GITA4CALAMITA - Evaluating the Physical Commonsense Understanding of Italian LLMs in a Multi-layered Approach: A CALAMITA Challenge

Giulia Pensa[1], Ekhi Azurmendi[2], Julen Etxaniz[2], Begoña Altuna[2] and Itziar Gonzalez-Dios[2]

[1]University of the Basque Country UPV/EHU
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract

In the context of the CALAMITA Challenge, we investigate the physical commonsense reasoning capabilities of large language models (LLMs) and introduce a methodology to assess their understanding of the physical world. To this end, we use a test set designed to evaluate physical commonsense reasoning in LLMs for the Italian language. We present a tiered dataset, named the Graded Italian Annotated dataset (GITA), which is written and annotated by a professional linguist. This dataset enables us to focus on three distinct levels of commonsense understanding. Our benchmark aims to evaluate three specific tasks: identifying plausible and implausible stories within our dataset, identifying the conflict that generates an implausible story, and identifying the physical states that make a story implausible. We perform these tasks using LLAMA3, Gemma2 and Mistral. Our findings reveal that, although the models may excel at high-level classification tasks, their reasoning is inconsistent and unverifiable, as they fail to capture intermediate evidence.

Keywords

Physical commonsense reasoning, large language models, Italian benchmark

1. Challenge: Introduction and Motivation

Physical commonsense understanding refers to the ability to comprehend the physical world and the events that transpire within it. This capability is a crucial component of human intelligence, enabling us to reason about our environment, anticipate future occurrences, and navigate our surroundings effortlessly, and recently there has been notable advancement in the development of large language models (LLMs) that can produce human-like language and execute a variety of language-related tasks.

LLMs have exhibited promising outcomes in grasping common sense in particular situations [1, 2]. Nevertheless, it is widely recognized that the most precise evaluation of their capabilities is attained when assessing their performance in specific end tasks [3, 4]. The evaluation often emphasizes the capacity of LLMs to replicate relatively straightforward tasks, rather than their authentic proficiency in reasoning and comprehending language [5, 6]. 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 contribute to this challenge developing an original Italian 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.

In this paper, we present GITA4CALAMITA, the Graded Italian Annotated dataset for the CALAMITA challenge [7]. GITA4CALAMITA is an adapted version of the GITA dataset proposed in [8]. In particular, we decided to revise the physical states annotation and adapt it to this challenge. The first version of GITA dataset is available in our repository under the license CC BY-NC-SA 4.0.[1]. The GITA4CALAMITA dataset is manually compiled by a professional linguist, which allows for this multi-layered evaluation of the reasoning process. 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.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 - 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
giulia.pensa.tr@gmail.com (G. Pensa); ekhi.azurmendi@ehu.eus (E. Azurmendi); julen.etxaniz@ehu.eus (J. Etxaniz); begona.altuna@ehu.eus (B. Altuna); itziar.gonzalezd@ehu.eus (I. Gonzalez-Dios)
https://github.com/GiuliaAPensa (G. Pensa)
0009-0008-4113-890X (E. Azurmendi); 0009–2099-7766 (J. Etxaniz); 0000-0002-4027-2014 (B. Altuna); 0000-0003-1048-5403 (I. Gonzalez-Dios)

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

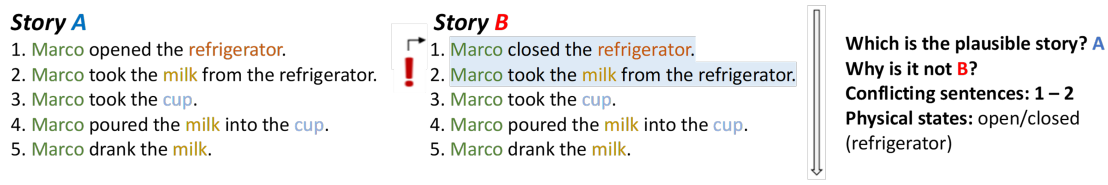

Figure 1: Representation of story pair from GITA

2. Challenge: Description

Our aim in this challenge is to assess the understanding of physical commonsense in LLMs for Italian. 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 that cause conflict in implausible stories.

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 GITA4CALAMITA 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.

We introduce a series of tasks that constitute a human-interpretable reasoning process, supported by a chain of evidence, reflecting the assessment methodology outlined above. To explain this approach, we present the tasks from the deepest to the shallowest, mirroring human reasoning:

Physical state classification: Leveraging our physical state annotations, systems must recognize the involved physical states in the conflicting sentences of implausible stories. If we look at the example in 1, we are able to identify the problematic physical state "open" as cause of implausibility.

Conflict detection: Next, the task of conflict detection entails identifying sentence pairs of the form $S_i \longrightarrow S_j$. Here, S_j represents the breakpoint, indicating the point at which the story becomes implausible based on the given context. S_i serves as the evidence that explains the breakpoint, typically causing a conflicting world state.

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.

3. Data description

The GITA4CALAMITA dataset is composed by plausible and 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". We created 5-sentence stories, giving context and requiring reasoning over multiple sentences. 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.

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 this dataset, we took into account 14 physical attributes that were included in the annotation phase, and we composed stories that contained those attributes. Following the work of [9] and [10], these are the 14 physical states that we wanted to have in our stories:

- location, conscious, dressed, wet, exist, clean, power, functional, in pieces, open, temperature, solid, occupied, edible.

3.1. Dataset creation

In the first two rows of Table 1 we can see an example of plausible story from the GITA4CALAMITA dataset

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

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 aimed to minimize subjectivity and limit potential confounding factors from complex language usage. By using simple language, we were able to shift our focus away from linguistic processing and semantic phenomena, allowing us to concentrate more on examining machines' reasoning abilities, particularly their physical commonsense understanding. Consequently, we created our simple sentences in a straightforward declarative structure, typically starting with the agent of the story, followed by a verb, a direct object, and optionally, an 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 stories respectively.

3.1.1. Order implausible stories

The plausible stories only work in the causal sequence that we created. 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. Origin of data

GITA4CALAMITA is a new version of [8], which is based on [11]. Our main objective was to create an Italian dataset, manually annotated, to assess a pre-trained language model on physical commonsense tiered tasks. To

create the stories, we took inspiration from the Story Cloze Test [12] and ROCStories Corpora [13]. The Story Cloze Test compiles four-sentence stories with a missing ending so that a system chooses the most appropriate conclusion; the ROCStories Corpora is composed of five-sentence stories about everyday life for story generation.

3.3. Annotation details

GITA4CALAMITA 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.

In each implausible story, we annotated the physical state that caused a conflict between two sentences. We annotated both Order and Cloze implausible stories according to the corresponding physical state involved. If we consider the stories in Table 1, both implausible stories (C and O) are annotated using the physical state "open", In fact, in both implausible stories the conflict is related to the openness of the refrigerator: in both cases the refrigerator appears closed when Marco tries to take the milk out of it. There are cases where for one plausible story there are two implausible stories that are implausible for two different reasons, hence the annotated physical state is different.

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. An example of a complete annotation can be found in Appendix B.

3.4. Data format

The GITA4CALAMITA dataset was created and annotated in a JSON format. The following example is story 0-C0 of our dataset, the first implausible Cloze story.


```
{
    "0-C0": {
        "story_id": 0,
        "worker_id": "GAP",
        "type": "cloze",
        "idx": 0,
        "aug": false,
        "actor": "Marco",
        "location": "cucina",
        "objects": "frigo, latte,
            tazza, cucchiaio",
        "sentences": [
            "Marco ha chiuso il frigo
                .",
            "Marco ha preso il latte
                dal frigo.",
            "Marco ha preso la tazza
                .",
            "Marco ha preso il
                cucchiaio.",
            "Marco ha messo il
                cucchiaio nella tazza
                ."
        ],
        "length": 5,
        "example_id": "0-C0",
        "plausible": false,
        "breakpoint": 1,
        "confl_sents": [0],
        "confl_pairs": [0, 1]
    }
}
```


3.5. Example of prompts used for zero or/and few shots

For each of the three proposed tasks we use a different prompt:

- **Task 1:** Please read the following story and answer if the story is plausible taking into account the order of the events. Please answer with true or false.
  Task 2: The following story is implausible. Identify the breakpoint, and then select the sentence

responsible for the implausibility. Please identify the breakpoint sentence and the conflicting sentence.
Task 3: The following story is implausible. Identify the physical state that causes the conflict in the story. These are the descriptions of each physical state: **Power**: Indicates whether an object is powered or not, relevant for electrical devices. **Location**: Refers to the spatial position of an entity, either human or object. **Exist**: Denotes whether an object is present or has disappeared. **Clean**: Refers to the cleanliness of an entity, indicating whether it is clean or dirty. **Edible**: Identifies whether an object is fit for consumption. **Wet**: Denotes whether an object or person is in a wet or dry state. **Functional**: Refers to whether an object is in working condition or broken. **Wearing**: Applies to humans, indicating whether they are dressed or not. **Open**: Refers to whether an object (e.g., a door or container) is open or closed. **Conscious**: Denotes whether a human is conscious or unconscious. **Temperature**: Refers to the relative temperature of an entity, e.g., hot or cold. **Solid**: Describes whether an object is in a solid state. **Occupied**: Indicates whether an object (e.g., a container) is occupied or contains something. **In pieces**: Refers to whether an object is intact or has been broken into pieces. Select one of them after reading the story.

We select some examples from our GITA4CALAMITA dataset to be used as few-shot examples. For some of the tests we randomly select the examples, for others, we base our choice on their variability. We select stories where all possible combination of conflicting sentences were happening; at the same time, within the selected stories we try to include most of the physical states annotated.

3.6. Detailed data statistics

The GITA4CALAMITA dataset is an Italian test composed by a total of 356 stories. The statistics of the GITA4CALAMITA dataset are in Table 2.

Measures	GITA4CALAMITA
plausible stories	117
implausible stories (ORDER)	122
implausible stories (CLOZE)	117
total stories	356

Table 2
Statistics of GITA4CALAMITA

4. Metrics

The metrics involved in our tasks for the GITA4CALAMITA benchmark are the following ones:

- **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 implausible 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 implausible story and the conflicting sentence pair for the implausible story are correctly identified, but also the underlying physical states 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.

Taking into consideration the three different metrics, in Table 3 we report the results in our test set. We perform experiments using the base and instruct Llama 3.1, Gemma 2 and Mistral models of various sizes. Each metric is obtained from a different task, where models are evaluated in the instances that are only guessed correctly in the previous tasks. All tasks are evaluated in a 3-shot setting, using random examples from the test set. For models that support system prompt (Llama3.1 models), the description of each task is included there, for models that do not support it (Gemma2 and Mistral models) the task description is included in the first user input. Each few-shot instance is formatted as a multiturn conversation between user and assistant. Next, we describe the main findings from these results.

Model Size and Performance: Generally, larger models (e.g., Llama-3.1 70B) outperform smaller models across the metrics. The 70B Llama-3.1 models show improvements over their 8B counterparts, particularly in consistency and verifiability. Gemma2 models also show improvements when bigger models are used. There are two exceptions in the case of the accuracy: Gemma2-Instruct 9B and Llama-3.1-Instruct 8B achieve better results than their bigger counterparts Gemma2 27B and Llama3 70B. They also outperform the base models.

Model	Size	Accuracy				Consistency			Verifiability		
		Overall	Cloze	Order	Plausible	Overall	Cloze	Order	Overall	Cloze	Order
Gemma-2 (base)	9B	72.75	86.96	70.49	61.34	32.35	45.22	20.66	12.18	16.52	8.26
Gemma-2-Instruct	9B	76.12	85.22	60.66	83.19	38.66	58.26	20.66	17.65	30.43	5.79
Gemma-2 (base)	27B	75.28	89.57	59.02	78.15	39.07	55.65	23.97	21.85	31.30	13.22
Gemma-2-Instruct	27B	73.88	80.00	54.10	88.24	39.08	60.87	19.00	24.79	40.87	9.92
Llama-3.1 (base)	8B	60.96	70.43	60.66	52.10	26.47	33.04	20.66	11.34	13.04	9.92
Llama-3.1-Instruct	8B	77.25	93.91	90.16	47.90	37.39	53.91	22.31	10.50	16.52	4.96
Llama-3.1 (base)	70B	82.02	94.78	92.62	58.82	57.14	66.96	47.93	28.99	36.52	21.49
Llama-3.1-Instruct	70B	74.16	99.13	98.36	25.21	68.07	82.61	54.55	18.07	25.22	11.57
Mistral-V0.3 (base)	7B	60.39	66.96	54.92	59.66	20.59	27.83	14.05	6.72	11.30	2.48
Mistral-Instruct-V0.3	7B	59.83	67.82	27.05	85.71	21.00	40.87	2.48	9.24	19.13	0.00

Table 3
Results of the base and instruct Llama 3.1, Gemma 2 and Mistral models of various sizes

Instruction Tuning Effects: Instruction-tuned versions (e.g., Gemma-2-Instruct, Llama-3.1-Instruct) typically outperform their base counterparts. There are exceptions such as order accuracy for LLama 3.1 70B and Gemma 2 9B. However, Mistral-V0.3-Instruct is very similar or worse than the base model and generally is more biased, it tends to classify as plausible the stories and it performs better in Cloze than in Order.

Cloze, Order and Plausible Most models perform generally better on Cloze examples compared to Order examples. This is consistent across models and metrics. Models are generally better in Cloze and Order than in Plausible. This could be explained by the bias of the models to answer true or false when they are asked if the story is plausible. Models also see double implausible few-shot examples, which could also cause models to give that answer more frequently.

5. Limitations

This study has some limitations that should be acknowledged. Firstly, only one prompt was tested for each task, which may not fully capture the potential variability in performance. Additionally, the models used were multilingual but not specifically tailored for the Italian language, potentially affecting the accuracy of the results for Italian-specific tasks. Furthermore, the dataset used in this study was limited to stories within the household domain, which may not generalize well to other contexts.

6. Ethical issues

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.

Acknowledgments

This work has been partially funded by:

- DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.
- Disargue (TED2021-130810B-C21) MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR.
- DeepKnowledge (PID2021-127777OB-C21) MCIN/AEI/10.13039/501100011033 and by FEDER, EU.
- Ixa group A type research group (IT1570-22) Basque Government
- IKER-GAITU project 11:4711:23:410:23/0808 by Basque Government

References

[1] J. Huang, K. C.-C. Chang, Towards Reasoning in Large Language Models: A Survey, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1049–1065. URL: https://aclanthology.org/2023.findings-acl.67. doi:10.18653/v1/2023.findings-acl.67.

[2] K. Sakaguchi, R. L. Bras, C. Bhagavatula, Y. Choi, WinoGrande: An Adversarial Winograd Schema Challenge at Scale, Commun. ACM 64 (2021) 99–106. URL: https://doi.org/10.1145/3474381. doi:10.1145/3474381.

[3] D. Pessach, E. Shmueli, A Review on Fairness in Machine Learning, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3494672. doi:10.1145/3494672.

[4] E. Davis, Benchmarks for Automated Commonsense Reasoning: A Survey, ACM Comput. Surv.

(2023). URL: https://doi.org/10.1145/3615355. doi:10.1145/3615355, just Accepted.

[5] T. Linzen, How Can We Accelerate Progress Towards Human-like Linguistic Generalization?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5210–5217. URL: https://aclanthology.org/2020.acl-main.465. doi:10.18653/v1/2020.acl-main.465.

[6] E. M. Bender, A. Koller, 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, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. URL: https://aclanthology.org/2020.acl-main.463. doi:10.18653/v1/2020.acl-main.463.

[7] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[8] G. Pensa, B. Altuna, I. Gonzalez-Dios, A Multilayered Approach to Physical Commonsense Understanding: Creation and Evaluation of an Italian Dataset, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 819–831. URL: https://aclanthology.org/2024.lrec-main.74.

[9] Q. Gao, M. Doering, S. Yang, J. Chai, 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), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1814–1824. URL: https://aclanthology.org/P16-1171. doi:10.18653/v1/P16-1171.

[10] A. Bosselut, O. Levy, A. Holtzman, C. Ennis, D. Fox, Y. Choi, Simulating Action Dynamics with Neural Process Networks, CoRR abs/1711.05313 (2017). URL: http://arxiv.org/abs/1711.05313. arXiv:1711.05313.

[11] S. Storks, Q. Gao, Y. Zhang, J. Chai, Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4902–4918. URL: https://aclanthology.org/2021.findings-emnlp.422. doi:10.18653/v1/2021.findings-emnlp.422.

[12] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, J. Allen, LSDSem 2017 Shared Task: The Story Cloze Test, in: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 46–51. URL: https://aclanthology.org/W17-0906. doi:10.18653/v1/W17-0906.

[13] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. Allen, 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, Association for Computational Linguistics, San Diego, California, 2016, pp. 839–849. URL: https://aclanthology.org/N16-1098. doi:10.18653/v1/N16-1098.

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 the stories.

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.