MRL 2025 Shared Task: Indonesian

Filbert Aurelian Tjiaranata, Vallerie Alexandra Putra,
Eryawan Presma Yulianrifat, Ikhlasul Akmal Hanif

Keywords: MRL Shared Task, Physical Reasoning Dataset

1 Introduction

Many languages lack culturally-specific evaluation datasets that are created by language community members themselves. The MRL 2025 Shared Task aims to contribute manually-annotated physical commonsense reasoning datasets for multiple languages, modeled after PIQA, a benchmark in which each example consists of a prompt and two candidate completions (“solutions”).

The purpose of this paper is to report the details of our contribution to the creation of the MRL dataset for Indonesian. We constructed and reviewed a dataset of 120 examples, designed to reflect physical commonsense reasoning patterns in our language. Each example was created and validated manually by members of our team, all of whom are native speakers of Indonesian.

2 Methods

Our dataset was created by four annotators, all of whom are co-authors of this report. Each annotator contributed 30 examples, resulting in 120 examples in total. To ensure quality and reduce bias, we implemented a peer-review mechanism:

- Author 1 supervised the contributions of Authors 2 and 3.
- Author 2 supervised the contributions of Authors 3 and 4.
- Author 3 supervised the contributions of Authors 1 and 4.
- Author 4 supervised the contributions of Authors 1 and 2.

This structure ensured that each instance was independently reviewed by at least two annotators, allowing us to identify and resolve errors, improve clarity, and enforce consistency across the dataset.

The prompts and solutions were created manually, inspired by the authors’ general knowledge, past experiences, and daily life activities. In addition, some prompts incorporated culturally specific Indonesian elements, such as food and traditional music instruments. In cases where we were not fully certain about the plausibility of a prompt and its solution, we cross-checked the information with reliable sources such as articles and reports from the internet. This process helped to ensure that both the prompt and solutions were valid.

3 Results

The final dataset consists of 120 examples. The average word count per example (prompt plus one candidate solution) is 25.13, with a standard deviation of 12.28. Out of the total, 63 examples contain more than 25 words, reflecting a diverse range of prompt complexity and detail.

These statistics suggest that our dataset captures a balanced distribution of short and long reasoning tasks, making it suitable for evaluating physical commonsense reasoning across different levels of linguistic complexity.

4 Conclusion

We presented our contribution to the MRL 2025 Shared Task: a manually-constructed dataset of 120 culturally-specific physical commonsense reasoning examples. Through a structured peer-review process, we aimed to ensure both quality and reliability of the data. We hope this resource will support future research in multilingual commonsense reasoning and highlight the importance of community-driven dataset creation.