Physical Reasoning Datasets for Estonian, Persian and Swedish

Kätriin Kukk, Jenny Kunz, Romina Oji and **Amin Bajand**
Linköping University
{katriin.kukk,jenny.kunz,romina.oji,amin.bajand}@liu.se

Abstract

We present small-scale physical commonsense reasoning datasets for three typologically diverse languages: Estonian, Persian and Swedish. The datasets were created for the MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets.

1 Introduction

When it comes to multilingual settings, Large Language Models are often evaluated on machine-translated data. This is an issue, not only because machine-translated text may show traces of the source language and feel unnatural to a native speaker of the target language, but also because translated examples preserve the cultural context of the source language. In an attempt to help increase the amount of culturally adapted evaluation data created by native speakers, we contribute 105 samples in Estonian, 123 in Persian and 100 in Swedish to the MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets. The rest of this paper describes the creation process of these datasets.

2 Method

2.1 Estonian

The Estonian part of the dataset was created by one native Estonian speaker. The annotator first brainstormed culturally relevant topics for each example with themes ranging from Estonian food, companies, places and the distances between them, cultural events and holidays, as well as typical activities and phenomena during different seasons of the year. In the majority of the cases, the theme of the example was such that the annotator felt confident about knowing all the relevant details based on experience, in which case no external sources were consulted for creating the example. Otherwise, the annotator used the search results on Google to read up on the theme – for example, to look up the recipe of a certain food item or the rules of a game – before constructing the sample. Once the minimum number of examples was reached, the annotator manually checked each question one more time to verify that the language was correct, the sample was about physical commonsense reasoning and that the label of the correct answer was accurate. Lastly, another Estonian native speaker verified the examples after which the samples that were unclear were modified.

2.2 Persian

The Persian subset of the dataset was created and annotated by two native Persian speakers to ensure linguistic naturalness and cultural relevance. It covers six thematic categories: cooking and food, housekeeping and cleaning, daily life and social customs, driving and travel, health and safety, life hacks and tools. The dataset focuses on scenarios grounded in everyday life, emphasizing practical physical reasoning that requires cultural and contextual knowledge. The source material was collected from publicly available online articles from the following websites: namnak.com, ninisite.com, and fa.wikipedia.org.

For each question, the annotators created one correct answer, either extracted or paraphrased from the source material, and one incorrect answer, which differed minimally from the correct answer while remaining grammatically well-formed but semantically implausible. This design ensures that models must rely on physical commonsense knowledge, rather than superficial lexical cues, to select the correct answer. All samples were systematically reviewed to verify that they represent physical reasoning tasks. In cases where the annotators did not agree on whether the question involved a physical aspect, the item was discarded and replaced with a new one. This iterative process guarantees that the dataset remains focused on physical, prac-

tical scenarios while retaining cultural authenticity.

2.3 Swedish

The annotator of the Swedish subset drew inspiration from a diverse set of web sources that cover everyday physical activities. These included sports-related Wikipedia pages and categories, as well as various Swedish websites such as villalivet.se (focused on house and garden life), naturvårdsverket.se (the Swedish Environmental Protection Agency), and krisinformation.se (a government portal for crisis information). From these sources, the annotator derived questions spanning a broad range of topics: sports and outdoor activities, household and gardening tasks, cooking and baking, physical activities associated with traditional festivities, and traffic-related scenarios. The primary annotator wrote questions and correct and wrong answers, while a second annotator who is a native speaker of Swedish verified them.

All questions are related to activities or things that are common in Sweden. A few are Sweden-*specific* in the sense that it is very helpful to be familiar with the Swedish context to answer them, but almost all questions concern things that are more popular in Sweden than in many other countries.

3 Conclusion

In this work, we have described our contribution to the MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets which comprises datasets in Estonian, Persian and Swedish. It is of great importance that datasets created for evaluating Large Language Models in multilingual settings are culturally adapted rather than machine-translated from English which makes this a valuable endeavor.