ARIV[1]: Annotated Reasoning dataset in Vernacular Malayalam

Introduction

Malayalam is a low-resource Dravidian language primarily spoken by the native people of the Indian state of Kerala, as well as the union territories of Lakshadweep and Puducherry. It is one of the 11 officially recognised classical languages of India, a status that reflects its rich literary heritage and historical significance. Malayalam has approximately 46.4 million native speakers [1], making it one of the most widely spoken Dravidian languages. It is widely noted by scholars that the geographical isolation of Kerala, due to the Western Ghats, protected it from major invasions. This allowed the Malayalam language to develop mostly on its own, and helped Kerala evolve its own way of life and social institutions without much outside interference [2].

In this work, we developed a Malayalam common sense reasoning dataset, which consists of 101 samples. The dataset is manually created and verified by two native Malayalam speakers.

Annotators

The annotation was performed by two native Malayalam speakers from different regions of Kerala: one from Muvattupuzha, which reflects a mix of Idukki and Kottayam language styles, and the other from Ottappalam, who is familiar with the dialects of Palakkad and Thrissur. This regional diversity allowed the annotators to include distinct language styles and a wide range of perspectives, thereby ensuring that the dataset (ARIV) reflects a broader and richer representation of natural Malayalam language use.

Annotation Procedure

The dataset consists of approximately 101 manually created Malayalam samples similar to the PIQA dataset. While some items follow a question–answer format, others adopt different formats or are open-ended. To make the dataset culturally rich and relevant, we included topics specific to Kerala, such as local weather, traditional food recipes, regional flora and

[1] Ariv (അറിവ്) in Malayalam means Knowledge/Wisdom

fauna, cultural flair, and religious traditions. Each sample was created by one annotator and then verified and refined by the other to ensure clarity, correctness, and naturalness.

	Min #words	Max #words	Average #words
prompt	2	16	5.17
solution0	1	63	9.10
solution1	1	65	9.15
Correct solution	1	63	9.10
Incorrect solution	1	65	9.15

Table: **Minimum, maximum, and average number of words in *ARIV* dataset**

To avoid any bias arising from the position of correct answers, we carefully balanced the dataset so that the number of times the first option (solution0) is correct is almost equal to the number of times the second option (solution1) is correct. This step helps prevent any models from exploiting positional bias during training or evaluation. The answer choices were carefully created to maintain similar length and complexity, avoiding superficial cues like length or wording that could bias the model. We also included challenging distractors—plausible but incorrect answers—to ensure the model relies on genuine reasoning. Both annotators regularly checked the data to catch and fix any accidental patterns or mistakes, making the dataset fair and high quality.

Why is the number of words per sample lower in Malayalam compared to English?

In our dataset, the average number of words per sample is 5.17 for prompt, 9.10 for solution0 and 9.15 for solution1, which is lower than the average in the PIQA dataset. We think this is natural because Malayalam sentences usually have fewer words than their English equivalents. This is due to the morphological complexity and agglutinative nature of Malayalam, which means that in Malayalam, multiple morphemes often combine into a single word. Malayalam words also undergo complex inflections, derivations, and compounding, which together create an almost limitless vocabulary [3].

Because of the morphological complexities, even shorter Malayalam sentences can be more challenging for large language models compared to longer English sentences. Its compact structure requires models to extract more meaning from fewer words, making accurate comprehension and inference more difficult. Therefore, resources such as the *ARIV* dataset are crucial for evaluating and advancing the ability of language models to process and understand morphologically rich languages like Malayalam.

Should this work be accepted, we are committed to extending the *ARIV* dataset with additional examples to further support research in Malayalam language understanding.

References

1. *https://www.worlddata.info/languages/malayalam.php*
2. *https://sde.uoc.ac.in/sites/default/files/sde_videos/history%20of%20kerala%20PDF.pdf*
3. *Manohar, Kavya & Jayan, A. & Rajan, Rajeev. (2020). Quantitative Analysis of the Morphological Complexity of Malayalam Language. 10.1007/978-3-030-58323-1_7.*