A Multi-layered Approach to Physical Commonsense Understanding: Creation and Evaluation of an Italian Dataset

***Giulia Pensa, **Begoña Altuna, **Itziar Gonzalez-Dios**
*University of the Basque Country, **HiTZ Center - Ixa, University of the Basque Country UPV/EHU
Manuel Lardizabal, 1, 20018 Donostia-San Sebastian, Gipuzkoa
giulia.pensa.tr@gmail.com,
{begona.altuna, itziar.gonzalezd}@ehu.eus

Abstract

In this paper, we explore physical commonsense reasoning of large language models (LLMs) and propose a specific methodology to evaluate low-level understanding of the physical world. Specifically, the goal is to create a test set to analyze physical commonsense reasoning in large language models for Italian and focus on a trustworthy analysis of the results. To that end, we present a tiered Italian dataset, called Graded Italian Annotated dataset (GITA), written and thoroughly annotated by a professional linguist, which allows us to concentrate on three different levels of commonsense understanding. Moreover, we create a semi-automated system to complete the accurate annotation of the dataset. We also validate our dataset by carrying out three tasks with a multilingual model (XLM-RoBERTa) and propose a qualitative analysis of the results. We found out that, although the model may perform at high-level classification tasks, its reasoning is inconsistent and unverifiable, since it does not capture intermediate evidence.

Keywords: Physical commonsense reasoning, large language models, multilingual model

1. Introduction

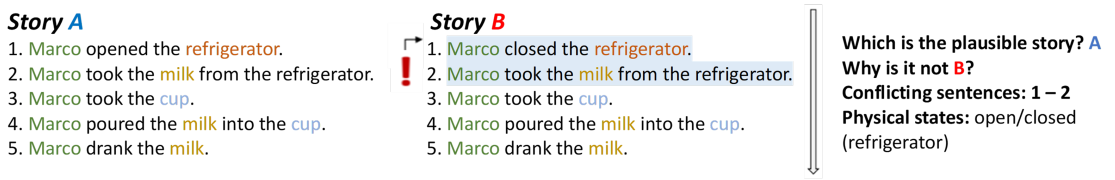
Physical commonsense understanding is the ability to make sense of the physical world and the events that occur in it. It is a fundamental aspect of human intelligence, allowing us to reason about the world, predict future events, and navigate our surroundings with ease. In recent years, there has been significant progress in developing large language models (LLMs) that can generate human-like language and perform a wide range of language-related tasks. LLMs have exhibited promising outcomes in grasping common sense in particular situations (Huang and Chang, 2023; Sakaguchi et al., 2021). Nevertheless, it is widely recognized that the most precise evaluation of their capabilities is attained when assessing their performance in specific end tasks (Pessach and Shmueli, 2022; Davis, 2023). The evaluation often emphasizes the capacity of LLMs to replicate relatively straightforward tasks, rather than their authentic proficiency in reasoning and comprehending language (Linzen, 2020; Bender and Koller, 2020). As a result, there remains uncertainty regarding machines' ability to truly perform reasoning and whether the existing issues in this regard have been sufficiently addressed.

In this context, our aim is to develop an original dataset suitable for an evaluation benchmark that can be used to assess the ability of language models to understand physical commonsense in a more truthful way, focusing not only on end tasks, but also on intermediate layer tasks. Moreover, with the creation of an Italian dataset we gain the linguistic and cultural perspective of Italian, while commonsense research in Natural Language Processing (NLP) has largely been focused on the English language. In this paper, we present GITA, the Graded Italian Annotated dataset, manually compiled by a professional linguist, which allows for a multi-layered evaluation of the reasoning process. The aim of our benchmark is to evaluate three different tasks: 1) the end task consists of identifying the plausible and implausible stories in our dataset, 2) the second-level task is to identify the conflict that generates an implausible story, and 3) the deepest one consists of identifying all the physical states that make a story plausible or implausible. Moreover, to prove the validity of our dataset, we experiment with a multilingual language model, and we perform an analysis of the experimental results.

The main contributions of this paper are the development of an Italian dataset, built by a professional linguist, which can be used to assess similar tasks in LLMs, and the creation of a semi-automated environment to deeply annotate a dataset. To the best of our knowledge this is the first Italian dataset of this kind.

2. Related Work

Commonsense reasoning involves the ability to make accurate predictions, infer missing information, understand cause-and-effect relationships, and draw logical conclusions from incomplete or ambiguous data. In recent years, there has been growing interest in developing LLMs that can



Story A

1. Marco opened the refrigerator.
2. Marco took the milk from the refrigerator.
3. Marco took the cup.
4. Marco poured the milk into the cup.
5. Marco drank the milk.

Story B

1. Marco closed the refrigerator.
2. Marco took the milk from the refrigerator.
3. Marco took the cup.
4. Marco poured the milk into the cup.
5. Marco drank the milk.

Which is the plausible story? A
Why is it not B?
Conflicting sentences: 1 – 2
Physical states: open/closed (refrigerator)

Figure 1: Representation of story pair from GITA

mimic the ability of drawing conclusions based on our understanding of the world (Du et al., 2023). Yet, despite their capabilities, LLMs still struggle with commonsense reasoning tasks. This is because much of our knowledge about the world is implicit and difficult to formalize (Marcus and Davis, 2020). To address this problem, evaluation benchmarks that can be used to test LLMs' performance on commonsense reasoning tasks (Storks et al., 2019) and that can identify areas for improvement, as well as new techniques, have been created.

Physical commonsense is the understanding of how the physical world works, including object properties, spatial relations, and cause and effect relationships. A well-known example of physical commonsense is the concept of gravity - we know that if we drop an object, it will fall to the ground.

Regarding the evaluation of its understanding, Bisk et al. (2020b) developed a dataset for assessing progress in physical commonsense understanding. The main task involved answering multiple-choice questions, where a question *q* was presented along with two possible solutions, *s1* and *s2*, and the model or human evaluator had to choose the most suitable option, with only one correct answer available. Though humans found the dataset easy (95% accuracy), large pretrained models struggled but scored consistently over 75%.

Singh et al. (2021) introduced a new commonsense reasoning benchmark dataset comprising natural language true/false statements, with the aim of reliably measuring an agent's ability to perform commonsense reasoning over a given situation. After fine-tuning, the model achieved 71% accuracy, while human performance reached 95%.

Storks et al. (2021) introduced a novel commonsense reasoning dataset called Tiered Reasoning for Intuitive Physics (TRIP) that enables a multitiered evaluation of machines' reasoning process. They showed that this type of benchmarks can motivate a verifiable evaluation of commonsense reasoning and facilitate future research in this field. The TRIP dataset comprises human-written stories that depict sequences of physical actions. To determine which of two stories, composed of individually plausible sentences and differing by only one sentence, is more credible, the task requires knowledge of verb causality, precondition, and intuitive physics. The TRIP training set is composed by 370 plausible stories, and 799 implausible stories; the validation set includes 152 plausible stories and 322 implausible ones; and the test sets comprises 153 plausible stories and 351 implausible ones. A characteristic of this dataset is the presence of a rich physical states annotation.

Ponti et al. (2020) introduced a new multilingual dataset for causal commonsense reasoning. This dataset covers 11 different languages and allows for the evaluation of commonsense reasoning across multiple languages. They showed that models trained on multiple languages can achieve strong performance on commonsense reasoning end tasks, demonstrating the importance of multilingual data for improving models' generalization.

Regarding the multilingual approaches to commonsense reasoning, Lin et al. (2021) evaluated multilingual language models for commonsense reasoning in multiple languages and they showed that these models are effective at capturing commonsense knowledge across different languages and that multilingual pre-training can improve performance on tasks in low-resource languages. Finally, Fang et al. (2022) proposed a multilingual approach to commonsense reasoning that leverages knowledge from related languages to improve performance on low-resource languages. They demonstrated that this approach can significantly improve performance on multilingual commonsense reasoning end tasks and highlights the importance of considering related languages in multilingual NLP research. The previous works show that multilingual data, pretraining, and knowledge transfer can be effective strategies for improving models' generalization and end task performance across multiple languages.

3. Building GITA

In order to build the Graded Italian Annotated dataset (GITA), we base our experimentation on Storks et al.'s (2021) work. Our main objective is to create an Italian dataset, manually annotated, to assess a pre-trained language model on physical commonsense tiered tasks. We configure our assessment proposal in the following terms:

1. given an original dataset of plausible/implausible stories related to physical commonsense, systems must identify the plausible and implausible stories;
2. systems must recognize the conflicting sentences that generate the conflict in implausible stories;
3. systems must spot the underlying physical states in those sentences causing the conflict.

The recognition of plausible/implausible stories is the end task envisaged in this benchmark, which must be justified by the second-level and third-level steps. In Figure 1 we present a story pair from the GITA dataset and the relation between the layers of annotation. Story A is a plausible story, Story B is the corresponding implausible story where the first and the second sentences are in conflict: Marco closes the refrigerator and cannot take the milk out of it. In the right part of the figure we can see the reasoning steps that the system must follow and resolve. This example is presented in English for clarity, but our entire dataset is in Italian.

3.1. Dataset creation

We created a dataset of 356 stories in Italian which compiles 117 plausible and 239 implausible stories. To compose the dataset, we focused on concrete actions that could be visualized in the physical world, avoiding mental actions such as "to think" or "to like". Several benchmarks focusing on plausible reasoning offer limited context, often consisting of just one sentence, accompanied by similarly brief options to complete the context (Gordon et al., 2012; Zellers et al., 2018; Bisk et al., 2020a). Instead, we created 5-sentence stories, giving more context and requiring reasoning over multiple sentences with associated physical state changes. In all the stories, we avoided nonsensical sentences, in fact, each sentence is plausible alone, but could be implausible if associated with another specific sentence in an implausible story. With these characteristics, the task requires reasoning over the entire context. To create the stories, we took inspiration from the Story Cloze Test (Mostafazadeh et al., 2017) and ROCStories Corpora (Mostafazadeh et al., 2016). The Story Cloze Test is composed by four-sentence stories with a missing ending, requiring a system to choose the most appropriate conclusion. The ROCStories dataset has two specific characteristics: 1) it includes a diverse range of causal and temporal commonsense relations between daily events, and 2) it offers a high-quality collection of everyday life stories that can be utilized for story generation.

An essential part of our evaluation process is constituted by the presence of physical state annotation. Systems must identify the underlying physical states that make a story not plausible in our physical world. During the creation of the dataset, we took into account 20 physical attributes that were included in the annotation phase, and we composed stories that contained those attributes. Following the work of Gao et al. (2016) and Bosselut et al. (2017), these are the 20 physical states that we wanted to have in our stories:

- for human actors: location, hygiene, conscious, dressed, wet;
- for objects: location, exist, clean, power, functional, in pieces, wet, open, temperature, solid, occupied, running, movable, mixed, edible.

In the first two rows of Table 1 we can see an example of plausible story from the GITA dataset together with the English translation. In this example, the human actor is Marco, and the five sentences are ordered in the required way: the action of opening something, picking something up and using it. We can see that some of the previously listed physical states appear: Marco is *conscious* because he is doing something, the refrigerator is *open* because the actor can take something out of it, the cup is not *occupied* by anything and can be *functional*.

We tried to avoid subjectivity and restrict potential confounding factors arising from complex language usage. The use of simple language enabled us to shift our focus away from linguistic processing and semantic phenomena, directing more attention towards investigating machines' reasoning abilities, specifically their physical commonsense understanding. Therefore, we constructed our simple sentences in a straightforward declarative structure. This usually involves beginning with the agent of the story, followed by a verb, a direct object, and an optional indirect object.

Implausible stories are built upon the plausible ones, preserving the same actor and objects; in doing so we ensured that implausible variations remained coherent and believable, and we avoided nonsensical information. To create implausible stories, we implemented two different methods:

1. we switched the order of two sentences;
2. we substituted a plausible sentence with an implausible one.

These two methods resulted in two different partitions of our dataset: the *Order dataset* of implausible stories, and the *Cloze dataset* of implausible

	sentence 1	sentence 2	sentence 3	sentence 4	sentence 5
T	Marco ha aperto il frigo.	Marco ha preso il latte dal frigo.	Marco ha preso la tazza.	Marco ha versato il latte nella tazza.	Marco ha bevuto il latte.
	Marco opened the refrigerator.	Marco took the milk from the refrigerator.	Marco took the cup.	Marco poured the milk into the cup.	Marco drank the milk.
F (order)	Marco ha preso il latte dal frigo.	Marco ha aperto il frigo.	Marco ha preso la tazza.	Marco ha versato il latte nella tazza.	Marco ha bevuto il latte.
	Marco took the milk from the refrigerator.	Marco opened the refrigerator.	Marco took the cup.	Marco poured the milk into the cup.	Marco drank the milk.
F (cloze)	Marco ha chiuso il frigo.	Marco ha preso il latte dal frigo.	Marco ha preso la tazza.	Marco ha versato il latte nella tazza.	Marco ha bevuto il latte.
	Marco closed the refrigerator.	Marco took the milk from the refrigerator.	Marco took the cup.	Marco poured the milk into the cup.	Marco drank the milk.

Table 1: Example of a plausible story, an implausible story from the Order dataset, and an implausible story from the Cloze dataset.

stories respectively. In the Order dataset there are 122 implausible stories, while in the Cloze dataset there are 117 implausible stories.

3.1.1. Order implausible stories

The plausible stories only work in the causal sequence that we created. In Table 1, the plausible story sees Marco as the protagonist, and physical states such as *open* and *occupied* as critical for the unfolding of the events.

In the first row of Table 1, there is an example of a plausible story. In the third row, we see the corresponding implausible story for the order dataset, in which Marco, first, takes the milk out from the refrigerator and then open the refrigerator, generating a physically impossible situation: it is not possible to take something out of a closed refrigerator. By switching the first and the second sentences, we created an implausible story. In the entire dataset, we decided to generate implausible stories changing the order of only two sentences for story.

3.1.2. Cloze implausible stories

The second approach involves the substitution of a sentence from the plausible story with a new sentence. Although the new sentence itself is not inherently implausible, its placement within the sequence renders it implausible.

In Table 1, the first sentence of the line F (Cloze), in the fifth row, was changed: Marco closes the refrigerator before taking out the milk. Again, the action is physically impossible: if the refrigerator is closed, nothing can be taken out from it.

3.2. Dataset annotation

GITA is annotated on three levels. In the first level, we annotated the plausibility/implausibility of a story with TRUE or FALSE. In the second level, in implausible stories we indicated between which sentences the conflict was, and in the third level we labelled the involved physical states in each sentence.

In the dataset, a plausible story is identified using a story number, while implausible stories are identified using the same story number as the plausible version, but with an additional **C** or **O** after the story number, where the letter C refers to the Cloze dataset, and the letter O refers to the Order dataset. Each story has been annotated using these elements: story id, worker id, actor of the story, objects of the story, physical states, sentences of the story, as well as number of sentences, and conflicting sentences, among others. The complete list and the specific meaning of each element are in Appendix A, while an example of a complete annotation can be found in Appendix B. The annotation was conducted on a JSON file, and the complete JSON file is available in our repository under the license CC BY-NC-SA 4.0.[1]

The annotation of physical states is organized by sentence: for each sentence in each story we see the 20 states. The physical states were annotated following Table 2, based on the attribute space features framework proposed by Gao et al. (2016) and Bosselut et al. (2017). Each physical state can be annotated using a number from 0 to 8, ex-

[1] https://github.com/GiuliaAPensa/GITAdataset

cept from human physical states that only present three dimensions. The dimension of the *location* state is adapted to the features of location itself; at the same time, the *human location* state presents three specific dimensions. These two states are unique in that each number in our classification corresponds to a specific physical change. For instance, in the case of *location* number 7 signifies ”taken out of a container”. Conversely, for all other attributes, except *human location*, it indicates that the attribute was false in the previous step, with uncertainty about its status in the current one. The other physical states are annotated with more generic labels.

Label	Human Location	Object Location	Other Attributes
0	irrelevant	irrelevant	irrelevant
1	disappeared	disappeared	*false* → *false*
2	moved	picked up	*true* → *true*
3	-	put down	*true* → *false*
4	-	put on	*false* → *true*
5	-	removed	____ → *no*
6	-	put in container	____ → *true*
7	-	taken our of container	*false* → ____
8	-	moved	*true* → ____

Table 2: Label space and meanings for physical states.

After concluding the annotation of all physical states in our stories, we applied two different classifiers (like the ones that Storks et al. (2021) presented in their work), a precondition classifier and an effect classifier. The precondition classifier extracted the label referred to the precondition from each physical state and the effect classifier extracted the label of the effect state. To better understand this feature, this is the first sentence of the implausible story (Cloze) in Table 1:

- Marco ha chiuso il frigo. (Marco closed the refrigerator.)

In this sentence, the entity *frigo* (refrigerator), was annotated for the *open* physical state as *true* → *false* with the number 3 of our Table 2. Each classifier classified this number with a label for the precondition and another for the effect. In this way, from that specific state label 3 (*true* → *false* or, in our example, open refrigerator → closed refrigerator) the classifiers extracted *open*: 2 (*true* → *true*) for the precondition and 1 (*false* → *false*) for the effect, meaning that the refrigerator was first open and then closed. During the analysis of our results (see Section 4.4), we take into consideration the precondition-effect labels when comparing the actual labels assigned to each physical state with the predictions of the model.

To ensure consistency and reduce human effort, we developed a custom environment and a Python script to streamline the annotation process. This semi-automated annotation process helped us process sentences from different story types, extract entities and actors, and organize them for manual annotation. The script provided a user-friendly terminal interface, and it is available in our repository. In terms of annotation efficiency, manually annotating one plausible story and two implausible ones typically took around 50 minutes. However, using our semi-automated annotation interface, we were able to complete the same task in approximately 20 minutes. Consequently, instead of the estimated 100 hours for annotating the entire dataset, we reduced the time to around 40 hours. Additionally, some annotations required review and occasional revisions, hence we estimated that the overall effort was of approximately 50-55 hours.

Table 3 lists the statistics of the resulting dataset: number of plausible/implausible stories and number of physical state labels.

Measure	GITA test set
# plausible stories	117
# implausible stories (ORDER)	122
# implausible stories (CLOZE)	117
# physical state labels (ORDER)	2,911
# physical state labels (CLOZE)	2,819

Table 3: Statistics of the GITA dataset.

4. Validation Experiments

In this section we present the experiments that serve to validate the effectiveness and reliability of the proposed dataset.

4.1. Tasks

Based on the GITA dataset and following the steps of Storks et al. (2021), we propose a series of tasks that form a human-interpretable reasoning process, supported by a chain of evidence.

1. **Physical state classification:** Leveraging our physical state annotations, we propose two subtasks for each entity within every story choice: precondition and effect state classification. For instance, if we consider the sentence “John cut the cooked potato in half” with the entity “potato”, the first subtask involves predicting that the potato is solid as a precondition for being cut (e.g., the precondition label for the *solidity* attribute is true). The second subtask entails predicting that the potato is in pieces as an effect resulting from the cutting action (e.g., the effect label for the *in pieces* attribute is true).

2. **Conflict detection:** Next, the task of conflict detection entails identifying sentence pairs of the form S_i → S_j. Here, S_j represents the breakpoint, indicating the point at which

the story becomes implausible based on the given context. Si serves as the evidence that explains the breakpoint, typically causing a conflicting world state.

3. **Story classification:** The end task revolves around determining the plausibility of two stories. This determination is based on the conflicts detected within the two stories. By considering the presence of conflicts, the model can assess the viability and coherence of each story, facilitating the classification of the more plausible one.

By incorporating physical state classification, conflict detection, and story classification, we analyze the aspects of coherent reasoning, supported by evidence-driven analysis.

4.2. Model and experimental set-up

We trained our model on the English training set provided by Storks et al. (2021) and tested it on our Italian dataset. For this reason, we ran our experiments using XLM-RoBERTa (Conneau et al., 2020), a multilingual LLM based on RoBERTa's architecture, which is a variant of the transformer neural network architecture, and it is designed to handle multilingual data.
Following the steps of Storks et al. (2021), the architecture's parameters were trained through gradient descent on the overall loss L:

$$L = \lambda_p L_p + \lambda_f L_f + \lambda_c L_c + \lambda_s L_s \quad (1)$$

L sums individual cross-entropy loss functions Lp for precondition classification, Lf for effect classification, Lc for conflict detection, and Ls for story choice classification, each balanced by respective weights λp, λf, λc, λs summing to 1.

4.3. Evaluation metrics

In order to evaluate the model, we employ the following evaluation metrics:

- **Accuracy** assesses the traditional measure of end task accuracy, which quantifies the proportion of testing examples where plausible stories and implausible stories are accurately identified.

- **Consistency** measures the proportion of testing examples where not only the plausible story is correctly identified, but also the conflicting sentence pair for the implausible story is accurately identified. The aim is to demonstrate the model's consistency in recognizing conflicts when reasoning about plausibility.

- **Verifiability** evaluates the proportion of testing examples where not only the plausible story and the conflicting sentence pair for the implausible story are correctly identified, but also the underlying physical states (i.e., preconditions and effects) that contribute to the conflict are accurately identified. This demonstrates that the detected conflict can be validated through a correct understanding of the underlying implausible change of physical states. For the notion of verifiability, we refer to Storks et al. (2021): to be verifiable, a story needed to have at least one physical state label predicted in the preconditions of the breakpoint sentence and one physical state label predicted in the effects of the evidence sentence, and all such predictions must be correct.

4.4. Results

Taking into consideration the three different metrics, in Table 4 we report the results in our test set.

	Accuracy	Consistency	Verifiability
Cloze	72.6	19.6	2.5
Order	58.1	1.6	0.8

Table 4: Results of GITA on the Italian test set.

The model performed better in the Cloze dataset than the Order dataset. The end task was correctly predicted in the 72.6% of the cases for the Cloze dataset, whereas the intermediate tasks fell short in terms of predictability. This disparity between the accuracy of the end task and the lack of predictability in the intermediate tasks demonstrates the incongruity that stands between the capacity of predicting the high-level classification task for which the system was fine-tuned, and the capacity of the same system to justify the steps in the reasoning process that brought to the final decision.
We examine the performance of the system in the Cloze dataset according to the three key metrics:

- the accuracy of 72.6% reflects the 84 pairs of stories out of 117 that were correctly identified as plausible and implausible;

- the consistency of 19.6% refers to the 23 implausible stories, where the model was able to spot the conflict sentences;

- and the 2.5% verifiability refers to the three instances when our model was able to predict the correct physical states present in our stories.

From this analysis of the data, we can recognize the difficulty encountered by the model to spot the right physical states in our 23 consistent implausible stories.

	sentence 1	sentence 2	sentence 3	sentence 4	sentence 5
T	Olivia è in camera da letto.	Olivia ha aperto la valigia.	Olivia ha messo in valigia due camicie.	Olivia ha chiuso la valigia.	Olivia ha trovato il passaporto.
	Olivia is in the bedroom.	Olivia has opened the suitcase.	Olivia has packed two shirts in the suitcase.	Olivia has closed the suitcase.	Olivia found her passport.
F	Olivia è in camera da letto.	Olivia ha aperto la valigia.	Olivia ha messo in valigia due camicie.	Olivia ha chiuso la valigia.	Olivia ha preso una camicia dalla valigia.
	Olivia is in the bedroom.	Olivia has opened the suitcase.	Olivia has packed two shirts in the suitcase.	Olivia has closed the suitcase.	Olivia has taken a shirt from the suitcase.

Table 5: Story No. 102 (102-C0)

4.5. Analysis of the results

In this section we present an analysis of the consistent stories and the cases we encountered in the predictions. We decided to have a closer look at the 23 consistent stories of the Cloze dataset, in order to understand the behavior of the model. For each physical state in one story we have both precondition and effect predictions. In these 23 consistent stories, out of 414 physical states in both precondition and effect predictions, only 112 were correctly predicted, reaching a 27% of correctly predicted physical attributes.

Looking at the stories where most of the attributes were correctly identified by the model, we focus on stories number 102-C0, and 94-C0 in the Cloze partition. In story 102 (Table 5), we reached a 60% of correct physical states, out of 15 states 9 were well predicted. The 9 identified states were *location*, *open*, and *contain*. In story 94 (Table 6), 6 out of 12 states were well predicted, all of them were *location* attributes that the model had no issue in predicting.

In our dataset, three stories appeared as verifiable: they constituted the proportion of testing examples where not only the plausible story and the conflicting sentence pair for the implausible story were correctly identified but also the underlying physical states (i.e., preconditions and effects) that contributed to the conflict were correctly identified. As we said, in order to be verifiable, a story needed to have at least one physical state label predicted in the preconditions of the breakpoint sentence, and one physical state label predicted in the effects of the evidence sentence, and all such predictions had to be correct. Among the total of 23 consistent stories, these three stories accounted for 13%. In these particular stories, the accuracy of predicting the total physical states reached 25%, 28%, and 39% respectively. This finding suggests that the identification of the total number of physical states did not play a significant role in determining the plausibility of the story.

An example of one of the three verifiable stories in the dataset is story 66-C0 (Table 7). In Story 66-C0, the evidence sentence is sentence No. 3 ("Giusy closed the shower door"), and the breakpoint is sentence No. 4 ("Giusy steps into the shower"). These two sentences generate the conflict in the story: if the shower door is closed, Giusy cannot step in. As anticipated, one correct prediction in the precondition physical states of the breakpoint, in this example sentence No. 4, and one correct prediction of effect physical states in the evidence sentence, in this example sentence No. 3, were correct. In this specific case, the precondition label in sentence No. 4 referred to *human location* was the only one correctly predicted by the model in the sentence (the correct label was *human location*: 2, which in our classification means that the human agent *moved*). While in sentence No. 3, the effect label *open*: 1, referred to the entity *doccia* (shower), was correctly identified, meaning that the shower was annotated as *closed* in that sentence.

These are, in general, some of the specific behaviours of the model:

- **Best predicted physical states:** the model recognized the *location* attribute many times, while attributes such as *clean*, and *function* were recognized in only a few stories.
- **Ratio of true/false predictions for each state:** we calculated the number of actual physical attributes in each consistent story. About 55% of the *location* and *human location* states were correctly predicted by our model, 50% of the *wearing* states were foreseen by the model (although we only see a total of 2 actual cases for this attribute), and *pieces* and *solid* states also reached 40% of positive predictions.
- **Location and human location attributes:**

	sentence 1	sentence 2	sentence 3	sentence 4	sentence 5
T	Giusy ha preso un pennello.	Giusy ha preso le tempere.	Giusy ha preso la tavolozza.	Giusy ha messo le tempere sulla tavolozza.	Giusy ha iniziato a dipingere un quadro.
	Giusy has taken a paintbrush.	Giusy has taken the tempera paints.	Giusy has taken the palette.	Giusy has put the tempera paints on the palette.	Giusy has started painting a picture.
F	Giusy ha preso un pennello.	Giusy ha preso le tempere.	Giusy ha preso la tavolozza.	Giusy ha fatto tutto a pezzi.	Giusy ha iniziato a dipingere un quadro.
	Giusy has taken a paintbrush.	Giusy has taken the tempera paints.	Giusy has taken the palette.	Giusy has put the tempera paints on the palette.	Giusy has broken everything into pieces.

Table 6: Story No. 94 (94-C0)

these two states are the only ones that have a full specific classification (Table 2) and at the same time are those states that appear to be better predicted.

- **Incoherence of the LLMs:** these are the cases where a state could be well predicted in a sentence but wrongly guessed in other sentences of the same story. For example, a *human location* attribute can be identified in a sentence, and then predicted again in another sentence of the same story, where no attributes were associated to a human actor. In these cases, the model seemed stuck with the correct prediction of *human location* and, in its over-confidence, it repeated the prediction for the following sentences. Again, it looks like no reasoning method was applied in the assignation of the correct predictions of physical changes and that the model was not consistent.

- **Random guesses:** the presence of numerous randomly assigned attributes in many stories aligns with the suspicion of an unpredictable and irrational categorization of physical states. An example is the confusion between *open* and *power* attributes throughout the dataset, or the correct prediction of a state but the wrong guess of the number related to it. There are cases in which the prediction looked completely random: for example, the *functional* and the *contain* attributes were predicted as *power*, while the *running* attribute, which did not appear in a story, was instead predicted in the precondition classification.

The lower-level task appears challenging to learn, highlighting the need for further investigation and refinement in our approach.

5. Discussion

In this discussion section, we delve deeper into the significance of our findings. The presented analysis highlights the complexities inherent in training LLMs to comprehend and predict physical states within narrative contexts. An area of further exploration is the extension of the structured annotation approach employed in this study to encompass a broader range of attributes. This expansion of the annotation space may yield additional insights into the performance of LLMs and their ability to understand diverse aspects of physical states.

A hypothesis arises regarding the model's proficiency in predicting *location* states: our hypothesis is that this proficiency may be linked to potential biases within the training dataset, where a significant number of examples featuring the *location* attribute might exist. This observation emphasizes the importance of careful dataset curation and balance to avoid an over-representation of specific attributes, which could skew the model's performance.

In contrast to our approach, one of Storks et al.'s (2021) experiments achieved an accuracy of 97%, when considering only the end task sentence classification, avoiding the intermediate tasks. This shows that reasonable supporting evidence is not required in order to achieve high accuracy of the end task. This outcome invites scrutiny of the validity of prevailing state-of-the-art results in the domain of commonsense benchmarks, which often lack the manifestation of coherent reasoning beyond end classification tasks (Davis, 2023). This inherent simplification of tasks underscores the existing skepticism surrounding the capacity of current methodologies to fully capture and comprehend commonsense knowledge. Furthermore, although some researchers have only focused on

	sentence 1	sentence 2	sentence 3	sentence 4	sentence 5
F	Giusy entra in bagno.	Giusy si toglie i vestiti.	Giusy chiude la porta della doccia.	Giusy entra in doccia.	Giusy apre il rubinetto della doccia.
	Giusy enters the bathroom.	Giusy undresses.	Giusy closes the shower door.	Giusy steps into the shower.	Giusy turns on the shower faucet.

Table 7: Story No. 66 (66-C0)

the end tasks (Bisk et al., 2020b; Yang et al., 2021), the efforts aimed at shedding light on the undergoing linguistic phenomena that condition commonsense information (Storks et al., 2021) seem more productive in advancing towards effective commonsense understanding. Prioritizing objective evaluations and adopting more pragmatic benchmarks in NLU holds the potential for substantial advancements in our understanding of LLMs' capabilities and limitations in managing commonsense knowledge.

6. Conclusion

In this work, we focused on physical commonsense evaluation in LLMs. We described the manual creation and annotation of GITA in Italian, as well as its automation, we evaluated a pre-trained language model's performance on it, and provided insights into its capabilities in physical commonsense reasoning. The findings from our study indicate that when it comes to acquiring commonsense language understanding, supervising LLMs solely through high-level classification tasks often yields inconsistent and unverifiable reasoning. These models struggle to capture intermediate evidence that contributes to the completion of the end task. We consider advisable to shift the research focus to truthful benchmarks, and emphasize the importance of evaluating pre-trained language models' ability to understand physical commonsense in a more realistic way.

In the short term, we strongly advise to direct future efforts towards a novel form of analysis that instills greater trust and specificity. This analysis should prioritize either the successful completion of the end task or the evaluation of tiered tasks. By doing so, we can establish a more reliable and targeted approach that addresses the immediate concerns and limitations we have encountered.

In future work, we aim to further refine the annotation process and to add new stories to the Italian dataset. We will also try to extend the annotation task to at least another professional, since it can be very interesting to incorporate multiple perspectives to gain a more comprehensive analysis of the dataset. Moreover, we aim to apply our benchmark to generative systems to evaluate their performance and adaptability. This analysis has highlighted challenges in low-level task performance, and assessing GenAI systems using our test set can provide valuable insights for improvement. Making the dataset available will allow for further work on physical commonsense understanding in Italian.

7. Ethics Statement

The dataset contains stories that may prototypically occur in Italian households. While most of these narratives are likely to be familiar to a broad audience, people from different cultural backgrounds may find some of the stories less frequent.

8. Acknowledgements

We acknowledge the funding from the following projects: a) Ixa group A type research group (IT-1805-22) funded by the Basque Government; b) DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe; c) AWARE: Commonsense for a new generation of natural language understanding applications (TED2021-131617B-I00) funded by MCIN/AEI /10.13039/501100011033 by the European Union NextGenerationEU/ PRTR; and d) DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

9. Bibliographical References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

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

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2017. Simulating Action Dynamics with Neural Process Networks. *CoRR*, abs/1711.05313.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ernest Davis. 2023. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Comput. Surv.* Just Accepted.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut Learning of Large Language Models in Natural Language Understanding.

Yuwei Fang, Shuohang Wang, Yichong Xu, Ruochen Xu, Siqi Sun, Chenguang Zhu, and Michael Zeng. 2022. Leveraging Knowledge in Multilingual Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3237–3246, Dublin, Ireland. Association for Computational Linguistics.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical Causality of Action Verbs in Grounded Language Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1814–1824, Berlin, Germany. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. 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.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Gary Marcus and Ernest Davis. 2020. Insights for AI from the Human Mind. *Commun. ACM*, 64(1):38–41.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.*, 55(3).

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Commun. ACM*, 64(9):99–106.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A Commonsense Reasoning Benchmark with Complementary Sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches. *CoRR*, abs/1904.01172.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4902–4918, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual Goal-Step Inference using wikiHow. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A. Annotations in the dataset

These are the attributes that encode the metadata and linguistic information in the GITA dataset:

- **story_id:** refers to the number of the story for both plausible and implausible stories.
- **worker_id:** refers to the name assigned to a specific worker during the creation of the story.
- **type:** refers to *cloze* or *order* and it is a label used only in implausible stories.
- **idx:** refers to the implausible dataset, where there is more than one implausible story for a given story number; for example, if we have more than one implausible version of a plausible story (we created more than an implausible story changing the order of our sentences more than once), the index number indicates to which implausible example we are referring.
- **aug:** refers to possible automatic data augmentation techniques that can be taken into account for future works to resolve an overfitting problem.
- **actor:** refers to the human agent of the story.
- **location:** refers to the room where the story takes place.
- **objects:** refers to all the inanimate entities that we find into each story.
- **sentences:** includes the 5 sentences in the story.
- **length:** refers to the number of sentences in each story.
- **example_id:** corresponds to the story number and includes letters for implausible stories.
- **plausible:** is TRUE when the story is plausible and FALSE when it is implausible.
- **breakpoint:** refers to the sentence where the story becomes implausible, where the conflict becomes evident; in plausible stories the breakpoint is always -1.
- **conlict_sents:** refers to the other sentence in the story that together with the breakpoint sentence makes the story implausible; in plausible stories this field is blank.
- **conlict_pairs:** refers to the conflict pair of sentences, gathering the two previous labels; in plausible stories this field is blank.
- **states:** includes all the physical states annotations for all entities in all sentences; we will look into this in the following section.

B. Annotation environment

```
actor:
Marco
objects:
frigo latte tazza %cucchiaio
story_number (same as story_id in
   quotes):
'0'
story_id (NO quotes, NO letter, only
    number):
0
worker_id (in quotes):
'GAP'
type (null for positive, order, or
   cloze, in quotes):
null
idx (null, or same as NUMBER in
   story number):
null
aug (false):
false
location (in quotes):
'cucina'
sentences:
Marco ha aperto il frigo.   Marco ha
    preso il latte.    Marco ha
   preso la tazza.    Marco ha preso
    il cucchiaio.    Marco ha messo
   il cucchiaio nella tazza.
length:
5
example_id (same as story number, in
    quotes):
'0'
plausible:
true
breakpoint:
-1
confl_sents (type only []):
[]
```

Listing 1: Annotation environment.

C. Example of annotated sentence

```
{ "0" :
{"story_id": 0,
"worker_id": "GAP",
"type": null,
"idx": null,
"aug": false,
"actor": "Marco",
"location": "cucina",
"objects": "frigo, latte, tazza,
   cucchiaio",

"sentences":
["Marco ha aperto il frigo.",
"Marco ha preso il latte.",
"Marco ha preso la tazza.",
"Marco ha preso il cucchiaio.",
"Marco ha messo il cucchiaio nella tazza
    ."],

"length": 5,
"example_id": "0",
"plausible": true,
"breakpoint": -1,
"confl_sents": [],
"confl_pairs": [],

"states":
[{"h_location": [[" Marco ", 0]],
"conscious": [[" Marco ", 2]],
"wearing": [[" Marco ", 0]],
"h_wet": [[" Marco ", 0]],
"hygiene": [[" Marco ", 0]],
"location": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"exist": [
["frigo", 2],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"clean": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"power": [
["frigo", 2],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"functional": [
["frigo", 2],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"pieces": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"wet": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"open": [
["frigo", 4],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"temperature": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"solid": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"contain": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"running": [
["frigo", 2],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"moveable": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"mixed": [
["frigo", 0],
["latte", 0],
["tazza", 0],
["cucchiaio", 0]],
"edible": [
["frigo", 0],
["latte", 0],
["tazza", 0],
"cucchiaio", 0]]}
```

Listing 2: Example of annotated sentence.