Dataset Description for Sindhi Commonsense Reasoning Benchmark

1. Introduction

We present a manually constructed dataset of commonsense reasoning questions in **Sindhi**, a low-resource Indo-Aryan language spoken primarily in Sindh (Pakistan) and parts of India. The dataset is inspired by **PIQA-style physical and commonsense reasoning tasks**, but localized to Sindhi culture, daily life, history, literature, folklore, and traditions.

The goal of this dataset is to enable the evaluation of language models on **commonsense reasoning grounded in Sindhi culture and everyday practices**. Unlike translation-based datasets, all examples were **manually authored** in Sindhi to reflect realistic usage and cultural specificity.

2. Language and Dialect

Sindhi is an Indo-Aryan language spoken primarily in the Sindh province of Pakistan and by Sindhi communities in India and the diaspora worldwide. It is one of the oldest documented languages of South Asia, with a rich literary tradition dating back over a thousand years. Around 40 million people speak this language. Sindhi is written in two major scripts: the **Perso-Arabic script** (used in Pakistan) and the **Devanagari script** (used mainly in India). For this dataset, we adopt the **Perso-Arabic script**, which is the standard writing system in Sindh, Pakistan.

Sindhi has multiple **regional dialects** (e.g., *Siroli, Vicholi, Lari, Thari, Kachhi*), as well as **diaspora varieties** influenced by Hindi and Gujarati in India. Among these, **Vicholi Sindhi**—the central variety spoken in urban Sindh (e.g., Hyderabad, Sukkur, Larkana)—is considered the **standard dialect** for education, literature, and media.

For our dataset, we use **Standard Sindhi** in the Perso-Arabic script. This choice ensures that the data is:

- **Mutually intelligible** to Sindhi speakers across regions.
- **Culturally grounded**, as the examples include folklore, foods, festivals, and traditions specific to Sindh, Pakistan.
- **Accessible** for use in natural language processing tasks, since most digital Sindhi resources are also in this standard variety.

This dialectal choice reflects the most widely understood and formally recognized register of Sindhi, while maintaining cultural authenticity.

3. Dataset Design

3.1 Format

The dataset is provided in **.TSV format** with four columns:

- **prompt**: 1–2 sentences in Sindhi posing a scenario or question.
- **solution0**: A candidate answer (can be correct or incorrect).
- **solution1**: A candidate answer (can be correct or incorrect).
- **label**: Binary value, `0` if solution0 is correct, `1` if solution1 is correct.

Example:

```
prompt    solution0    solution1    label
جيڪڏهن مٽيءَ جو ٿانو زمين تي ڪري پوندو،    ٿانو ٽٽي ٽڪرا ٿي ويندو.    ٿانو
نرم ٿي وڌيڪ لچڪدار ٿيندو.    0
```

3.2 Construction Process

- All examples were **manually authored** by a Sindhi native speaker (no machine translation).
- Prompts were written in **varied styles**: questions, cause–effect statements, cultural explanations.
- Solutions were designed to be **minimally different** (e.g., one or two words changed, order swapped) to make the task non-trivial for models.
- Incorrect answers were chosen to be **plausible but wrong**, avoiding absurd or overly obvious distractors.
- Labels were balanced between 0 and 1.

3.3 Domains Covered

- **Local foods:** e.g., ساڳ ۽ ماني, ٻاڄڙي, ڳڙ, انب.
- **Places:** e.g., ڀٽ شاهه, موئن جو دڙو, هالا.
- **Customs & Traditions:** e.g., شاديءَ جون رسمون (مهندي، ساٿرڻ، ٻانوري), ثقافتي ميلا, چنڊ رات.
- **Religion & Festivals:** شاهه لطيف جو اُرس, جهولي لال جو ميلو, عيدالاضحی.
- **Literature & Folklore:** شاهه لطيف, شيخ اياز, ڪهاڻيون (سسي پنهون, مومل راڻو, مارئي, سورٺ راڻي).
- **Everyday Objects:** مٿيءَ جا ٿانوَ, چوڙيون, رليون.
- **Music & Art:** هو جمالو, مالهاڻ, دهل ۽ شهنائي.

4. Scale and Statistics

- **Number of examples:** 150 **examples**.
- **Prompt length:** Most prompts are **>25 words**, with multiple sentences in some cases.
- **Cultural coverage:** Each example is tied to Sindhi lifestyle, folklore, or practical knowledge.

5. Quality Control

- All examples were **checked by native speakers** for grammatical correctness and cultural validity.
- Incorrect solutions were validated to ensure they are **plausible but not correct**.
- Special care was taken not to reuse the same topic repeatedly.