A Dataset for Physical Commonsense Reasoning in Kosta Andhra Telugu

Author: Gopi Naga Sai Ram Challagolla, Vishnu Vardhan Reddy Kummitha
Native Speakers and Dataset Curators

Abstract

we introduce a dataset for physical commonsense reasoning in Kosta Andhra Telugu, a dialect of Telugu spoken in the coastal regions of Andhra Pradesh, India. This dataset contributes to the MRL 2025 Shared Task on Multilingual Physical Commonsense Reasoning at EMNLP. It consists of 112 manually crafted examples in the PIQA format, each comprising a prompt, two candidate solutions (one correct and one incorrect), and a label indicating the correct solution. The examples focus on physical properties of objects and actions, incorporating culturally specific elements from Andhra Pradesh, such as local festivals (e.g., Sankranti, Dasara, Vinayaka Chavithi) and traditional foods (e.g., idli, pongal, ugadi pachadi). All examples were manually curated and verified by a native speaker to ensure accuracy and cultural relevance. No machine translation or LLMs were used in construction.

Introduction

The MRL 2025 Shared Task aims to build multilingual datasets for physical commonsense reasoning, following the PIQA format: a prompt describing a scenario, followed by two similar solutions where only one is physically correct. This dataset targets Kosta Andhra Telugu, a vibrant dialect used by millions in coastal Andhra Pradesh, India. It emphasizes everyday physical interactions, blending universal commonsense with regional cultural contexts to reflect real-world usage in this dialect.

The dataset is submitted as a TSV file with columns: prompt, solution0, solution1, label (0 or 1 indicating the correct solution). It meets the task requirements, with a minimum of 100 examples, all manually checked.

Dataset Construction

The dataset was entirely manually constructed by native speakers of Kosta Andhra Telugu with expertise in dataset creation for LLMs. Examples were written to cover diverse physical reasoning scenarios, such as material properties (e.g., heating metal expands it), fluid dynamics (e.g., milk overflows without a lid), and object interactions (e.g., rubber bounces, glass shatters).

Heuristics used:

- **Variety in Prompts**: Prompts vary in structure and length to avoid repetition. Some are single sentences (e.g., "ఒక గాజు తలుపు బలంగా మూసేస్తే"), while others are multi-sentence or procedural (e.g., descriptions of cooking or festivals).

- **Similar Solutions**: Candidate solutions are designed to be structurally similar, differing by one or two key words or phrases (e.g., "వేడెక్కుతుంది" vs. "చల్లబడుతుంది"). The incorrect solution is plausible but physically wrong, not absurd, to challenge models.

- **Cultural Relevance**: Incorporated Andhra-specific elements, such as pongal preparation during festivals, mango pickles, or clay idols in Vinayaka Chavithi, requiring regional commonsense.

- **Length Variation**: Most examples aim for combined prompt and solutions over 25 words (approximated via character count in Telugu script), with some shorter for diversity.

- **Label Balance**: Labels are assigned based on physical accuracy, with a natural skew toward 0 due to construction patterns, but including some 1 for variety.

- **Quality Assurance**: All examples were double-checked for grammatical correctness, physical accuracy, and commonsense (verifiable by an average native speaker). No translated PIQA examples were included.

Dataset Statistics

- **Number of Examples**: 112

- **Label Distribution**: 112 examples with label 0 (solution0 correct), 4 with label 1 (solution1 correct).

- **Dialect Specificity**: All text uses Kosta Andhra Telugu orthography and vocabulary, reflecting coastal Andhra colloquialisms (e.g., regional terms for foods and festivals).

A speaker of Kosta Andhra Telugu could reconstruct a similar dataset by brainstorming everyday physical scenarios, incorporating local customs, and ensuring solutions differ minimally while one aligns with physics.

Conclusion

This dataset advances multilingual physical commonsense benchmarks by representing Kosta Andhra Telugu. It is suitable for training and evaluating LLMs on low-resource languages with cultural nuance. Future expansions could include more balanced labels and additional dialects. The dataset is available in TSV format for the shared task.