

# Dataset Description: Telugu Physical Reasoning Dataset (MRL 2025)

---

## 1. Introduction

This dataset contributes Telugu-language examples to the MRL 2025 Shared Task on Multilingual Physical Reasoning. The objective is to create culturally grounded, non-English reasoning data that challenges models on commonsense physical reasoning tasks. Telugu is a widely spoken Dravidian language with over 80 million speakers, yet remains underrepresented in multilingual NLP benchmarks. This dataset fills that gap by offering original PIQA-style examples relevant to Telugu culture and daily life.

## 2. Dataset Design

The dataset consists of 132 manually authored examples in Telugu. Each entry follows the PIQA-style format: a 'prompt' describing a physical reasoning situation, two candidate 'solutions' (solution0, solution1), and a 'label' indicating the correct solution. Examples were crafted to reflect realistic scenarios encountered in Telugu households, agriculture, cooking, transportation, and daily problem-solving. We avoided direct translation from English datasets; instead, prompts were designed organically with local context.

## 3. Annotation Process

All examples were authored and verified by native Telugu speakers with domain knowledge of daily activities. Each question was double-checked for clarity, cultural appropriateness, and correctness of the labeled answer. Edge cases and ambiguous situations were discarded to ensure high quality.

## 4. Analysis

The dataset covers multiple categories of physical reasoning:

- Cooking and household reasoning (e.g., boiling milk, frying spices).
- Agricultural and environmental reasoning (e.g., storing grains, watering plants).
- Physics-inspired scenarios (e.g., heat expansion, condensation, light reflection).
- Everyday object use and safety (e.g., electrical safety, furniture stability).

The balance ensures coverage of both common and nuanced situations relevant to Telugu speakers.

## 5. Ethical Considerations

We ensured examples are culturally respectful, non-stereotypical, and free from harmful or biased language. No sensitive personal or political content is included. The dataset is intended for research use only and does not include personally identifiable information.

## 6. Conclusion

This dataset represents one of the first Telugu contributions to multilingual physical reasoning. By including it in the MRL 2025 shared task, we expand representation of Dravidian languages and offer a benchmark for future multilingual commonsense reasoning research. The dataset and description are prepared in compliance with the shared task guidelines.