

---

# SamayaPCR: Tamil Culinary Physical Commonsense Reasoning Dataset

---

**Ashvanth S**  
Cohere Labs Community  
ashvanthml@gmail.com

**SenthilNathan M**  
Independent  
msenth.connect@gmail.com

**Suba S**  
Independent  
subasenthil1@gmail.com

September 15, 2025

## Abstract

We present **SamayaPCR**, a manually constructed physical commonsense reasoning dataset for Tamil that addresses the lack of culturally-specific evaluation resources for non-English languages. The dataset is grounded in traditional Tamil culinary practices: it contains **235 task instances** (binary-choice questions in the PIQA-style format) covering **49 distinct recipes** and culinary preparations. Each instance describes a short cooking scenario and candidate solutions, where one demonstrates correct physical reasoning about cooking processes (heat transfer, consistency, timing, ingredient interactions). The dataset targets reasoning about traditional Indian food preparation, incorporating culture-specific cooking techniques, ingredients, and terminology that reflect the lived experiences of Tamil speakers.

## 1 Introduction

Physical commonsense reasoning requires understanding how objects behave in the physical world, including properties like temperature, texture, causality, and temporal sequences. While datasets like PIQA [1] have advanced research in this area for English, there remains a significant gap in culturally-grounded evaluation resources for other languages.

We contribute SamayaPCR, a Tamil physical commonsense reasoning dataset that addresses this gap by focusing on traditional culinary practices familiar to Tamil speakers. Our approach ensures cultural authenticity by grounding reasoning scenarios in well-established cooking traditions rather than direct translations of existing English datasets.

## 2 Dataset Construction

### 2.1 Domain Selection

We selected traditional Tamil cooking as our domain for several reasons: (1) cooking involves rich physical reasoning about heat, consistency changes, timing, and ingredient interactions; (2) traditional recipes represent shared cultural knowledge among Tamil speakers; (3) the domain allows for natural incorporation of Tamil-specific vocabulary and techniques.

### 2.2 Construction Methodology

Each example follows the PIQA format with a prompt and two candidate solutions labeled as correct (1) or incorrect (0). We employed the following systematic approach:

**Prompt Creation** Prompts describe specific cooking scenarios that require physical reasoning, such as preventing milk from curdling, achieving proper consistency, or determining optimal cooking stages

**Solution Generation** For each prompt, we generated two plausible solutions where one demonstrates correct physical understanding while the other contains subtle but critical errors in reasoning about cooking processes.

**Physical Reasoning Types** Our examples target various types of physical reasoning including:

- Heat transfer and temperature control
- Phase changes and consistency detection
- Ingredient interaction and timing
- Surface tension and stirring mechanics
- Cooling and storage processes

### 2.3 Cultural Authenticity

To ensure the dataset reflects genuine lived experience, every instance is rooted in the Tamil language, incorporating vocabulary that is second nature to a native speaker. This includes ingredients like பால் (paal - milk), சர்க்கரை (sarkarai - sugar), நெய் (nei - clarified butter), மாந்திரி (mundhiri - cashew), and ஏலக்காய் (elakkai - cardamom); cooking implements such as the கடாய் (kadai - a type of wok) and கரண்டி (karandi - ladle); and culinary actions like கிளறா (kilaru - to stir continuously), வறுப்பது (varuppathu - to roast or fry).

## 3 Dataset Analysis

The dataset contains 235 manually constructed examples with an even distribution across different types of cuisines and cooking processes. Each example requires an understanding of physical principles such as:

- **Temperature dynamics:** Understanding how heat affects milk proteins, sugar caramelization, and ingredient textures
- **Consistency changes:** Recognizing visual and tactile cues for proper cooking stages
- **Ingredient behavior:** Knowing how different components react under various conditions
- **Timing sequences:** Understanding the importance of when to add ingredients for optimal results

The binary classification task requires models to demonstrate understanding of both the physical processes and the cultural context in which these cooking practices occur.

## 4 Potential Applications

SamayaPCR can serve multiple purposes: (1) evaluating multilingual models' ability to perform physical reasoning in Tamil; (2) assessing cultural knowledge transfer in language models; (3) benchmarking Tamil language understanding capabilities; (4) supporting development of culturally-aware cooking assistance applications.

## 5 Limitations

Our dataset has several limitations that should be considered:

- **Domain-limited:** The dataset primarily focuses on culinary scenarios, with many examples seeded from sweets and household recipes
- **Size constraints:** With 235 examples, the dataset is adequate for evaluation probes but not large-scale training
- **Cultural specificity:** Items are intentionally tied to Tamil culinary practices, which may reduce cross-cultural generalization

## 6 Conclusion

We present SamayaPCR, the first Tamil physical commonsense reasoning dataset grounded in traditional culinary practices. By focusing on culturally authentic scenarios rather than translations, our dataset provides a valuable evaluation resource for assessing multilingual models’ ability to perform physical reasoning in Tamil while respecting cultural context.

## References

- [1] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, 2020.