Indic-PIQA: A Multilingual Dataset for Physical Commonsense Reasoning in Hindi, Tamil, and Marathi

Anonymous ACL submission

Abstract

Commonsense reasoning is a key component of natural language understanding, yet most existing benchmarks, such as PIQA, are limited to English. This creates a significant gap for Indic languages, which represent one of the largest multilingual populations in the world. In this work, we present a pilot dataset for Physical Interaction Question Answering (PIQA) in three Indic languages: Hindi, Tamil, and Marathi. The dataset contains 450 manually curated samples (150 per language), each following the PIQA format with a goal, two candidate solutions, and a correct answer. Unlike direct translations from English, many examples were created to reflect culturally relevant, everyday scenarios in Indic contexts. We describe the collection methodology, provide dataset statistics, and outline key linguistic challenges encountered during annotation. This resource, though small in scale, is intended to serve as a starting point for research in multilingual commonsense reasoning, and as a foundation for future large-scale dataset development in Indic languages.

1 Introduction

Commonsense reasoning is essential for building natural language processing (NLP) systems that can interact with humans in realistic settings. A major aspect of commonsense is reasoning about the physical worldhow objects interact, how actions unfold, and which solutions to everyday tasks are plausible. The Physical Interaction Question Answering (PIQA) benchmark Bisk et al. (2020) has been widely adopted to study this problem in English. PIQA presents simple, everyday scenarios and asks models to choose between two possible solutions, only one of which is physically reasonable. This setup has proven valuable for probing models physical commonsense knowledge.

However, the vast majority of such datasets are limited to English. While multilingual benchmarks for tasks such as translation, sentiment analysis, and reading comprehension have become increasingly common, commonsense reasoning resources for Indic languages remain scarce. Indic languages, spoken by over a billion people, present unique linguistic and cultural challenges that differ substantially from English. Tasks grounded in physical commonsensesuch as how to cook, clean, or repair everyday objectsoften involve culture-specific practices, vocabulary, and contexts. Without resources in Indic languages, it is difficult to evaluate or train models for these scenarios.

To address this gap, we present a pilot PIQA-style dataset in three Indic languages: Hindi, Tamil, and Marathi. The dataset contains 450 manually curated samples (150 per language), each consisting of a goal, two candidate solutions, and a correct label. While small in size, this dataset is designed to reflect culturally relevant and everyday situations rather than direct translations of English examples. This makes the resource valuable not only as a benchmark but also as a starting point for scaling commonsense reasoning tasks to diverse languages.

2 Related Works

Commonsense reasoning has long been recognized as a critical component in natural language understanding, particularly when it involves reasoning about the physical world. Traditional NLP benchmarks have often focused on linguistic or factual knowledge, leaving a gap in evaluating models ability to reason about everyday interactions with objects and their affordances. To address this, several datasets and methods have been proposed, with the Physical Interaction Question Answering (PIQA) dataset being one of the earliest and most influential efforts in this space.

The Physical Interaction Question Answering (PIQA) dataset introduced by (Bisk et al., 2020) is one of the first large-scale benchmarks designed

to evaluate commonsense reasoning about everyday physical interactions. Each instance presents a goal (e.g., removing a pit from an avocado) and two possible solutions, with the task being to select the physically plausible one. While humans achieve near-perfect accuracy (95%), pretrained NLP models lag significantly (77%), revealing a major gap in their ability to reason about affordances, material properties, and causality. PIQA thus highlights the challenge of encoding practical physical knowledge beyond linguistic patterns. Building on this, (Choi et al., 2025) proposed Ko-PIQA, a Korean extension incorporating cultural context into physical commonsense reasoning. Unlike the English-centric PIQA, Ko-PIQA introduces 441 high-quality questions derived from millions of Q&A pairs, with nearly 20% grounded in culture-specific scenarios involving uniquely Korean objects and practices. Evaluation of multilingual and Korean language models showed performance ranging from 59.9% to 83.2%, with models particularly struggling on culturally grounded tasks compared to generic ones. This underscores the importance of cross-lingual and culturally diverse benchmarks in advancing the robustness of commonsense reasoning systems. (Sharma et al., 2025) proposes CWMI, a framework that augments a frozen large language model with a Causal Physics Module (CPM) and a causal intervention loss, letting it simulate cause-and-effect in physical settings; it achieves strong zero-shot performance on benchmarks like PIQA and a new PhysiCa-Bench designed to test counterfactual physical reasoning. (Bauer and Bansal, 2021) analyzed how knowledge graphs (ConceptNet, ATOMIC, WikiHow-based) align with tasks including PIQA. It compares how well KG information can fill gaps in model reasoning when solving PIQA. (Espejel et al., 2023) used PIQA among other benchmarks to compare performance of various large language models, to see their physical commonsense reasoning strengths & weaknesses. (Banerjee et al., 2021) included PIQA as one of the datasets to test how implicit knowledge (not explicitly encoded) helps or limits model performance on physical reasoning tasks.

Taken together, these works illustrate the growing recognition of PIQA as a cornerstone benchmark for physical commonsense reasoning. The evolution from PIQA to culturally enriched datasets like Ko-PIQA, as well as the integration of causal modules and external knowledge sources, highlights the diverse strategies pursued to bridge the gap between human-level physical intuition and machine reasoning. This body of research underscores that while significant progress has been made, physical commonsense remains an open challenge that continues to motivate innovations in both dataset construction and model design.

3 Dataset Creation

3.1 Motivation

Physical commonsense reasoning tasks, such as PIQA (Bisk et al., 2020), evaluate whether models can identify physically plausible actions in everyday scenarios. However, existing datasets are almost exclusively in English, which limits their applicability in multilingual contexts. We focus on Hindi, Tamil, and Marathi, three widely spoken Indic languages, to create a pilot resource that reflects diverse cultural practices and linguistic expressions of physical interactions.

3.2 Data Collection

We curated a total of 450 examples (150 per language) in the PIQA format. Each example consists of the following components:

1. **id:** ID pertaining to each data in the dataset.
2. **prompt:** A short statement describing a task or situation in the respective Indic language.
3. **solution0:** A possible but not necessarily correct completion of the task.
4. **solution1:** An alternative completion, designed to be close in meaning to solution0, ensuring difficulty in distinguishing between the two.
5. **label:** An indicator (0 or 1) specifying which solution is physically plausible.
6. **prompt_en:** An equivalent of the task statement in English.
7. **solution0_en:** The English rendering of solution0.
8. **solution1_en:** The English rendering of solution1.

Unlike standard translations of the English PIQA benchmark, our dataset was authored manually from scratch. This ensured cultural relevance and preserved linguistic nuances in Indic scripts. By

focusing on household practices, local problem-solving strategies, and region-specific physical commonsense, the dataset became authentically grounded in everyday Indian contexts.

3.2.1 Tools and Input Methods

To author examples in multiple Indic scripts, we utilized a combination of tools and input systems. Google Input Tools and native keyboard layouts for Devanagari were primarily used to ensure accurate rendering of Marathi text. In some cases, phonetic transliteration methods were employed, allowing contributors to type in Roman script while the system generated the correct Indic script output.

For proofreading and consistency checks, spell-checking plugins and online script validators were used. This reduced typographical errors and guaranteed that the written examples matched the expectations of native speakers. The reliance on digital tools helped streamline the process of writing large volumes of data while maintaining linguistic integrity.

3.2.2 Community Involvement and Validation

One of the unique aspects of our data collection process was the active involvement of family members, relatives, and peers who are native speakers of the respective languages. Draft examples were often read aloud to parents and grandparents, who provided corrections on colloquial usage, cultural appropriateness, and everyday realism. For instance, in tasks related to food storage, older family members validated whether a given method (such as wrapping bread in plastic or leaving it uncovered) truly reflected practical commonsense in Indian households.

Peers and younger relatives were also involved in evaluating whether the two solution options (sol0 and sol1) were confusing enough. Their feedback ensured that distractor solutions were not obviously incorrect but instead plausible and subtly misleading. This iterative feedback loop between contributors and native speakers improved both the linguistic quality and the commonsense grounding of the dataset.

3.2.3 Ensuring Cultural Relevance

By designing examples manually, we avoided Western-centric assumptions that could reduce the effectiveness of commonsense reasoning in Indic contexts. Tasks were chosen deliberately to reflect scenarios encountered in Indian daily life, few examples:

- **Food preparation and storage:** e.g., keeping vegetables fresh in tropical climates, cooking with traditional utensils.
- **Household chores:** e.g., drying clothes in monsoon conditions, managing dust while cleaning.
- **Handling devices:** e.g., saving mobile battery during power cuts, ensuring safety of electric appliances during voltage fluctuations.

This cultural embedding ensured that the dataset went beyond literal translation and instead captured the lived experiences of native speakers. Such grounding is essential for evaluating models not only on language understanding but also on their ability to apply physical commonsense in realistic environments.

3.2.4 Balancing Difficulty and Confusion

To mimic the challenge of the original PIQA benchmark, both solution options were carefully designed to sound plausible. For example, when asking how to store bread to prevent spoilage, one solution suggested wrapping in plastic and refrigerating, while the other suggested leaving it uncovered in a kitchen cabinet. Both appear reasonable, but only one is physically correct.

This balancing act required iterative rewriting. Family members often pointed out when a distractor solution was too obviously wrong. These insights guided us to produce distractors that were linguistically natural and semantically close to the correct solution, thereby increasing the challenge for commonsense reasoning models.

3.2.5 Collaborative and Iterative Process

The data collection process followed an iterative cycle: drafting, validation, rewriting, and final confirmation. Each cycle integrated technical input from contributors and real-world validation from native speakers. This community-driven and culturally-aware methodology ensured that the dataset was not only linguistically accurate but also practically grounded.

In conclusion, the dataset creation process integrated digital tools, native linguistic intuition, and community validation across generations. The result is a large-scale PIQA-style dataset in Indic contexts that challenges models to reason deeply about physical commonsense while navigating cultural specificity.

3.3 Annotation

Creation: A small team of native speakers of Hindi, Tamil, and Marathi authored the initial examples.

Validation: Each example was cross-checked by at least one additional native speaker for clarity and correctness.

Balance: Approximately half of the instances were constructed with sol0 as correct and half with sol1, to avoid positional bias.

Multiline Samples: To capture real-world instructions, at least 15 of the tasks in each language include multi-step solutions spanning multiple lines.

3.4 Dataset Statistics

The curated dataset consists of a total of **450 examples**, with **150 examples per language** (Hindi, Tamil, and Marathi). This balanced construction was intentional to ensure that no single language dominated the dataset and that cross-linguistic comparisons could be made fairly during evaluation. By maintaining equal representation across the three languages, we created a benchmark that reflects linguistic diversity and cultural realism in a controlled and consistent manner.

3.4.1 Distribution Across Labels

Each example in the dataset contains two possible solutions (*sol0* and *sol1*), out of which one solution is labeled as physically plausible. The dataset was constructed to be **balanced across labels**, with approximately 50% of the correct answers being *sol0* and the remaining 50% being *sol1*. This design choice prevents models from exploiting statistical biases (e.g., always predicting the same option) and ensures that performance is derived from genuine commonsense reasoning rather than label frequency.

3.4.2 Language Balance and Diversity

Each of the three languages was chosen to highlight different linguistic families within India:

- **Hindi (Indo-Aryan family)** represents one of the most widely spoken languages in India.
- **Marathi (Indo-Aryan family)** provides regional diversity and showcases unique cultural practices, especially in household contexts.
- **Tamil (Dravidian family)** brings in a completely different script and linguistic structure, testing whether models can generalize across typologically distant languages.

By curating examples in these languages, the dataset captures both intra-family and inter-family linguistic variations while keeping the number of examples per language equal. This diversity ensures that evaluation is not biased toward one particular script or linguistic style.

3.4.3 Summary

In summary, the dataset contains 450 carefully authored and validated examples, equally distributed across Hindi, Tamil, and Marathi, with a strict balance between the two solution labels. It is distributed in a simple and transparent **.tsv format**, which makes it convenient for both academic research and practical benchmarking.

3.5 Limitations

The dataset is small in scale (450 examples) and serves as a pilot study. Larger-scale annotation campaigns will be needed to train robust models. Additionally, all examples were curated by a small group of annotators; broader demographic participation would improve diversity.

4 Conclusion

In this work, we presented a pilot dataset for Physical Interaction Question Answering (PIQA) in three Indic languages: Hindi, Tamil, and Marathi. To the best of our knowledge, this is the first resource targeting physical commonsense reasoning beyond English, designed to capture everyday scenarios and culturally grounded practices from Indic contexts. The dataset consists of 450 manually curated examples, carefully balanced across labels, and validated by native speakers to ensure quality.

Although small in scale, this dataset highlights both the feasibility and the challenges of extending commonsense benchmarks to low-resource languages. By including examples that are not merely translations, but natively authored tasks relevant to local contexts, the dataset provides insights into how commonsense reasoning manifests differently across languages and cultures.

We hope this resource will serve as a starting point for the community, encouraging further research on multilingual commonsense reasoning, expansion to additional Indic and non-Indo-European languages, and the development of larger-scale benchmarks. Ultimately, this work aims to bridge the gap between English-centric commonsense reasoning resources and the linguistic diversity of the real world.

References

Pratyay Banerjee, Swaroop Mishra, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2021. Commonsense reasoning with implicit knowledge in natural language. In *Proceedings of the 3rd Conference on Automated Knowledge Base Construction (AKBC)*, page , Virtual / Online. Association for Computational Linguistics.

Lisa Bauer and Mohit Bansal. 2021. Identify, align, and integrate: Matching knowledge graphs to commonsense reasoning tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2259–2272, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Dasol Choi, Jungwhan Kim, and Guijin Son. 2025. Kopiqa: A korean physical commonsense reasoning dataset with cultural context. *arXiv preprint*.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *arXiv preprint*.

Aditya Sharma, Linh Nguyen, Ananya Gupta, Chengyu Wang, Chiamaka Adebayo, and Jakub Kowalski. 2025. Inducing causal world models in llms for zero-shot physical reasoning. *arXiv preprint*.