Dataset Description of the Tamil Reasoning Dataset

Varsha Jeyarajalingam, Menan Velayuthan, Kengatharaiyer Sarveswaran

varujeya@gmail.com, vmenan95@gmail.com, sarves@univ.jfn.ac.lk

Department of Computer Science, University of Jaffna. Sri Lanka.

The dataset has been manually curated to provide a diverse range of reasoning contexts that fall into three broad categories: physical reasoning, cultural reasoning, and common sense reasoning. Each category is represented through multiple subdomains of knowledge, ensuring the coverage of Sri Lankan context alongside general reasoning skills.

- Physical Reasoning
    - This category covers knowledge of the natural world, scientific principles, and environmental understanding. Subcategories include natural phenomena, material properties, agricultural and environmental knowledge, and concepts related to energy and electricity.
- Cultural Reasoning
    - This category captures the cultural and traditional dimensions of Sri Lankan life. Subcategories include food practices, health and safety, religious traditions, rituals and customs, literature and arts, and traditional dress and identity.
- Common Sense Reasoning
    - This category encompasses practical day-to-day knowledge and socially shared wisdom. Subcategories include everyday routines, transport and rules, social practices, and logical consistency.

Note: We added an additional column, *category*, to indicate whether a datapoint falls under physical, common sense, or cultural categories, in addition to the standard columns (prompt, solution0, solution1, label).

Statistics of the data

We measured the average word count (prompt + solution) for each category (physical, cultural, reasoning) along with their average character count (prompt + solution) per data point. The statistics are as follows,

	Physical	Commonsense	Cultural
Number of data points	30	55	35
Average word count	8.87	7.25	7.49
Average character count	77.70	69.63	69.61

Although we have not reached the 25-word count limit, we strongly believe that different languages pack information in a varied manner. As you can see by the average word count and the average character count.

Rejection Criteria

During the curation process, we applied clear rejection criteria to ensure the dataset remained focused and unambiguous. In total, 15 data points were removed:

- Ambiguous items: We excluded items that could vary significantly across cultural contexts. For instance, the drink typically served to visitors in Sri Lanka is most often tea or coffee, but in some households it could also be wine or other beverages, making the answer culturally inconsistent.
- Overly factual questions: We removed questions that required memorization rather than reasoning. Such items often depended on specialized knowledge that not all participants would possess, and thus did not align with the reasoning-focused design of the dataset.

Qualification of the dataset creators

Dr. Kengatharaiyer Sarveswaran: Supervised the project. He is a Tamil native speaker. He also played the role of evaluator. He is a computational linguist mainly focusing his research on Tamil and other low-resource languages. He is also qualified as a Tamil Junior Pundit (Bala Pundit), making him the perfect fit for evaluation. He is currently a Senior Lecturer in Computer Science and leads the Language Technology Group at the University of Jaffna, Sri Lanka. His complete publication list can be found in his Google Scholar.

Menan Velayuthan: One of the dataset creators. He is a Tamil native speaker. He recently completed a research-based Master's in Computer Science, with a focus on low-resource languages such as Sinhala and Tamil. He previously worked on parallel dataset creation for Sinhala, Tamil and English funded by the Google Award for Inclusion Program. His complete publication list can be found in his Google Scholar.

Varsha Jeyarajalingam: One of the dataset creators. She is a native Tamil speaker and a third-year undergraduate student in the Department of Computer Science at the University of Jaffna. She is currently working with Menan under the supervision of Dr. Sarveswaran on a Tamil evaluation project of large language models (LLMs).