Towards Multilingual Physical Commonsense Reasoning: A Dataset Contribution for Urdu

Mahnoor Malik[1]

1 Dataset Overview

The Multilingual Physical Commonsense Reasoning Shared Task at the 5th MRL Workshop aims to create culturally diverse evaluation datasets in multiple languages. In line with this objective, this paper presents the contribution of an Urdu dataset that captures physical commonsense reasoning grounded in everyday scenarios relevant to Urdu speakers. Urdu, spoken by millions worldwide, lacks standardized benchmarks for physical reasoning tasks, which limits the evaluation of multilingual and culturally sensitive language models. The dataset follows the PIQA format, where each example consists of a prompt paired with two candidate completions: one correct and one incorrect. This contribution not only broadens the linguistic diversity of the shared benchmark but also provides a resource for evaluating and improving multilingual models on culturally specific reasoning tasks in Urdu.

2 Methodology

The dataset was constructed following the guidelines of the Multilingual Physical Commonsense Reasoning Shared Task. It was manually developed in Romanized Urdu, reflecting the way many Pakistanis communicate on social media platforms. People are accustomed to typing in Romanized Urdu when messaging on WhatsApp and similar applications. The dataset also contains English words. Many of these English borrowings are so common that native speakers are often unaware of their Urdu equivalents, making them a natural part of the dataset. In total, the dataset comprises 100 examples. Table 1 presents the fields in the dataset.

Once completed in Romanized Urdu, the dataset was converted into standard Urdu script using the Gemini-2.5-Flash model. After automatic conversion, each example was manually verified. During this process, several noteworthy observations were made:

- **Error correction during conversion:** In many cases, spelling or grammatical errors in Romanized Urdu were automatically corrected by the model in Urdu script. In the following example, ***pe*** was automatically converted to ***per*** and ***hawsbay*** was automatically converted to ***hawksbay*** in Urdu script.

  sahili Patti pe toofan ka khadsha hai. Jiski Waja se hukumat ne khas tor per sea view or hawsbay par dafa 144 nafiz krdi hai.

  ساحلی پٹی پر طوفان کا خدشہ ہے۔ جس کی وجہ سے حکومت نے خاص طور پر سی ویو اور ہاکس بے پر دفعہ 144 نافذ کر دی ہے۔

- **Correction of incorrect solutions:** When converting deliberately incorrect solutions, the model sometimes "repaired" them into correct solutions. In the following example, ***dabanay*** was automatically converted to ***milnay*** in Urdu script.

  baray ijtimaat mai shamil honay se guraiz karen. aapas mai hath milanay or galay dabanay se parhaiz karain

  بڑے اجتماعات میں شامل ہونے سے گریز کریں۔ آپس میں ہاتھ ملانے اور گلے ملنے سے پرہیز کریں۔

[1] Lecturer, NED University of Engineering and Technology, Pakistan, mahnoormalik@neduet.edu.pk

Table 1: Fields in the Urdu Physical Commonsense Reasoning Dataset.

Field	Description
prompt_latn	Prompt in Romanized Urdu
sol0_latn	Candidate solution 0 in Romanized Urdu
sol1_latn	Candidate solution 1 in Romanized Urdu
prompt_urdu	Prompt in native-script Urdu
sol0_urdu	Candidate solution 0 in native-script Urdu
sol1_urdu	Candidate solution 1 in native-script Urdu
label	Correct solution (0 or 1)
{model-name}-output-romanized-urdu	Output by the model (0 or 1) for Romanized Urdu
{model-name}-output-urdu	Output by the model (0 or 1) for native-script Urdu
{model-name}-output-reason-romanized-urdu	Reason by the model for Romanized Urdu
{model-name}-output-reason-urdu	Reason by the model for native-script Urdu
{model-name}-incorrect-reason-romanized-urdu	Manual verification of reason for Romanized Urdu
{model-name}-incorrect-reason-urdu	Manual verification of reason for native-script Urdu

3 Evaluation of Models

To evaluate accuracy, structured prompts [1,2] were designed for both Romanized Urdu and native-script Urdu . Each prompt consists of (i) a short input (a question or incomplete sentence), and (ii) two candidate solutions. The model was required to choose the more appropriate solution and provide a brief justification for its choice.

```
You are given a prompt in Romanized Urdu along with two possible solutions.
The prompt may be a question or a sentence that needs completion.
Your task is to decide which solution is more correct or appropriate.

Prompt:
{prompt_latn}

Options:
solution0: {sol0_latn}
solution1: {sol1_latn}

Instructions:
- Select only one correct/appropriate solution (0 or 1).
- Follow the exact output format below:

output: 0 or 1
reason: <brief explanation why this solution is correct>
```

Figure 1: Prompt used for evaluating accuracy across different models for Romanized Urdu.

Table 2 presents the accuracy of different large language models on both Romanized Urdu and Urdu in its native script. Overall, the results indicate that model performance varies significantly across the two writing systems. The gemini-2.5-flash model achieves the highest accuracy, reaching 99% on Romanized Urdu and 100% on the native script, showing its strong multilingual capability. In contrast, models such as Mixtral-8x7B-Instruct perform relatively poorly, with accuracies of 54% on both Romanized Urdu and Urdu in its native script. Interestingly, some models (e.g., Llama-3.1-8B-Instruct) show higher performance on the native script (77%) compared to Romanized Urdu (62%), suggesting a stronger alignment with the standard orthography. Meanwhile, gpt-4o-mini demonstrates strong performance across both forms, particularly excelling in the native script

```
You are given a prompt in Urdu along with two possible solutions.
The prompt may be a question or a sentence that needs completion.
Your task is to decide which solution is more correct or appropriate.

Prompt:
{prompt_urdu}

Options:
solution0: {sol0_urdu}
solution1: {sol1_urdu}

Instructions:
- Select only one correct/appropriate solution (0 or 1).
- Follow the exact output format below:

output: 0 or 1
reason: <brief explanation why this solution is correct>
```

Figure 2: Prompt used for evaluating accuracy across different models for Urdu in its naitve script.

(98%). These results highlight that while some models generalize well across both orthographies, others display a bias toward the standard Urdu script. Note that max_tokens was set to 128 incase of all the models except for gemini-2.5-flash.

Table 2: Accuracy of different models on Urdu(Romanized) and Urdu(Native Script)

Model	**Urdu(Romanized)**	**Urdu(Native Script)**
gemini-2.5-flash	99%	100%
Qwen2.5-7B-Instruct	61%	66%
aya-expanse-8b	69%	60%
Mixtral-8x7B-Instruct-v0.1	54%	54%
Llama-3.1-8B-Instruct	62%	77%
gpt-4o-mini	89%	98%

Table 3 reports the reasoning accuracy of different large language models when evaluated on Urdu in both Romanized and native script forms. Reasoning accuracy is defined as the proportion of cases in which a model produced a correct explanation for its decision, conditioned on the prediction (0 or 1) being correct. The results show a wide variation across models. The reasons generated by the models were expressed in both Romanized Urdu and standard Urdu script, with occasional instances of Hindi words also appearing. For all examples in which a model predicted the correct output (0 or 1), the corresponding reasoning was manually verified. A reason was judged as incorrect under several conditions: if it was entirely wrong or only partially correct, if it lacked coherent structure (regardless of whether written in Urdu script or Romanized Urdu), if it contained major grammatical errors, if the sentence was incomplete, or if the model simply restated the correct solution without providing any justification.

Among all models, Gemini-2.5-Flash achieves perfect reasoning accuracy (100%) in both scripts, indicating consistent and reliable explanation generation. GPT-4o-mini also demonstrates high reasoning fidelity, particularly in the native script (94.90%), outperforming its Romanized counterpart (79.78%). In contrast, smaller models such as Qwen2.5-7B-Instruct, aya-expanse-8B, Mixtral-8×7B-Instruct, and Llama-3.1-8B-Instruct show significantly lower reasoning accuracy, with performance dropping further on native script text. For instance, aya-expanse-8B performs relatively poorly, with only 1.67% reasoning accuracy on native script compared to 27.54% on Romanized input.

Overall, these results highlight (i) that proprietary frontier models provide far more reliable reasoning compared to open-source models, and (ii) that native script Urdu remains more challenging for most models, except in the case of GPT-4o-mini and Gemini-2.5-Flash.

Table 3: Reasoning Accuracy of different models on Urdu(Romanized) and Urdu(Native Script)

Model	**Urdu(Romanized)**	**Urdu(Native Script)**
gemini-2.5-flash	100%	100%
Qwen2.5-7B-Instruct	36.07%	10.61%
aya-expanse-8b	27.54%	1.67%
Mixtral-8x7B-Instruct-v0.1	44.44%	38.89%
Llama-3.1-8B-Instruct	41.94%	36.36%
gpt-4o-mini	79.78%	94.90%

4 Conclusion

This paper introduced a novel Urdu dataset for physical commonsense reasoning, designed to expand the linguistic and cultural coverage of the Multilingual Physical Commonsense Reasoning Shared Task. By providing examples in both Romanized and native-script Urdu, the dataset reflects real-world usage patterns of Urdu speakers, particularly in digital communication. Evaluation results show that while frontier models like Gemini-2.5-Flash and GPT-4o-mini achieve high task and reasoning accuracy, open-source models often struggle, particularly in producing reliable explanations. These findings highlight the gap between prediction accuracy and reasoning quality, especially in low-resource languages.

5 Dataset

The dataset is publicly available on GitHub[2].

[2] https://github.com/Mahn00rMalik/Shared-Task-for-MRL-Workshop-at-EMNLP-2025