Italian Physical Commonsense Reasoning Dataset for MRL 2025

Francesca Grasso **Rachele Mignone** **Ivan Spada**
Department of Computer Science, University of Turin
Corso Svizzera 185, 10149 Turin (Italy)
`{fr.grasso, rachele.mignone, ivan.spada}@unito.it`

1 Introduction

The present document provides a detailed description of a dataset created for the MRL 2025 Shared Task on Multilingual Physical Commonsense Reasoning. The dataset consists of **105 instances** in **Italian**, each following the format:

`[prompt, solution0, solution1, label]`

where exactly one solution is correct according to the physical properties of the matter involved in the given instance. All items have been manually constructed by three native speakers of Italian, without the use of Large Language Models (LLMs) or other automated generation tools.

The focus of the dataset is on *physical commonsense reasoning*, covering properties such as weight, durability, fragility, solubility, and temperature effects. In addition to general everyday scenarios, we included culturally specific examples relevant to Italian contexts, such as local foods (pasta, olive oil), products of artisanal excellence (Murano glass, Venetian brocade), domestic practices, and folklore-related beliefs. These latter include traditional or ancient beliefs about the properties of objects which have no official confirmation from the Western scientific community but are still conceptually linked to physical properties (e.g., alchemy, popular beliefs).

The variety of examples, the attention to similarity between correct and incorrect solutions, and the balance between short and long prompts as well as between different styles and syntactic variations, are intended to make the dataset both challenging for models and representative of natural commonsense reasoning in Italian.

2 Dataset Construction

2.1 Annotators

The dataset has been built by three native speakers of Italian, corresponding to the three authors of this document, from now on called “annotators”. All annotators are doctoral students, two with a computer science background and one with a linguistics background. They are aged between 25 and 32 years old, two are female and one is male.

2.2 Annotation Process

All examples were manually written. We used a shared spreadsheet to collect and eventually align and merge the instances. At the beginning, we followed the guidelines already provided by the organizers of this task. We first performed a pilot creation and annotation of 10 instances each (i.e., 30 in total), working independently without being influenced by the others to avoid bias and excessive homogeneity, after which a thorough group discussion was held.

Once the pilot annotation was done, we carried out a choral discussion with the following steps:
(i) joint reading of the 30 pilot items, discussing each example together and correcting those where we were not fully aligned;
(ii) rewriting of improper items;
(iii) discussion on how to refine the guidelines.

After this stage, each annotator proceeded with the creation and annotation of 25 more instances. The final construction of the dataset thus followed the official shared task guidelines, slightly adapted to our observations. None of us used Large Language Models or automated generation tools to create the instances.

2.3 Guideline Refinement

Prompt writing. Each example begins with a natural language prompt describing a physical scenario involving one or more objects. To avoid triviality, most prompts exceed 25 words, and some consist of multiple sentences. Shorter prompts were also included for variety.

Solution design. For each prompt, we created two candidate solutions that differ only slightly (e.g., by a few words or reordering phrases). One solution is unambiguously correct, while the other is physically implausible but still plausible enough not to appear absurd. This ensures that the task requires genuine commonsense reasoning.

Labeling. Each item is provided with a label “0” or “1”, corresponding to the index of the correct answer.

3 Dataset Description

3.1 Knowledge Types

We included a wide range of physical properties, including, for example:

- **Cooking and food transformations** (e.g., changes in texture or color when food is overheated or overcooked).
- **Weight and balance** (e.g., how objects of different masses behave when dropped).
- **Fragility and durability** (e.g., differences in how materials break or resist damage).
- **Solubility and reactions** (e.g., substances dissolving or reacting when mixed with liquids).
- **Temperature and heat transfer** (e.g., materials expanding, melting, or becoming brittle under temperature changes).
- **Electricity and electrostatics** (e.g., effects of static charge or electrical conductivity in everyday contexts).
- **Cultural beliefs and practices** (e.g., traditions or popular sayings that connect to physical processes).

3.2 Cultural Specificity

We tried to design most of the examples around Italian contexts, such as preparing pasta, handling olive oil bottles, or using moka coffee makers. This provides cultural grounding beyond generic physical reasoning. Many examples were also inspired by regional specificities (e.g., Murano glass, Sardinian *culurgiones*) and by personal world experience (personal world experience does not mean that it is not commonly shared).

3.3 Diversity of Examples

Besides mixing long (>25 words) and short items, we aimed to diversify the dataset in terms of personal world knowledge, syntax, style, and register, in order to make it as heterogeneous as possible. In terms of content, the topics spanned across different dimensions of variation (e.g., diatopic, diastratic, and stylistic).

3.4 Verification and Quality Control

At the end of the creation and annotation process, all examples were double-checked with a choral inspection for appropriateness, grammaticality, naturalness, correctness, and consistency with the guidelines. Incorrect solutions were revised to avoid being either too obvious or too ambiguous. We also removed duplicates (e.g., two annotators had independently included an example about mayonnaise curdling).

In particular, disagreements or corrections arose mostly for the following reasons:

- duplicate examples;
- sentences of commonsense not linked to physical properties of objects or matter;
- wrong solution being too obviously implausible;
- both solutions actually being correct.

4 Final Dataset Release

Once all 105 instances were ready, we merged the examples from the three annotators into a single file and performed a random shuffle. This ensured that the final dataset does not preserve any ordering bias and reflects a balanced mixture of contributions.

The resulting dataset is thus composed of **105 manually written and verified Italian examples**, ready for submission to the shared task.