation by grounding prompts and solutions in culturally authentic contexts, ensuring that the reasoning task reflects real-world practices and traditions. The use of minimal-pair solutions further strengthens the evaluation, making it a meaningful challenge for current models. We hope this dataset will serve as a foundation for future work on building culturally inclusive and linguistically diverse reasoning systems.

References

A. Arora et al. 2025. Indicsquad: A comprehensive multilingual question answering dataset for indic languages. In *arXiv preprint arXiv:2505.03688*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Aniket Bhattacharyya, et al. 2020. Indicglue: A benchmark dataset for indic natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Anoop Kumar, Divyanshu Kakwani, Satish Golla, et al. 2022. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*.

Xiaodong Liang, Zhiguo Wang, Mo Shen, Zhixing Zhang, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Edoardo Maria Ponti, Goran Glavaš, Ivan Vulić, Roi Reichart, 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*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: A question answering benchmark for artificial social intelligence. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.