Methodology for Constructing a Tamil Physical Commonsense Benchmark Dataset

This paper outlines the methodology for creating a manually constructed dataset of 741 Tamil language items designed to evaluate physical and cultural commonsense reasoning. The primary objective was to generate a challenging benchmark where models must discern subtle differences between a correct and an incorrect solution to a physical everyday problem.

Item Construction and Heuristics

Each data item consists of a prompt and two candidate solutions. The prompt describes a common physical scenario or goal rooted in the daily life and culture of Tamil speakers in Sri Lanka (e.g., preventing dampness, cooking local foods, handling monsoonal weather, traditional practices). Prompts were crafted to be multi-sentence or sufficiently detailed (over 25 words where possible) to avoid triviality. To ensure diversity, writers were instructed to vary the opening phrases and topics across items, drawing from domains such as domestic chores, culinary practices, agriculture, and traditional artifacts.

The core challenge lay in generating the pair of candidate solutions (solution0 and solution1). The guiding heuristic was to make them as minimally different as possible while ensuring one was unambiguously correct and the other plausibly incorrect. This was achieved through several techniques:

- Swapping Key Words: Replacing a correct material or action with an incorrect but physically plausible one (e.g., "ரப்பர் விரிப்புகள்" (rubber mats) vs. "களிமண் விரிப்புகள்" (clay mats) for preventing dampness).
- Phrase Reordering: Flipping the order of two steps in a procedure where the sequence is critical for success.
- Introducing Subtle Errors: Adding or removing a single word that invalidates the entire solution without making it absurdly obvious.

The incorrect solution was designed to be compelling enough that a model without robust commonsense might select it, thereby increasing the benchmark's difficulty.

Quality Assurance and Validation

Rigorous validation was paramount. A pool of annotators, all native Tamil speakers from Sri Lanka who ranked in the top 5% of the national GCE AL Examination, was recruited. Each of the curated samples was independently checked by four annotators. Their task was to evaluate each item based on three criteria:

- Linguistic Accuracy: The language is natural, fluent, and culturally appropriate.
- Commonsense Grounding: The correct solution is indeed right, and the incorrect solution is plausibly wrong based on shared cultural and physical knowledge.

- Adherence to Constraints: Each item meets the predefined requirements for length, difficulty, and minimal contrast between solutions.
- And only one solution is correct.

Items were only accepted into the final dataset upon achieving unanimous consensus from all annotators. To ensure dataset integrity, items were also screened for offensive content, unsafe instructions, and harmful stereotypes. Furthermore, a deduplication pipeline employing character n-grams and semantic similarity checks using SBERT embeddings (paraphrase-multilingual-mpnet-base-v2) indexed with FAISS was implemented to make the manual quality assurance and validation process faster with exact, near-duplicate, and semantically equivalent entries. After this careful process, the dataset ended with 741 items.

Inter-Annotator Agreement

To quantify the reliability of the annotations, we measured inter-annotator agreement on a subset of items evaluated by four annotators. We employed two metrics:

1. Percent Agreement: This measured the proportion of items for which all annotators provided identical labels. The observed unanimous agreement was 95%.
2. Fleiss' Kappa: To account for chance agreement, we computed Fleiss' Kappa, which yielded a score of 0.88, indicating an "almost perfect" level of agreement according to standard interpretation scales.

This high degree of consistency demonstrates the robustness of the annotation guidelines and the clarity of the items, ensuring the reliability of the final benchmark dataset.