Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese

Rifki Afina Putri[1], **Faiz Ghifari Haznitrama**[1], **Dea Adhista**[2], **Alice Oh**[1]
[1]KAIST, Republic of Korea [2]Independent Researcher
`{rifkiaputri,haznitrama}@kaist.ac.kr`, `deadhista@gmail.com`, `alice.oh@kaist.edu`

Abstract

Large Language Models (LLMs) are increasingly being used to generate synthetic data for training and evaluating models. However, it is unclear whether they can generate a good quality of question answering (QA) dataset that incorporates knowledge and cultural nuance embedded in a language, especially for low-resource languages. In this study, we investigate the effectiveness of using LLMs in generating culturally relevant commonsense QA datasets for Indonesian and Sundanese languages. To do so, we create datasets for these languages using various methods involving both LLMs and human annotators, resulting in ∼4.5K questions per language (∼9K in total), making our dataset the largest of its kind. Our experiments show that automatic data adaptation from an existing English dataset is less effective for Sundanese. Interestingly, using the direct generation method on the target language, GPT-4 Turbo can generate questions with adequate general knowledge in both languages, albeit not as culturally 'deep' as humans. We also observe a higher occurrence of fluency errors in the Sundanese dataset, highlighting the discrepancy between medium- and lower-resource languages.[1]

1 Introduction

The development of Large Language Models (LLMs) is significantly impacting NLP, leading to an increasing trend in the automated generation of datasets, particularly for question answering (QA) tasks. However, a major challenge arises with underrepresented languages like Indonesian and Sundanese due to the need for cultural context. For the generated data to be fully useful, it must not only be linguistically accurate, but it also needs to reflect the cultural nuances, historical references, and social norms. It is not yet clear whether current LLMs can create QA data that adequately includes the cultural nuances specific to certain languages.

Another common method for constructing non-English datasets is by using machine translation. Although more scalable, this method cannot be straightforwardly applied due to the contextual irrelevancy of the data, primarily due to geographical differences (which could also influence cultural differences). For example, many English CommonsenseQA questions (Talmor et al., 2019) include concepts such as *snow* or any *winter sports*, which are irrelevant in Indonesia due to its tropical climate with only two seasons. This dataset also often includes English-centric names and locations, primarily limited to the US, which are not considered "commonsense" for Indonesians.

Therefore, in this study, we aim to investigate how well current LLMs can adapt and generate a commonsense QA dataset that is both linguistically accurate and culturally relevant to Indonesia. We focus on Indonesian, the lingua franca of Indonesia, and Sundanese, one of the regional languages in Indonesia with 34 million speakers, yet still considered low-resource (Aji et al., 2022). To address dataset scarcity, especially in Sundanese, we also manually constructed datasets by involving annotators from various regions in Indonesia, ensuring a representation of diverse cultural perspectives. To sum up, our contributions are as follows:

- We create a new Indonesian and Sundanese CommonsenseQA dataset using various methods (Figure 1), including adapting existing English data (LLM_ADAPT) and generating new datasets manually (HUMAN_GEN) and automatically (LLM_GEN). The dataset contains ∼3K LLM-generated and ∼1.5K human-generated question-answer pairs per language, which, to our knowledge, is the largest culturally nuanced commonsense QA dataset in both Indonesian and, particularly, Sundanese.

---

[1] All datasets and codes in this work are available at `https://github.com/rifkiaputri/id-csqa`.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20571–20590
November 12-16, 2024 ©2024 Association for Computational Linguistics

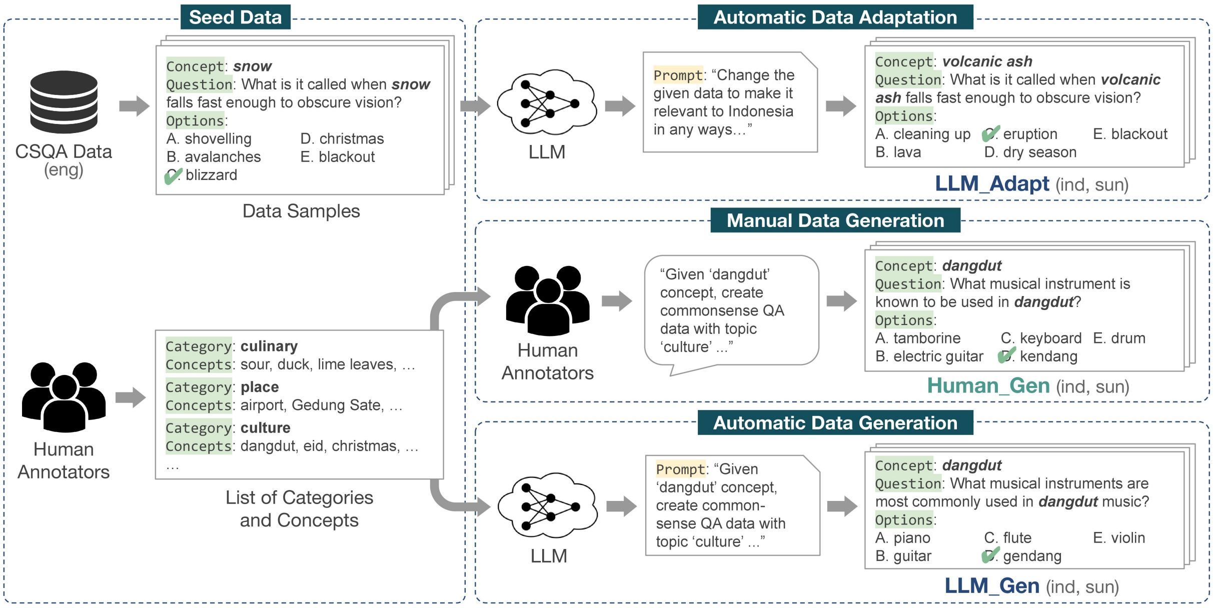
Seed Data
CSQA Data (eng)
Concept: snow
Question: What is it called when snow falls fast enough to obscure vision?
Options:
A. shovelling
B. avalanches
C. blizzard
D. christmas
E. blackout
Data Samples
Human Annotators
Category: culinary
Concepts: sour, duck, lime leaves, ...
Category: place
Concepts: airport, Gedung Sate, ...
Category: culture
Concepts: dangdut, eid, christmas, ...
...
List of Categories and Concepts
Automatic Data Adaptation
LLM
Prompt: "Change the given data to make it relevant to Indonesia in any ways..."
Concept: volcanic ash
Question: What is it called when volcanic ash falls fast enough to obscure vision?
Options:
A. cleaning up
B. lava
C. eruption
D. dry season
E. blackout
LLM_Adapt (ind, sun)
Manual Data Generation
Human Annotators
"Given 'dangdut' concept, create commonsense QA data with topic 'culture' ..."
Concept: dangdut
Question: What musical instrument is known to be used in dangdut?
Options:
A. tamborine
B. electric guitar
C. keyboard
D. kendang
E. drum
Human_Gen (ind, sun)
Automatic Data Generation
LLM
Prompt: "Given 'dangdut' concept, create common-sense QA data with topic 'culture' ..."
Concept: dangdut
Question: What musical instruments are most commonly used in dangdut music?
Options:
A. piano
B. guitar
C. flute
D. gendang
E. violin
LLM_Gen (ind, sun)

Figure 1: Our dataset generation methods. The examples of LLM_ADAPT, HUMAN_GEN, and LLM_GEN datasets are shown in English for clarity. The original versions of these datasets are in Indonesian and Sundanese.

- We perform a comprehensive analysis to assess the effectiveness of LLMs in creating a culturally relevant commonsense QA dataset. We find that adaptation from English data is less effective, especially for Sundanese. However, GPT-4 Turbo can generate questions with some basic local knowledge in Indonesian and Sundanese when provided with human-created categories and concepts.

- We evaluate several LLMs on our dataset and find that they perform better on LLM-generated data than on human-generated data, indicating that the former is less challenging, especially for proprietary models like GPT-4. Nevertheless, many open-source LLMs still struggle to answer LLM-generated questions, highlighting significant room for improvement in these datasets.

2 Related Work

Commonsense Dataset Datasets for commonsense reasoning are crucial for developing models that understand and reason about real-world complexities. Datasets like COPA (Roemmele et al., 2011), X-COPA (Ponti et al., 2020), and The Winograd Schema Challenge (Levesque et al., 2012) evaluate causal reasoning in real-world scenarios. Other datasets like ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and Mcscript (Ostermann et al., 2018) test commonsense reasoning via QA, but some questions require grade-school science knowledge. CommonsenseQA (Talmor et al., 2019) presents the type of "purely" commonsense QA in the form of multiple-choice questions based on knowledge from ConceptNet (Speer et al., 2017), built through crowdsourcing. However, the crowdworkers bring their cultural background as part of their common knowledge in building the data, resulting in many questions that are "commonsense" only in the Western culture.

Cultural Evaluation Dataset Various datasets are used to assess language models' ability to understand cultural nuances specific to a language. FORK (Palta and Rudinger, 2023) explores culinary cultural biases and assumptions of US, Indian, and Chinese customs. IndoMMLU (Koto et al., 2023) includes questions from Indonesian exams and covers regional cultural topics, such as the Minangkabau or Sundanese cultures. COPAL-ID (Wibowo et al., 2023) is a COPA-style dataset written by native speakers, thus incorporating more Indonesian cultural nuances compared to X-COPA. COPAL-ID is composed of approximately 500 questions, with a primary emphasis on the cultural aspects of the Jakarta region. In contrast, our dataset is much bigger and covers a broader range of annotators from various regions of Java and Bali, in addition to Jakarta. We also include Sundanese, addressing the gap in this low-resource language.

3 Background

3.1 Commonsense QA: Definition and Scope

In our dataset, we focus on building questions that probe common or cultural knowledge in daily life within Indonesian and Sundanese contexts. Each data point is a triple of *concept*, *question*, and *options*, with one correct answer. To enhance the relevance of the dataset to local contexts and reduce Western cultural bias, we incorporate both existing question concepts from the English CommonsenseQA (Talmor et al., 2019) and novel question concepts that we manually created.

Unlike English CommonsenseQA, our dataset includes question category metadata, covering five categories: *culinary*, *place*, *culture*, *history*, and *activity*. Each category has 150 unique question concepts, ensuring broad domain and knowledge coverage.[2] Another key difference is that English CommonsenseQA concepts are sourced from a knowledge base, and they divide the concept sets into questions and options. In contrast, our dataset provides only a question concept and its category as input. The options, including the correct answer, are entirely generated by the data creator.

3.2 Languages in Indonesia

Indonesia is one of the most culturally and linguistically diverse countries, with more than 700 languages spoken across the country (Aji et al., 2022; Eberhard et al., 2021). Among the many languages, Indonesian is a unifying language used nationally. It utilizes the Latin script and was developed from literary "Classical Malay" (Sneddon, 2003), with regional variations. Over 80% of Standard Malay's vocabulary is similar to Indonesian.

Apart from Indonesian, regional languages like Sundanese are spoken by people of the same ethnicity. It is primarily spoken in West Java and is the second-largest regional language in Indonesia, with 34 million speakers (Eberhard et al., 2021). Regional languages, including Sundanese, have influenced the formation and development of the Indonesian language. Both languages share similarities, such as their grammatical structure, but also differ significantly in aspects like the number of vowels (i.e., Sundanese has 2 additional vowels: é, eu) and morphological features, including affixes.

We chose to study the national language and one regional language to illustrate the differences in the commonsense QA data generated via LLMs. It is also important to note that despite its large number of speakers, Sundanese has very limited data, and to date, there are no commonsense QA datasets available in Sundanese.

Dataset Version	**Indonesian**		**Sundanese**	
	Train / Valid / Test	**Total**	**Train / Valid / Test**	**Total**
LLM_ADAPT	1,506 / 191 / 158	**1,855**	1,506 / 191 / 158	**1,855**
HUMAN_GEN	0 / 0 / 1,498	**1,498**	0 / 0 / 1,499	**1,499**
LLM_GEN	0 / 0 / 1,063	**1,063**	0 / 0 / 1,183	**1,183**
	1,506 / 191 / 2,719	**4,416**	1,506 / 191 / 2,840	**4,537**

Table 1: Statistics of our generated Indonesian and Sundanese CommonsenseQA dataset. We retained the original English CommonsenseQA splits in LLM_ADAPT to avoid data contamination.

4 Data Generation Methods

To investigate whether LLMs can generate culturally relevant commonsense QA data in Indonesian and Sundanese, we build a dataset using various methods with LLMs as data generators. We also employ humans to generate data for comparison.

As illustrated in Figure 1, we apply three dataset generation methods: (1) **Automatic Data Adaptation** (LLM_ADAPT), where we leverage LLMs to automatically adapt English CommonsenseQA data to our target languages; (2) **Manual Data Generation** (HUMAN_GEN), where we ask native-speaker human annotators to manually construct the dataset; and (3) **Automatic Data Generation** (LLM_GEN), where we use LLMs to generate data based on the list of categories and concepts used in HUMAN_GEN. Table 1 shows the statistics of our final dataset, containing a total of **8,953** QA pairs (4,416 Indonesian and 4,537 Sundanese).

While all prompts given to the LLM are in English, for the LLM_ADAPT data, we specifically instruct the model to produce responses directly in Indonesian. Subsequently, we translate the resulting Indonesian data into Sundanese. In contrast, for the LLM_GEN data, we instruct the model to give the output directly in Indonesian and also Sundanese, to closely replicate the data generation process used for HUMAN_GEN. The generation methods for each dataset are detailed below.

4.1 Automatic Data Adaptation

We build the first LLM-generated data by adapting English CommonsenseQA dataset (Talmor et al., 2019) to make it culturally relevant to Indonesian and Sundanese.

---

[2] More details on how we curate our categories and concepts can be seen in Appendix B.

Data Selection To select the data to be adapted, we first remove data containing offensive keywords and those with duplicate or similar options in different tenses. Unlike English, Indonesian and Sundanese do not change form to indicate tenses (e.g., 'passed' and 'passing' both translate to *'lulus'*).

Then, we sample the data by assessing three elements: *concept*, *name*, and *location*. Data that are considered irrelevant in at least one of the three elements are selected to be adapted. We take question concepts from the existing CommonsenseQA data and use Stanza (Qi et al., 2020) and ConceptNet API[3] for name and location extraction.

To determine concept relevance, we utilize GPT-3.5 Turbo with five different prompts to ensure reliable results, asking whether a concept is relevant in Indonesia or West Java.[4] For example, given the concept 'snow', one prompt is, *"Can snow be found in West Java? Answer with only 'yes' or 'no'."* A 'yes' means the concept is relevant. Final relevancy is determined by majority voting. Data containing person names and/or locations are regarded as irrelevant by default to save the relevancy classifier API cost, as our preliminary experiment showed all such data are Western-centric. From the total of $\sim$12K QA pairs in English CommonsenseQA, $\sim$2K are selected to be adapted.

QA Pairs Adaptation The next step is to transform the selected irrelevant data. First, we prompt GPT-4 Turbo[5] to rephrase the sampled questions and options to align with Indonesian cultures. Subsequently, for data flagged with Western-centric names, we use GPT-3.5 Turbo to replace all person names. The fully adapted data are then translated from Indonesian to Sundanese using Google Translation API. We choose machine translation due to the unreliability of direct adaptation, as the concepts are sourced from English data. Our preliminary run on `eng` to `sun` adaptation shows that 90% of the samples contained errors, with half due to hallucinations. This occurred less frequently in `eng` to `ind` adaptation. For example, for the 'bald eagle' concept, GPT-4 Turbo generated *'elang Jawa'* (Javan hawk-eagle) in Indonesian, but it generated the non-existent concept *'garuda puspa'* (literally translates to eagle flower) in Sundanese. Although not ideal, translation from the adapted Indonesian QA pairs generally provides better data quality.

Data Filtering Finally, we eliminate low-quality data by removing instances where the concepts do not appear in the questions. We also utilize a back-translation method to filter out poor translations of Indonesian (`ind`) to Sundanese (`sun`) data. We discard QA pairs if the similarity between the original Indonesian (`ind`) and the back-translated version (`ind'`) is less than 0.9. We measure similarity using an embedding-based metric with multilingual MiniLM (Reimers and Gurevych, 2019) and LaBSE (Feng et al., 2022) models, and keep the data if either similarity score is above 0.9.

4.2 Manual Data Generation

To build the human-generated dataset, we recruited 12 experienced annotators from diverse regions across Java and Bali. All are native speakers of Indonesian (and Sundanese), with a minimum of 15 years residing in their particular target language regions. Among them, 5 annotators have a degree in linguistics. These criteria ensure they possess both language proficiency and cultural familiarity. As for the data collection process, we have two main phases: (1) creating commonsense question-answer pairs and (2) answering commonsense questions. We also perform quality control to ensure the data quality.[6]

QA Pairs Creation We first instruct the annotators to create commonsense question-answer pairs based on the given category and question concept that we had newly created (§3.1). We also ask them to rely on their existing knowledge when making questions and avoid using internet search or LLMs as much as possible. Each concept is annotated by 2 annotators, with each annotator covering all 5 categories to ensure consistency. Since we have 150 unique concepts per category and 6 annotators per language, each annotator is assigned to create QA pairs for 50 concepts per category. This results in 300 QA pairs per category (150 concepts × 2 annotators) and a total of 1,500 QA pairs per language (300 QA pairs × 5 categories).

Answering Question After passing the quality assurance, the 1,500 QA pairs are then redistributed among all annotators for further review, where they are tasked with answering 1,250 commonsense

---

[3] `https://github.com/commonsense/conceptnet5/wiki/API`

[4] All data generation prompts are detailed in Appendix A.1.

[5] Our initial experiment (Appendix A.2) reveals that GPT-4 Turbo significantly outperforms Merak-v4, an open Indonesian LLM, leading us to select GPT-4 Turbo for our work.

[6] More details on annotators' demographic information and annotation guidelines are presented in Appendix C and D.

questions each (excluding the set of data they have made in the first phase). Therefore, in this phase, each question is answered by five different annotators to capture the consistency and variance in the commonsense knowledge among annotators from various cultural backgrounds. Annotators are instructed not only to provide answers to the questions but also to comment on any ambiguities in the questions or options, or if they have any uncertainties when answering the questions.

Quality Control We conduct Quality Control (QC) through manual validation to ensure data accuracy and maintain its "commonsense" nature. QC annotators manually review data for errors and provide feedback, which is then corrected by the annotator concerned. Once corrected, QC annotators re-check the data to determine whether it can be considered complete or still requires revision. An evaluation meeting with all annotators is also held to convey a more comprehensive evaluation regarding all types of errors and other findings, ensuring annotators do not repeat similar errors and have a better understanding of the notion of commonsense. After the second phase (answering questions), we also excluded some questions if more than three annotators marked them as ambiguous.

4.3 Automatic Data Generation

For generating the second type of LLM-generated data, rather than adapting questions from the English dataset (§4.1), we use the same set of categories and question concepts as the human-generated data (§4.2). We also utilize GPT-4 Turbo and instruct it to generate questions, options, and answers. To ensure the generated dataset aligns closely with the intended cultural context, we explicitly incorporated the categories and question concepts in the prompt. Additionally, we instruct the model to strictly include the question concepts in the generated questions.

We adopt a batching approach to streamline the data generation process, providing the model with a maximum of 5 distinct question concepts from the same categories in one API call. We chose 5 concepts because this number is optimal; larger batches would exceed the API's maximum length, while smaller batches would result in longer processing times.

To maintain dataset quality and uniqueness, we filter out duplicate entries and questions that do not explicitly contain the question concepts.

Status	**Num (%)**	**Concept Example**	
		Orig (`eng`)	**Modified (`ind`)**
Correct (major)	32 (20.25%)	beaver	komodo
		snow	hujan abu vulkanik (*volcanic ashfall*)
Correct (minor)	119 (75.32%)	tower	menara (*tower*)
		grape	anggur (*grape*)
Wrong	7 (4.43%)	orchestra pit	sumur orkestra (*orchestra well*)
		skate	ice skating

Table 2: `eng` to `ind` concept adaptation result.

Status	**Num (%)**	**Concept Example**	
		Orig (`ind`)	**Modified (`sun`)**
Correct	122 (77.22%)	hujan abu vulkanik (*volcanic ashfall*)	hujan lebu vulkanik (*volcanic ashfall*)
		menara (*tower*)	munara (*tower*)
Wrong	36 (22.78%)	cicak (*house gecko*)	kadal imah (*house lizard*)
		klinik gigi (*dental clinic*)	klinik dental (*dental clinic*)

Table 3: `ind` to `sun` concept adaptation result.

5 Data Analysis

5.1 LLM-Generated Data

To evaluate the quality of our LLM-generated data, we manually reviewed all 158 samples from the LLM_ADAPT test set (see Table 1). Based on the evaluation, we then calculate the accuracy of the generated concepts, questions, and options. We also evaluate 300 randomly selected samples from the LLM_GEN using the same procedure.

5.1.1 Concept Analysis

Concept Quality To evaluate the quality of the adapted concepts, we regard the concept as correct if it is a real, existing concept (not a hallucination) and relevant to the Indonesian or Sundanese context. As shown in Table 2, 95.57% of English to Indonesian adaptations are correct, including 16 out of 19 concepts needing major adaptation, such as 'snow' to '*hujan abu vulkanik*' (volcanic ashfall). This result indicates that the LLM can adequately adapt some English concepts to Indonesian, although most involve direct translations without major alterations. However, for Indonesian to Sundanese (Table 3), the correct adaptation drops to 77.22%, reflecting weak machine translation (MT) performance in Sundanese.

Concept Variation Despite the high accuracy of concept adaptation, as shown in Figure 2, the

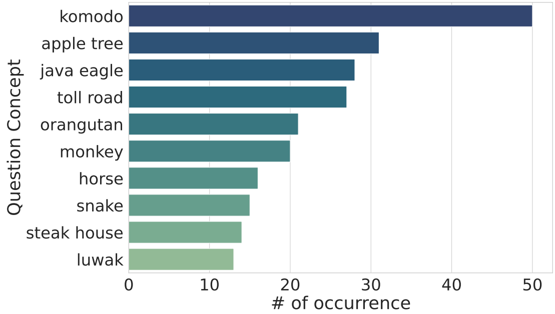

Figure 2: Top-10 adapted question concepts taken from train, validation, and test set of LLM_ADAPT data.

adapted concepts are skewed towards *'komodo,'* indicating a bias towards a specific entity within a category. This could be due to the model being trained on data with insufficient knowledge or the absence of a direct equivalent of the English concept in Indonesian, leading it to default to one standard concept. This finding highlights the limitations of concept adaptation from existing English data. To improve dataset diversity and coverage, manual concept creation in the target language is needed.

5.1.2 Question Analysis

Question Quality To evaluate the quality of the generated questions, we apply a *strict* criterion: any errors, even minor, are marked as incorrect. As shown in Table 4, the Indonesian datasets have a high percentage of error-free questions, ranging between 68–75%. However, for Sundanese, the accuracy decreases significantly. In particular, the weak performance of the MT system in Sundanese is evident from its very low accuracy in LLM_ADAPT, suggesting synthetic data adaptation/translation from English is not an optimal method for low-resource languages. In contrast, the LLM_GEN dataset, which involves generating synthetic data directly in the target language, shows a significantly higher number of correct questions, particularly for Sundanese. This indicates that direct generation in the target language is a more promising approach than adaptation or translation.

Common Mistakes We manually reviewed the questions to identify common errors, detailed in Table 5. We observe that most of the errors in LLM_ADAPT come from translation errors. For Indonesian LLM_GEN, despite lower question generation accuracy, most of the errors are actually minor typos. In Sundanese LLM_GEN, however, the predominant issues relate to sentence fluency, indicating ongoing challenges in automatically generating smooth and fluent sentences in Sundanese. We also found that the model occasionally outputs wrong languages, particularly using Indonesian phrases or words instead of Sundanese. While Indonesian and Sundanese share many linguistic features, they differ significantly in morphological features (§3.2). This difference, along with the smaller amount of Sundanese data used for training LLMs, results in the models being less exposed to Sundanese words, suffixes, or prefixes compared to Indonesian. This leads to the occasional appearance of Indonesian words in generated Sundanese data, even though the sentences may be grammatically correct.

Dataset	% of correct questions	
	`ind`	`sun`
LLM_ADAPT	**75.32%**	15.19%
LLM_GEN	68.67%	**51.00%**

Table 4: Question generation acc. of LLM-generated datasets, measured by the % of error-free questions.

Error Type	% of questions			
	LLM_ADAPT		LLM_GEN	
	`ind`	`sun`	`ind`	`sun`
No error	*75.32%*	*15.19%*	*68.67%*	*51.00%*
Translation	**8.23%**	**41.14%**	0.00%	0.00%
Wrong language	0.00%	0.00%	0.00%	6.67%
Sent. structure	3.16%	10.13%	0.00%	0.00%
Sent. fluency	6.96%	23.42%	11.33%	**18.00%**
Sent. context	1.90%	1.90%	3.00%	8.00%
Subjectivity	0.63%	0.63%	0.00%	0.00%
Typo/mechanics	3.80%	7.59%	**17.00%**	16.33%

Table 5: Distribution of question generation error types of LLM-generated datasets.

5.1.3 Options and Answer Analysis

Options Quality We also assess the options quality using a similar method to the question evaluation. From Table 6, we find LLM_GEN generates higher quality options compared to LLM_ADAPT. However, there is still a significant gap in performance between Indonesian and Sundanese, once again highlighting the performance discrepancy between medium- and lower-resource languages.

Common Mistakes As detailed in Table 7, for LLM_ADAPT, the most common issues are minor typos and mechanical errors, particularly with capitalization (e.g., "indonesia" instead of "Indonesia"). For Sundanese, the errors are mainly due to major translation errors. As for LLM_GEN, the

Dataset	% of correct options	
	`ind`	`sun`
LLM_ADAPT	62.66%	38.61%
LLM_GEN	**93.00%**	**58.67%**

Table 6: Options generation acc. of LLM-generated datasets, measured by the % of error-free options.

Error Type	% of options			
	LLM_ADAPT		LLM_GEN	
	`ind`	`sun`	`ind`	`sun`
No error	*62.66%*	*38.61%*	*93.00%*	*58.67%*
Translation	3.80%	**47.47%**	0.00%	0.00%
Wrong language	0.00%	0.00%	0.00%	**31.67%**
Sent. fluency	0.63%	0.63%	0.67%	1.33%
Sent. context	0.63%	0.63%	0.00%	0.00%
Invalid options	1.27%	1.27%	**5.67%**	8.33%
Typo/mechanics	**31.01%**	11.39%	0.67%	0.00%

Table 7: Distribution of options generation error types of LLM-generated datasets.

most common errors involve the presence of invalid options, particularly when no correct answers are among the options. For Sundanese, alongside invalid options, the model also produces some options in the wrong language. Still, it is important to note that the proportion of invalid options is very low for both languages, indicating that the model generates a substantial number of valid options.

5.2 LLM vs. Human: Lexical Diversity

To compare the lexical diversity between LLM and human-generated data, we analyze the proportion of shared tokens between the LLM_GEN and HUMAN_GEN, calculated by dividing the number of unique shared tokens by the total number of unique tokens. We find that the unigram overlap percentage is 39.75% for HUMAN_GEN and 65.48% for LLM_GEN. A similar trend is observed for the bigram overlap percentage, with 12.41% for HUMAN_GEN and 15.98% for LLM_GEN. This shows that many tokens present in HUMAN_GEN also exist in LLM_GEN, but the reverse is not equally true. We also find that HUMAN_GEN dataset has 8,596 unique tokens, higher than LLM_GEN with 6,677 unique tokens.

Upon sample-level analysis, we find that given the same set of categories and question concepts, humans generate more token variations that are not produced by LLMs, such as some unique terms like *kalis* or *cimol*.[7] Although LLM can still generate some valid data, it tends to return more popular concepts or general questions. For instance, given *'kerupuk'* (crackers) concept, human annotators can create questions tied to their cultural background, such as asking about *'kerupuk rambak'* (rambak crackers). In contrast, LLM tends to give more general (but still relevant) questions, like *"What is the common primary ingredient of crackers?"* More examples are shown in Table 12 in the Appendix.

6 Benchmark Result

6.1 Experiment Setup

We conduct a zero-shot evaluation of various LLMs to assess their performance on our datasets.

English-centric LLMs We include LLaMA-2 7B and 13B (Touvron et al., 2023), a widely used open LLM, and MistralOrca-7B (Lian et al., 2023).

Multilingual LLMs We include PolyLM-13B (Wei et al., 2023), which trained on a multilingual dataset (mostly English and Chinese); BLOOMZ-7B (Muennighoff et al., 2022), which fine-tuned on xP3; SeaLLM-7B (Nguyen et al., 2023), which covers Southeast Asian languages; and Aya (Üstün et al., 2024), recent open LLM trained on 101 languages, including Indonesian and Sundanese.

Monolingual LLMs We include Merak-v4 (Ichsan, 2023), an Indonesian LLM based on MistralOrca and MalayMistral-7B (Zolkepli et al., 2024), a Malaysian LLM extended from Mistral.

Proprietary LLMs We include GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo (OpenAI, 2023).

We use three prompt variations for all models.[8] For a fair comparison between open and proprietary LLMs, we extract the answer key from the text generation result instead of the next token probability, using a rule-based and regex. Evaluations are performed on RTX A6000 48GB. We use accuracy as the evaluation metric.

6.2 Overall Performance

We first benchmark all selected LLMs on our combined datasets to measure the overall performance. As shown in Figure 3, GPT models outperform other LLMs, with an average ∼80% accuracy. Among open models, Merak-v4 scores highest in Indonesian but does not surpass GPT-3.5. Interestingly, the score difference between Merak-v4 and

---

[7] In the context of culinary, *'kalis'* means a state of dough that is well-kneaded and ready to be processed further. *'Cimol'* is an Indonesian street food made from tapioca flour.

[8] Please refer to Appendix F for the prompts details.

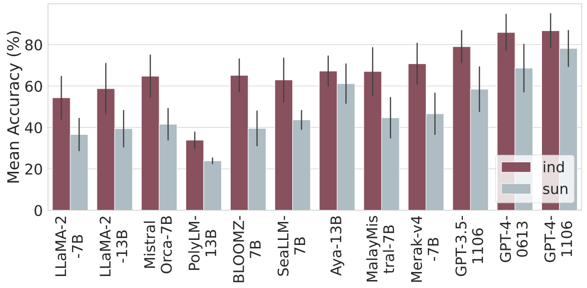

Figure 3: LLMs' performance on our combined test set.

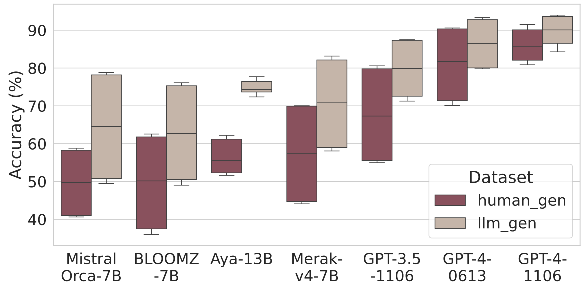

Figure 4: LLMs' performance on LLM_GEN vs. HUMAN_GEN in Indonesian and Sundanese. We combined data points from both languages for visualization, with lower quartiles typically representing Sundanese data.

MalayMistral is small, possibly due to the high lexical similarities between Indonesian and Standard Malay. This may also be due to some instruction data used to train MalayMistral is generated using GPT-4, which tends to produce texts in Indonesian rather than Standard Malay.

We also observe a substantial gap between Indonesian and Sundanese (10–20% accuracy drop), suggesting that current LLMs struggle with Sundanese questions, even in multiple-choice settings. This gap, particularly in Merak-v4, highlights the limitations of training LLMs solely on Indonesian texts, which does not ensure transferable performance across other local languages due to morphological differences. Including Sundanese texts in the training data, as shown by the improved performance in Aya, effectively narrows the performance gap between Indonesian and Sundanese.

6.3 LLM vs. Human-Generated Data

To assess LLMs' ability to answer LLM and human-generated data, we compare their performances on our LLM_GEN and HUMAN_GEN datasets. Figure 4 shows that LLMs generally perform higher on LLM_GEN, especially on MistalOrca, which is trained on English-centric data. This suggests the model can handle LLM_GEN questions despite not being specifically trained on Indonesian or Sundanese data. Still, it is important to note that MistalOrca's average score on LLM_GEN is still lower than other LLMs that include Indonesian texts in their training.

Interestingly, the performance gap tends to narrow as models improve, particularly for GPT models. For Aya, despite relatively strong performance compared to other open-source LLMs, shows significantly dropped performance in HUMAN_GEN data. This decline is likely due to Aya being a 'translation-heavy' model (Üstün et al., 2024), with a large proportion of its training data derived from translations rather than human annotations.

6.4 Performance by Question Category

Figure 5 shows the LLMs' performance across different question categories. It shows that LLMs perform better in the *activity* and *place* questions, which benefit from a lot of information readily available on the internet. However, they struggle with *culinary* questions, where specialized or cultural knowledge is often required. Interestingly, a significant improvement is seen from GPT-4 (`0613` ver.) to GPT-4 Turbo (`1106` ver.), particularly in Sundanese HUMAN_GEN. This suggests that the model is "acquiring" more knowledge, possibly due to the interactions on the ChatGPT web. However, some categories still score below 80%, showing there is still room for improvement.

7 Discussion

7.1 Multiple-Choice vs. 'Free' Generation

Since our dataset is in a multiple-choice format, LLMs might look better than they actually are because they can just pick one answer from the given options. To see LLMs' actual capability, we test GPT-4 Turbo on 100 randomly sampled questions from Indonesian HUMAN_GEN in open-ended settings, i.e., asking the question directly without showing answer options or providing any extra instructions. Our manual evaluation shows an accuracy of 75%, a huge drop from the ∼90% accuracy in multiple-choice settings. This highlights the limitation of the model in open-ended settings.

Among the wrong answers, most cases come from overly general answers not specific to Indonesia, especially for questions in *culinary*, *culture*, and *activity* categories. For instance, given unique *activity* questions such as: *"What is the mandatory song to be sung during the 'mengheningkan cipta' (moment of silence) in the flag ceremony?"*

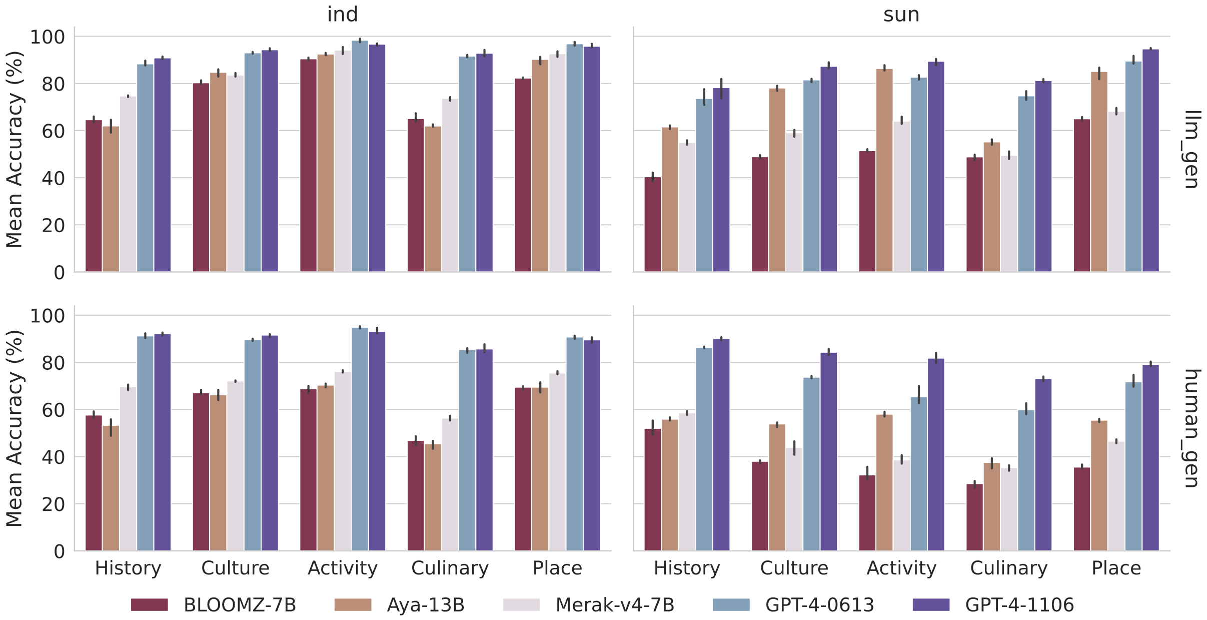

Figure 5: LLMs performance by question category in LLM_Gen and Human_Gen for Indonesian and Sundanese.

the model gives a wrong answer in open-ended generation settings even if it could identify the correct answer in the multiple-choice setup. More failure examples can be seen in Table 13 in the Appendix.

7.2 LLM-generated Data Quality

Due to the high performance of LLMs in answering LLM-generated data, one might argue that such data is less valuable. Ideally, creating a larger human-generated dataset is the most straightforward method to ensure data quality. However, this approach may not always be feasible in the context of underrepresented languages, particularly given the limited resources. Building a dataset manually can be costly and challenging (we spent more than $3,000 in total to build Human_Gen), and leveraging LLMs could be a practical and cheaper alternative, especially for scaling up the data.

To address the limitation of LLM-generated data, we can apply additional processes to reduce the noise. For instance, a collaboration between humans and models can be conducted to fix and revise the synthetic questions (Liu et al., 2022; Putri and Oh, 2022). This noise can also be reduced automatically, especially in Indonesian LLM_Gen questions, where errors typically involve minor typos.

It is also important to note that the quality of LLM-generated data can vary. Our experiments show LLM_Gen has significantly fewer errors than LLM_Adapt, particularly in Sundanese. We also compared the performance between the cleaned vs. raw version of the data (Appendix G) and found that cleaning the data had a relatively minor impact on performance in LLM_Gen. Still, it remains crucial not to depend solely on LLMs, as they still have limitations, such as producing convincing yet incorrect outputs (Pan et al., 2023).

8 Conclusion

In this study, we introduced new Indonesian and Sundanese CommonsenseQA datasets built using various methods: automatic dataset adaptation, direct generation with LLMs, and manual dataset creation by human annotators from diverse regions. Our thorough analysis reveals that adapting existing English data is less effective for Sundanese. In contrast, direct generation in the target languages by GPT-4 Turbo produces relevant data for both languages, although the cultural depth has not yet matched the human-generated ones. Nevertheless, given the limited resources, combining LLMs and humans for dataset creation can be a practical solution, allowing for more efficient dataset creation by reducing the over-reliance on a single data source.

Our goal is for LLMs to serve as beneficial tools for diverse communities, not solely those from higher-resource languages, cultures, or regions. This is especially important since LLMs are continuously being used to generate data. We hope our work represents a crucial step toward achieving this broader objective.

Limitations

Language and Region Coverage In terms of language coverage, we were only able to cover Indonesian and Sundanese due to the available resources and the authors' familiarity with these languages. Additionally, the annotators we recruited were mostly from Java island, with one annotator from Bali island. Despite our effort to include a range of question concepts from different regions, including those beyond Java and Bali islands, it is possible that some bias may exist, especially in the Indonesian dataset. This is because the questions were generated primarily by annotators from Java and Bali, and their perspectives and cultural backgrounds may have influenced the content. Nonetheless, we have taken measures to eliminate potentially harmful or stereotypical questions.

Extension to Other Local Languages Besides Indonesian, our study focuses on one Indonesian local language, Sundanese. As previously discussed, Indonesia has many local languages; however, we cannot cover all of them due to resource constraints. We aim for our findings in Sundanese to act as a starting point for other languages. We anticipate that LLMs might perform worse than Sundanese for extremely low-resource languages, like Buginese or Toba Batak. Javanese, on the other hand, is expected to have comparable performance to Sundanese (Winata et al., 2023; Bang et al., 2023). Generating synthetic data could be particularly beneficial for extremely low-resource languages, given the difficulty of finding native speakers. Yet, our case study in Sundanese indicates that LLMs might struggle even more with generating fluent sentences and/or culturally unique concepts in such languages. Still, we believe that the combination or even collaboration between LLM and humans may serve as a starting point.

Multiple-Choice Format We use a multiple choice question format, following English CommonsenseQA data format (Talmor et al., 2019) to facilitate a more straightforward and robust evaluation process. Although open-ended generation may offer a more challenging benchmark for LLMs, as also discussed in Section 7.1, evaluating LLMs in such settings poses its own set of challenges, especially in low-resource languages where 'LLM-as-a-judge' approach may not be as effective as in English. Nevertheless, our dataset can serve as a starting point for this line of research direction.

Usage of GPT-4 for LLM_GEN LLM_GEN can initially be seen as a dataset to "only" close the gap between local LLMs and GPT-4 since we solely use GPT-4 as the data generator. At the time of the experiment (October–December 2023), using GPT-4 was our only sufficient option. Moving forward, with the rise of better LLMs that support Indonesian, such as Command-R+[9] or Claude 3.5 Sonnet[10], we can enrich the existing data by including the generation of multiple "strong" LLMs. This will make LLM_GEN a more comprehensive testing tool. Additionally, we would like to emphasize that even if LLM_GEN is seen "only" as a tool for measuring the gap between local models and GPT-4, we believe its value should not be underestimated. Considering the lack of strong-performing open-source LLMs in low-resource languages like Sundanese, achieving comparable performance to GPT-4 in these languages would be highly valuable for the community.

Dataset Usage HUMAN_GEN and LLM_GEN datasets should be used exclusively as test data. When utilizing the training data portion in the LLM_ADAPT dataset, special caution is required, particularly for the Sundanese language, where numerous errors have been identified. Training models on this data have a risk of error propagation of incorrect information. Additionally, the test data portions in both LLM_ADAPT and LLM_GEN should be used carefully, as they contain "silver" labels and may include some inaccuracies. We strongly advise using the cleaned version of the data available on our GitHub repository and accompanying the evaluation results of the LLM-generated data with those from the HUMAN_GEN dataset. This is because high performance of the evaluated model, especially on the LLM-generated data, does not necessarily indicate the robustness of the model. One should be careful when making claims.

Ethical Consideration

All human-generated datasets have been manually validated to ensure that harmful or offensive questions are not present in the dataset. We also excluded potentially harmful questions in the LLM-generated datasets through automatic filtering. Our work has been reviewed by KAIST Institutional

---

[9] `https://docs.cohere.com/docs/command-r-plus`

[10] `https://www.anthropic.com/news/claude-3-5-sonnet`

Review Board (IRB). All recruited annotators were paid above the minimum wage based on the standard in Republic of Korea. Our datasets are publicly available under the Creative Commons Non-Commercial (CC BY-NC 4.0) license.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00509258, AI Guardians: Development of Robust, Controllable, and Unbiased Trustworthy AI Technology).

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World*, 24 edition. SIL International, Dallas, Texas.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Ichsan. 2023. Merak-7b: The llm for bahasa indonesia. *Hugging Face Repository*.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Mistralorca: Mistral-7b model instruct-tuned on filtered openorcav1 gpt-4 dataset. `https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca`.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms – large language models for southeast asia.

OpenAI. 2023. Gpt-4 technical report.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Rifki Afina Putri and Alice Oh. 2022. IDK-MRC: Unanswerable questions for Indonesian machine reading comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6918–6933, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

James Neil Sneddon. 2003. *The Indonesian Language: Its History and Role in Modern Society*. UNSW Press, Sydney.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. 2024. Large malaysian language model based on mistral for enhanced local language understanding.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

A LLM-Generated Data

A.1 Prompts Details

A.1.1 Concept Relevancy Classifier

To classify whether a concept is relevant to the Indonesian and Sundanese context, we prompt GPT-3.5 Turbo[11] using an ensemble of five prompts. We condition the prompts based on whether the question concept is a verb or not, determined using the ConceptNet API. We set the location context as *'Indonesia'* for Indonesian data and *'West Java'* for Sundanese data. The examples of concept relevancy classifier results can be seen in Table 11.

Below are the prompts if the question concept is a verb (e.g., *skiing*, *studying*):

Prompt Variation 1

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Can one '{QUESTION CONCEPT}' in {Indonesia/West
Java}? Answer with 'yes' or 'no'.
```

---

[11] We used `gpt-3.5-turbo` API checkpoint, accessed in September 2023.

Prompt Variation 2

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Do people in {Indonesia/West Java} familiar with
 '{QUESTION CONCEPT}' concept? Answer with 'yes'
 or 'no'.
```

Prompt Variation 3

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Is '{QUESTION CONCEPT}' concept exist in {
Indonesia/West Java}? Answer with 'yes' or 'no'.
```

Prompt Variation 4

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

In {Indonesia/West Java}, can people '{QUESTION
CONCEPT}'? Answer with 'yes' or 'no'.
```

Prompt Variation 5

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Can '{QUESTION CONCEPT}' be done in {Indonesia/
West Java}? Answer with 'yes' or 'no'.
```

Below are the prompts if the question concept is *not* a verb (e.g., *snow*, *bald eagle*):

Prompt Variation 1

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Can one find '{QUESTION CONCEPT}' in {Indonesia/
West Java}? Answer with 'yes' or 'no'.
```

Prompt Variation 2

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Do people in {Indonesia/West Java} familiar with
 '{QUESTION CONCEPT}' concept? Answer with 'yes'
 or 'no'.
```

Prompt Variation 3

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Is '{QUESTION CONCEPT}' concept exist in {
Indonesia/West Java}? Answer with 'yes' or 'no'.
```

Prompt Variation 4

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

In {Indonesia/West Java}, can people find '{
QUESTION CONCEPT}'? Answer with 'yes' or 'no'.
```

Prompt Variation 5

```
Text: {QUESTION}
Concept: {QUESTION CONCEPT}

Can '{QUESTION CONCEPT}' be found in {Indonesia/
West Java}? Answer with 'yes' or 'no'.
```

A.1.2 Automatic Data Adaptation

There are two prompts used to adapt the CommonsenseQA into LLM_ADAPT: Adapt All and Adapt Name. Below are the details of both prompts.

Adapt All Prompt

```
Change the given data to make it relevant to
Indonesia in any ways. Make all elements
relevant to each other, and the concept always
appear explicitly in the question. Return in
Indonesian language with JSON format where
question is string, concept is string, options
is dictionary where label is the keys and option
 text is the values, and question_answer is
string contain one label from the options.

Data:
###
Question: {QUESTION}
Concept: {QUESTION CONCEPT}
Options:
{CHOICES}
Question Answer: {ANSWER}
###
```

Adapt Name Prompt

```
Change all names in the given question to
Indonesian names. Change only the names, keep
all other phrases in the question the same and
keep it all in Indonesian.

Question: {QUESTION}
Changed Question:
```

A.1.3 Automatic Data Generation

For LLM_GEN, we directly generate the data given a set of question concepts. Below are the prompt used for data generation.

Data Generation Prompt

```
Given a list of {LANGUAGE} concepts [QUESTION
CONCEPTS}], create one {LANGUAGE} commonsense QA
 data with topic "{CATEGORY}" for each concept,
that consists of three components: "question", "
choices", and "answer_creator". The "question"
must contains the concept explicitly. The "
choices" consist of 5 different choices marked A
 to E where one should be the "answer_creator".
All data should be in {LANGUAGE}, return only
your answer in JSON data format, and add the
concept of the data as "question_concepts".

JSON Data:
```

Model	Win Rate		
	Concept	Question	Choices
Automatic Data Adaptation			
Merak-v4	28.0%	8.5%	4.0%
GPT-4 Turbo	**72.0%**	**91.5%**	**96.0%**
Automatic Data Generation			
Merak-v4	-	10.5%	9.5%
GPT-4 Turbo	-	**89.5%**	**90.5%**

Table 8: Win rate comparison of Merak-v4 (open Indonesian LLM) and GPT-4 Turbo (best-performing proprietary LLM).

Error Type	Num (%) of errors	
	`eng` → `ind`	`ind` → `sun`
Translation	**4 (57.14%)**	**31 (86.11%)**
Phrase structure	1 (14.29%)	1 (2.78%)
Typo/mechanics	2 (28.57%)	4 (11.11%)

Table 9: Summary of errors in concept adaptation.

A.2 Choosing Model for Data Generation: Indonesian LLM vs. GPT-4 Turbo

Table 8 shows the win rate accuracy of concept, question, and options generation of Indonesian LLM, Merak-v4, and best-performing proprietary LLM, GPT-4 Turbo. The results demonstrate that GPT-4 Turbo significantly outperforms Merak-v4, with win rates ranging from 72% to 96%. Our sample-level analysis indicates that a significant number of questions generated by Merak-v4 tend to have obvious answers or are formulated as yes/no questions, such as *"Apakah jagung dapat dimakan?"* (`eng`: Is corn edible?). Additionally, despite being trained on Indonesian texts, Merak-v4 occasionally produces questions with US-centric knowledge. For instance, it generates *"Siapa yang memimpin kampanye pemilihan presiden pertama di Amerika Serikat?"* (`eng`: Who led the first presidential election campaign in the United States?).

A.3 Additional Analysis of the LLM-generated Data: Common Mistakes in Concept Adaptation

The details of common mistakes in concept adaptation for Indonesian and Sundanese are shown in Table 9. In the case of adapting concepts from `eng` to `ind`, many of the errors are translation errors, resulting from awkward phrasing of the translated concept or the concept remaining in English instead of being translated to Indonesian. However, the number of errors is relatively small. In the case of adapting concepts from `ind` to `sun`, similar to `eng` to `ind`, the majority of errors also arise from translation errors, with a larger number of errors.

B Details on Categories and Concepts

Categories Our LLM_GEN and HUMAN_GEN datasets include five selected categories, detailed as follows:

1. **Culinary**: Concepts in this category include everything related to culinary, starting from culinary types, cooking ingredients, cooking tools & methods, to etiquette or eating habits.
2. **Places**: Concepts in this category include everything related to places, starting from public facilities, landmarks, buildings, and various other concepts related to places.
3. **Culture**: Concepts in this category include everything related to culture, starting from cultural elements, cultural tools, cultural actors, to customs and habits that exist in Indonesia.
4. **History**: Concepts in this category include everything related to history, starting from historical events, historical actors, historical findings, and various other concepts related to history.
5. **Activities**: Concepts in this category include everything related to activities, starting from sports, hobbies, household work, and various other concepts related to activities.

Concepts Creation To create the concepts, we had a human annotator manually list concepts in the selected categories, ensuring they were culturally relevant but general enough for both Indonesian and Sundanese. For example, in the culinary category, concepts might include 'duck' or 'lime leaves,' avoiding specific dish names. We reviewed the concepts and provided feedback for revisions to ensure quality. This process results in 150 unique concepts per category in Indonesian. The annotators then manually translated these concepts into Sundanese during the QA pairs creation phase (§4.2). We aimed for parallel concepts to ensure a fair comparison between the languages.

C Annotators Demographics

In accordance with the aim of constructing the dataset, we involved 6 Indonesian and 6 Sundanese

Data Team	Ethnicity	Domicile	Num
Indonesian	Sundanese	Sukabumi, West Java	1
		Depok, West Java	1
	Javanese	Magelang, Central Java	1
		Bojonegoro, East Java	1
	Betawi	Tangerang, Banten	1
	Balinese	Denpasar, Bali	1
Sundanese	Sundanese	Bandung, West Java	2
		Bogor, West Java	1
		Majalengka, West Java	1
		Sukabumi, West Java	2
Total			**12**

Table 10: Demographic information of the annotators from each dataset team. Note that even though Depok is included in West Java Province and Tangerang is included in Banten Province, both are geographically closer to Jakarta and considered as part of the Greater Jakarta area (*Jabodetabek*).

native-speaker annotators. We chose annotators who have prior experience in building NLP datasets in Indonesian and/or Sundanese through close recruitment. All the annotators who worked on the Indonesian dataset were people from several regions on the islands of Java and Bali. Meanwhile, the annotators involved in building the Sundanese dataset were Sundanese people who come from several different regions in West Java. The detailed annotators' demographics are shown in Table 10.

D Human Annotation Guideline

To ensure a high-quality and standardized format for question-answers annotation, we provide a specific guideline during the annotation process. The process of creating question-answer pair data is carried out using Google Spreadsheets. Each person will get a Google Sheets document that will be their worksheet. The explanation of each field is described below.

1. **ID**: This column contains the ID of each data.
2. **Category**: This column contains the overarching category of the data.
3. **Question Concept**: This column contains concepts from categories that need to be used in creating questions.
4. **Question**: This column is used to write commonsense questions that contain the given concept and match the category.
5. **Choices**: This column is used to write 5 choices for the questions given. Of the 5 choices given, the annotator needs to ensure there is 1 correct answer and 1 distractor. Distractor is an answer that could potentially be considered the correct answer.
6. **Answer**: This column only needs to be filled in when the row contains the correct answer.
7. **Distractor**: This column only needs to be filled in when the row contains the distractor.

General Rules There are several rules in the data creation process, detailed as follows:

1. The questions asked are commonsense questions (not factual questions) related to Indonesian/Sundanese culture. Especially for the History category, questions can also be in the form of factual questions, but they must be general facts (commonly known).
2. The questions created must be related to the given category and must contain the given concept.
3. Annotators are allowed to change the morphological form of concepts as long as they do not change the categories and basic words. (i.e. *ber**kunjung*** → ***kunjung**an*, *meng**unjungi***, *di**kunjungi***).
4. The priority is that the questions asked are always related to general daily life or Indonesian/Sundanese culture.
5. Both language groups will use the same list of categories and concepts. However, when creating data, annotators are expected to incorporate perspectives that align with their respective cultural backgrounds.
6. Each person will get 50 concepts from each category. The total number of concepts that will be accepted is 250 concepts, equivalent to the workload for each person, where the expectation is to create one question per concept.

E Manual Data Generation Findings

During the manual data generation process, we discovered several interesting findings from the data created by the annotators. Note that the common errors within these findings occur before the quality control meeting (§4.2). In our final dataset, we have ensured that such problems have been fixed.

Questions We found quite a lot of interesting findings in the process of creating commonsense questions. We found that there were questions created using very general contexts, while we required the data generation to be scoped within the Indonesian and Sundanese contexts (specifically for the Sundanese dataset). For instance, an annotator produced *"Apa jenis restoran yang paling terkenal di seluruh dunia?"* (eng: Which type of restaurant is most famous worldwide?) In that question, the annotator developed the concept question *“restoran”* (eng: restaurant) into a question with a very general context: *“di seluruh dunia"* (eng: all over the world). We also found subjective questions, so the answers that emerged also had subjective value (not commonsense). These subjective questions are usually characterized by the use of superlative adjectives. Another type of finding related to the question category is the creation of logical/causal questions. This seems to be based on confusion from annotators regarding the boundaries of commonsense and logic/causation.

Options The findings related to options that are most often encountered are making choices that have the same value as each other so that it is difficult to determine the answer and distractor. For instance, the question *"Kue apa yang biasanya disajikan pada momen lebaran?"* (eng: What cakes are usually served during Eid?). All the options given are types of cakes that are generally served during Eid. This finding is also related to the issue of subjectivity.

Answers-Distractors Answers and distractors are another category that also has many findings in the manual data generation process. This issue is not much different from the findings in Options: answers and distractors are equivalent. Apart from that, in this category, it was also found that personal experience was used in determining answers so that more common answers were determined as distractors. For instance, the question *"Apa yang biasanya orang lakukan di stasiun?"* (eng: What do people usually do at the train station?). The annotator specified *"mengantar teman/kerabat"* (eng: accompany friends/relatives) as the answer. While the *"naik-turun kereta"* (eng: get on and off the train) option was chosen as a distractor.

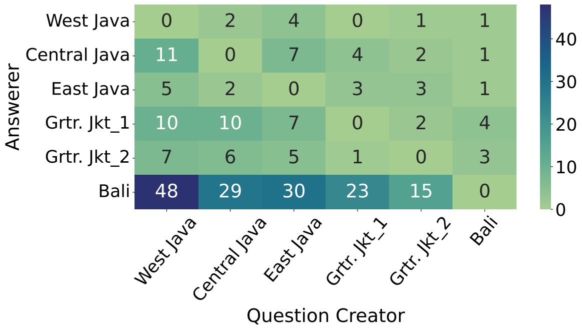

Figure 6: Answer conflict across Indonesian annotators.

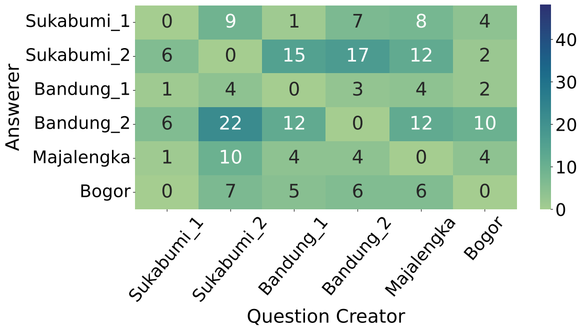

Figure 7: Answer conflict across Sundanese annotators.

Knowledge Variations between Annotators As we employ annotators from different regions, we can analyze variations in the data they generate. We examine this by calculating the number of answer conflicts that arise during the "answering question" phase of our data generation pipeline (§4.2). The results from Indonesian and Sundanese annotators are detailed in Figure 6 and 7, respectively.

Our analysis revealed that, out of the Indonesian annotators, the one from Bali has the highest number of conflicting answers. However, the number of questions generated by the Bali annotator does not seem to have a lot of conflicts, suggesting that this annotator tends to generate easier questions, some of which have an obvious answer. This finding contrasts with the results from West Java annotators, who, despite generating questions that lead to a higher number of conflicts, do so mainly due to the creation of more challenging option distractors.

Interestingly, in the case of Sundanese annotators, the variation in answer conflicts across regions is not significant. This lack of variation can likely be attributed to the Sundanese language’s narrower geographic distribution which primarily spoken in West Java, unlike the Indonesian language, which serves as a lingua franca within the country. However, it is important to note that given the small number of annotators, these trends cannot be conclusively linked to the annotators' regional characteristics. The observed differences may also be due to individual differences rather than regional ones.

F Zero-Shot Benchmark Prompts

We apply three prompt variations to test LLMs' performance on our CommonsenseQA datasets. Each prompt is described below.

Prompt Variation 1

```
The following are multiple choice questions (
with answers) about "{CONCEPT}".
{QUESTION}
A. {CHOICE_A}
B. {CHOICE_B}
C. {CHOICE_C}
D. {CHOICE_D}
E. {CHOICE_E}
Answer:
```

Prompt Variation 2

```
Question: {QUESTION}
Choices:
A. {CHOICE_A}
B. {CHOICE_B}
C. {CHOICE_C}
D. {CHOICE_D}
E. {CHOICE_E}
Answer:
```

Prompt Variation 3

```
The following are multiple choice questions (
with answers) about "{CONCEPT}".
Question: {QUESTION}
A. {CHOICE_A}
B. {CHOICE_B}
C. {CHOICE_C}
D. {CHOICE_D}
E. {CHOICE_E}
Answer:
```

G Effect of Synthetic Dataset Cleaning

To check how much cleaning the noise in synthetic (LLM-generated) data affects LLMs performance in answering our questions, we manually correct the errors in all 158 test sets of LLM_ADAPT and 300 randomly sampled data from the LLM_GEN. The results are presented in Figure 8. The performance variance appears smaller on the cleaned dataset for LLM_ADAPT, leading to more consistent performance. However, the performance improvement is not as significant for the LLM_GEN dataset, likely due to the lesser noise in this dataset compared to LLM_ADAPT. Thus, a robust model capable of producing cleaner data can minimize the need for extensive dataset cleaning. However, caution is still necessary as errors can still occur.

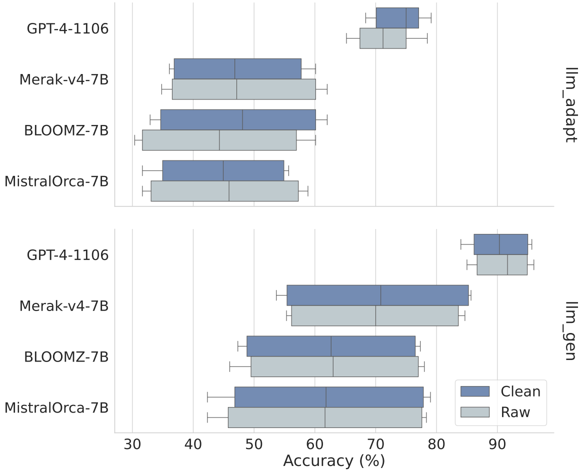

Figure 8: Performance comparison of raw vs. cleaned version of LLM-generated data.

H LLM vs. Human-Generated Data Examples

We provide some GPT-4 Turbo generation samples compared with human-generated data in Table 12.

I Examples from Multiple-Choice vs. 'Free' Generation Experiment

We provide some GPT-4 Turbo generation samples when given the question in 'free' or open-ended generation settings in Table 13.

Classification	**Original Data (`eng`)**	**LLM Adapted Data (`ind`)**

concept=irrelevant
name=irrelevant
option=irrelevant
	Concept: koala
	Question: James looked for koalas, but misunderstood and went to the wrong environment where is it impossible to find koalas?
	Options:
	A. jungle
	B. great outdoors
	C. siberia
	D. queensland
	E. wilderness
		Concept: orangutan
		Question: Budi mencari orangutan, tetapi salah paham dan pergi ke lingkungan yang salah. Di mana mustahil untuk menemukan orangutan?
		Options:
		A. hutan hujan tropis
		B. alam bebas
		C. gurun sahara
		D. kalimantan
		E. belantara

concept=relevant
name=irrelevant
option=irrelevant
	Concept: hotel room
	Question: John rented a hotel room while on a business trip. It was a company sponsored event. Where might it have been?
	Options:
	A. convention center
	B. business interviews
	C. michigan
	D. resort
	E. city
		Concept: kamar hotel
		Question: Andi menyewa kamar hotel saat dalam perjalanan bisnis. Ini adalah acara yang disponsori perusahaannya. Di mana acara tersebut mungkin diadakan?
		Options:
		A. pusat konvensi
		B. wawancara bisnis
		C. jakarta
		D. resor
		E. kota

concept=irrelevant
name=relevant
option=irrelevant
	Concept: blue ridge mountains
	Question: Which state is the blue ridge mountains roanoke city in?
	Options:
	A. appalachians
	B. north carolina
	C. virginia
	D. pennsylvania
	E. montana
		Concept: Gunung Bromo
		Question: Di provinsi mana Gunung Bromo terletak?
		Options:
		A. jawa barat
		B. jawa tengah
		C. jawa timur
		D. yogyakarta
		E. bali

concept=relevant
name=relevant
option=relevant
	Concept: teacher
	Question: What does a kindergarten teacher do before nap time?
	Options:
	A. lower expectations
	B. encourage
	C. fear
	D. time test
	E. tell story
		N/A

Table 11: Examples of the concept, name, and option relevancy classification result. The English data instance that is classified as irrelevant in any of these three elements will be adapted.

Category	**Concept**	**Data Creator**	**Question**	**Choices**	**Answer**
activity	panen (*harvest*)	GPT-4 Turbo	Kegiatan apa yang biasanya dilakukan ketika musim panen tiba? (*What activity is usually done when the harvest season arrives?*)	A. Makan bersama keluarga (*eating together with family*) B. Berlibur (*going on a holiday*) C. Memetik hasil pertanian (*picking agricultural produce*) D. Memancing (*fishing*) E. Berbelanja (*shopping*)	C
		Human	Kegiatan apa yang masyarakat desa lakukan setelah panen mereka berhasil? (*What activity do village communities do after their harvest succeeds*?)	A. Bercocok tanam (*farming*) B. Memotong rumput (*cutting grass*) C. Syukuran (*having a 'syukuran' event*) D. Jalan-jalan bersama (*going out together*) E. Wortel (*carrots*)	C
culinary	kerupuk (*crackers*)	GPT-4 Turbo	Kerupuk biasanya terbuat dari bahan dasar apa? (*What is the common primary ingredient of crackers?*)	A. Tepung terigu (*wheat flour*) B. Tepung beras (*rice flour*) C. Tepung tapioka (*tapioca flour*) D. Tepung jagung (*corn flour*) E. Tepung kentang (*potato flour*)	C
		Human	Terbuat dari apakah kerupuk rambak? (*What is 'rambak' crackers made of?*)	A. Kulit hewan (*animal skin*) B. Lidah hewan (*animal tounge*) C. Usus hewan (*animal intestines*) D. Nasi (*rice*) E. Tepung beras (*rice flour*)	A
culture	tarian (*dance*)	GPT-4 Turbo	Tarian apa yang merupakan tarian tradisional dari Bali? (*Which dance is a traditional dance from Bali*?)	A. Tari Pendet (*Pendet dance*) B. Tari Saman (*Saman dance*) C. Tari Kecak (*Kecak dance*) D. Tari Jaipong (*Jaipong dance*) E. Tari Yapong (*Yapong dance*)	C
		Human	Manakah yang termasuk tarian dari Bali? (*Which includes a dance from Bali?*)	A. Tari Piring (*Piring Dance*) B. Tari Pendet (*Pendet Dance*) C. Tari Topeng (*Topeng Dance*) D. Tari Sajojo (*Sajojo Dance*) E. Tari Melasti (*Melasti Dance*)	B
history	era (*era*)	GPT-4 Turbo	Era apa yang ditandai dengan berakhirnya Perang Dingin dan runtuhnya Uni Soviet? (*Which era is marked by the end of the Cold War and the collapse of the Soviet Union*?)	A. Era Globalisasi (*globalization era*) B. Era Informasi (*information era*) C. Era Reformasi (*reformation era*) D. Era Pasca-Perang Dingin (*pasca-cold war era*) E. Era Industrialisasi (*industrialization era*)	D
		Human	Berapa lama era Orde Baru berlangsung di Indonesia? (*How long did the New Order era last in Indonesia*?)	A. 4 tahun (*4 years*) B. 5 tahun (*5 years*) C. 20 tahun (*20 years*) D. 25 tahun (*25 years*) E. 32 tahun (*32 years*)	E
place	gunung (*mountain*)	GPT-4 Turbo	Di manakah biasanya gunung berada? (*Where are mountains usually located?*)	A. Di kota (*in the city*) B. Di dataran tinggi (*in highlands*) C. Di dalam laut (*in the sea*) D. Di padang pasir (*in the desert*) E. Di pinggir sungai (*at the riverbank*)	B
		Human	Apa nama gunung tertinggi di Indonesia? (*What is the name of the highest mountain in Indonesia?*)	A. Gunung Jaya Wijaya (*Mt. Jaya Wijaya*) B. Gunung Bromo (*Mt. Bromo*) C. Gunung Merapi (*Mt. Merapi*) D. Gunung Semeru (*Mt. Semeru*) E. Gunung Sumbing (*Mt. Sumbing*)	A

Table 12: Examples of data generated by GPT-4 Turbo (`gpt-4-1106-preview`) compared to human annotators.

Prompt	**Response**	**Failure Reason**
Apa judul lagu yang wajib dinyanyikan pada sesi mengheningkan cipta di setiap upacara bendera? *(What is the mandatory song to be sung during the 'mengheningkan cipta' (moment of silence) in every flag ceremony?)*	Pada sesi mengheningkan cipta di setiap upacara bendera di Indonesia, biasanya tidak dinyanyikan lagu. ... *(During the moment of silence in every flag ceremony in Indonesia, a song is usually not sung. ...)*	The mandatory song to be sung during the *'mengheningkan cipta'* (moment of silence) session in flag ceremonies in Indonesia is *'Hymne Pahlawan (Mengheningkan Cipta).'*
Manakah menu yang dikonsumsi menggunakan sumpit? *(Which dishes are consumed with chopsticks?)*	Sumpit adalah alat makan yang biasanya digunakan dalam tradisi kuliner Asia Timur, seperti Jepang, Korea, dan China. Berikut adalah beberapa contoh menu yang sering dikonsumsi menggunakan sumpit: Sushi dan Sashimi (Jepang), Ramen (Jepang), ... *[proceed with other dish examples in Japan, Korea, and China]* *(Chopsticks are utensils that are commonly used in East Asian culinary traditions, such as Japan, Korea, and China. Here are some examples of dishes often consumed with chopsticks: Sushi and Sashimi (Japan), Ramen (Japan), ... [proceed with other dish examples in Japan, Korea, and China])*	Many Indonesian foods are commonly eaten with chopsticks, such as *mie ayam* (chicken noodles), *bakmi*, *mie yamin*, and many more.
Di daerah mana yang terdapat banyak vila? *(In which area can we find many villas?)*	... beberapa daerah yang terkenal dengan banyaknya vila adalah: 1. Bali, Indonesia 2. Tuscany, Italia 3. Provence, Prancis *[proceed with other location in Spain, US, Thailand, Portugal, US]* *(... some areas known for having many villas include: 1. Bali, Indonesia 2. Tuscany, Italy 3. Provence, France [proceed with other locations in Spain, US, Thailand, Portugal, US])*	Apart from Bali, there are many other areas in Indonesia famous for having many villas, such as Bandung, Lembang, Puncak, Batu, and many more.

Table 13: Examples of incorrect responses by GPT-4 Turbo (`gpt-4-1106-preview`) in 'free' or open-ended generation settings.