Extending Physical Interaction Question Answering to Under-Resourced Languages : Nepali, Telugu, and Hindi @ MRL 2025

Ram Mohan Rao Kadiyala *[1], Jebish Purbey *[1], Siddhant Gupta *[1,2],

[1]Cohere Labs Community, [2]IIT Roorkee,

Correspondence: contact@rkadiyala.com

Datasets & Code: Datasets

Abstract

This paper presents a multilingual Physical Interaction Question Answering (PIQA)-style dataset for Nepali (ne), Telugu (te), and Hindi (hi), developed as part of the MRL 2025 (MRL, 2025) on Multilingual Physical Reasoning Datasets. The dataset comprises 576 manually curated examples designed to evaluate physical commonsense reasoning. Each example includes a prompt with two candidate solutions differing minimally in wording, along with corresponding english translations. We describe the dataset creation process, emphasizing native speaker involvement and rigorous quality checks to ensure uniqueness and accuracy.

1 Introduction

Physical commonsense reasoning is essential for natural language processing systems to understand and predict outcomes of everyday physical interactions. The Physical Interaction Question Answering (PIQA) dataset (Bisk et al., 2020) provides a robust English-language benchmark for evaluating such capabilities, featuring prompts with two near-identical candidate solutions that test subtle physical reasoning. However, comparable datasets are scarce for non-English languages, particularly under-resourced ones, limiting the development of inclusive NLP models. This paper introduces a multilingual PIQA-style dataset for Nepali, Telugu, and Hindi, three languages with significant speaker populations but limited resources. We contribute original examples crafted to reflect culturally specific scenarios, such as cooking traditional dishes or handling region-specific tools. To ensure quality, we implemented several validation steps, including native speaker reviews and uniqueness checks to confirm that examples are distinct from each other and the original PIQA dataset. Our dataset aims to support the development of a comprehensive multilingual physical reasoning benchmark with additional languages.

* Equal Contribution.

2 Dataset

The dataset comprises 576 PIQA-style examples, evenly distributed across Nepali, Telugu, and Hindi, with 192 examples per language. Each example follows the PIQA format, consisting of a prompt describing a physical scenario and two candidate solutions differing very minimally. The characteristics of the samples can be seen below in Table 1.

3 Quality Verification

Initial checks were performed using Gemini 2.5 Flash to assess translation quality and sample accuracy. Subsequently, a manual review by a native speaker for each language evaluated translation quality, accuracy, grammar, and punctuation. Only samples passing both automated and manual checks were retained, with all others discarded.

4 Similarity Checks

To ensure the originality of our dataset, we conducted similarity checks between the English translations of our annotated samples and the 21,000 samples in the PIQA benchmark across all 3 splits. Using a Jaccard similarity metric based on character and word n-grams, we discarded any sample with a similarity score above 0.05 against any PIQA sample to preserve dataset novelty. Additionally,

Part	Field	Average word count			
		Hindi	Nepali	Telugu	Overall
English (Original	question	18.96	19.27	19.49	19.24
	correct option	13.54	14.18	14.07	13.93
	incorrect option	13.52	14.11	14.04	13.89
Multilingual	question	22.15	16.02	13.72	17.30
	correct option	15.57	10.93	10.14	12.21
	incorrect option	15.47	10.83	10.16	12.15

Table 1: Average word counts by language, grouped by *Original* vs. *Multilingual* parts.

Feature	Data Type	Description
prompt	Str	The Question i.e the Input
solution0	Str	Option labeled 0
solution1	Str	Option labeled 1
label	Int	The correct label i.e, 0 or 1
self_index	Int	Original Index of the sample prior to several stages of filtering
sim_indices	List	List of indices of the 3 samples that are the most similar from PIQA
sim_scores	List	List of similarity scores (.5f) of the top 3 most similar samples from PIQA
self_indices	List	List of indices of the top 3 most similar samples from our own data (self_index)
self_scores	List	List of similarity scores (.5f) of the top 3 most similar samples from our own samples
question	Str	The question (prompt) in English
correct_option	Str	The correct option in English
incorrect_option	Str	The incorrect option in English
question_m	Str	The question (prompt) in Target language (lang)
correct_option_m	Str	The correct option in the Target language (lang)
incorrect_option_m	Str	The incorrect option in the Target language (lang)
Subset	Int	Subset number (initial samples :1 , later extended :2)
op_sim	Float	Similarity scores between the two options (multilingual)
Lang	Str	Target Langauge

Table 2: The features available in the dataset along with their data types and descriptions

we computed pairwise similarity scores among our samples to ensure intra-dataset diversity. Using a graph-based approach, we identified and removed samples with similarity scores exceeding 0.05, retaining only those in an independent set to maximize uniqueness within the collection. Samples with options whose similarity score is lower than 0.6 were rewritten.

5 Conclusion

This dataset developed for the MRL 2025 Task fills a critical gap in physical commonsense reasoning resources for under-resourced languages and is available under MIT[1] license. Rigorous quality and uniqueness checks ensure its cultural relevance and distinctiveness, making it valuable for multilingual NLP research.

[1] https://opensource.org/license/mit

Limitations

The dataset, while carefully curated, may not encompass the full range of physical commonsense reasoning domains relevant to Nepali, Telugu, and Hindi speakers, potentially limiting its coverage of specialized scenarios. Finally, the Jaccard similarity threshold of 0.05 for filtering samples, while effective, may exclude some valid but unique cases.

References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

MRL. 2025. Mrl 2025 shared task on multilingual physical reasoning datasets. https://sigtyp.github.io/st2025-mrl.html.