Culturally Nuanced Multilingual Indic Physical Commonsense Dataset (MRL 2025 Shared Task)

Mausami Narayan*
Independent Researcher, India
mausami464.mn@gmail.com

Pramit Sahoo*
Independent Researcher, India
pramitsahoo.gnipst@gmail.com

Abstract

We describe a community-authored, culturally nuanced PIQA-style dataset in eight Indian languages for the MRL 2025 Shared Task on Multilingual Physical Reasoning. Examples are written natively (no translation), with minimal lexical differences between candidate solutions and emphasis on locally grounded physical commonsense. We report per-language sizes, token lengths, and label balance, and outline our native-speaker creation protocol and ethical safeguards.

1 Introduction

The MRL 2025 Shared Task(The MRL 2025 Workshop Organizers, 2025)invites manually constructed, culturally grounded PIQA-style datasets in non-English languages (Bisk et al., 2019). High-quality, community-authored resources remain scarce for many languages, especially low-resource ones. This gap is acute for culture-sensitive commonsense: many everyday physical practices (materials, tools, food preparation, rituals, climate-specific heuristics) are language–region specific and resist direct translation.

We present a culturally nuanced PIQA-style dataset spanning eight Indian languages, authored by native speakers following shared guidelines emphasizing local practices and terminology. The resource targets two needs: (i) native-language physical commonsense evaluation and (ii) coverage of under-resourced scripts and varieties.

2 Dataset

2.1 Languages and Scripts

Our release covers the following languages and scripts (Table 1).

2.2 Size and Basic Statistics

The dataset contains **2182** items. Mean token counts (whitespace) are: prompts **7.86** (chars 47.53), solution0 **5.06** (chars 30.53), solution1 **4.76** (chars 28.30). Label distribution: **1291** for label 0 and **891** for label 1.

2.3 Cultural Coverage

Following a 17-facet cultural template inspired by prior culturally grounded datasets like DIWALI[1], annotators were asked to write prompts anchored in everyday practices across food, rituals, tools, climate, and household heuristics, among others. Items emphasize locally salient materials (e.g., clay, brass, coconut husk), climate-driven tactics (e.g., drying, fermentation), festival-specific actions, and region-specific cookware or scripts.

3 Creation Process

All items were written by native speakers. Two primary authors contributed as follows: one native in Bengali and Odia created those sets; another native in Hindi and Maithili created those sets. For the remaining languages (Meitei/Manipuri, Assamese, Malayalam, Telugu), native-speaker friends contributed under the same guidelines. Each language had at least one native-speaker authoring pass; minor spot checks were done by the coordinating authors.All the annotator details are avialble in Table 2. We did not translate PIQA items; all original items were written directly in the target language and script.

Authoring Guidelines. Annotators were asked to (i) ground prompts in local culture and physical practice; (ii) keep candidate solutions minimally different but semantically contrastive; (iii) avoid encyclopedic trivia; (iv) vary length and surface forms; and (v) avoid code-switching unless natural

* Equal contribution

[1] https://huggingface.co/datasets/nlip/DIWALI

Language	Script	#Items	Avg Prompt Tok	Avg Sol0 Tok	Avg Sol1 Tok	Labels (0/1)
Assamese	Bengali	195	8.49	6.51	6.25	99/96
Bengali	Bengali	500	8.27	3.91	3.66	251/249
Hindi	Devanagari	495	7.98	4.82	4.62	390/105
Maithili	Devanagari	227	8.45	7.44	7.07	115/112
Malayalam	Malayalam	215	6.17	5.43	5.22	159/56
Manipuri	Meitei	114	8.23	5.33	4.96	57/57
Odia	Odia	181	7.65	3.54	3.19	92/89
Telugu	Telugu	255	7.26	5.17	4.62	128/127

Table 1: Per-language size and average token lengths (whitespace tokenization).

to the language variety. Labels denote the physically/plausibly correct choice.

4 Analysis

Length. We compute whitespace-token means per field and report them in Table 1; we also track mean character counts to capture script-dependent surface length.

Balance. Label balance varies by language (Table 1). Future rounds will target tighter balance and controlled lexical confounds.

Scripts. All items are authored in native scripts per language; no systematic Latin transliteration/code-mixing remains after curation.

5 Ethical Considerations

Cultural sensitivity. Annotators avoided stereotyping and favored neutral, practice-oriented descriptions; we welcome post-release community feedback.

Attribution and consent. Primary annotators (named authors) consent to release; community contributors opted to remain unnamed and are acknowledged collectively ("helpful community").

Content scope. Items avoid personal data, identity attributes, or harmful instructions; all examples are everyday, benign practices.

Intended use. Research/benchmarking of physical commonsense in multilingual settings; not for safety-critical deployments without further validation.

6 Reporting Checklist (per Shared Task)

- **Languages:** Assamese, Bengali, Hindi, Maithili, Malayalam, Manipuri, Odia, Telugu.
- **Scripts:** Included per language in Table 1.
- **Original items:** All items written natively; no English-PIQA translations counted toward the minimum.
- **Native checks:** At least one native speaker per language.
- **Dataset format:** `.tsv` with columns `prompt`, `solution0`, `solution1`, `label`.
- **Construction details:** Native authoring under unified cultural-facet guidelines; minimal lexical differences between candidates.

7 Limitations

Coverage across the 17 facets is uneven; some domains (e.g., specialized crafts) remain underrepresented. Label balance varies by language. Whitespace tokenization underestimates morphological complexity in some scripts.

8 Conclusion

We release a community-authored, culturally nuanced PIQA-style dataset covering multiple Indian languages and scripts. We hope it serves as a strong, native-language benchmark for physical commonsense and spurs similar community efforts.

Acknowledgments

We thank our helpful community contributors for language expertise and cultural insights.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

The MRL 2025 Workshop Organizers. 2025. MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets. https://sigtyp.github.io/st2025-mrl.html. Shared Task at the MRL Workshop at EMNLP 2025. Accessed: 2025-09-16.

A Cultural Facets Used

1. Food & cooking methods
2. Utensils & materials
3. Festivals & rituals
4. Clothing & textiles
5. Architecture & household objects
6. Tools & crafts
7. Agriculture & livelihoods
8. Markets & trade practices
9. Music & dance
10. Games & sports
11. Transport & travel
12. Weather & climate practices
13. Health remedies & hygiene
14. Social etiquette & greetings
15. Folklore, beliefs & taboos
16. Places, geography & environment
17. Education, work & daily routines

B Annotator Details

Evaluator	Location (Sub-Sub-Region/Sub-Region/Country)	YoR†	Educational Qualification
A	Kishanganj / Bihar / India	22	Post-Graduate
B	Contai / West Bengal / India	23	Post-Graduate
C	Dibrugarh / Assam / India	21	Graduate
D	Imphal / Manipur / India	24	Post-Graduate
E	Hyderabad / Telangana / India	26	Post-Graduate
F	Kochi / Kerala / India	25	Post-Graduate

Table 2: Demographic information of human evaluators.
†YoR = years of residence in the sub-region.