# Sinhala Physical Common Sense Reasoning Dataset

**Nisansa de Silva**
Dept. of Computer Science & Engineering
University of Moratuwa, 10400, Sri Lanka
NisansaDdS@cse.mrt.ac.lk

**Surangika Ranathunga**
School of Math. & Comp. Sciences
Massey University, New Zealand
S.Ranathunga@massey.ac.nz

## Abstract

This paper presents the first-ever Sinhala physical common sense reasoning dataset. It contains 110 human-created and verified data samples, where each sample consists of a prompt, the corresponding correct answer, and a wrong answer. Most of the questions refer to the Sri Lankan context, where Sinhala is an official language.

## 1 Sri Lankan Culture and Sinhala

Sri Lanka is a multi-ethnic and multilingual country, with Sinhala and Tamil being the official languages. Sinhala is an Indo-Aryan language spoken by more than 17 million people in Sri Lanka (de Silva, 2025). It possesses a unique alphabet and script. According to Ranathunga and de Silva (2022)'s language categorisation, which uses the category definition established by Joshi et al. (2020), Sinhala is a low-resource language. The scarcity of digital language resources and the lack of collaborative research have been the main reasons for the slow progress of Sinhala language computing research (Ranathunga and de Silva, 2022; de Silva, 2025). Given that Sri Lanka is geographically isolated from its nearest neighbour - India, Sinhala does not get included in Indian data creation initiatives or Language Models such as Indic Gemma[1], despite being the Indo-Aryan language family.

Sri Lanka has an unbroken written history of its civilisation for more than 2500 years (Mahanama and Geiger, 1908). Being an island nation that has been on the busiest naval trade routes for millennia (Sudharmawathie, 2017), it boasts a unique culture, which is both different to but also linked to other South Asian countries (Roberts, 2021). It

also had influence from its colonial rulers - Portuguese, Dutch and the British. Given that Sri Lanka is a multi-ethnic and multi-religious country, the presence of different sub-cultures is evident (e.g. Sinhala-Buddhist, Tamil-Hindu, etc).

## 2 Corpus Creation

The corpus was prepared by two Sri Lankan nationals who have lived in Sri Lanka for more than 30 years. They are the authors of this paper. Each author had their school education in Sinhala. They both hold PhDs and are NLP researchers. Each data creator created at least 50 samples, and the other verified the quality of the samples.

The dataset follows the format of the PIQA dataset (Bisk et al., 2020). All the questions were manually created (i.e., they were not translated from PIQA or any other dataset). Some samples from the corpus are shown in Figure 1 along with the description of the notation that is used in this paper. The wrong answer for a given prompt was prepared by either 1. changing one letter of a word, 2. changing 1-3 words of a sentence, or 3. swapping words/phrases in a sentence.

## 3 Corpus Statistics

Out of the data samples, 67 are in the form of sentence/paragraph completion (Shown in the left column of Figure 1). The remainder is in the form of Question-Answer pairs (Shown in the right column of Figure 1).

Table 1 shows how the samples are distributed across different domains and the aforementioned two forms. Here, the 'other' category mostly refers to common sense reasoning in day-to-day life (e.g. the first example in the left column of Figure 1). Some data samples refer to general common sense applicable to any culture (e.g. how to open a tightened bottle lid?), some are very specific to the Sri Lankan culture (e.g. rituals related to Sri

---

[1] https://dataloop.ai/
library/model/telugu-llm-labs_
indic-gemma-7b-finetuned-sft-navarasa-20/

| Paragraph Completion | | | Question Answer Pairs | | |
|---|---|---|---|---|---|
| [Goal] | Si | ඔළුවේ සිටින උකුණන් අයින් කිරීමට | [Goal] | Si | මදුරුවන් කෑම වළක්වා ගැනීමට කුමක් කළ යුතුද? |
| | En | To remove lice from the head | | En | What should be done to stop mosquitoes from biting? |
| [Sol1] | Si | පොඩි දැති පනාවකින් ඔලුව පීරන්න ✓ | [Sol1] | Si | එළඟිතෙල් ටිකක් ගා ගත යුතු ය. ✗ |
| | En | Use a <u>small</u> toothed comb | | En | Apply some <u>ghee</u>. |
| [Sol2] | Si | ලොකු දැති පනාවකින් ඔලුව පීරන්න ✗ | [Sol2] | Si | පැහිරිතෙල් ටිකක් ගා ගත යුතු ය. ✓ |
| | En | Use a <u>large</u> toothed comb | | En | Apply some <u>citronella</u>. |
| [Goal] | Si | උස පොල් ගසකට පහසුවෙන් නැගීමට | [Goal] | Si | හංසයාට දිය මිශ්‍ර කිරි දුන් විට ඌ කුමක් කරන්නේ යැයි කියයිද? |
| | En | To climb a tall coconut tree | | En | What is it said that a swan does when given milk mixed with water? |
| [Sol1] | Si | ඉණිමගක් භාවිතා කරන්න ✗ | [Sol1] | Si | දියෙන් කිරි වෙන් කර බොනවා යැයි කියති ✓ |
| | En | Use a <u>ladder</u> | | En | It is said to drink the <u>milk</u> leaving the <u>water</u> |
| [Sol2] | Si | වළල්ලක් භාවිතා කරන්න ✓ | [Sol2] | Si | කිරෙන් දිය වෙන් කර බොනවා යැයි කියති ✗ |
| | En | Use a <u>loop</u> [of rope]. | | En | It is said to drink the <u>water</u> leaving the <u>milk</u> |

Figure 1: A few examples from the Sinhala physical common sense reasoning dataset. `[Goal]` is the question (prompt) while `[Sol1]` and `[Sol2]` are the two possible solutions, of which a human or a trained model must choose the most appropriate solution, of which exactly one is correct. The correct answer is denoted by a ✓ and the incorrect answer is denoted by a ✗. Short phrases that differ between `[Sol1]` and `[Sol2]` are shown in <u>underline</u>. In addition to the original entry in Sinhala (`Si`), in this table, we also provide an English (`En`) translation even though our dataset itself does not carry English translations.

| Domain | Number of Samples | | |
|---|---|---|---|
| | Paragraph Completion | Question Answer Pairs | Total |
| Buddhism | 4 | 6 | 10 |
| Literature | 2 | 2 | 4 |
| Mythology | 2 | 2 | 4 |
| Sports and games | 4 | 1 | 5 |
| Food | 11 | 4 | 15 |
| Farming/Agri/Fishery | 4 | 4 | 8 |
| Proverbs | 3 | 2 | 5 |
| History | 9 | 10 | 19 |
| Other | 28 | 12 | 40 |
| Total | 67 | 43 | 110 |

Table 1: Distribution of samples across different domains and types



Figure 2: Sentence length distributions for both correct solutions and incorrect solutions

Lankan Sinhala Buddhist wedding ceremonies), while some refer to the Sri Lankan version of concepts found in other cultures (e.g. the children's game *'batto paneema'* is the Sri Lankan version of *'hopscotch'*).

Figure 2 shows a plot of the sequence lengths of the correct and incorrect solutions, as tokenised by the `SinLLaMA` tokeniser (Aravinda et al., 2025). While there are minor differences, the two distributions are reasonably similar for a dataset of 110 samples.

We then compared the distribution of words in the data set created by us against the general domain Sinhala word distribution calculated by Wickramasinghe and De Silva (2023). This yielded a Pearson correlation of 0.109404. This *weak* (Shi and Conrad, 2009) correlation with the general domain is expected, as this data set is specialised for the physical reasoning domain.
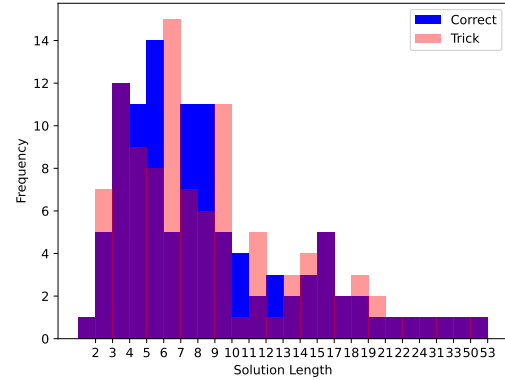
## 4 Experiments

When introducing the PIQA dataset, Bisk et al. (2020) conducted an experiment on the task of *Multiple Choice Question Answering* using RoBERTa (Liu et al., 2019) and some other language models. We also conducted an experiment of the same task using the dataset created by us. We used `SinBERT` (Dhananjaya et al., 2022) as our model, given that it is the closest Sinhala language model to the original experiment. However, one crucial difference between our experiment and the original experiment is the fact that, in the original experiment with the PIQA dataset, Bisk et al. (2020) fine-tuned the RoBERTa model and then reported the testing and validation accuracy, but given the small size (110 entries) of our dataset, we

| | | Examples where SinBERT succeeded | | | Examples where SinBERT failed |
|---|---|---|---|---|---|
| [Goal] | Si | ගුත්තිල කාව්‍යයට අනුව සුරඟනන් නටත්දී අත් ලෙල දුන්නේ කුමක් සේ ද? | [Goal] | Si | ආච්චි ගමන් බිමන් යන්නට භාවිතා කරන හැරමිටිය අස්ථානගත වී ඇත. එය සොයා ගන්නකන් ඇගේ පාවිච්චියට |
| | En | According to the epic poem *Guththilaya*, how did the angels wave their hands while dancing? | | En | The walking stick grandma uses to walk about has gone missing. Until it is found for her to use |
| [Sol1] | Si | විදුලිය සේය ✓ | [Sol1] | Si | මඤ්ඤොක්කා ගසක් කපා හැරමිටියක් සාදා දෙන්න ✗ |
| | En | Like <u>lightning</u> | | En | Cut a <u>cassava</u> tree and make a walking stick |
| [Sol2] | Si | ගිගිරුම් සේය ✗ | [Sol2] | Si | ගිනිසීරියා ගසක් කපා හැරමිටියක් සාදා දෙන්න ✓ |
| | En | Like <u>thunder</u> | | En | Cut a <u>gliricidia</u> tree and make a walking stick |
| | | | | | |
| [Goal] | Si | අල්මාරිය උඩ තිබූ පිත්තල හෙප්පුව, පූසාගේ ඇඟේ වැදී බිමට වැටුණි. | [Goal] | Si | පායන කාලයේ ⁣ රෑ දවල් මහන්සි වී කූඹියෝ කුමක් කරති ද? |
| | En | The brass *heppuwa* that was on the almirah, fell on the ground after the cat bumped into it. | | En | What do the ants do with much effort day and night in the dray season? |
| [Sol1] | Si | "අයියෝ, හෙප්පුව, බිඳුණා", ආච්චි කීවාය ✗ | [Sol1] | Si | කන්න දේ රැස් කරති ✓ |
| | En | "Alas, the *heppuwa*, <u>shattered</u>", grandma said. | | En | Collect things to <u>eat</u> |
| [Sol2] | Si | "අයියෝ, හෙප්පුව, තැළුනා", ආච්චි කීවාය ✓ | [Sol2] | Si | නාන්න දේ රැස් කරති ✗ |
| | En | "Alas, the *heppuwa*, <u>dented</u>", grandma said. | | En | Collect things to <u>bathe</u> |

Figure 3: Qualitative analysis of SinBERT's predictions with. **Left:** Two examples that SinBERT gets right. **Right:** two examples that SinBERT gets incorrect. Blue shading shows SinBERT succeeding while the Red shading shows SinBERT failing.

opted for a zero-shot experiment.

| Domain | Paragraph Completion | | Question Answer Pairs | |
|---|---|---|---|---|
| | SinBERT | GPT-5 | SinBERT | GPT-5 |
| Buddhism | 25.00 | 75.00 | 83.33 | 50.00 |
| Literature | 50.00 | 100.00 | 0.00 | 50.00 |
| Mythology | 50.00 | 100.00 | 50.00 | 0.00 |
| Sports and games | 75.00 | 0.00 | 100.00 | 0.00 |
| Food | 36.36 | 81.82 | 25.00 | 25.00 |
| Farming/Agri/Fishery | 50.00 | 50.00 | 25.00 | 75.00 |
| Proverbs | 66.67 | 100.00 | 0.00 | 50.00 |
| History | 55.56 | 66.67 | 40.00 | 70.00 |
| Other | 57.14 | 71.43 | 50.00 | 66.67 |

Table 2: Accuracy of SinBERT and GPT-5 across different domains and types

The accuracy achieved by SinBERT across different domains and sample types is shown in Table 2. In most domains, SinBERT seems to do better when the sample is stated as an instance of *paragraph completion* rather than a direct *question answer pair*. Further, we show some of the qualitative analysis of this experiment in Figure 3.

What we can observe from this is that while SinBERT may have some success on the examples where general common sense might reasonably arrive at the correct answer, the deeper the question and answers are embedded into the Sri Lankan culture, the more it starts to struggle to provide answers. In fact, the overall accuracy landed at a dismal 49.09, which is even lower than the expected value of a uniform random guess.

In Figure 4 we show the distribution of the examples against the edit distance between the two solution choices. We compute edit distance over tokenised strings with a cost of 1 for edits, insertions, and deletions. We also show how the zero-shot results of SinBERT (Dhananjaya et al., 2022) change
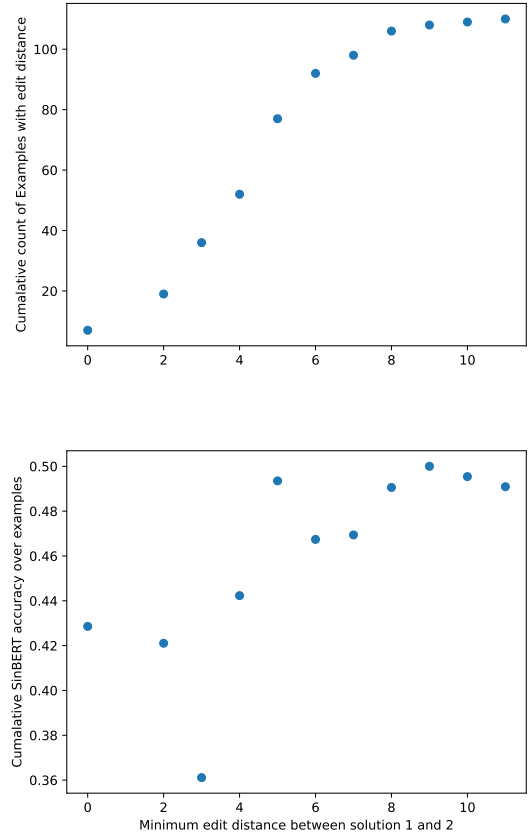




Figure 4: Breaking down the data set by edit distance between solution choices. **Top:** Cumulative histogram of examples in the data set, in terms of minimum edit distance d between the two solution choices. The majority of the dataset consists of small tweaks between the two solution pairs; nevertheless, this is enough to confuse NLP models. **Bottom:** Zero-Shot SinBERT accuracy over examples with a minimum edit distance. Dataset difficulty increases somewhat as the two solution pairs are allowed to drift further apart.
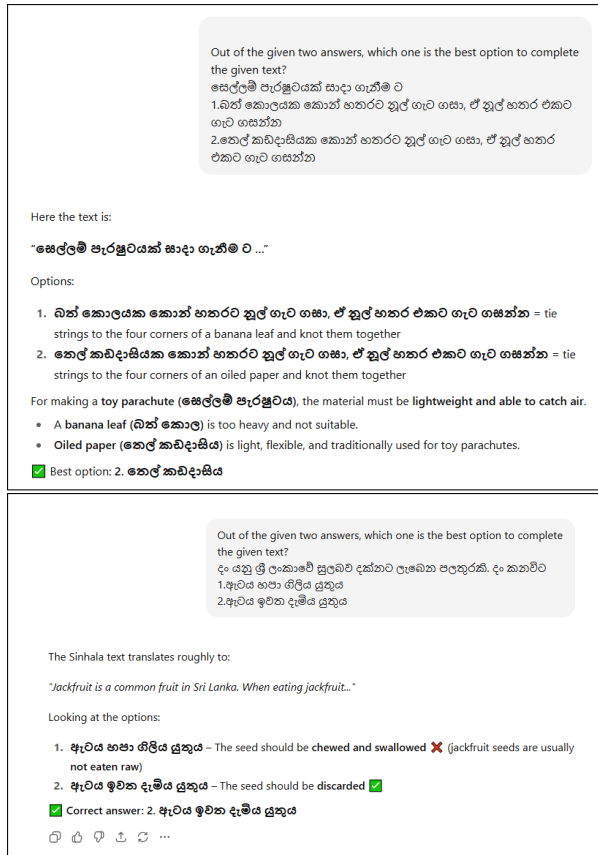
Figure 5: Two sample outputs from GPT-5 mini.

over the edit distance.

We also tested GPT-5 mini free version (Accessed on 17/09/2025) with the 110 data samples. It managed to answer 71/110 (64.5%) samples correctly. The accuracy achieved by GPT-5 across different domains and sample types is shown in Table 2. In most domains, GPT-5 seems to do better than SinBERT. However, specifically in *question answer pair* type under *Buddhism*, *Sports and games* and *Mythology* domains SinBERT succeeds over GPT-5.

We noted that GPT-5 mini translates the Sinhala text to English before reasoning on it. Sometimes, translation errors result in the model providing wrong answers. One example is shown in the top part of Figure 5. The Sinhala text *'bath kolaya'* refers to a thin polythene that is used to wrap rice to make a packet of rice. GPT-5 mini incorrectly translates this term to *banana leaf*. This is a very interesting case, because before polythene was introduced to Sri Lanka, banana leaves were primarily used to wrap rice. However, in Sinhala, a banana leaf is refereed to as *'kesel kolaya'*, while *'bath kolaya'* refers to the polythene. Due to this wrong translation, GPT-5 mini produces the wrong answer

to the question. Interestingly, the opposite happens in the second (bottom) example of Figure 5. It refers to *'dan'* /dʌŋ/ (*Syzygium caryophyllatum (L.) Alston*), a wild fruit endemic to Sri Lanka. It is the size of a blue berry. While eating it, the seed has to be thrown away. However, GPT-5 mini translates *'dan'* to the jackfruit seed, which is not eaten raw. The final answer of GPT-5 mini results in the correct answer, because of this wrong translation, despite GPT-5 not knowing about *dan*.

## 5    Conclusion

This paper presented the first-ever physical common sense reasoning dataset written in Sinhala. We evaluated two models - SinBERT and GPT-5 with this dataset. While GPT-5 outperformed, its overall accuracy is just 64.5%. This underscores the challenges faced by modern-day Large Language Models, when working with data from different cultures, written in languages other than English.

## Limitations

Given that both data creators are Sinhala Buddhists, the dataset has a bias towards Sinhala Buddhist culture. There can be minor spelling mistakes in the dataset (while there has been some research on implementing Sinhala spell correctors, none is at a production-ready level (Sonnadara et al., 2021)). We tested GPT-5 mini only once via its GUI interface. Due to its non-deterministic nature, it may produce different results if tested multiple times with the same question.

## References

H W K Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Surangika Ranathunga, and Rishemjit Kaur. 2025. SinLlama-A Large Language Model for Sinhala. *arXiv preprint arXiv:2508.09115*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, 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.

Nisansa de Silva. 2025. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358v25*.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification. In *Proceedings of the Thirteenth Language Resources*

*and Evaluation Conference*, pages 7377–7385. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Mahanama and Wilhelm Geiger. 1908. *Mahavamsa: Great Chronicle of Ceylon*, volume 63. Pali text society.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848. Association for Computational Linguistics.

Michael Roberts. 2021. *Exploring confrontation: Sri Lanka: politics, culture and history*. Routledge.

Runhua Shi and Steven A Conrad. 2009. Correlation and regression analysis. *Ann Allergy Asthma Immunol*, 103(4):S34–S41.

Charana Sonnadara, Surangika Ranathunga, and Sanath Jayasena. 2021. Sinhala spell correction: A novel benchmark with neural spell correction.

J Sudharmawathie. 2017. Foreign trade relations in Sri Lanka in the ancient period: With special reference to the period from 6th century BC to 16th century AD. *Humanities and Social Sciences Review*, 7(2):191–200.

Kasun Wickramasinghe and Nisansa De Silva. 2023. Sinhala-English Parallel Word Dictionary Dataset. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 61–66.