Thinnai-PIQA: A Benchmark for Cultural Commonsense Reasoning in Tamil

Kamesh R Kamesh V
Sathyabama Institute of Science and Technology
kameshrajeshkanna@gmail.com kameshvedaviyas@outlook.com

Abstract

Existing common sense reasoning benchmarks for AI predominantly focus on universal physical knowledge, failing to capture deep, nuanced procedural understanding embedded within specific cultural contexts. This limitation makes current models brittle and culturally unaware. To address this gap, we introduce **Thinnai-PIQA**, a new benchmark to test an AI's physical common sense reasoning within the rich tapestry of Tamil culture, literature, and tradition. The data set was crafted using a rigorous human-driven methodology. Initial attempts to use large language models produced outputs that were often generic, obvious, or culturally inaccurate. Consequently, a team of native speakers conducted an intensive process of creation and curation. Each question was meticulously handcrafted or substantially rewritten to feature our "Logical Trap" principle, where the incorrect solution is a logical but culturally inept action. Thinnai-PIQA is therefore a product of deep human expertise, designed to serve as a formidable new benchmark to drive the development of truly culturally intelligent AI systems.

1 Introduction

Modern Large Language Models (LLMs) have demonstrated impressive capabilities in commonsense reasoning. Benchmarks like PIQA (Physical Interaction: Question Answering) have pushed the boundaries of AI's understanding of the physical world. However, this understanding remains largely generic and culturally sterile. Can a model that knows how to open a jar also know the correct physical procedure for preserving a 400-year-old palm-leaf manuscript?

This gap between generic logic and embedded cultural wisdom is the challenge we address with **Thinnai-PIQA**. This paper details the creation of this unique benchmark and analyzes the specific categories of deep Tamil nuances it is designed to test.

2 Dataset Creation Methodology

The creation of Thinnai-PIQA was a demanding, multi-stage process driven by human expertise, necessitated by the failure of initial automated approaches to capture the required level of cultural subtlety.

Our preliminary attempts involved using state-of-the-art Large Language Models (LLMs) to generate question candidates. However, we found these outputs to be consistently inadequate. The generated questions were often superficial, fact-based rather than process-based, or featured "incorrect" solutions that were nonsensical rather than plausible "logical traps." The AI-generated content fundamentally lacked the lived experience and deep contextual understanding that this dataset required.

Faced with these limitations, we adopted a rigorous, human-centric methodology:

1. **Principle Definition:** First, our team of native Tamil experts defined the core principles of the dataset. The central axiom was that every question must test **procedural reasoning** (the "how-to" of a physical action) and not declarative knowledge. The second principle was the development of the **"Logical Trap,"** ensuring that the incorrect choice would be alluring to a system relying on generic, context-free logic.

2. **Manual Creation & Intensive Curation:** A team of authors then embarked on the intensive, overnight process of creating the dataset. While prompted LLMs were used as a preliminary brainstorming tool to generate initial concepts, a vast majority of these candidates were discarded. Each of the 100+ questions in the final dataset was either **created entirely from scratch by the human authors** or **substan-**

tially rewritten and refined from a primitive AI-generated concept.

3. **Expert Review:** Every question underwent a strict validation process. The authors debated the subtlety of each choice, ensuring the "expert" solution was authentically the most apt according to deep cultural and traditional norms, and that the "plausible" solution formed a powerful and fair challenge.

Therefore, while we utilized LLMs as a tool in the initial phase, Thinnai-PIQA is fundamentally a **handcrafted dataset**. It is the product of extensive human labor, cultural knowledge, and a commitment to creating a benchmark that truly reflects the complexity of human commonsense.

3 Principles for Replicable Question Construction

To ensure that Thinnai-PIQA can serve as a foundation for future work, we present the specific framework used by our team of human authors to create and curate each question. Adherence to these principles is crucial for any researcher aiming to reconstruct a comparable dataset or contribute new examples.

3.1 Core Construction Principles

Each question in the dataset was required to satisfy the following four principles:

1. **The Primacy of Physical Process:** Every question must be fundamentally about a physical action, procedure, or interaction. The goal must describe a tangible task, and the solutions must be two distinct physical methods to achieve it. This tests procedural reasoning (how to *do* something) rather than declarative knowledge (what something *is*).

2. **The "Logical Trap" Mechanism:** The incorrect answer must be a highly plausible "logical trap." This means it should be an action that appears correct based on generic logic, modern convenience, or a surface-level understanding, but is subtly wrong or damaging within the specific cultural, material, or historical context.

3. **Deep Cultural Grounding:** The physical action must be situated within a specific, nuanced domain of Tamil culture. The reasoning required should not be universal but should instead depend on an understanding of the traditions, literature, history, or material science specific to the Tamil context.

4. **Enforced Thematic Diversity:** To prevent thematic bias and ensure a broad test of cultural reasoning, questions were intentionally diversified across a wide palette of domains, including:

    - Ritualistic and Ceremonial Processes
    - Literary and Historical Artifacts
    - Agrarian and Folk Wisdom
    - Material and Artistic Nuances
    - Traditional Beliefs and Omens

3.2 Exclusion Criteria

During the curation process, candidate questions were rigorously vetted and discarded if they violated any of the following negative constraints:

- **No Simple Factual Recall:** Questions that could be answered by looking up a simple fact (e.g., "What is the main ingredient in Sambar?") were excluded.

- **No Universal Commonsense:** Questions testing generic commonsense applicable to any culture (e.g., "How do you put out a small fire?") were excluded.

- **No Absurd Distractors:** The incorrect solution could not be nonsensical or physically impossible. Both choices had to be plausible actions.

- **No "Signaling" Words:** The phrasing of the solutions could not contain "giveaway" words that explicitly signal the correct answer.

By following this detailed framework, we believe other researchers can contribute to this benchmark and develop comparable datasets for other languages, enriching the field of culturally-aware AI evaluation.

4 Analysis of Tamil Cultural Nuances in Thinnai-PIQA

Thinnai-PIQA tests a model's reasoning across a diverse palette of deep cultural domains. The questions are designed to be impenetrable to those without contextual understanding, forcing models to

reason about physical processes within a cultural framework. Below are the key categories, with examples drawn directly from the dataset.

4.1 Ritualistic and Ceremonial Processes

This category tests the unwritten, physically-enacted rules of Tamil ceremonies and rituals. The "logical trap" is often a modern or convenience-based action that violates a specific, deeply-held tradition. A modern, efficiency-minded system would choose the logical trap, but a culturally aware one would know the specific, non-negotiable placement rule of the banana leaf feast.

4.2 Literary and Historical Process Knowledge

These questions require reasoning about physical actions related to historical artifacts or literary contexts. The knowledge is not about the story itself, but about the physical world in which the story is set.

4.3 Agrarian and Folk Wisdom

This category includes questions about traditional farming, home remedies, and nature-based beliefs (omens) where the correct action is passed down through generations and often has a subtle, non-obvious practical basis.

4.4 Material and Artistic Nuances

This tests the correct procedure for handling traditional materials, foods, and art forms. The distinction is often between a method that produces an acceptable result and one that produces an authentic, superior result.

5 Conclusion

In conclusion, Thinnai-PIQA represents a new frontier in commonsense reasoning evaluation. By focusing on the intersection of physical interaction and deep cultural context, it challenges the AI community to build models that are not just knowledgeable but truly worldly and wise.

Acknowledgements

We express our sincere gratitude to Astalakshmi and Sindhu for providing invaluable insights into the nuances of Tamil knowledge and for their dedicated assistance in the curation of this dataset.