Dataset Description Indonesia

Salsabila Zahirah Pranida[1], **Jauza Akbar Krito**[2]
[1]MBZUAI, [2]Universitas Gadjah Mada

Both authors contributed equally.

1 Dataset

The dataset is written in Bahasa Indonesia, the national variety used in formal contexts, and does not include regional dialects (will be considered in future work). It covers 12 cultural topics and 63 subtopics adapted from IndoCulture. All items were independently reviewed by two native speakers for cultural relevance, ambiguity, correctness, physical commonsense, and naturalness, with ambiguous or unclear cases removed. The final dataset consists of 228 validated items.

1.1 Dataset Creation

In constructing our dataset, we followed a structured pipeline as illustrated in Figure 1. The process began with the generation of raw items using LLM-assisted prompts with GPT-4o under strict constraints. To maintain consistency, we defined a minimum word length of 30 words for each generated item, which includes both the prompt and its corresponding solutions. Each solution was required to contain at least seven words, and any response shorter than this threshold was discarded. In addition, we placed emphasis on minimizing variation across solutions, ensuring that paired solutions were of comparable length and exhibited a similar level of complexity. This restriction was crucial for preventing overly short or trivial solutions that could compromise the overall quality of the dataset.

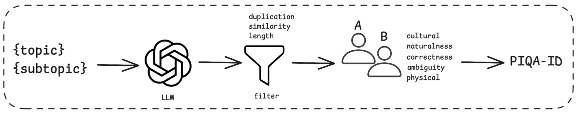

Figure 1: Pipeline of the dataset creation process.

Another key requirement was the integration of cultural elements. The generated items were designed to reflect cultural relevance with an emphasis on Indonesian identity and diversity. To ensure broad thematic coverage, each prompt was assigned a specific topic and subtopic. In total, the dataset was constructed around 12 topics encompassing 63 diverse subtopics adapted from IndoCulture (Koto et al., 2024). These topics include themes such as *Agriculture, Art, Daily Activities, Death, Family Relationships, Fisheries and Trade, Food, Pregnancy and Childcare, Religious Holidays, Socio-religious Aspects of Life, Traditional Games*, and *Wedding traditions*.

Across all topic and subtopic combinations, the generated items underwent multiple filtering stages. Only about 25% of the initial outputs satisfied the predefined constraints. Following this stage, we applied a near-duplication check using SentenceBERT (Reimers and Gurevych, 2019), with a similarity threshold of 0.9. This process revealed that ≈19% of the items, or around 88 entries, were highly similar. To preserve variety, we dropped these near-duplicates. The overall statistics are summarized in Table 1.

Stage	**Count**
Generated (raw)	1810
Validated (pass)	466
Near-duplicate items detected	88
Final dataset size	**378**

Table 1: Dataset statistics across the stages of generation, validation, and deduplication.

As an illustration, Table 2 shows a pair of prompts with a SentenceBERT similarity score of 0.93. The highlighted overlaps in the harvesting context, along with the parallel solution structures.

After all filtering and refinement procedures, the final dataset consisted of 417 valid items, corresponding to a success rate of approximately 23% of the originally generated data. These entries represent only those that fully met the requirements. To further characterize the dataset, we analyzed the mean word counts of the prompts and solutions, as

Prompt A	Prompt B
Prompt: Ketika memanen padi di sawah , mana yang lebih efektif untuk mengumpulkan butiran padi setelah dipotong? *(When harvesting rice in the field, which is more effective for collecting rice grains after cutting?)*	**Prompt:** Ketika memanen padi di sawah , petani biasanya menggunakan ani-ani untuk memotong batang padi. Mana metode yang lebih efisien untuk mengumpulkan hasil panen dengan cepat ? *(When harvesting rice in the field, farmers commonly use the ani-ani to cut the stalks. Which method is more efficient for quickly collecting the harvest?)*
Sol0: Gunakan tampah bambu untuk menjemur dan mengumpulkan padi. *(Use a bamboo tray to sun-dry and collect the rice.)*	**Sol0:** Menggunakan karung plastik untuk menampung padi yang sudah dipotong. *(Use a plastic sack to hold the cut rice.)*
Sol1: Gunakan kantong plastik besar untuk menjemur padi. *(Use a large plastic bag to sun-dry the rice.)*	**Sol1:** Menggunakan keranjang bambu untuk menampung padi yang sudah dipotong. *(Use a bamboo basket to hold the cut rice.)*

Table 2: Overlapping parts highlighted in yellow show why these two prompts are near-duplicates. Both share the same harvesting-rice context and use similar solution structures (bamboo vs. plastic container).

Text Type	Mean Word Count
Prompt	21.10
Solution0	9.84
Solution1	9.60

Table 3: Mean sentence lengths (in words) for prompts and solutions in the final dataset.

summarized in Table 3. On average, prompts were more elaborate with longer word counts, while solutions were shorter yet maintained sufficient length to ensure meaningful variation within the dataset.

Beyond average statistics, the distribution of sentence lengths across text types is shown in Figure 2. Prompts exhibit a wider spread. In contrast, both solutions display a more compact distribution.

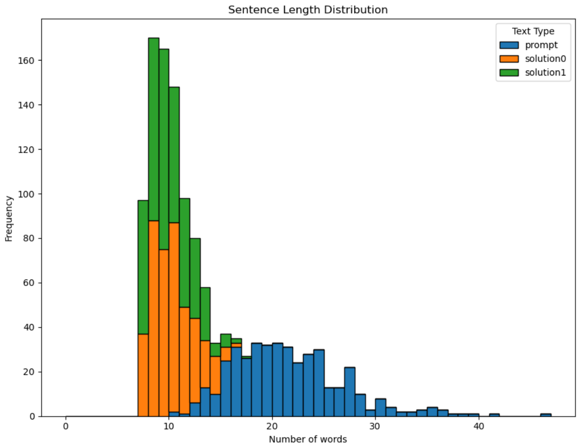

Figure 2: Sentence length distribution for prompts and solutions.

Meanwhile, the distribution of dataset items across topics is shown in Figure 3. The topics are unevenly represented, with *Food, Wedding*, and *Socio-religious Aspects of Life* are the most frequent. In contrast, categories such as *Agriculture, Traditional Games*, and *Fisheries and Trade* contain fewer than 20 items each. This distribution shows a certain themes naturally dominate due to their higher cultural relevance in Indonesian daily life.

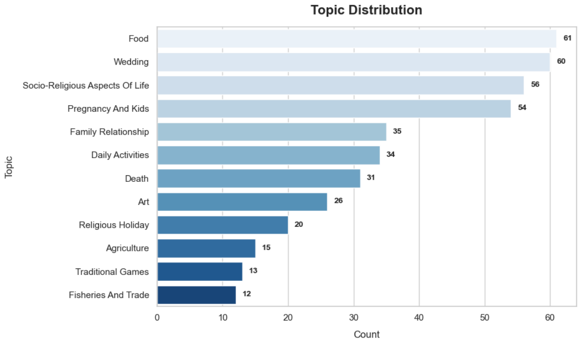

Figure 3: Distribution of dataset items across the 12 cultural topics.

1.2 Dataset Validation

After generation, all dataset items underwent a careful validation process. To ensure quality, we conducted both post-editing and systematic checking across multiple dimensions, including cultural relevance, ambiguity, correctness, physical common sense, and overall naturalness. Each item was independently reviewed by **two native speakers** for cross-validation.

During the review, we carried out targeted post-editing to improve the quality of the data. This involved **(i)** fixing cases where cultural grounding was inaccurate, **(ii)** where solutions were too easy or not distinct enough, **(iii)** where the length or phrasing needed adjustment for natural flow, and **(iv)** where the label was wrong. In some cases, solutions were rewritten entirely.

Step	**Count**
Initial validated data	378
Ambiguity	-113
Remaining after exclusion	265
Physical + Correctness	228
Cultural relevance	45
Naturalness	95%
Final dataset used	228

Table 4: Stepwise dataset validation and filtering results

After validation, we removed items that both annotators marked as ambiguous, and we only kept items where both agreed that the answer was correct and demonstrated physical common sense. We also identified and marked the items that had cultural relevance. As shown in Table 4, about 17% of the final data reflects cultural value.

References

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.