MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets: Galician Dataset

Laura Castro Sánchez, Silvia Paniagua Suárez, Pablo Rodríguez Fernández, Pablo Gamallo, Marcos García
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
{laura.castro, silvia.paniagua.suarez, pablorodriguez.fernandez
pablo.gamallo, marcos.garcia.gonzalez}@usc.gal

1 Introduction

The goal of this work is to contribute a manually annotated Galician dataset for physical commonsense reasoning aimed at increasing the datasets for the MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets. We adopt the PIQA format (Bisk et al., 2020), releasing original, from-scratch entries authored and curated by native speakers. Our contribution combines general physical knowledge with Galician culture-specific contexts so that evaluation considers both general physical plausibility and culturally specific use cases.

2 Methodology

This section outlines the dataset format and details its construction, including prompt and solution design, topic and physical phenomena coverage, sources and composition.

2.1 Dataset format

Following the PIQA instance schema, each QA entry consists of a prompt (goal) that poses a practical goal or question about objects, materials, tools, actions, or their outcomes, followed by two candidate solutions (solution0 and solution1) that are almost identical except for one or two words. Only one solution is correct, indicated by a binary label (0 or 1), where 0 refers to solution0 and 1 refers to solution1.

2.2 Dataset design and composition

The dataset was created from scratch with entirely original examples authored and curated by native Galician speakers. Each entry was manually written and reviewed by at least two native speakers to ensure linguistic accuracy and logical coherence.

Entry example

goal: Por que non consigo deixar as claras de ovo en punto de neve?
(Why can't I whip the egg whites to stiff peaks?

solution0: O truco está en bater os ovos de forma constante e progresiva ata conseguir una consistencia firme.
(The trick is to whisk the eggs steadily and gradually until you achieve a firm consistency.)

solution1: O truco está en mesturar os ovos de forma constante e progresiva ata conseguir una consistencia firme.
(The trick is to mix the eggs steadily and gradually until you achieve a firm consistency.)

label: 0

topic: cooking

Prompt design Each QA entry includes a prompt (goal) plus two candidate solutions and a label indicating the correct option (0 or 1). The prompt is a short, practical query about how to perform an action or achieve a result. Wording is intentionally varied to avoid templated patterns; forms include *"How can X be done?"*, *"How to do X:"*, *"What should I use for. . . ?"*, *"What should one do if. . . ?"*, *"The best way to achieve X is:"*, *"To do X, it is essential that you. . . "*, and *"Remedies/Hacks for. . . "*. Prompts vary in length, from very brief formulations to longer ones of up to 17 words, and they provide no additional context. This variation in length and phrasing is meant to reduce superficial cues and encourage models to rely on physical reasoning.

Solution design The two solutions address the prompt directly and, like the prompts, vary in wording and length. To avoid overly simplified entries, we targeted at least 25 words per entry and included longer examples for variety, with some entries reaching approximately 170 words. Some solutions are a single long sentence; others comprise multiple independent or subordinate clauses. Within each entry, the two solutions are almost identical but differ in at most one or two words, either by substitution or deletion. These minimal edits are chosen to be meaningful and clearly interpretable for human readers, requiring non-trivial reasoning about objects, materials, tools, actions, and outcomes, while avoiding distractors that are obviously absurd. Only one solution is correct, and the label records the correct option.

Coverage The dataset represents physical commonsense in both general and Galician culture-specific contexts. For general knowledge, entries target a single, consolidated set of phenomena: (i) material properties of foods, substances, materials, and tools; (ii) interactions between objects; (iii) action constraints and implications, including the feasibility of movements, positions, or operations; and (iv) relational reasoning over spatial relations, quantities, and temporal/duration aspects. Within this scope, examples draw on everyday domains such as cooking, cleaning, gardening, climate conditions, health and exercise, transportation, and entertainment.

For culture-specific knowledge, entries are framed around contexts distinctive to Galician life, focusing on traditions and seasonal festivities (e.g., *San Xoán* bonfires), local customs and folklore, and traditional instruments (e.g., *zanfona*), including the role of specific tools within these practices. We use these contexts since they provide common, well-known scenarios and terms for native speakers, making entries clear in Galician and less likely to map one-to-one into other languages.

Sources For general-knowledge entries, we drew inspiration from tutorial-style websites, news outlets, official public-safety pages, and sites with everyday tips and hacks. For culture-specific entries, we consulted the Galician edition of Wikipedia and local websites covering traditions, customs, recipes, and traditional instruments. Due to Galician being a low-resource language, suitable written references are sometimes scarce; in those cases, culture-specific scenarios were abstracted from the annotators' knowledge of Galician culture. Source material was used only to inspire prompts and plausible alternatives, not to translate or copy existing text.

Composition The dataset contains 109 entries, with 57 general-knowledge and 52 Galician culture-specific entries. Labels are balanced across the two options (54 entries with label 0 and 55 with label 1).

To facilitate analysis and clarify the types of commonsense situations addressed, the entries were grouped into broad topical categories, as shown in Table 1. This table also illustrates the topical coverage, highlighting both Galician-specific entries and general-knowledge items related to traditions, cooking, or games.

Balance remarks Labels are essentially even (54 with 0 vs. 55 with 1), which helps avoid position bias toward solution0 or solution1. Topic coverage concentrates in a few areas, namely objects, cooking, materials, tools, and traditions. This concentration is expected: these common topics provide many everyday situations, allow small but meaning-changing edits for minimal pairs, are well covered by accessible sources, and are easy for native speakers to judge. Less frequent topics (e.g., myths, cleaning, animals, exercise) broaden coverage, but results by these topics should be interpreted with care or aggregated into broader groups. Galician-specific entries concentrate in traditions and cooking because these domains are widely shared and rich in physical reasoning: festivities and customary practices provide common scenarios, tools, and rules that most speakers know, and cooking engages materials, heat, timing, and safety.

2.3 Representative examples

The following cases illustrate the kinds of physical knowledge described above. In general-knowledge entries, questions about cleaning weigh the behavior of different agents: when deciding how to clean a window, one must distinguish detergents, which break greasy films and leave glass clear, from oils, which may lift some residue but tend to smear and are not ideal for a smooth finish. The subtleties of everyday actions appear in kitchen and home settings: to obtain stiff peaks, the right choice is whipping rather than simple mixing, since whipping incorporates air and sets structure; to remove water rings on wooden furniture, sanding mechanically lifts the stained surface, whereas wiping does not resolve the mark. Comparable reasoning is required for materials and their properties, as in the entry where resin and glue are contrasted to decide which is appropriate for tuning a hurdy-gurdy:

resin increases friction at the relevant interface, while glue would inappropriately bond surfaces. Purpose-driven action logic is tested when lighting a fire in a wood stove, where keeping the airflow open supplies oxygen and aids ignition, while closing it impedes combustion. In cooking, choosing between sugar and sweeteners for a low calorie apple pudding is mainly about reducing calories, since sugar adds many whereas sweeteners do not, while still achieving the desired sweetness and maintaining the pudding's texture and setting. Some entries call for basic kinematics, such as identifying that a somersault entails a full 360-degree rotation rather than 180. Safety judgments about objects are captured in travel scenarios, where a tire's readiness is indicated by tread pattern and depth rather than color. In Galician culture-specific entries, the same forms of reasoning are anchored in familiar Galician contexts: choosing the appropriate tool for a local horse-shearing and deworming event calls for scissors rather than a sickle; moving safely when jumping San Xoán bonfires (a traditional festivity in Galician culture) depends on simple movement logic about posture and take-off; and during the New Year's twelve-grapes tradition, selecting grapes without peel and seeds reduces choking risk under time pressure.

In short, general-knowledge entries foreground universal physical properties and processes in everyday tasks, whereas Galician culture-specific entries place the same kinds of decisions in shared Galician settings, using locally familiar tools, actions, and customs to make the reasoning clear and meaningful for native speakers.

Topic	**Total**	**General**	**Galician-specific**
Objects	17	11	6
Cooking	13	5	8
Materials	11	7	4
Tools	11	7	4
Traditions	11	0	11
Spatial relations	8	3	5
Health	7	4	3
Time/Duration	7	5	1
Climate/Weather	4	4	0
Games	3	0	3
Travel	3	1	2
DIY/Handmade	2	0	2
Transportation	2	2	0
Gardening	2	2	0
Actions	2	2	0
Music	2	0	2
Myths	1	0	1
Cleaning	1	2	0
Animals	1	1	0
Exercise	1	1	0
Total	109	57	52

Table 1: Distribution of topical categories by knowledge type.

References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.