EA-PIQA: Swahili, Dhuluo, Luganda, Lingala

Anonymous EMNLP submission

Abstract

Physical commonsense reasoning allows people to anticipate outcomes, select tools, and adapt to daily environments, yet most benchmarks remain English-only, limiting their cultural and linguistic coverage. We introduce a multilingual dataset of PIQA-style physical commonsense reasoning items in Swahili, Dholuo, Lingala, and Luganda, four widely spoken but underrepresented African languages. The dataset combines original examples authored by native speakers to capture culturally grounded practices such as cooking on charcoal stoves, farming, cassava preparation, and river transport with adapted translations of PIQA items. In total, it contains over 29,000 examples, including 544 original prompts. All items were validated by native speakers to ensure linguistic accuracy and physical plausibility, with adaptations made to address cultural mismatches, morphological constraints, and code-mixing. Our contributions are: (1) the first PIQA-style dataset for African languages, (2) a documented methodology for community-driven dataset construction, and (3) a resource release through the MRL 2025 Shared Task to enable systematic multilingual evaluation of physical commonsense reasoning.

1 Introduction

Physical commonsense reasoning is an essential part of everyday life: it enables people to anticipate the outcomes of actions, select appropriate tools, and adapt to changing environments. In language, such reasoning appears when speakers describe how to cook, build, repair, or carry out tasks with the resources around them. Capturing this knowledge in computational benchmarks is crucial for building language technologies that can reason in ways aligned with human experience, from voice assistants to safety-critical systems. Commonsense reasoning has been studied formally as the "skeleton" of reasoning processes, unifying different modes of logic such as Boolean and fuzzy reasoning (Trillas et al., 2022).

Over the past decade, numerous benchmarks have been developed to measure commonsense reasoning in AI systems (Davis, 2023). Among these, physical commonsense reasoning benchmarks have played a central role in testing how well models understand the properties, uses, and interactions of everyday objects (Bisk et al., 2020). Yet researchers have shown that these datasets often contain annotation artifacts or superficial cues that models can exploit as shortcuts, overstating their true reasoning ability (Mok and Kim, 2023). More recent work has emphasized the importance of multimodal physical commonsense reasoning, incorporating audio and visual signals to move beyond text-only settings (Yu et al., 2022; Zong et al., 2025). Despite these advances, nearly all resources remain focused on English or on multimodal reasoning in globalized contexts.

This focus overlooks the cultural and linguistic diversity that shapes how people conceptualize everyday actions. What counts as the most practical or safe solution often depends on locally available materials, tools, and customs—for example, choosing between a charcoal stove and an electric cooker, or carrying water in a jerrican rather than a bucket. For African languages in particular, the gap is striking: everyday practices such as cassava preparation, small-scale farming, and household repairs rarely appear in English-centric datasets. Without culturally grounded benchmarks, evaluation risks producing a narrow and misleading picture of commonsense reasoning that does not generalize across contexts.

To address this challenge, we introduce a new multilingual dataset of PIQA-style physical commonsense reasoning items in **Swahili, Dholuo, Lingala, and Luganda**. These four languages were selected because they are widely spoken across East and Central Africa, represent distinct linguistic families, and remain underrepresented in existing NLP resources. Our dataset combines:

- **Original examples** written by native speakers to capture culturally specific practices, and

- **Translated PIQA examples** carefully adapted to each language.

All items were manually reviewed and labeled by native speakers to ensure both linguistic accuracy and physical plausibility.

Our contributions are threefold:

1. Introduce the first PIQA-style dataset for four African languages, grounded in East and Central African contexts.

2. Document the methodology for generating, translating, and validating items with native speakers.

3. Release the dataset as part of the MRL 2025 Shared Task, enabling systematic evaluation of multilingual models on culturally grounded physical commonsense reasoning.

This work highlights the importance of creating benchmarks that go beyond English and reflect the physical and cultural realities of diverse language communities. By centering African languages, we broaden the scope of multilingual NLP evaluation and contribute to more inclusive approaches to commonsense reasoning.

2 Dataset Construction

2.1 Language Selection

We selected four African languages spoken across East and Central Africa: **Swahili, Dholuo, Lingala, and Luganda**. These languages represent distinct linguistic families and cultural contexts, while also being widely used in education, media, and everyday life. By spanning both Bantu and Nilotic families, the dataset captures differences not only in vocabulary but also in how everyday actions are conceptualized.

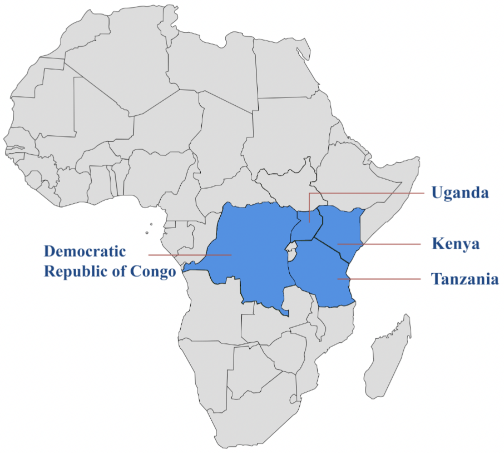

Figure 1: Geographic distribution of the languages in our dataset. Highlighted countries (Tanzania, Kenya, Uganda, and the Democratic Republic of Congo) correspond to the primary regions where Swahili, Dholuo, Lingala, and Luganda are spoken and where contributors are based.

- **Swahili** is a Bantu language with more than 100 million speakers across East Africa and serves as a lingua franca in Kenya, Tanzania, Uganda, and the Democratic Republic of Congo. While Swahili is structurally similar across the region, our dataset distinguishes between Tanzanian and Kenyan varieties. Tanzanian contributions emphasize domestic and rural practices, while Kenyan contributions highlight more urbanized contexts. This ensures representation of not only linguistic similarity but also regional diversity in cultural preferences, food names, household practices, and phrasing conventions.

- **Dholuo** (also known as Luo or Dhuluo) is a Nilotic language spoken by the Luo community in western Kenya, with over 5 million speakers.

- **Lingala** is a Bantu language spoken by an estimated 40 million people in the Democratic Republic of Congo and the Republic of Congo, widely used in music, trade, and urban communication.

- **Luganda** is spoken by over 10 million people in Uganda and serves as a major language of commerce, education, and media.

2.2 Prompt and Solution Generation

Following the PIQA format, each item consists of a *prompt* describing an everyday physical task and two candidate *solutions*, which differ minimally in wording but diverge in physical plausibility. The correct solution is marked with a binary label (0 or 1).

Dataset construction combined both **original** and **translated** examples:

- **Swahili:** Original prompts were developed across ten thematic categories, including *Food & Cooking, Agriculture, Household Care, Construction, Health & Hygiene, Transport, and Technology*. Contributions were collected from both Tanzania and Kenya. Translations from PIQA were also adapted to local contexts using neural machine translation and manual review. Tanzanian examples often referenced staple food preparation and rural household routines, such as cooking *ugali* on charcoal stoves. By contrast, Kenyan examples included additional scenarios around urban services (e.g., paying electricity bills) and public transport. This contrast illustrates how a shared language encodes region-specific practices, giving the dataset broader cultural grounding.

- **Dholuo:** Original prompts were created through brainstorming and localization heuristics. Examples covered agriculture (e.g., harvesting groundnuts), food preparation (e.g., cooking millet porridge), household activities, childcare, and paying electricity bills. PIQA items were translated and localized for cultural relevance.

- **Lingala:** Prompts reflected a "villager's perspective" on Central African daily life, including cassava preparation, fishing, river transport, market trading, and hut construction. Culturally specific practices (e.g., preparing flying termites as food) were included. Code-mixing with French required additional linguistic consultation.

- **Luganda:** PIQA examples were translated using a high-quality Luganda NMT system (the "Crane model"). These translations were then curated for grammatical correctness and naturalness. No original items were authored in Luganda.

2.3 Curation and Validation

Each example was manually checked by at least one native speaker of the target language. Reviewers ensured that:

1. Prompts were grammatically correct and natural.

2. Candidate solutions were syntactically similar but diverged in physical plausibility.

3. The labeled correct solution reflected unambiguous physical reasoning.

Table 1: Dataset statistics by language

Language	Original	Translated	Total
Swahili (TZ)	175	14,444	14,619
Swahili (KE)	100	–	100
Dholuo (KE)	100	∼200	∼300
Lingala (DRC)	169	26	195
Luganda (UG)	0	14,444	14,444
Total	544	∼29,114	∼29,658

Ambiguous or inconsistent items were revised collaboratively. For translated items, reviewers adapted meanings to fit local contexts. Specific challenges included the lack of standardized technical vocabulary in Dholuo and pervasive French code-mixing in Lingala. Following recommendations in prior work (Mok and Kim, 2023), distractor solutions were designed to avoid annotation artifacts and shortcut patterns.

3 Dataset Statistics

Table 1 summarizes the dataset contributions across languages. We report both original and translated items, highlighting the balance between community-authored examples and adapted PIQA translations.

4 Linguistic and Cultural Considerations

A central motivation for this dataset is that physical commonsense reasoning is not culture-neutral. The ways in which people cook, transport, build, or repair objects are deeply shaped by their environments, available resources, and linguistic traditions. For example, in many African contexts, carrying water in a jerrican, preparing food on a charcoal stove, or building with mud and thatch are everyday practices that differ fundamentally from the assumptions embedded in English-centric resources. Benchmarks designed only in English therefore fail to capture many distinctions that are obvious to speakers of other languages but invisible in translated datasets.

In constructing this dataset, contributors intentionally incorporated culturally grounded practices specific to their communities:

- **Swahili and Dholuo:** Items frequently reference staple foods such as maize flour, *ugali*, and *sukuma wiki*, alongside household tasks like cooking on charcoal stoves and collecting water in jerricans. These activities reflect East African daily life and create reasoning challenges around heat, fire safety, water transport, and food preparation that do not map neatly to English scenarios. For Swahili specifically, we observed that Tanzanian contributions emphasized rural domestic practices, whereas Kenyan contributions more frequently reflected urban and service-oriented contexts. This highlights how a single language can encode region-specific commonsense knowledge, enriching evaluation beyond purely linguistic variation. In Dholuo, prompts also drew from agriculture (e.g., groundnut harvesting), childcare, and first aid practices, reflecting the lived experiences of rural and urban households alike.

- **Lingala:** Examples highlight practices from Central Africa, including cassava preparation, fishing techniques, and river-based transport. Physical commonsense in these contexts involves tools such as dugout canoes, woven baskets, or traditional cooking implements rarely represented in global datasets. Contributors also incorporated culturally specific practices such as preparing flying termites as food, demonstrating the importance of embedding local ecological knowledge into evaluation tasks.

- **Luganda:** Although all items were translated from PIQA, the process revealed subtle cultural and linguistic adaptations. Certain prompts referenced objects or environments unfamiliar to Luganda speakers (e.g., “snow boots” or “microwave ovens”). These were replaced with more contextually relevant items such as rubber boots or charcoal stoves. Even translated items thus illustrate how commonsense reasoning must be anchored in local realities rather than assumed to be universally shared.

We also observed that linguistic structure itself shaped how prompts and solutions could be framed. In Swahili and Lingala, verb morphology allowed fine-grained distinctions in outcomes, such as whether an action was completed correctly, safely, or improperly. Dholuo’s Nilotic grammar facilitated rich causal structures, enabling contributors to express why one action leads to success while another fails. Luganda’s agglutinative morphology required careful word choice to preserve syntactic similarity between solutions, ensuring that differences reflected only physical plausibility rather than grammatical structure.

These linguistic and cultural dimensions underscore that evaluation datasets must extend beyond direct translation. To be meaningful, they must encode the lived practices, ecological contexts, and linguistic resources of the communities they represent. Without such grounding, evaluation risks reinforcing a narrow view of commonsense that fails to generalize across the multilingual and multicultural realities of human life.

5 Discussion

Our dataset represents one of the first systematic efforts to extend physical commonsense reasoning benchmarks to African languages. Through its construction, several key insights emerged that point to both the opportunities and challenges of multilingual benchmark creation.

- **Value of original examples:** Community-authored items provided stronger cultural grounding than direct translations of PIQA. For example, Swahili and Dholuo contributors introduced prompts about local foods, farming tools, and daily survival tasks that are absent from English datasets. These examples demonstrate the irreplaceable value of native speakers’ knowledge in ensuring that evaluation reflects real-world practices rather than abstracted, Western-centric scenarios.

- **Challenges of translation:** Direct translations often produced cultural mismatches or introduced irrelevant objects. Adapting PIQA items required replacing or re-contextualizing references while preserving the structure of the original task. This adaptation process highlights the difficulty of assuming universality in physical commonsense benchmarks and reinforces the importance of localization heuristics.

- **Annotation reliability:** Despite careful review, some prompts remained ambiguous or allowed multiple plausible solutions depending on interpretation. For instance, whether food should be dried in the sun or covered might depend on seasonal weather conditions. Collaborative discussion among annotators was essential for resolving these cases and ensuring consistency.

- **Resource implications:** Creating original items is resource-intensive, requiring both time and cultural expertise. Our multilingual approach distributed the workload across contributors in different regions, but scaling up will demand broader community participation and potentially structured annotation frameworks. At the same time, large-scale machine translation enabled rapid dataset expansion for Swahili and Luganda, illustrating the potential of hybrid approaches that balance manual authorship with automated support.

Looking ahead, this dataset provides an initial but significant step toward broader multilingual evaluation resources. Future directions include increasing the number of examples per language, expanding to additional African languages, and incorporating multimodal contexts such as images or audio that better capture the multisensory nature of physical reasoning. Importantly, this project demonstrates that community-driven dataset creation is both feasible and impactful. By foregrounding the perspectives of underrepresented language communities, we not only expand the coverage of commonsense reasoning benchmarks but also contribute to a more inclusive and globally relevant foundation for NLP research.

6 Conclusion

In this paper, we presented a multilingual dataset for physical commonsense reasoning in Swahili, Dhuluo, Lingala, and Luganda, created as part of the MRL 2025 Shared Task. Unlike prior benchmarks that focus primarily on English, our dataset combines original items grounded in culturally specific practices with adapted translations of PIQA examples. This work highlights the importance of evaluation resources that reflect not only linguistic diversity but also the everyday activities and reasoning patterns of different communities. By situating prompts in domains such as food preparation, farming, household repair, and transport, we ensure that the dataset is both relevant and authentic to the experiences of African language speakers. Looking forward, we aim to expand the dataset to include more items and additional languages, conduct systematic evaluations of multilingual models, and explore multimodal extensions that capture aspects of physical reasoning beyond text. We hope that this collaborative, community-driven effort will inspire further research into culturally grounded benchmarks and strengthen the representation of African languages in global NLP.

References

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

E. Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Computing Surveys*, 56:1 – 41.

Wing-Lam Mok and SungHo Kim. 2023. Mitigating annotation artifacts in physical commonsense reasoning benchmark. *2023 IEEE International Conference on Big Data (BigData)*, pages 6236–6238.

E. Trillas, S. Termini, and M. Tabacchi. 2022. A formal skeleton of commonsense reasoning. *Studies in Computational Intelligence*.

Samuel Yu, Peter Wu, Paul Pu Liang, R. Salakhutdinov, and Louis philippe Morency. 2022. Pacs: A dataset for physical audiovisual commonsense reasoning. pages 292–309.

Daoming Zong, Chaoyue Ding, Kaitao Chen, Yinsheng Li, and Shuaiyu Wang. 2025. Counterfactual debiasing for physical audiovisual commonsense reasoning. pages 15265–15273.