Urdu Physical Commonsense Reasoning Dataset Submission for MRL 2025

Anonymous ACL submission

Abstract

Commonsense remains a cornerstone of natural language understanding. For AI systems to work effectively in everyday context, we need to consider reasoning as a key capability, particularly the ability to draw upon physical commonsense knowledge. This involves how objects behave, how they interact, and what physical consequences result from their common actions. Although benchmarks in English have driven significant progress, resources for underserved languages such as Urdu remain scarce. In this paper, we introduce the first Urdu dataset for physical common sense reasoning, developed as a part of MRL 2025 Shared Task. The dataset contains 100 manually constructed examples in PIQA format, covering scenarios such as cooking, religion, weather, science, and household activities. We also considered culturally specific contexts unique to Urdu speakers. The contribution fills a gap in multilingual commonsense research, providing a foundation for reasoning abilities in low-resource settings.

1 Introduction

In the real world, humans are able to reason about the physical world with ease. Experience has taught us that damp clothing takes longer to dry in humid climates, that ceramic cups are more brittle than glass ones and that rainwater that is left on soil will not evaporate right away. What is commonly referred to as common sense is reflected in these minor but trustworthy assessments. This type of knowledge development is far from being simple for artificial intelligence, but it is necessary if AI systems are to communicate with humans.

Languages such as English have advanced thanks to benchmarks like PIQA (Bisk et al., 2020) and associated resources. However, there aren't any specific datasets for physical commonsense reasoning other than English. This disparity is particularly noteworthy for low-resource languages, as the lack of such metrics can hinder the methodical assessment and enhancement of multilingual models. One of these underrepresented languages is Urdu, which is spoken by over 230 million people (Chaman, 2022).

In this paper, the first dataset for physical commonsense reasoning in Urdu is presented, created for MRL 2025 Shared Task.The dataset contains 100 manually created examples in PIQA format, complete with prompt, two potential solutions, and the appropriate label. To ensure linguistic and cultural validity, a native speaker wrote and reviewed each example. In addition to culturally relevant contexts that might be difficult to convey in English, the dataset includes a variety of scenarios.

Our contribution fills a significant void in multilingual commonsense reasoning. We make it possible to assess reasoning skills in a widely spoken language with limited resources by introducing physical commonsense benchmarks to Urdu . Future research on cross-lingual transfer, culturally sensitive reasoning, and the more general objective of creating AI systems that can reliably operate across languages and societies is what we hope this resource will support.

2 Dataset Construction

The Urdu Physical Commonsense Reasoning Dataset follows the Physical Interaction Question Answering (PIQA) format which was presented by (Bisk et al., 2020).In each instance, a natural language prompt that describes a physical scenario is presented alongside two possible solutions, only one of which is possibly correct based on common sense.A model must select the correct response to show that they understand fundamental physical commonsense.

To create the dataset, we manually authored 100 examples covering diverse everyday scenarios.To discourage cursory pattern recognition, the majority of the prompts were longer than 25 words and varied in complexity. The wording of the candidate solutions was purposefully designed to be very similar, with only a few important terms or phrases separating them. In one example, the solution claims that a clay cup feels less hot to the hand because it releases heat gradually, while the distractor claims that a clay cup would feel colder because it does not retain heat. This ensures that only one of the answers is physically possible, even though they appear to be similar.

During the construction of the dataset, cultural and linguistic specificity were given top priority. To achieve this, examples were taken from situations and examples that Urdu speakers know well, like regional cuisine (chapati or biryani), domestic items (charpai, matka), and daily practices (e.g. wearing a dupatta for shade). This ensures that instead of being a direct translation of the standards and benchmarks in English, the dataset is based on common sense reasoning based on the lived experiences of communities that speak Urdu.

Each example was manually reviewed by a native speaker of Urdu, to verify grammar, naturalness, and correctness of the labels. Often, while large language models were consulted during the brainstorming phase, all final examples were curated and validated manually to maintain quality and prevent repetitive or trivial patterns.

3 Dataset Statistics

Our Urdu Physical Commonsense Reasoning dataset contains a total of **100 examples**. Each instance consists of a natural language prompt, two nominee solutions and a binary label marking the correct answer.

3.1 Prompt and solution lengths

The prompts are moderately detailed, with an **average length of 24.5 tokens**, often spanning multiple clauses or sentences to provide sufficient reason out context. Candidate solutions are shorter and more concise, **averaging 15.5 tokens** for both solution0 and solution1. The solutions are intentionally designed to be similar in form and length, often differing by only a few words, ensuring that the task requires true common sense reasoning rather than superficial cues.

3.2 Label Balance

The dataset is well-balanced: 50.5 % of examples have, more or less, solution0 as correct and 49.5 % have solution1 as correct. This balance prevents models from relying on positional heuristics and ensures that successful performance postulates genuine reasoning ability.

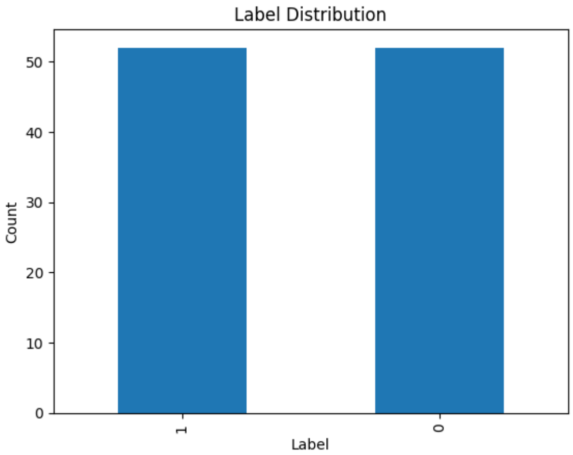

Figure 1: Distribution of Correct Answers Across Solution0 and Solution1

3.3 Domain coverage

The dataset spans a wide range of domains, including:

- **Science and the physical world** (e.g., reasoning about water, fire, plants, pressure, and light),
- **Daily life activities** (e.g., cooking, cleaning, clothing, and household objects),
- **Culture and religion** (e.g., traditional foods, mosques, religious practices), and
- **Society, politics, and history** (e.g., national identity, poetry, governance, and historical events).

This distribution reflects our goal of combining general physical commonsense with culturally specific contexts relevant to Urdu speakers. Such diversity broadens the benchmark's applicability and increases the difficulty of the task for both human annotators and AI models.

Overall, the dataset is carefully constructed to be balanced, diverse, and linguistically grounded, making it a challenging resource for evaluating multilingual physical commonsense reasoning.

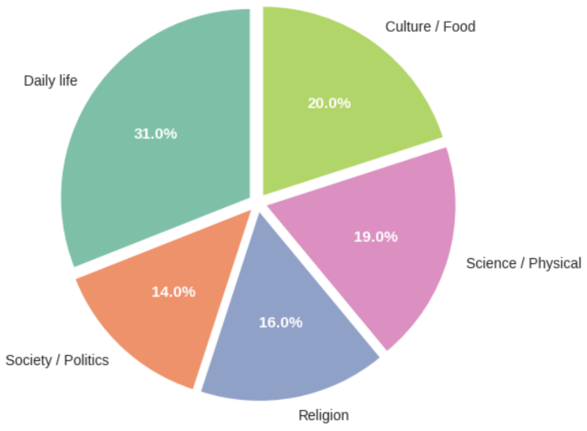

Figure 2: Domain-wise Distribution of Dataset Examples

3.4 Scientific Categorical Distribution

Within the scientific and physical world domain, examples are further divided into specific scientific categories, including Physics, Biology, Chemistry, and General Science. This finer-grained categorization highlights the diversity of scientific reasoning required by the dataset.

Figure 3 shows the distribution of examples across these categories.

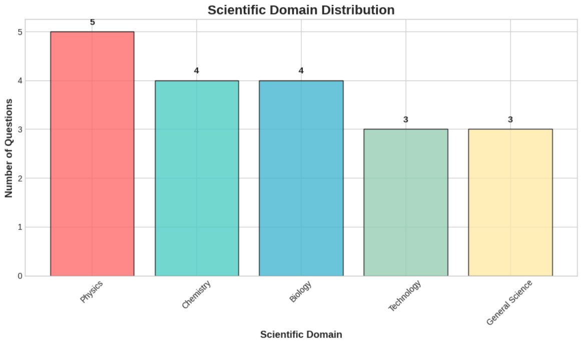

Figure 3: Distribution of examples across specific scientific categories (Physics, Biology, Chemistry, and General Science) within the scientific and physical world domain

4 Potential Applications

The Urdu Physical Commonsense Reasoning Dataset has several potential applications in NLP and AI research:

- **Multilingual Commonsense Benchmarking:** It enables evaluation of cross-lingual transfer capabilities of pre-trained models, particularly for low-resource languages like Urdu

- **Culturally Aware AI systems:** By incorporating culturally relevant scenarios, the dataset allows models to reason in ways that align with local knowledge and practices.

- **Educational Tools:** The dataset can serve as a basis for AI-assisted tutoring systems, where physical reasoning and problem-solving in real-world contexts are tested.

- **Robotics and Embodied AI:** Physical common sense reasoning is essential for robots operating in human environments. This dataset can help train and evaluate models that guide robotic actions.

5 Conclusion

We present **Urdu Physical Commonsense Reasoning Dataset**, the first manually curated dataset for Urdu physical commonsense reasoning. This resource, based on culturally relevant scenarios, is a much needed tool for evaluating reasoning skills of multilingual models in a low-resource language.

In the future, we plan to significantly expand both the size of the dataset and the diversity of authors. Ultimately, we hope that this work will encourage the development of similar resources for other underrepresented languages, resulting in AI systems that are more equitable and capable of operating globally.

6 References

References

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Hussain Chaman. 2022. Language politics in pakistan: Urdu as official versus national lingua franca. *Annals of Human and Social Sciences*, 3(2):82–91.