Multilingual Physical Commonsense Reasoning Datasets (Farsi)

Hamidreza Saffari[*1], **Mohammadamin Shafiei**[*2],
[1]Politecnico di Milano, [2]University of Milan
hamidreza.saffari@mail.polimi.it
m.shafieiapoorvari@studenti.unimi.it

Abstract

We present a dataset for physical commonsense reasoning in Farsi. To construct the dataset, we first used a large language model (LLM) to suggest broad categories, examples, and factual statements. Two native speakers then refined these categories and created or edited instances to ensure naturalness, cultural alignment, and coverage of diverse reasoning phenomena. The final dataset contains 100 high-quality samples, all human-authored or human-edited, with the LLM serving only in the initial brainstorming phase.

Methodology

To construct our dataset for physical commonsense reasoning in Farsi, we followed a hybrid process combining initial guidance from large language models (LLMs) with careful human curation. Since the instructions raised concerns about relying exclusively on LLMs for data creation, we used LLMs only as a supporting tool to bootstrap the design of categories and seed examples.

Category and Seed Generation. We prompted an LLM to propose high-level categories relevant to physical commonsense reasoning. For each category, the model was further asked to provide illustrative examples and candidate factual statements. Most of these outputs were not used directly in the dataset but rather served as a starting point for structuring the data collection. The categories spanned both everyday knowledge and culturally specific practices. For example, in the category *Food Preparation (Culturally Relevant Foods)*, one sample was: “To prepare mild pomegranate paste, boil pomegranate juice in a copper pot over low heat until it thickens evenly vs. boil pomegranate juice in a steel pot over low heat until it thickens evenly” (originally written in Farsi). In the category *Household Tasks (Everyday Objects)*, one sample was: “To clean an ink stain from a silk carpet, use isopropyl alcohol with a white cloth to preserve the carpet’s color vs. use ethyl alcohol with a white cloth to preserve the carpet’s color” (originally written in Farsi). In *Traditional Practices (Customs and Traditions)*, a sample was: “During the Samanu cooking ceremony for Nowruz, stir wheat in a copper pot with a wooden spoon to release starch evenly vs. stir wheat in a copper pot with a metal spoon to release starch evenly” (originally written in Farsi). Other categories included *Physical Interactions with Nature*, *Everyday Objects (Tools and Materials)*, and *Folklore and Traditional Art*, each with corresponding examples.

Human-Centered Curation. Two native Farsi speakers then thoroughly reviewed the generated examples and facts. They either (i) created new samples from scratch inspired by the proposed categories, or (ii) edited the model-generated seeds to ensure grammaticality, naturalness, and cultural appropriateness in Farsi. During this process, annotators paid particular attention to aligning the samples with the intended reasoning challenges, avoiding translation artifacts, and ensuring diversity in phrasing.

Quality Control. Two annotators went through all the samples to guarantee consistency and adherence to the dataset’s design criteria. The final dataset consists of 100 samples, all of which are human-authored or human-edited, with the LLM serving only in the initial brainstorming phase.

This methodology allowed us to combine the efficiency of automated category suggestion with the cultural and linguistic reliability of human expertise, resulting in a dataset that is both systematic and faithful to the nuances of Farsi physical commonsense reasoning.

[*] Equal contribution.