Burushaski [bsk] PIQA-style Dataset for Physical Commonsense Reasoning (MRL 2025 Shared Task)

1. Introduction

This paper describes our contribution to the MRL 2025 Shared Task: a dataset of physical commonsense reasoning examples in Burushaski, an under-resourced language spoken primarily in Gilgit-Baltistan, Pakistan. Many languages lack culturally-specific evaluation datasets, and this work provides a resource created by a native speaker to support multilingual commonsense reasoning research.

2. Language and Dialect

The dataset is written in Burushaski[bsk] (Yasin dialect). Burushaski is a language isolate spoken mainly in Hunza, Nagar, and Yasin valleys of Gilgit-Baltistan. The Yasin dialect was chosen because it is the variety spoken natively by the author.

3. Dataset Construction

Format: The dataset follows the PIQA format, with four columns: prompt, solution0, solution1, label.

Prompts: Each item describes a physical situation, instruction, or question relevant to daily life and culture.

Solutions: Two candidate completions are provided (solution0 and solution1). They differ minimally (only one or two words).

Labels: A binary label indicates the correct solution (0 = solution0 correct, 1 = solution1 correct).

Size: The dataset contains 100 original items (not translations of PIQA).

4. Example Items

prompt	solution0	solution1	label
After placing ice in the sun,	It will melt	It will become colder	0
Which tool is used to cut paper?	A seasor	A spoon	0

When salt is added to water, It dissolves It burns 0

5. Quality Control

All examples were authored and verified by a native speaker of Burushaski (Yasin dialect). Items were checked for grammatical correctness, cultural relevance, and physical commonsense validity.

6. Dataset Availability

The dataset is provided as a .tsv file named: mrl2025_burushaski_dataset.tsv. It is UTF-8 encoded to ensure Burushaski script compatibility.

This dataset is the first PIQA-style resource for Burushaski, and we hope it supports future research in multilingual commonsense reasoning.

Here is the link of dataset which is available in google sheet : https://docs.google.com/spreadsheets/d/1-DowUcJWy3UJhlqscNF_E42lJG02oXF-mRatXwkdZEE/edit?usp=sharing

With deep thanks Sardar Ali

sardaralikhamosh@gmail.com

+923415336669

LinkedIn Profile: https://www.linkedin.com/in/sardaralikhamosh/