# Vietnamese Dataset Containing Physical Common Sense Choices

Andrew Tran
EleutherAI
andrew.tran117@outlook.com

September 9, 2025

**Abstract**

We introduce a Vietnamese Language Dataset containing 100 manually constructed pairs of choices on physical commonsense. Each instance contains a short prompt and two plausible options with exactly one correct answer. The dataset emphasizes everyday physical and social reasoning. Prompts contain both Vietnamese cultural contexts and commonplace everyday situations. In this paper, we describe data construction process and dataset.

## 1 Introduction

Physical commonsense reasoning benchmarks are underrepresented among non-English languages, especially with culturally grounded scenarios. We present this dataset to target physical and social commonsense in Vietnamese and provide data items to facilitate multi-lingual studies.

## 2 Task and Data Format

Each example (row) consists of: one **Prompt**, and two **Options**, that are both plausible in isolation, and a single **Label** indicating the correct "Common Sense" choice given the situational context.

## 3 Dataset Construction

All 100 entries in this dataset were brainstormed and generated manually by the author. The author used their imagination as well as past life experiences and memories to construct the prompts.

### 3.1 Manual Authoring Process

- **Domain Topics** Everyday physical interactions that includes dealing with objects, the weather, clothing, routines, safety, school, and simple social norms. Included are also specific cultural situations such as holidays.
- **Prompt heuristics.** Concise sentences establishing context. Context is provided to help "sensically" decide between two options. Avoid niche knowledge that pertains to the authors household and does not generalize across their culture or other people.
- **Option design.** Two plausible options; exactly one is correct and makes "Common Sense" in the context. The distractors should be reasonable and not absurd.

- **Negative patterns to avoid.** Options that are too absurd or trivial without context.

### 3.2 Translation

The author is bilingual and grew up speaking Vietnamese in their household. The author curated each prompt situation from scratch. The author used Google Translate as a secondary verification to confirm the translation for each entry before adding to the dataset. In addition to reading the text, the author used the sound feature in Google Translate to hear the text to confirm whether each translation was correct and naturally sounding. No Generative AI was used to create prompts or solution choices.

## 4 Ethical Considerations

While there was no sensitive or personal data collected, there may be potential biases in the data, such as cultural stereotypes.

## 5 Limitations

Scope restricted to short, everyday scenarios. Answers are provided in binary-choice format. There is limited domain coverage.

## 6 Acknowledgments

Thank You Catherine Arnett for helping provide feedback!

## 7 Dataset File Format

CSV with columns: `prompt`, `solution0`, `solution1`, `label`, and their English parallels.