Physical Commonsense Reasoning For West African Languages

Esther Adenuga[1], Akintunde Oladipo[1], Odunayo Ogundepo[1],
Ismail Daud[3], Deborah Popoola[2], Adeyemi Praise[2]
Okechukwu God'spraise[2], Olasoji Akindejoye[3], Abolade Daud[5]
Rapheal Afolayan[4], Sharon Ibejih[2], Cynthia Amol[2]
[1]The African Research Collective [2]Tonative [3]University of Ibadan [4]University of Ilorin [5]Masakhane
Correspondence:{akintunde.oladipo, odunayo.ogundepo}@taresco.org

Abstract

We present a physical commonsense reasoning test collection for two West African languages, Yoruba and Nigerian Pidgin, developed for the MRL 2025 Shared Task on Multilingual Physical Reasoning. Our test collection contains 400 instances (200 per language), each consisting of a short goal paired with two candidate solutions and a gold-labeled correct answer.

We employ three complementary methods to construct these instances: **instruction-based** (following explicit directions), **activity-based** (completing everyday tasks), and **object-based** (interacting with physical items). All instances were manually created and annotated by native speakers and then validated for grammatical correctness and cultural relevance. The dataset provides a challenging benchmark for evaluating multilingual and culturally grounded physical commonsense reasoning, and is publicly released to support research on multilingual LLMs.

1 Introduction

As frontier large language models (LLMs) push the boundaries of performance on existing benchmarks for African languages (Ojo et al., 2025; Adebara et al., 2025), there is a pressing need for new test collections not only probe surface-level understanding but also evaluate more complex reasoning capabilities. Current evaluations, however, broadly focus on classification (Adelani et al., 2023; Yu et al., 2025; Muhammad et al., 2025), translation (Goyal et al., 2022; Adelani et al., 2022) and factual question answering (Ogundepo et al., 2023; Bandarkar et al., 2024; Bayes et al., 2024). While these tasks are valuable, they do not fully capture deep reasoning or sophisticated understanding of language-specific and culturally grounded contexts. Recent efforts such as AfriMGSM and AfriMMLU (Adelani et al., 2024) advance this landscape by translating existing benchmarks targeting mathematical reasoning. However, commonsense reasoning, particularly reasoning about everyday physical interactions, object properties, and practical problem-solving remains critically underexplored. in African languages. Such capabilities are essential for AI systems to understand and assist with real-world tasks in African contexts, and datasets like this help us truly understand how well existing LLMs may perform.

In this work, we introduce a physical commonsense reasoning test collection for Yoruba (yor) and Nigerian Pidgin (pcm) — two languages with over 100M speakers in West Africa. Our dataset was sourced entirely from native speakers using a systematic three-step approach. First, we collected comprehensive lists of physical items that people interact with daily in both language communities, sourcing these items from YouTube videos, language dictionaries, Facebook discussions, and other culturally relevant sources. Second, for each physical item, we created realistic scenarios where speakers would likely use or interact with the object in their everyday lives. Finally, we framed each scenario as a practical goal accompanied by two candidate solutions that require physical reasoning to identify the correct choice.

We believe this approach offers a reproducible pipeline that can be extended to create similar datasets for any language, provided reliable sources of physical objects and their cultural descriptions are available for that linguistic community.

2 Methodology

This section outlines our methodology for developing the physical commonsense reasoning dataset, covering the creation, annotation, and review process, and the final dataset statistics.

2.1 Dataset Creation

We developed a reproducible approach to guide dataset creation while providing contributors the

Language	Object-Based	Activity-Based	Instruction-Based
yor	**g:** Ataalẹ́ jẹ́ èròjà fún oúnjẹ wo? **s0:** Iyán tí a fi iṣu ṣe **s1:** Ògì tí a fi bàbà tàbí àgbàdo ṣe	**g:** Tí mo bá fẹ́ dìbò, ìka wo ni o màá lò? **s0:** Àtàǹpàkò ni ìka tí ó yẹ láti dìbò **s1:** Ìka àárín ni ó yẹ láti dìbò	**g:** Láti se ẹyin, èwo ninú àwọn igbẹ́sẹ̀ wọ̀nyì ni ó yẹ kí o tẹ̀lẹ́? **s0:** fi ẹyin sínu omi, gbé e ka ná, sì jẹ́ ki o sè fún bi iṣẹ́jú mẹ́wàá **s1:** fi ẹyin sinu òróró, gbé e ka na, sì jẹ́ ki ó sè fún bi iṣẹ́ju mẹ́wàá
pcm	**g:** How you go use dustbin well for compound? **s0:** I go scatter di dirty all around compound and leave di dustbin empty **s1:** I go gather dirty inside dustbin so refuse collector fit carry am	**g:** You wan take selfie, wetin you go on? **s0:** I go on my camera use take selfie **s1:** I go on my torchlight use take selfie	**g:** If you wan send mail for Gmail, how you go run am? **s0:** Open Gmail app, make you just click for di place wey you see compose, afta that one put di email of di pesin wey you wan send di mail to, type your message, make you now click on send **s1:** Open Gmail app, make you just click for di place wey you see trash, afta that one put di email of di pesin wey you wan send di mail to, type your message, make you now click on send

Table 1: **Qualitative Analysis** Samples of the PIQA task in Yorùbá and Nigerian Pidgin

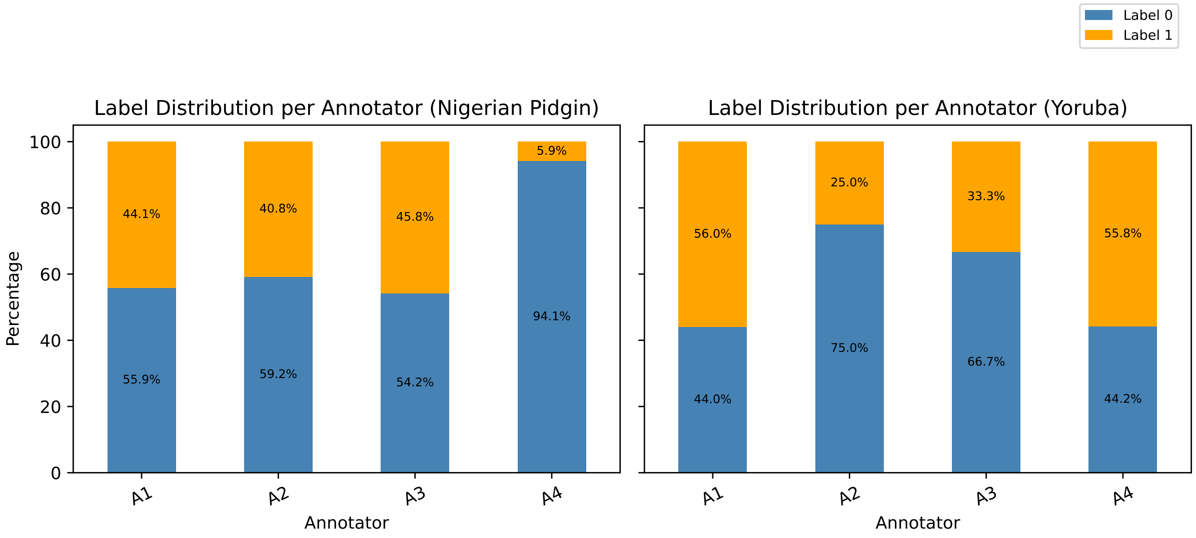

Figure 1: Label distribution per annotator for Yoruba and Nigerian Pidgin. Percentages indicate the proportion of each label assigned by individual annotators.

flexibility to adapt and document additional methods as needed. Our systematic approach consisted of three core steps:

1. **Reference Object Collection:** We collected a list of reference objects and activities from diverse, publicly available sources including language dictionaries, YouTube videos, social media platforms (e.g., X, Facebook), and other culturally relevant materials. These collection include everyday objects, routine activities, occupation-specific actions, and descriptive elements such as places, colors, and shapes. A complete list of sources is provided in Table 2.

2. **Scenario development:** Contributors created realistic, culturally grounded scenarios for each compiled object or activity. To guide this process, We encouraged contributors to consider two key prompts such as "What is this object typically used for?" and "What approach might seem plausible but would actually fail in this context?" This systematic questioning helped identify effective distractors—incorrect but believable alternatives that require genuine physical reasoning and understanding to distinguish from correct solutions.

3. **Instance structuring:** Each scenario was framed as a practical goal accompanied by two candidate solutions (`solution0` and `solution1`), with the correct solution explicitly annotated in the `label` field. This binary choice format ensures consistent evaluation while maintaining the cognitive complexity required for physical reasoning assessment.

To facilitate creation and ensure variety across the dataset, we designed three templates: object-based, activity-based, and instruction-based. These templates provided structured guidance while allowing flexibility in implementation

Object-based template: Entries revolve around everyday items and their typical functions. An entry begins with a physical object linked to a realistic everyday use case or scenario, followed by two candidate solutions, one correct and one plausible but incorrect.

Object → ***Scenario*** → ***Options/Distractor*** → ***Example***

Language	Resource Type/Name	Source
Yoruba	Dictionary	Dictionary of Yoruba Language
	X/Twitter	The Yoruba Nation CH
	Facebook	Eating Utensils in Yoruba Language
		Àwon Orisi ohun-ise -tools
	Language Drops	Home Appliances in Yoruba
		Fruits and Vegetables in Yoruba
	Youtube(Yoruba Gbode)	Common Things In The Kitchen And What They Are Called In Yoruba
	Youtube(Diaspora Speaks Yoruba)	Yoruba Finger Names | Oruko awon ika ni ede Yoruba
	Wikidot	Types of Tools
		Yoruba Science and Technology Wikipedia
	Brainscape	Yoruba Words
Nigerian Pidgin	Dictionary	Pidgin Dictionary
	Language Varieties	Naija (Nigerian Pidgin)
	Naija Lingo	Nigerian Pidgin Words & Slangs

Table 2: Sources used for compiling lexical items and activities.

Activity-based template: This template focuses on routine or simple tasks, such as cooking, cleaning, washing, or do-it-yourself (DIY) activities. Each task is contextualized with the tools, materials, or methods needed to complete it successfully. The correct solution may highlight the right tool(s) for the task, the right step to achieve the activity, or the appropriate use of the tools or materials.

Activity → Tools/Materials → Distractor → Example

Instruction-based template: This template evaluates procedural knowledge and attention to critical details. Candidate solutions present two nearly identical sets of instructions (such as recipes, assembly directions, or step-by-step procedures). One solution maintains the correct sequence and details, while the other introduces a subtle but consequential error that would lead to failure.

Activity → Two near-identical instructions → 1 incorrect detail (distractor) → Example

While these templates were designed to guide dataset creation, entries were not required to adhere strictly to a single template. Most entries aligned with at least one template, either directly or through adaptation. Furthermore, templates frequently overlapped in practice—for instance, an activity-based entry like *washing clothes* naturally incorporates objects such as soap or buckets, creating alignment with the object-based template. Similarly, some instruction-based examples resembled activity-based entries but featured additional procedural steps or technical details. We deliberately preserved such overlaps as they reflect the interconnected nature of physical reasoning in real-world contexts.

2.2 Annotation and Review

Each entry was manually created and annotated by native speaker contributors who identified the correct solution for every example. To minimize bias, we maintained clear role separation—annotators who created entries did not participate in the review process.

For Yoruba, the dataset was created by four (4) annotators and reviewed by two reviewers. The Nigerian Pidgin dataset was also created by four (4) annotators and reviewed by one (1) reviewer. Contributors collaborated via Google Meet to brainstorm and refine examples, while reviewers independently assessed each entry for grammatical correctness, orthography, cultural grounding, and label accuracy. Any inconsistencies or ambiguities identified during review were systematically documented and resolved through discussion.

To maintain high quality and consistency, all examples adhered to the following key principles: (i) scenarios must be realistic, culturally grounded, and reflective of contemporary life; (ii) solutions must be grammatically parallel and differ subtly (by 1–2 words); and (iii) examples could vary in length to accommodate different types of physical reasoning tasks.

Language	Total Entries	label(0)	label(1)
yor	200	106	94
pcm	200	125	75
Total	400	231	169

Table 3: Label Distribution per Language

2.3 Dataset Statistics

The dataset comprises **400** entries, evenly divided between two languages: Yoruba (200) and Nigerian Pidgin (200). Each entry follows the PIQA-style format, consisting of a goal, two candidate solutions (`solution0` and `solution1`), and a label indicating the correct solution.

Table 3 summarizes the overall label distribution for each language. Label assignments are generally balanced, with a slight skew toward `solution0` in Nigerian Pidgin. Additionally, Figure 1 illustrates per-annotator label distributions, highlighting variations in annotation patterns across contributors. Together, the table and figure provide a comprehensive view of the dataset's label composition, both at the aggregate language level and at the individual annotator level.

As our dataset is binary, uniform random guessing yields an expected accuracy of 50%. When guesses are weighted according to the label distribution, the expected accuracy increases slightly to 51%. Considering the languages individually, weighted random guessing results in 50.2% accuracy for Yorùbá and 53.1% for Nigerian Pidgin. These numbers provide a lower-bound reference for model performance and help contextualize the challenge posed by the label distributions in our test collection.

3 Conclusion

In this work, we introduce a reproducible and innovative pipeline for creating physical commonsense reasoning datasets for low-resource languages without relying on translation. By combining instruction-based, activity-based, and object-based methods, our approach produces high-quality, realistic test instances grounded in culturally relevant contexts. We demonstrate this pipeline by constructing a dataset for Yorùbá and Nigerian Pidgin, comprising 400 instances that challenge models to reason about everyday goals and actions.

4 Limitations

4.1 Domain and Dataset Size

Our dataset is intended primarily as a benchmark for evaluating culturally grounded physical commonsense reasoning, rather than as a comprehensive resource for all reasoning or NLP tasks in Yorùbá and Nigerian Pidgin. The dataset size (200 instances per language) is sufficient to probe model performance and establish baseline comparisons, but it does not cover the full breadth of everyday physical scenarios, activities, or objects in these languages. Researchers should interpret results with these limitations in mind, and future work could expand the dataset to improve coverage and diversity.

4.2 Annotation Consistency and Subjectivity

Although all instances were created and validated by native speakers, some scenarios inherently involve subjective judgments about what constitutes a reasonable or correct physical action. Different annotators may have slightly different interpretations of how an object is used or what the most practical solution to a goal is. Regional or cultural variations within Yorùbá and Nigerian Pidgin-speaking communities may also influence these judgments. While we employed review and consensus procedures to minimize discrepancies, the dataset may still contain subtle biases or inconsistencies that reflect annotator perspectives rather than universal truths about physical commonsense. Researchers should consider this subjectivity when interpreting model performance.

References

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating LLM performance on African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32704–32731, Vienna, Austria. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, et al. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Zhuang Yun Jian, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *ArXiv*, abs/2406.03368.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Edward Bayes, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *ArXiv*, abs/2412.00948.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, Samuel Rutunda, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Seid Muhie Yimam, and Nedjma Ousidhoum. 2025. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Afrobench: How good are large language models on african languages?

Hao Yu, Jesujoba O. Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson K. Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Juliet W. Murage, Dietrich Klakow, and David Ifeoluwa Adelani. 2025. Injongo: A multicultural intent detection and slot-filling dataset for 16 african languages.