Culturally Grounded Physical Commonsense Reasoning in Italian and English: A Submission to the MRL 2025 Shared Task

Marco De Santis
University of Udine
desantis.marco001@spes.uniud.it

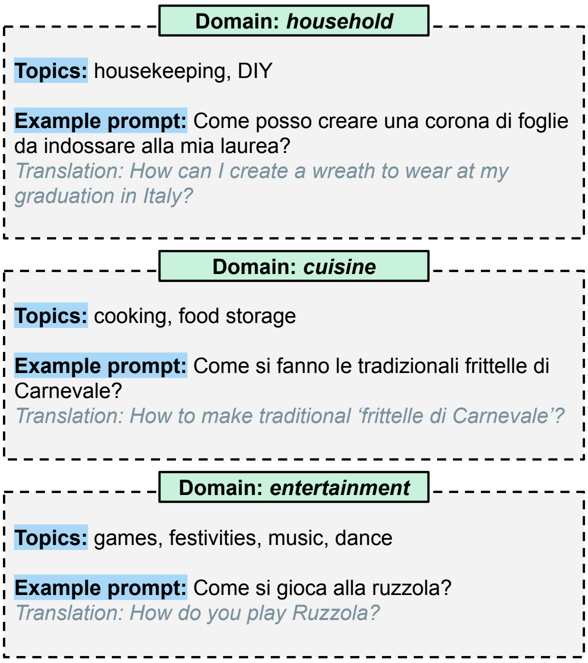
Lisa Alazraki
Imperial College London
lisa.alazraki20@imperial.ac.uk

Abstract

This paper presents our submission to the MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets. The objective of the shared task is to create manually-annotated evaluation data in the physical commonsense reasoning domain, for languages other than English, following a format similar to PIQA (Bisk et al., 2020). Our contribution, FormaMentis, is a novel benchmark for physical commonsense reasoning that is grounded in Italian language and culture. The data samples in FormaMentis are created by expert annotators who are native Italian speakers and are familiar with local customs and norms. The samples are additionally translated into English, while preserving the cultural elements unique to the Italian context.

1 Introduction

Commonsense reasoning in language models is an active area of research (Kavumba et al., 2019; Talmor et al., 2021; Gupta et al., 2023; Li et al., 2025), with recent work extending beyond English to other languages (Ghosh et al., 2025). Existing multilingual datasets, however, have primarily targeted causal reasoning (Ponti et al., 2020), sentence completion, and general question answering (Lin et al., 2021), leaving other aspects of commonsense underexplored. In this work, we highlight physical commonsense as a particularly important yet overlooked dimension for multilingual research. Unlike causal or textual reasoning, physical commonsense requires reasoning over objects and actions that are often grounded in culture-specific practices, and may even lack direct English equivalents (Anacleto et al., 2006; Zou et al., 2009).



Domain: ***household***

Topics: housekeeping, DIY

Example prompt: Come posso creare una corona di foglie da indossare alla mia laurea?
Translation: How can I create a wreath to wear at my graduation in Italy?

Domain: ***cuisine***

Topics: cooking, food storage

Example prompt: Come si fanno le tradizionali frittelle di Carnevale?
Translation: How to make traditional 'frittelle di Carnevale'?

Domain: ***entertainment***

Topics: games, festivities, music, dance

Example prompt: Come si gioca alla ruzzola?
Translation: How do you play Ruzzola?

Figure 1: FormaMentis domains. For each, we show an example prompt requiring cultural knowledge. E.g., a graduation wreath is a specific type of wreath worn by Italian university students, *frittelle di Carnevale* is a local holiday recipe, and *Ruzzola* is a traditional Italian country sport. Completing these prompts correctly thus requires an understanding of the specific characteristics and rules of each item or practice. It is also worth noting that the English translations aim to preserve these culture-specific characteristics, which involves leaving words in their original Italian form where necessary.

Building on the above observations, we introduce FormaMentis, a new Italian-language benchmark manually constructed by expert annotators who are native speakers. FormaMentis focuses on everyday scenarios involving objects and actions rooted in Italian customs. The samples in the dataset are distributed among three domains (*household*, *cuisine*, and *entertainment*) that require reasoning about physical items and practices presupposing cultural knowledge. To broaden accessibility and facilitate cross-lingual evaluation, all samples are also manually translated into English, with careful attention to preserving the Italian-specific cultural references embedded in the original.

2 The FormaMentis Benchmark

2.1 Data Format

The FormaMentis benchmark adopts a format similar to the PIQA dataset (Bisk et al., 2020): each sample consists of a prompt paired with two candidate completions, only one of which is correct. The completions are closely matched, typically differing by just one or two words, with the incorrect choice designed to be clearly wrong but not so implausible as to make the task trivial.

2.2 Data Collection

The data samples in FormaMentis are manually created by expert annotators who are native Italian speakers, with detailed guidelines serving as a reference (see Appendix A). Every sample must be novel (i.e., not translated from other sources) and must belong to one of three domains reflecting local customs: *household*, *cuisine*, and *entertainment*. Figure 1 presents an overview of these domains, their associated topics, and representative prompts.

Annotators are instructed to create samples that demand physical commonsense reasoning and incorporate cultural references rooted in Italian everyday culture. These include activities tied to local traditions as well as linguistic expressions that do not translate directly into other languages.

2.3 Data Validation

Each data sample in FormaMentis undergoes evaluation through a multi-step validation questionnaire, carried out by a native speaker other than the sample author. The validation steps are detailed in Table 1. For step #5, minor adjustments to determiners, verbs, adjectives, or pronouns are not counted as substantive changes when they arise solely as a consequence of substituting other words. This exception reflects the rules of Italian grammar, which requires agreement in gender (feminine/masculine) and number (singular/plural).

If any validation step receives a negative outcome, the sample must be revised or rewritten, after which the full assessment is repeated. Only those samples that successfully pass every step of the questionnaire are included in the final benchmark.

2.4 Data Statistics

Number of samples. FormaMentis contains 120 test samples, a size consistent with other high-quality, human-written reasoning benchmarks (Eleftheriadis et al., 2023; Press et al., 2023; Santos et al., 2024; Alazraki et al., 2025). The samples are evenly distributed across the three domains – household, cuisine, and entertainment – with 40 samples per domain.

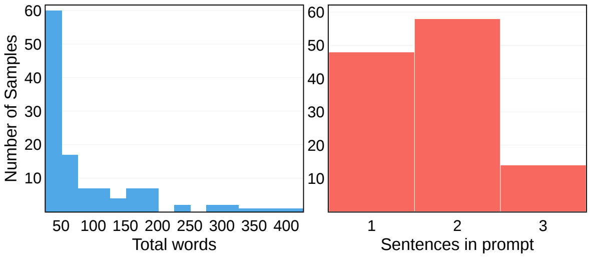


Figure 2: Sample distribution by total number of words (left) and number of sentences in the prompt (right).

Step	Validation Question
#1	Does the prompt require physical reasoning and knowledge beyond what is stated in it? [Yes/No]
#2	Is the physical knowledge required common among native Italian speakers? [Yes/No]
#3	Do the prompt and/or completions contain cultural references, linguistic expressions or colloquialisms that are specific to Italian culture? [Yes/No]
#4	Is the prompt unambiguous, and is only one of the two candidate completions correct? [Yes/No]
#5	Are the two candidate completions mostly similar, differing only by one or two words? [Yes/No]
#6	Is the incorrect completion plausible enough not to appear absurd? [Yes/No]

Table 1: Validation steps for all samples in FormaMentis. Only samples that obtain a positive answer at all steps are included in the benchmark.

Sample length. The total sample length (Italian-language prompt + solution) ranges from 26 to 452 words, with half of the samples falling in the 26–50 bin, as shown in Figure 2. Sixty percent of the samples feature multi-sentence prompts (58 contain two sentences and 14 contain three), while the rest (40%) are single-sentence prompts. Additional data statistics are provided in Appendix B.

3 Conclusion

We introduced FormaMentis, a manually curated benchmark for physical commonsense reasoning in Italian, focused on everyday scenarios reflecting local practices. Our benchmark is carefully validated for quality and includes English translations that preserve contextual nuances. By providing this resource, we aim to support research on multilingual physical reasoning and facilitate the evaluation of language models on culturally grounded commonsense tasks beyond English.

Limitations

FORMAMENTIS is created for the text-based evaluation of culturally grounded physical commonsense reasoning in language models. As such, the benchmark does not contain any images or videos, though we acknowledge that such multimodal inputs can support physical reasoning. We leave the study of multimodal physical reasoning that integrates culture-specific knowledge to future work.

Ethical Considerations

All samples in FORMAMENTIS are manually written by human experts, which guarantees originality and independence from third-party licensed sources. We have carefully reviewed the dataset to avoid any offensive or harmful content, and all samples focus on neutral, everyday scenarios.

References

Lisa Alazraki, Lihu Chen, Ana Brassard, Joe Stacey, Hossein A. Rahmani, and Marek Rei. 2025. Agent-CoMa: A compositional benchmark mixing commonsense and mathematical reasoning in real-world scenarios. *Preprint*, arXiv:2508.19988.

Junia Anacleto, Henry Lieberman, Aparecido de Carvalho, Vania Neris, Muriel Godoi, Marie Tsutsumi, Jose Espinosa, Américo Talarico, and Sílvia Zem-Mascarenhas. 2006. Using common sense to recognize cultural differences. In *Advances in Artificial Intelligence*, pages 370–379.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Petros Eleftheriadis, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2023. Evaluating deep learning techniques for natural language inference. *Applied Sciences*, 13(4).

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind : A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Kumar Pal, Saurabh Arjun Sawant, Kevin Scaria, Siddharth Goyal, and Chitta Baral. 2023. "John is 50 years old, can his son be 65?" Evaluating NLP models' understanding of feasibility. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 407–417, Dubrovnik, Croatia. Association for Computational Linguistics.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever Hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Xiaoyuan Li, Moxin Li, Rui Men, Yichang Zhang, Keqin Bao, Wenjie Wang, Fuli Feng, Dayiheng Liu, and Junyang Lin. 2025. HellaSwag-pro: A large-scale bilingual benchmark for evaluating the robustness of LLMs in commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9038–9072, Vienna, Austria. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Henrique Santos, Ke Shen, Alice M. Mulvehill, Mayank Kejriwal, and Deborah L. McGuinness. 2024. A theoretically grounded question answering data set for evaluating machine common sense. *Data Intelligence*, 6(1):1–28.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Xi Zou, Kim-Pong Tam, Michael W. Morris, Sau-Lai Lee, Ivy Yee-Man Lau, and Chi-Yue Chiu. 2009. Culture as common sense: perceived consensus versus personal beliefs as mechanisms of cultural influence. *Journal of personality and social psychology*, pages 579–597.

A Annotator Guidelines

The expert annotators follow the guidelines below when creating the samples in FORMAMENTIS. All samples are written in Italian and are subsequently manually translated into English, with care taken to preserve the cultural cues and linguistic expressions of the original.

FormaMentis: A Culturally Grounded Physical Commonsense Reasoning Benchmark in Italian.

Instructions (see also the original Shared Task instruction slides).

Note: Samples must be manually created. The use of LLMs to write the samples is *not* allowed. You may refer to the PIQA dataset for the format of questions/answers required, but please *do not* translate the samples from PIQA (or elsewhere).

The prompt+completions samples need to be about:
- Physical reasoning: For each example, the solution must relate to physical properties of one or more objects. You may include common physical tasks or actions.
- Common sense: For each example, an average person who speaks your language natively should know the answer.

They also need to be culturally specific:
For example, some items may not be easily translatable into English, or may require regional and/or cultural commonsense.

Other requirements:

Use items of variable length. Try not to include too many short items, as they may be too easy for larger models. If possible:

- Examples (prompt+solution) should be over 25 words long.

- There should be some prompts that are multiple sentences long.

- The two candidate solutions should be as similar as possible (e.g. differing only by one or two words, or just flipping the order of two phrases). One solution should be unambiguously correct and the other incorrect.

- To ensure that the benchmark is not too "easy", the incorrect solution should not be so absurd that it is extremely obvious.

- Try not to start all examples the same way.

Please create an equal number of samples in each of the three categories:

- Household (includes housekeeping and DIY)
- Cuisine (includes cooking and food storage)
- Entertainment (includes games, festivities, music and dance)

B Fine-grained Data Statistics

B.1 Length of Prompts

The distribution of prompt lengths in FORMAMENTIS is shown in Figure 3. Lengths range from 4 to 61 words, with most prompts containing fewer than 25. Word counts are obtained by splitting text sequences at whitespace. In Italian, words may be abbreviated and joined with an apostrophe. Although such pairs are grammatically distinct words, we treat them as a single unit in our analysis.

B.2 Length of Completions

Figure 4 presents the distribution of completion lengths in FORMAMENTIS. These cover a wide range from 2 to 223 words, with approximately half (62 completions) containing fewer than 15 words. Because the two completion options are always identical or nearly identical in length, we measure the aggregate length of both options for each sample and divide this value by two to obtain the length of an individual option.

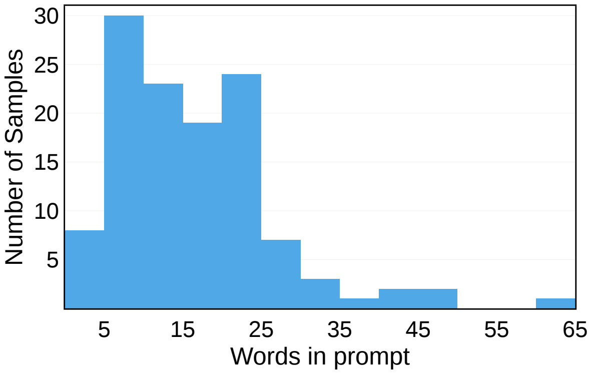


Figure 3: Sample distribution in FORMAMENTIS by number of words in the prompt.

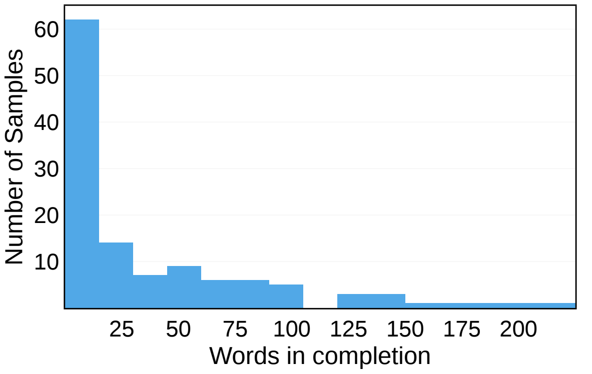


Figure 4: Sample distribution in FORMAMENTIS by number of words in a completion. We measure the completion length as the average number of words between both completions in a sample.