Physical Commonsense Reasoning for Lower-Resourced Languages and Dialects: a Study on Basque

Jaione Bengoetxea, Itziar Gonzalez-Dios, Rodrigo Agerri
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{jaione.bengoetxea,itziar.gonzalezd,rodrigo.agerri}@ehu.eus

Abstract

Physical commonsense reasoning represents a fundamental capability of human intelligence, enabling individuals to understand their environment, predict future events, and navigate physical spaces. Recent years have witnessed growing interest in reasoning tasks within Natural Language Processing (NLP). However, no prior research has examined the performance of Large Language Models (LLMs) on non-question-answering (non-QA) physical commonsense reasoning tasks in low-resource languages such as Basque. Taking the Italian GITA as a starting point, this paper addresses this gap by presenting BasPhyCo, the first non-QA physical commonsense reasoning dataset for Basque, available in both standard and dialectal variants. We evaluate model performance across three hierarchical levels of commonsense understanding: (1) distinguishing between plausible and implausible narratives (accuracy), (2) identifying the conflicting element that renders a narrative implausible (consistency), and (3) determining the specific physical state that creates the implausibility (verifiability). These tasks were assessed using multiple multilingual LLMs as well as models pretrained specifically for Italian and Basque. Results indicate that, in terms of verifiability, LLMs exhibit limited physical commonsense capabilities in low-resource languages such as Basque, especially when processing dialectal variants.

Keywords: Physical Commonsense Reasoning, Multilingualism, Less-Resourced/Endangered Languages, Italian, Basque, Dialects

1. Introduction

Commonsense reasoning represents the human capacity to understand and manipulate real-world objects and their interactions. This domain has attracted considerable attention in Artificial Intelligence research in recent years (Davis, 2023; Sun et al., 2025). Physical commonsense reasoning, a specific subdomain, addresses events occurring in the physical world by capturing knowledge about everyday objects, their physical properties, and their potential uses and manipulations (Bisk et al., 2020; Pensa et al., 2024a).

As a fundamental component of human intelligence, physical commonsense reasoning enables individuals to reason about their environment, anticipate future events, and navigate their surroundings. Recent research has increasingly examined the reasoning capabilities of LLMs, though such investigations have been conducted predominantly in English (Bisk et al., 2020; Storks et al., 2021).

This paper focuses on Basque, as well as its Western dialect, and Italian, the source language of the dataset upon which our work is based on (Pensa et al., 2024a,b). These low-resource languages provide valuable insight into LLM performance on complex physical-world reasoning tasks under data-limited conditions.

We manually translated the Italian dataset GITA into standard Basque and automatically adapted it into the Western dialect. The Western dialect was selected due to its peripheral status and documented linguistic distance from other Basque varieties, as established in dialectological research (Mitxelena, 1981). This linguistic divergence is corroborated by several NLP studies: Estarrona et al. (2023) identified Biscayan (Western) and Souletin as the most distinct among historical Basque dialects, while Bengoetxea et al. (2025) attributed the negative impact of the Western dialect on Natural Language Inference (NLI) performance to its distance from Standard Basque.

We evaluate model performance across three hierarchical reasoning tasks: (i) distinguishing plausible from implausible narratives (accuracy), (ii) identifying conflicting sentences within implausible narratives (consistency), and (iii) determining the physical states that render narratives implausible (verifiability). Our evaluation uses two multilingual LLMs alongside two Italian-pretrained models and one Basque-pretrained model, thereby examining current LLM knowledge of the physical world and human-object interactions.

To our knowledge, this represents the first investigation combining physical commonsense reasoning with Basque dialectal variation. Data and code are publicly available[1]. Our investigation presents the following contributions:

- The first publicly available non-QA physical commonsense reasoning dataset in Basque,

[1] https://github.com/hitz-zentroa/BasPhyCo

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3343–3354
11–16 May 2026.

including the first such dataset in a Basque dialect (Western).

- The first evaluation of LLM performance on non-QA physical commonsense reasoning in a low-resource language such as Basque. Results indicate that, in terms of verifiability, LLMs exhibit limited physical commonsense capabilities in low-resource languages such as Basque, especially when considering dialectal variants.
- A comprehensive evaluation of LLMs' knowledge gaps when faced with physical commonsense reasoning for low-resource languages shows that this task is still challenging. Additionally, results with Basque language variation show that models pretrained for the target language seem to have a better ability to handle linguistic variation.
- Fine-grained evaluation of physical states indicates that models have a general difficulty in correctly predicting these labels, Location, Edible, and Conscious states being particularly challenging.

2. Related Work

Physical Commonsense Recent research has tried to test physical commonsense knowledge of current LLMs. To this end, researchers have developed various datasets and benchmarks, including textual information (Rajani et al., 2019; Bisk et al., 2020; Rajani et al., 2020; Storks et al., 2021; Aroca-Ouellette et al., 2021; Wang et al., 2023; Pensa et al., 2024a; Jeong et al., 2025), images (Bakhtin et al., 2019; Hong et al., 2021; Liu et al., 2022; Meng et al., 2024), and videos (Weihs et al., 2022; Yu et al., 2022; Motamed et al., 2025).

Datasets focusing on textual information have been generally presented as Question-Answering (QA) tasks, such as PIQA (Bisk et al., 2020). Some works have attempted to introduce other methodologies, such as TRIP (Storks et al., 2019), which is a physical commonsense reasoning benchmark composed of five-sentence stories. It evaluates models on three tasks: classifying stories as plausible or implausible, detecting the conflicting sentence, and identifying the physical states of objects involved.

The majority of the datasets in physical commonsense reasoning have been curated in English. Some exceptions include GITA (Pensa et al., 2024a) for Italian, a non-QA physical commonsense reasoning dataset based on TRIP, and EPiK (Jeong et al., 2025) for Korean, which follows the PIQA dataset, while culturally adapting it to Korean.

To our knowledge, the sole existing resource for physical commonsense reasoning in Basque is a professionally translated version of the PIQA dataset (Baucells et al., 2025), which provides only the validation subset. Consequently, no Basque-language dataset for physical commonsense reasoning exists beyond the question-answering (QA) format.

Dialects and Reasoning Regarding the use of dialects in commonsense reasoning, Lin et al. (2025) have very recently analyzed LLMs' dialect robustness and fairness with Standardized English (SE)[2] and African American Vernacular English (AAVE). They create the ReDial (Reasoning with Dialect Queries) dataset, a high-quality, end-to-end human-annotated SE-AAVE parallel dataset for reasoning tasks (algorithm, logic, math, and integrated reasoning) that contains over 1.2K parallel prompts in SE and in AAVE. An evaluation on LLM families (GPT, Claude, Llama, Mistral, Phi) shows lower performance when using dialectal prompts.

Pan et al. (2025) analyze dialectal bias on reasoning tasks through a multiple-choice question answering task, where they compare results in Standard American English (SAE)[2] with results in 5 English dialects, such as Chicano, African American, or Indian English. The dataset was generated by applying grammatical perturbations to the original SAE multiple-choice benchmark using the Multi-VALUE package (Ziems et al., 2023). Results demonstrate that dialectal variation was the main reason for accuracy reductions of up to 20%.

Variation in Basque Modern Basque dialects have been studied and categorized into a comprehensive representation of features by Zuazu (2008). In NLP, early works introduced a morpho-syntactically annotated corpus of Basque historical texts as an aid in the normalization process (Estarrona et al., 2020). Additionally, a corpus of syntactic variation of northern Basque dialects has been presented (Uria and Etxepare, 2012). More recently, Bengoetxea et al. (2025) presented the first manually created modern Basque dialect dataset for the evaluation of Natural Language Inference (NLI).

Finally, Basque dialects have also been considered in some dialectal benchmark works such as Alam et al. (2024) and Faisal et al. (2024), which presented benchmarks for Machine Translation (MT) with northern Basque dialects.

3. Data

This study examines physical commonsense reasoning in Italian and Basque. We employed GITA (Pensa et al., 2024a), an Italian dataset derived

[2]We use the terms the authors use in their papers.

Type	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
Plausible	George filled the glass with water.	George put the glass in the microwave.	George turned on the microwave.	The water got hot.	George put a tea bag in the water.
Implausible (order)	George put the glass in the microwave.	George filled the glass with water.	George turned on the microwave.	The water got hot.	George put a tea bag in the water.
Implausible (cloze)	George filled the glass with water.	George put the glass in the microwave.	George turned on the microwave.	The water got cold.	George put a tea bag in the water

Table 1: Example of a story with its plausible and implausible versions.

from TRIP (Storks et al., 2019), and manually translated it into Basque. GITA was selected as the foundation dataset due to its manual construction by a professional linguist with explicit attention to semantic coherence. Additionally, whereas TRIP incorporates compound sentences, GITA consists exclusively of simple sentences. This structural simplification reduces linguistic complexity and potential subjectivity, thereby isolating physical reasoning capabilities from confounding syntactic factors during model evaluation.

The following section introduces GITA and the process of its adaptation into both standard and dialectal Basque.

3.1. GITA

GITA (Pensa et al., 2024a) is an Italian physical commonsense evaluation dataset which consists of 356 5-sentence stories, out of which 117 are plausible, and 239 are implausible. The stories focus on concrete actions of the physical world, while mental actions such as “to think” or “to like” are avoided.

Two methods were used to create the implausible stories: (i) **Order**, where the order of two sentences was switched, and (ii) **Cloze**, where a plausible sentence has been substituted with an implausible one.

Consequently, individual sentences within the narratives are independently plausible but become implausible when placed with specific sentences in implausible narrative sequences. This design ensures that the reasoning task requires the use of the entire context.

In Table 1 we present an example translated into English. The plausible line contains the story with the logical sequence of events. In the implausible (order) example, the order of sentence 1 and 2 has been switched to make a non-logical and implausible story, and in the implausible (cloze) example, the adjective in sentence 3 has been changed from the original *hot* to *cold*, which makes no logical sense as the microwave heats water up.

3.2. BasPhyCo

BasPhyCo is the first non-QA physical commonsense reasoning dataset for Basque, available in both standard and dialectal variants. BasPhyCo has been created by manually translating GITA from Italian to Standard Basque.

The translation process included a localization phase in which two linguists adapted cultural elements of GITA to align with Basque contexts. These adaptations included proper names and references to local meteorological agencies, among other culturally-specific elements. The translations adhered closely to standard Basque conventions, specifically excluding lexical items characteristic of Basque dialectal variants.

3.3. BasPhyCo$_{\textbf{west}}$

The Standard Basque dataset was automatically converted to Western Basque using a few-shot prompting strategy implemented with the Latxa-3.1-Instruct model (Sainz et al., 2025). Western Basque was selected for two reasons: (1) as a peripheral dialect, it exhibits substantial linguistic distance from Standard Basque, making it a valuable subject for comparative analysis; and (2) preliminary experiments with LLM-based automatic adaptation of GITA revealed a consistent tendency towards Western Basque generation. This methodology leveraged Latxa’s perceived tendency to generate Western dialect while accounting for its perceived divergence from standard Basque. The conversion prompt is provided in Appendix A.

Given that plausible and implausible story pairs contain identical sentences (with the exception of one sentence in cloze implausible narratives),

Standard	Dialectal
Jonek lorategi handi bat dauka. Elur lorategian dago. Jonek lorategiko atea ireki du. Elurrek alde egin du. Lorategia hutsik dago.	Jonek lorategi handi bat dauko. Elur lorategixen dago. Jonek lorategiko atie zabaldu dau. Elurrek ospa egin dau. Lorategixe hutsik dago.

Table 2: Example of a story adapted from Standard Basque to Western Basque.

the adaptation process grouped each plausible story with all corresponding implausible variants. The conversion prompt explicitly instructed consistent adaptation of repeated sentences across variants. This methodology ensured uniformity in the adapted narratives.

An example of this adaptation can be found in Table 2, where words like *lorategia* (garden) have been adapted to its Western form *lorategixa*, as well as auxiliary verbs such as *du* have been adapted to *dau*.

A native professional linguist validated the automatic adaptations to assess overall quality (including minor formatting issues mitigated through prompt engineering) and identify dialectal adaptation errors. The subsequent subsections detail the findings from this initial manual inspection.

However, further evaluation of the quality of this automatic evaluation would be required in the future, in order to assess dialectal adaptation abilities of LLMs.

3.3.1. Correct Adaptations

During the manual evaluation step, different types of dialectal linguistic modifications were identified.

Lexical features Some lexical changes found to correspond to the Western dialect include *itzali* > *amatatu* (to switch off), *galtzak* > *prakak* (trousers) or *jolas egin* > *olgetan egin* (to play), to name a few. Not only that, but many words have also displayed Western phonology features, such as *ordulari**XE*** > *ordulari**A*** (clock) or *salda* > *saldea* (soup).

Morphosyntactic features Some common Western morphosyntactic characteristics include the comitative (*norekin*, with what/who) case marker, which in Standard Basque is marked with -KIN, while in the Western dialect this case is represented with the termination -GAZ, as in the following example: *aterkiare**KIN*** > *aterkixe**GAZ*** (with the umbrella).

In terms of auxiliary verb forms, the majority of them have been adapted into the Western dialect, such as *da* > *dau*, *nuen* > *neban*, *ditut* > *dodaz*, to name but a few.

3.3.2. Incorrect Adaptations

During the manual inspection of the adapted dialectal sentences, we found the following errors.

Lexical deviations Some sentences contained made-up words that looked like dialectal words, such as *mugikorra* (phone, standard) > **mobillora* or *tomate* (tomato, standard) > **totame*. These lexical adaptations are not part of the Western dialectal vocabulary and could be considered examples of model hallucination. However, they represent a very minimal part of the whole dataset.

Additionally, some words contained changes that mimic Western dialectal phonology (e.g. *baten* > **paten*, *sagar* > **saga*), but are not in fact a part of Western dialectal phonological changes.

Morphosyntactic deviations Although sentences generally follow dialectal morphosyntactic patterns, some outputs are not aligned with known dialectal features. For instance, some sentences with missing or additional ergative markers were found: the sentence **Teknikarixa**K** ez dau oraindiño etorri*[3], has an extra ergative marker -K, as intransitive verbs do not need this marker.

Additionally, some sentences contained verb concordance mismatches, such as *****indiolarrak*** *hartu* ***dau***[4], where the noun the verb is referring to is plural, but the verb form is singular. Thus, the preferred form would be ***indiolarrak*** *hartu* ***dauz***

The observed morphosyntactic divergences are not attested in dialectal corpora, suggesting they stem from the model's generalization errors rather than dialectal norms.

In Figure 1, we illustrate differences and similarities between the Standard and Western Basque datasets, highlighting both their lexical overlaps and divergences. While a portion of the vocabulary is shared between the two varieties, the analysis reveals that there is a substantial part of the lexicon that differs. We additionally present a series of contrastive examples that exemplify the most salient lexical and orthographic variations across the datasets.

4. Experimental Setup

This section presents the three evaluated tasks and their associated metrics, followed by a description

[3]Translation: the technician has not arrived yet.
[4]Translation: [someone] has taken the turkey.

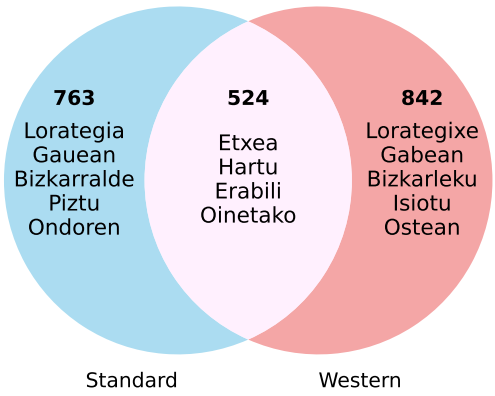

Figure 1: The number of unique words in BasPhyCo (left) and BasPhyCo$_{west}$ (right), as well as the overlap of both datasets (middle). Additionally, some examples from each dataset.

of the selected models and evaluation framework.

4.1. Task description

Our setup is based on GITA4CALAMITA, a GITA version which was adapted to work with generative LLMs for the CALAMITA shared task (Pensa et al., 2024b). This approach evaluated three different tasks, which were based on the mirroring of human reasoning, from the shallowest to the deepest. The evaluated tasks are the following:

- **Story classification** determines if the story is plausible or not. Continuing with the example in Table 1, the plausible story should be classified as plausible and the other two as implausible.
- **Conflict detection** involves identifying sentence pairs where the story becomes implausible. The conflicting sentences in the example of implausible-order in Table 1 are sentences 1 and 2, since once George puts the glass in the microwave, it is not logical to fill it with water.
- **Physical state classification** recognizes the involved physical states in the conflicting sentences of implausible stories. In the case of the example implausible-cloze in Table 1, the involved physical state is the temperature.

As in GITA4CALAMITA, we restrict the physical states to 14: location, conscious, dressed, wet, exist, clean, power, functional, in pieces, open, temperature, solid, occupied, and edible.

4.1.1. Data Annotation

We adopt the annotation from Pensa et al. (2024b), which was manually revised by a professional linguist. Some minor annotation errors were detected and corrected, such as occasional mislabeling between *cloze* and *order* story types.

The following is an example from the dataset and its annotation. Some relevant fields include *Type*, which can be *Null* for plausible stories, and *Order* or *Cloze* for implausible ones; *Confl_sents* and *Confl_pairs*, that indicate which sentences make the story implausible.


```
{
  "0-C0": {
    "story_id": 0,
    "type": "cloze",
    "sentences": [
     "Mikelek hozkailua ireki du
        .",
     "Mikelek esnea hartu du.",
     "Mikelek katilua hartu du.",
     "Mikelek goilara hartu du.",
     "Mikelek goilara katiluan
        sartu du."
    ],
    "length": 5,
    "example_id": "0-C0",
    "plausible": false,
    "breakpoint": 1,
    "confl_sents": [0],
    "confl_pairs": [0, 1]
  }
}
```


4.2. Metrics

To evaluate model performance, we adopt a tiered evaluation framework (Storks et al., 2021; Pensa et al., 2024b). In this setup, each task is evaluated conditionally on the success of the previous one, forming a crescendo of increasingly demanding reasoning requirements. Specifically, only the correctly solved instances from one level are used as input to the next. Accordingly, we adopt three complementary metrics for the three evaluated tasks:

- **Accuracy**: Quantifies the proportion of the correctly identified plausible and implausible stories. This metric will be used in the story classification task.
- **Consistency**: Measures the proportion of the correctly identified plausible sentences and the conflicting ssentencesin the implausible stories. This measure aims to check the models' consistency when recognizing conflicts. Thus, this metric will be used to evaluate the conflict detection task.
- **Verifiability**: Evaluates the proportion of the correctly identified plausible sentences, the conflicting sentence and the underlying physical states. This shows that the detected conflict can be validated because the underlying implausible change of physical states has

Test data	Model	Accuracy	Consistency	Verifiability
GITA	Llama-3.1-8B-It	72.47	29.29	14.23
	Llama-3.1-70B-It	**87.64**	**65.27**	36.40
	Gemma-2-9B-It	71.91	35.98	16.74
	Gemma-2-27B-It	67.42	30.96	15.06
	Latxa-3.1-8B-It	67.42	23.85	13.39
	Latxa-3.1-70B-It	85.96	60.25	**40.17**
	Minerva-7B-It	38.20	1.67	0.00
	LlaMAntino-3-8B-It	58.99	18.41	7.95
BasPhyCo	Llama-3.1-8B-It	57.30	11.30	5.02
	Llama-3.1-70B-It	**84.83**	47.70	26.78
	Gemma-2-9B-It	66.01	25.10	7.53
	Gemma-2-27B-It	62.36	24.27	8.37
	Latxa-3.1-8B-It	65.45	23.43	8.79
	Latxa-3.1-70B-It	81.46	**48.12**	**30.54**
	Minerva-7B-It	36.52	2.09	0.42
	LlaMAntino-3-8B-It	37.64	3.77	1.67
$\text{BasPhyCo}_{\text{west}}$	Llama-3.1-8B-It	51.12	9.62	4.18
	Llama-3.1-70B-It	74.16	35.56	17.57
	Gemma-2-9B-It	64.61	21.34	5.44
	Gemma-2-27B-It	57.87	18.41	6.69
	Latxa-3.1-8B-It	63.48	17.99	9.21
	Latxa-3.1-70B-It	**80.34**	**46.86**	**28.03**
	Minerva-7B-It	38.48	2.51	0.42
	LlaMAntino-3-8B-It	36.24	2.93	1.67

Table 3: Overall results for Story Classification, Conflict Detection and Physical State Classification, measured by accuracy, consistency and verifiability, respectively. GITA: original Italian data; BasPhyCo: manually translated Standard Basque data; $\text{BasPhyCo}_{\text{west}}$: automatically adapted data into the Western dialect.

been correctly understood. This last metric will be used to evaluate the physical state classification task.

4.3. Evaluation Setup

We have evaluated our task on generative models, as previous works that evaluated discriminative models (Storks et al., 2019; Pensa et al., 2024a) were outperformed by generative models (Pensa et al., 2024b). The evaluation for the three tasks is implemented on EleutherAI's Language Model Evaluation Harness framework v0.4.9 (Gao et al., 2024). This system enables the evaluation of generative LLMs and tasks in a reproducible, automated, and systematic way. The experiments were carried out in a 3-example few-shot setting specified by Harness. All code and prompts are publicly available[5].

We evaluated all tasks across the three test datasets representing Italian and Standard and Western Basque (Section 3). The evaluation employed four multilingual models, Llama-3.1 of 8B and 70B parameters (Dubey et al., 2024) and Gemma-2 9B and 27B parameters (Team et al., 2024), alongside language-specific models pretrained on Italian (Minerva-7B (Orlando et al., 2024) and LlaMAntino-3-8B (Polignano et al., 2024)) and Basque, namely, Latxa-3.1-8B and Latxa-3.1-70B (Sainz et al., 2025). All models were instruction-tuned variants.

[5] https://github.com/hitz-zentroa/BasPhyCo

5. Results

We present the results for the three tasks in Table 3, for Italian (GITA), Standard Basque (BasPhyCo) and Western Basque ($\text{BasPhyCo}_{\text{west}}$).

Italian The multilingual Llama-3.1-70B-It model achieved the highest performance in accuracy and consistency metrics, while Latxa-3.1-70B-It outperformed other models in terms of verifiability. Conversely, Italian-pretrained models (Minerva-7B-It and LlaMAntino-3-8B-It) yielded the lowest performance across all evaluated tasks, with Minerva-7B-It showing notably inferior results compared to

Language	Model	Accuracy			Consistency		Verifiability	
		Order	Cloze	Plausible	Order	Cloze	Order	Cloze
GITA	Llama-3.1-8B-It	64.75	81.20	71.79	14.75	44.44	4.92	23.93
	Llama-3.1-70B-It	**88.52**	**95.73**	**78.63**	**55.74**	74.36	**25.41**	47.01
	Gemma-2-9B-It	49.18	81.20	86.32	15.57	57.26	2.46	31.62
	Gemma-2-27B-It	34.43	78.63	90.60	9.02	53.85	4.10	26.50
	Latxa-3.1-8B-It	55.74	68.38	**78.63**	13.11	35.04	6.56	20.51
	Latxa-3.1-70B-It	84.43	94.87	**78.63**	45.90	**75.21**	**25.41**	**55.56**
	Minerva-7B-It	13.11	13.68	88.89	0.00	3.42	0.00	0.00
	LlaMAntino-3-8B-It	38.52	61.54	77.78	7.38	29.91	1.64	14.53
BasPhyCo	Llama-3.1-8B-It	46.72	50.43	75.21	8.20	14.53	2.46	7.69
	Llama-3.1-70B-It	**79.51**	**88.03**	**87.18**	31.97	**64.10**	**18.03**	35.90
	Gemma-2-9B-It	51.64	72.65	74.36	13.11	37.61	3.28	11.97
	Gemma-2-27B-It	40.16	70.94	76.92	9.84	39.32	2.46	14.53
	Latxa-3.1-8B-It	48.36	67.52	81.20	10.66	36.75	4.10	13.68
	Latxa-3.1-70B-It	76.23	87.18	81.20	**36.07**	60.68	**18.03**	**43.59**
	Minerva-7B-It	11.48	16.24	82.91	0.82	3.42	0.00	0.85
	LlaMAntino-3-8B-It	6.56	12.82	94.87	0.00	7.69	0.00	3.42
$\text{BasPhyCo}_{\text{west}}$	Llama-3.1-8B-It	34.43	44.44	75.21	4.10	15.38	1.64	6.84
	Llama-3.1-70B-It	67.21	76.92	**78.63**	20.49	51.28	6.56	29.06
	Gemma-2-9B-It	66.39	68.38	58.97	13.93	29.06	3.28	7.69
	Gemma-2-27B-It	42.62	59.83	71.79	6.56	30.77	1.64	11.97
	Latxa-3.1-8B-It	50.00	64.96	76.07	8.20	28.21	3.28	15.38
	Latxa-3.1-70B-It	**78.69**	**83.76**	**78.63**	**32.79**	**61.54**	**17.21**	**39.32**
	Minerva-7B-It	22.13	17.95	76.07	1.64	3.42	0.82	0.00
	LlaMAntino-3-8B-It	4.92	8.55	96.58	0.00	5.98	0.00	3.42

Table 4: Fine-grained examples for all three metrics. GITA: original Italian data; BasPhyCo: manually translated Standard Basque data; $\text{BasPhyCo}_{\text{west}}$: automatically adapted data into the Western dialect.

LlaMAntino-3-8B-It.

Notably, Basque-trained Latxa models outperformed Italian-specific models when evaluated on Italian data. Specifically, the smaller Latxa-8B-It model, despite being comparable in size to the Italian models, consistently surpassed LlaMAntino-3-8B-It across all tasks. This performance advantage can be attributed to Latxa's continual pretraining approach (Etxaniz et al., 2024), which effectively mitigates catastrophic forgetting from its base model, Llama-2.

Standard Basque While Llama-3.1-70B-It obtained the highest accuracy score for story classification (84.83 vs 81.46), the Basque pretrained model Latxa-3.1-70B-It had higher scores for the other two more fine-grained metrics, consistency (47.70 vs 48.12) and verifiability (26.78 vs 30.54), respectively.

Western Basque Latxa-3.1-70B-It obtained the highest results across all metrics. Llama's performance drop from standard to dialectal data is worth mentioning, as all three metrics undergo important drops (84.83 vs 74.16 for accuracy, 47.70 vs 35.56 for consistency, 26.78 vs 17.57 for verifiability). With Latxa, although there is a performance drop from standard to dialectal, the drop is not nearly as dramatic (81.46 vs 80.34, 48.12 vs 46.86, 30.54 vs 28.03). These results highlight the importance of pretraining in the target language, as it appears to facilitate more fine-grained linguistic competence and enhance robustness to language variation.

Overall LLMs demonstrate notably poor performance in predicting *verifiable* instances for low-resource languages, with performance degrading further when applied to dialectal data. Regarding task-specific performance, Llama-3.1-70B-It exhibited optimal results in story classification for Italian and Standard Basque, whereas Latxa-3.1-70B-It demonstrated superior consistency and *verifiability*, particularly for Standard and Western Basque. These results indicate that pretraining on target language data yields more substantial improvements in complex reasoning tasks. Additionally, Latxa-3.1-70B-It achieved the highest performance in *verifiability*, which is the most cognitively demanding reasoning task across all evaluated languages.

P. state	total	GITA	BasPhyCo	BasPhyCo$_{west}$	Avrg
Open	88	20.45	19.32	12.50	17.42
Functional	53	26.41	15.09	16.98	19.49
Exist	47	27.66	23.40	21.28	24.11
Power	36	41.67	36.11	13.63	30.47
In pieces	35	42.86	25.71	31.43	33.33
Location	33	15.15	6.06	6.06	9.09
Edible	22	4.54	0.00	4.54	3.03
Conscious	13	7.69	7.69	15.38	10.25
Temperature	12	50.00	50.00	41.67	47.22
Wet	7	42.86	58.57	14.28	38.57
Solid	5	80.00	60.00	40.00	60.00
Wearing	3	0.00	0.00	0.00	0.00
Clean	1	0.00	0.00	0.00	0.00
Occupied	1	100.00	100.00	100.00	100.00

Table 5: Verifiability results per physical state. These results are for Latxa-3.1-70B-It, the model with the highest verifiability results for Italian, Standard and Western Basque.

Finally, the drop from the shallowest to the deepest reasoning task for all models is to be highlighted. Table 3 shows substantial performance degradation, especially in the physical state classification task (verifiability). These findings indicate that, although some models are able to identify implausible stories, providing explanations for their implausibility presents a considerably more challenging task. This will be further discussed in Section 6.

6. Discussion

In this section, we focus on more fine-grained results, as the three metrics have been specifically computed for the different types of implausible stories (order and cloze). This analysis aims to identify any possible biases that the models could have towards implausible story types.

The results for all three metrics, as well as for the different types of implausible stories, are presented in Table 4. The main finding indicates that order implausible stories consistently yield lower scores than cloze implausible stories across all metrics, models, and languages. This pattern suggests that the models exhibit stronger reasoning capabilities when confronted with a conflicting sentence within a narrative sequence, compared to cases where implausibility arises solely from the reordering of sentences. These results are consistent with the findings reported by Pensa et al. (2024b).

Italian and Standard Basque seem to follow similar patterns. Llama-3.1-70B-It obtains the highest results in the majority of the tasks and story types, only being surpassed by Latxa-3.1-70B-It in consistency and verifiability cloze story types. This suggests that, for Italian and Standard Basque, while Llama obtains higher results in shallower reasoning tasks (story classification), Latxa seems to perform slightly better in reasoning tasks involving physical state classification (verifiability).

Regarding the results for Western Basque, Latxa-3.1-70B-It outperforms all other models, including both multilingual models and those pretrained for Italian, following general results in Table 3.

Furthermore, the general decrease in performance observed for Llama-3.1-70B-It compared to the standard Basque results highlights the need for multilingual language models that could better handle Basque dialectal variation.

Finally, Latxa-3.1-70B-It consistently obtains high verifiability results for both order and cloze types, which is the metric that measures how much physical states are predicted correctly. This suggests Latxa's capacity to deal with deeper reasoning tasks such as physical state classification.

Per Physical State Label Verifiability In Table 5, we report the verifiability results for each physical state label across Italian, as well as Standard and Western Basque. Labels represented by fewer than ten instances are excluded from the following analysis, due to potential sampling bias. Consequently, the subsequent analysis focuses exclusively on those physical states with sufficient representation (i.e., more than ten instances), ensuring more reliable and interpretable results.

Overall, the findings indicate that no particular physical state is consistently easier to predict than the others. In general, performance across categories remains relatively low, highlighting both the intrinsic complexity of this reasoning task and the current limitations of LLMs in capturing nuanced physical state distinctions.

The predictions of Location, Edible, and Conscious states appear to be particularly challenging, as reflected by their comparatively lower verifiability scores. These results suggest that such categories may involve subtleties that LLMs struggle to capture effectively, possibly due to their dependence on implicit world knowledge.

7. Conclusion

This paper introduces a novel dataset for evaluating physical commonsense reasoning in Basque and its Western dialect. The dataset was derived from GITA, a manually curated Italian corpus, which underwent manual translation and localization into Standard Basque. Subsequently, the Standard Basque data was automatically adapted to the Western Basque dialect, followed by manual post-editing to ensure accuracy and minimize errors.

We have carried out a suite of experiments to see how multilingual and language-specific LLMs perform on the tasks of physical commonsense reasoning. To our knowledge, this is the first evaluation of non-QA physical commonsense reasoning in low-resourced languages such as Basque and its dialectal varieties. To that end, we have followed a

tiered strategy with three tasks of different depth levels: story classification, conflict detection and physical state classification. The results show the LLMs ability to predict verifiable instances is generally low, which highlights the need for further research in the field of physical commonsense reasoning. Further analysis has indicated that identifying implausible instances is more complex when the only change is sentence order. Finally, physical state classification remains a particularly challenging task.

This work establishes a baseline evaluation framework for commonsense reasoning in low-resource languages and dialectal varieties. Future research directions include extending the dataset to additional languages and dialects.

Limitations

The physical commonsense reasoning dataset that we present in this work can be culturally localized, reflecting the norms and logic of certain communities, and may need to be adapted to other cultures in order to be applicable in other contexts.

Additionally, the size of our dataset is currently limited. Expanding this test data as well as building a training set, could alleviate this issue.

Finally, it is important to recognize the inherent bias of Basque LLMs toward Western Basque. Current models show a strong tendency to generate Western Basque features, indicating that their training data and modeling are heavily aligned with this dialect. Expanding this ability to other dialects could enable the analysis of additional variations.

Acknowledgments

This work has been supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22). Jaione Bengoetxea is funded by the Basque Government pre-doctoral grant (PRE_2024_1_0028).

We also acknowledge the following MCIN/AEI/10.13039/501100011033 project: (i) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR and (ii) DeepThought (PID2024-159202OB-C21) funded by ERDF, EU.

Bibliographical References

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical Reasoning about Objects through Space and Time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. PHYRE: A New Benchmark for Physical Reasoning. *Advances in Neural Information Processing Systems*, 32.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A Benchmark for LLM Evaluation in Iberian Languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2025. Lost in Variation? Evaluating NLI Performance in Basque and Spanish Geographical Variants. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 452–468, Vienna, Austria. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Ernest Davis. 2023. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Comput. Surv.*, 56(4).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano, and Ander Soraluze. 2020. Dealing with Dialectal Variation in the Construction of the Basque Historical Corpus. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 79–89, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Ainara Estarrona, Izaskun Etxeberria, Manuel Padilla-Moyano, and Ander Soraluze. 2023. Measuring Language Distance for Historical Texts in Basque. *Procesamiento del Lenguaje Natural*, 70:53–61.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972.

Fahin Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *ArXiv*, abs/2403.11009.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The Language Model Evaluation Harness.

Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. 2021. PTR: A Benchmark for Part-based Conceptual, Relational, and Physical Reasoning. *Advances in Neural Information Processing Systems*, 34:17427–17440.

Jihae Jeong, DaeYeop Lee, DongGeon Lee, and Hwanjo Yu. 2025. Everyday Physics in Korean Contexts: A Culturally Grounded Physical Reasoning Benchmark. *arXiv preprint arXiv:2509.17807*.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J Wooldridge, Janet Pierrehumbert, and Furu Wei. 2025. Assessing Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not Written in Text: Exploring Spatial Commonsense from Visual Signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. 2024. PhyBench: A Physical Commonsense Benchmark for Evaluating Text-to-Image Models. *CoRR*.

Luis Mitxelena. 1981. Lengua común y dialectos vascos. *Anuario del Seminario de Filología Vasca" Julio de Urquijo"*, 15:289–313.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. 2025. Do Generative Video Models Understand Physical Principles? *arXiv preprint arXiv:2501.09038*.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719. CEUR Workshop Proceedings.

Eileen Pan, Anna Seo Gyeong Choi, Maartje ter Hoeve, Skyler Seto, and Allison Koenecke. 2025. Analyzing Dialectical Biases in LLMs for Knowledge and Reasoning Benchmarks. *arXiv preprint arXiv:2510.00962*.

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. 2024a. A Multi-layered Approach to Physical Commonsense Understanding: Creation and Evaluation of an Italian Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 819–831.

Giulia Pensa, Ekhi Azurmendi, Julen Etxaniz, Begoña Altuna, and Itziar Gonzalez-Dios. 2024b. GITA4CALAMITA-Evaluating the Physical Commonsense Understanding of Italian LLMs in a Multi-layered Approach: A CALAMITA Challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1153–1160.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *ArXiv*, abs/2405.07101.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. ESPRIT: Explaining Solutions to Physical Reasoning Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7906–7917, Online. Association for Computational Linguistics.

Oscar Sainz, Naiara Pérez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, Germán Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for basque. *ArXiv*, abs/2506.07597.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60.

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.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. 2025. A Survey of Reasoning with Foundation Models: Concepts, Methodologies, and Outlook. *ACM Computing Surveys*, 57(11):1–43.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Larraitz Uria and Ricardo Etxepare. 2012. Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: Basyque aplikazioa. *Lapurdum. Euskal ikerketen aldizkaria| Revue d'études basques| Revista de estudios vascos| Basque studies review*, (16):117–135.

Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. 2023. NEWTON: Are Large Language Models Capable of Physical Reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9743–9758, Singapore. Association for Computational Linguistics.

Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Benchmarking Progress to Infant-level Physical Reasoning in AI. *Transactions on Machine Learning Research*.

Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. PACS: A Dataset for Physical Audiovisual CommonSense Reasoning. In *European Conference on Computer Vision*, pages 292–309. Springer.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A Framework for Cross-Dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

Koldo Zuazu. 2008. *Euskalkiak. Euskararen dialektoak*. Elkar.

A. Automatic Adaptation Prompt

Figure 2 provides the prompt that was used to obtain the standard-to-dialectal adaptations of the Basque dataset.

I will give you three versions of a story. Each version has five sentences. Some sentences are identical across versions. You need to adapt this text so that it includes Bizkaian dialectal features. You can use non-standard orthography. Try to make it as similar as possible to oral language.
Task:

1. First, list all unique sentences across all three stories.
2. Adapt each unique sentence exactly once into the Bizkaian dialect.
3. Then reconstruct the three stories with the translations, making sure that any identical source sentence always has the identical translation.
4. If there are more than three stories, repeat the same process for all of them.

Format:
This is an example of an standard (INPUT) instance and an example of the dialectal (OUTPUT) adaptation that you need to do:
Standard:
STORY1: ['Mikel lanera joan da', 'Mikelek ordenagailua piztu du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']
STORY2: ['Mikel lanera joan da', 'Mikelek mezuak erantzun ditu', 'Mikelek mezuak irakurri ditu', 'Mikelek ordenagailua piztu du', 'Mikel etxera joan da']
STORY3: ['Mikel lanera joan da', 'Mikelek ordenagailua itzali du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']
Dialectal:
STORY1: ['Mikel lanera jun de', 'Mikelek ordenagaillua piztu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']
STORY2: ['Mikel lanera jun de', 'Mikelek mesuek erantzun ditu', 'Mikelek mesuek irakurri dauz', 'Mikelek ordenagaillua piztu dau', 'Mikel etxera jun de']
STORY3: ['Mikel lanera jun de', 'Mikelek ordenagaillua amatatu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']
Output only the reconstructed stories in the exact same format as the input. Do not output explanations, steps, or commentary.

Figure 2: Dialectal adaptation prompt.