VN-TRTV: A Vietnamese Culturally-Aware, Inclusive Physical Commonsense Reasoning Dataset

Gia Nghia Ngo
True North International School
Hanoi 12110, Vietnam
nghia.g.ngo2008@truenorth.edu.vn

Thu Mai Do
The Dewey Schools, Hanoi
Hanoi 10000, Vietnam
dtmai@stu.dgs.org.vn

Abstract

Vietnamese, whilst being the native language for over a hundred million speakers in Vietnam and around the world, remain underrepresented in AI training data, with the vast majority of existing datasets predominantly focusing on scholarly and factual knowledge, leaving a gap in physical commonsense reasoning. Compounding this issue is the complete absence of training data regarding any of Vietnam's fifty-three minority ethnic groups due to widespread poverty, a lack of education, and a drought of written material in minority languages. We introduce VN-TRTV, a culturally-aware physical commonsense reasoning dataset that highlights both Kinh Vietnamese culture and minority ethnic culture, broadening the Vietnamese training corpus and providing a platform for the preservation of ethnic culture and tapestry.

1 Introduction

Vietnam frequently leads among the top countries in terms of digital engagement and Internet traffic, yet the majority of current Large Language Models incorporate very little, if any, original Vietnamese sources in their training data. (Alluru, 2024)

Much of the Vietnamese training corpus have so far been concentrated on academic, professional and formal contexts, prioritizing fact-based question-and-answer formats over commonsensical, everyday reasoning. (Bui et al., 2025) (Cuong et al., 2023) This reliance on academic and professional data is also partially attributed to the Saliance Bias, whereby individuals only feel compelled to take note of "special" or "important" phenomena, disregarding common sense in the process. (Lee et al., 2018)

Furthermore, there has been a complete lack of attention paid to the rich cultural and historical tapestry of the fifty-three minority ethnic groups cohabiting present-day Vietnam. Due to both their geographic isolation and structural inequality in resource partitioning, minority ethnic groups in Vietnam are considerably poorer than the two most dominant ethnic groups, the Kinh and the Hoa (Chinese). (Baulch et al., 2002) (Imai et al., 2011) This fundamental gap in income translate to education and academic contribution, with the illteracy rate for ethnic minorities five times that of the Kinh. (Do et al., 2020) This gap in academic corpus and training data is only exarcebated by the fact that twenty-three of the fifty-three ethnic minority groups in Vietnam do not have any kind of written language, and that many minority students no longer use their native language at school or work, instead reverting to Vietnamese. (Linh et al., 2020) In addition, many individuals belonging to minority groups migrate towards urban centres, where Vietnamese and Kinh culture is predominant, in search of work or education, diluting their culture and identity. (Nguyen et al., 2017)

In our paper, we will make the following contributions through introducing VN-TRTV, a comprehensive culturally-aware physical commonsense reasoning dataset that incorporates parts of Vietnamese culture:

- We present the first culturally-aware Vietnamese commonsense physical reasoning dataset, weaving in the often nuanced and distinct Vietnamese culture, using the PIQA-style. (Bisk et al., 2019)
- We introduce some of the first-ever training data specifically focused on under-resourced ethnic minority culture, tapestry and daily life, transliterated into the Vietnamese language.

2 Methodology

2.1 Heuristics

All of our prompts and solution pairs broadly fall into three main descriptors: "none", "cultural" and/or "indigenous." "None" prompts and solutions broadly follow universal physical actions and can readily be translated into English; they tend to include actions and objects commonly used around the world. "Cultural" prompts incorporate one or more aspects of Vietnamese culture, they often assume Vietnamese-specific cultural knowledge and experience to correctly solve, such as cooking and farming methods, common Vietnamese folklore, or well-known cultural events. Lastly, "Indigenous" prompts contain all criteria of the "cultural" tag, however "Indigenous" prompts exclusively focus on minority ethnic culture, traditions, and history. Weaving in traditional folklore and culture, the "Indigenous" prompts are designed to highlight unique knowledge but is completely solvable by the average reasonable Vietnamese.

All prompts and solutions were entirely manually crafted and cross-checked by a team of two native Vietnamese speakers. To differentiate between the two solutions, we mostly employed two very subtle techniques: (1) numerical differentiation, in which dimensions or counts were artificially inflated or deflated to the extent that an average reasonable individual could choose the correct answer with high certainty; and (2) topical differentiation, in which key actions, verbs or steps were omitted or altered such that the resulting solution is, while still appearing reasonable, completely impossible.

The vast majority of prompts and solution pairs broadly follow either one of two heuristics: (1) question-and-answer, where the LLM is asked to pick the more accurate of two answers in response to a written question presented by the prompt; and (2) fill-in-the-blank, where the prompt leaves mid-sentence, asking the LLM to choose the better of two solutions to finish the prompt, given the context of the situation.

The creation and inclusion of certain prompts in our dataset were largely influenced by our own daily lives and events. However, a significant minority of prompts, especially those carrying the "Indigenous" tag, were partly inspired from well-known Vietnamese novels, books, and other reading material.

2.2 Dataset-at-a-Glance and Statistical Analysis

We contributed N = 120 common sense physical reasoning examples as part of our dataset. The shortest prompt + solution pair clocked in at 11 words, whilst the 25th percentile, median and 75th percentile were at 58, 102 and 185 words, respectively. Our longest prompt & solution pair measured at 500 words, a very subtle difference for Large Language Models to spot.

Seventy prompt & solution pairs (58.33%) are tagged as "cultural", requiring specific cultural knowledge and Vietnamese common sense. Out of the seventy pairs tagged "cultural", twenty-three (32.86%) were tagged as "indigenous", showing our focus on indigenous culture, a facet often underrepresented or omitted entirely from previous Vietnamese training datasets, benchmarks and corpus data.

References

L. Alluru. 2024. How to make generative ai better for non-english speakers. *Analytics Magazine*.

Bob Baulch, Truong Thi Kim Chuyen, Dominique Haughton, and Jonathan Haughton. 2002. Ethnic minority development in vietnam: A socioeconomic perspective. Policy Research Working Paper 2836, The World Bank.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Cuc Thi Bui, Nguyen Truong Son, Truong Van Trang, Lam Viet Phung, Nhut Huy Pham, Hoang Anh Le, Quoc Huu Van, Phong Nguyen-Thuan Do, Van Le Tran Truc, Duc Thanh Chau, and Le-Minh Nguyen. 2025. Vmlu benchmarks: A comprehensive benchmark toolkit for vietnamese llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11495–11515, Toronto, Canada. Association for Computational Linguistics.

Le Anh Cuong, Nguyen Trong Hieu, Nguyen Viet Cuong, Nguyen Ngoc Quoc, Le-Minh Nguyen, and Cam-Tu Nguyen. 2023. Vlsp 2023 challenge on vietnamese large language models.

Ha Thi Hai Do, Nui Dang Nguyen, Anh Ngoc Mai, Duc Minh Phung, Nguyen Duy Vu, Hung Dinh Nguyen, Dung Manh Tran, and Thuy Minh Thu Phung. 2020. Ensuring basic education for ethnic minority groups in vietnam. *Management Science Letters*, 10(12):2805–2812.

Katsushi S. Imai, Raghav Gaiha, and Woojin Kang. 2011. Poverty, inequality and ethnic minorities in vietnam. *International Review of Applied Economics*, 25(3):249–282.

Ho Cheung Brian Lee, Sulin Ba, Xinxin Li, and Jan Stallaert. 2018. Salience bias in crowdsourcing contests. *Information Systems Research*, 29(2):401–418.

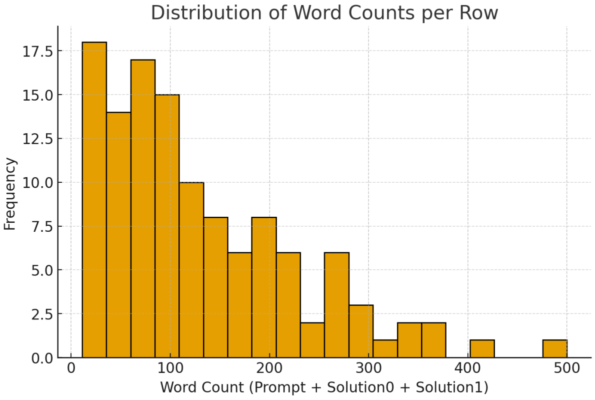

Figure 1: Word Count (Prompt + Solution0 + Solution1) Distribution

Ha Thi Kim Linh, Nguyen Thi Tinh, and Huynh Tan Hoi. 2020. Protection of ethnic language of ethnic minority students in schools. In *Proceedings of the 2nd World Symposium on Social Sciences and Education (WSSE 2020)*.

Loc Duc Nguyen, Ulrike Grote, and Rasadhika Sharma. 2017. Staying in the cities or returning home? an analysis of the rural-urban migration behavior in vietnam. *IZA Journal of Development and Migration*, 7(3):1–18.

A Dataset Length Distribution