MRL-Malay: Malaysian Commonsense and Math Reasoning Dataset

Zhi Rui Tam
National Taiwan University
d14922019@ntu.edu.tw

Abstract

We introduce MRL-Malay, a benchmark dataset of 122 two-choice questions for evaluating physical common sense and mathematical reasoning in Bahasa Melayu. Although Malay is compulsory for all Malaysians, it lacks comprehensive NLP evaluation resources in the LLM era. Our data set comprises: (i) human-written items testing culturally grounded common sense and social norms in Malaysian contexts, and (ii) culturally adapted GSM8K math problems. We incorporate authentic Malaysian usage including colloquial expressions, local entities, and internet abbreviations (e.g., "utk" for "untuk"), with 30% of common phrases in shortened forms. All elements undergo human verification for cultural suitability and translation quality. MRL-Malay addresses the critical need for localized benchmarks that evaluate both reasoning and cultural understanding in Malay language models.

1 Introduction

Bahasa Melayu (Malay) is a compulsory subject for all Malaysians throughout six years of primary education. While Indonesian shares extensive lexical and semantic overlap with Malay and is spoken by a larger population, Malay still lacks broad, standardized evaluation resources for natural language understanding and reasoning. In the era of large language models (LLMs)(Zolkepli et al., 2024), high-quality Malay benchmarks have only recently begun to emerge (Poh et al., 2024).

In this report we present **MRL–Malay**, a two-choice dataset in the format from PiQA (Bisk et al., 2020) targeting culturally grounded common sense and step-by-step reasoning in Malay. The current release contains **122** items curated as (i) human-written questions that reflect local knowledge and online writing conventions, and (ii) translated arithmetic/word-problem reasoning items. We emphasize translation fidelity, natural Malay phrasing, and colloquial register where appropriate (e.g., shortened forms like *"utk"* for *"untuk"*, *"nak"* for *"hendak"*).

1.1 Human Written Questions

Our human-authored items are designed to probe *local* common sense, physical causality, and social norms in Malaysian contexts. Rather than asking for definitions, each prompt frames a short scenario and requires choosing the more plausible consequence, explanation, or norm-consistent action. We intentionally use natural Malay, including colloquial particles and contractions, to mirror how people actually write online.

Authoring principles.

- **Local grounding.** Questions revolve around lived Malaysian settings: food and drink (e.g., *kopi*, *otak-otak*, *lemang*), religious life (*masjid*, *puasa*), and everyday routines (*kampung*, pasar, cuaca tropika).
- **Reasoning over recall.** Prompts are mini-narratives that require cause–effect or social reasoning (not keyword lookup). Distractors are plausible at a glance but violate a physical fact, constraint, or norm.
- **Clear decision boundary.** Exactly one option is intended to be correct. We avoid ambiguous readings and rewrite items until ties are eliminated.
- **Natural phrasing.** Register and lexicon follow contemporary usage (e.g., contractions, particles) while remaining respectful in cultural or religious contexts.

Example (from the dataset).

> **Prompt:** *"kalau saya ada kopi yang panas tapi saya letak bagi dia reda sidiki dan saya tumpah susu pekat. Kopi ini*

akan"
"If I have hot coffee, but I let it cool a little and I pour in condensed milk. This coffee will"

Options:

0. *menjadi manis*
become sweet (correct)

1. *masih pahit*
still be bitter

This item targets everyday physical/culinary causality framed in colloquial Malay. The two-choice format discourages surface pattern matching by requiring integration of several cues (*kopi panas* → dibiarkan *reda*, kemudian ditambah *susu pekat*). Similar items in this subset cover cultural knowledge (e.g., ingredients and practices), norm-sensitive scenarios (e.g., greetings, etiquette), and simple physics of daily life (e.g., heat, containers, weather).

2 Math Reasoning Translations

We select partial GSM8K(Cobbe et al., 2021) questions as our source of math questions, first we use a 5-shot english to malay translation prompt to ensures all translated words do not contain any potential Indonesia words. After the first round of translation alot of the names, monetary units are still in the original US form. So we added another cultural translation prompt which converts to Malaysia culture names and units. In this step, the name "Olivia" was found to be translated to "Aisyah" which is a common name found in malaysia malay ethics female. Finally we added a verification step which the prompt validates if the values and semantic meaning remain the same from the original form. Lastly we use weaker model (Gemma2-27B (Team, 2024)) to mine for incorrect model while use gemmini-2.5-pro(Comanici et al., 2025) to answer for the correct answer. For the correct and incorrect solution we ensures the final answer is both correct and incorrect when compared to the ground truth.

2.1 Short forms

To make the questions more local to how Malaysian types online, we also remap 30% of the commonly used phrases to their short form. We styled the data to reflect the current use of Malaysian Internet while preserving mathematical fidelity. To that end, we apply a lightweight colloquialization pass that substitutes a subset of high-frequency function words with widely intelligible short forms used in Malaysian social media and messaging.

Design. A curated lexicon of abbreviation→full-form pairs was compiled by native speakers (e.g., *dgn→dengan*, *yg→yang*, *utk→untuk*). The procedure operates at the word *type* level: for each full-form type that appears in an instance and is covered by the lexicon, a Bernoulli trial with probability p=0.3 decides whether that type is rendered in its short form. If selected, all whole-word, case-insensitive occurrences of the type are replaced. This yields realistic mixtures of formal and colloquial tokens without over-abbreviating any single item. A fixed random seed ensures reproducibility.

Illustration. *Formal*: *Ali membeli 3 buku dengan harga RM12 setiap satu. Berapakah jumlah yang perlu dibayar?*
Colloquialized (target 30%): *Ali beli 3 buku dgn harga RM12 setiap satu. Brp jumlah yg perlu dibayar?*

Limitations and decisions. We intentionally include several shorthands on social networks that can be visually compact (e.g. *x* for *tidak*) to match the usage in the wild. While whole-word matching avoids altering algebraic variables or operators, such forms may still increase reader ambiguity in math-heavy contexts; we therefore report style-normalized results in addition to aggregate scores. We avoid Indonesian-leaning variants where Malaysian usage diverges, and exclude ambiguous patterns that could clash with titles or abbreviations (e.g., medical honorifics) via strict word-boundary matching.

2.2 Human verification

Finally we manually validate a subsets of 100 questions which is challenging enough that gemini-2.0-flash fails to answer correctly and inspect the translation and answer quality. During this process, we also substitue some of the location entity into real names found in Malaysia. Only 69 questions are selected with the rest filtered due to culture mismatch or low quality translation.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Soon Chang Poh, Sze Jue Yang, Jeraelyn Ming Li Tan, Lawrence Leroy Tze Yao Chieng, Jia Xuan Tan, Zhenyu Yu, Foong Chee Mun, and Chee Seng Chan. 2024. MalayMMLU: A multitask benchmark for the low-resource Malay language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 650–669, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. 2024. Mallam–malaysia large language model. *arXiv preprint arXiv:2401.14680*.