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

Anonymous ACL submission

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. Our experiments show that the current best-performing LLM, GPT-4 Turbo, is capable of generating questions with adequate knowledge in Indonesian but not in Sundanese, highlighting the performance discrepancy between medium- and lower-resource languages. We also benchmark various LLMs on our generated datasets and find that they perform better on the LLM-generated datasets compared to those created by humans.[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 non-English languages, especially those that are under-represented, 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to be fully useful, it must not only be linguistically accurate, but it also needs to reflect the cultural nuances, historical references, and social norms.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 languages such as Indonesian and Sundanese.

In this study, we aim to investigate how well current LLMs 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 local languages in Indonesia. To sum up, our contributions are as follows:

- We create a new Indonesian and Sundanese CommonsenseQA dataset using various methods, including adapting existing English data (Talmor et al., 2019) and generating new datasets manually and automatically as illustrated in Figure 1. The dataset contains approximately 3K LLM-generated and 1.5K human-generated question-answer pairs per language, making it the largest of its kind. For human-generated data, we ensure cultural diversity by involving annotators from various regions, including Java and Bali.

- 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, particularly for Sundanese. However, GPT-4 Turbo can generate questions with some basic local knowledge in Indonesian when provided with human-created categories and concepts. Yet, for Sundanese, questions often have major errors.

- We evaluate several LLMs using our dataset and find that they perform better on the dataset generated by LLMs than humans. This finding reveals the limitation of the synthetic dataset that it is not as challenging for LLMs, especially for larger models such as GPT-4.

---

[1] All datasets and codes in this work will be publicly available. The dataset will be available under the Creative Commons Non-Commercial (CC BY-NC 4.0).

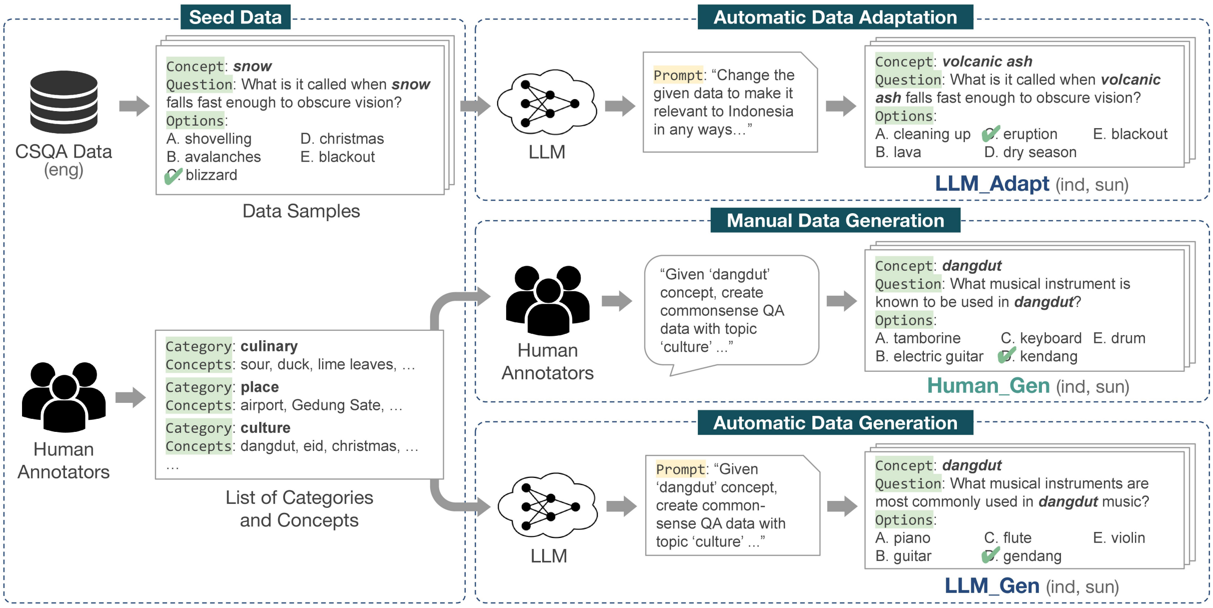
Seed Data
CSQA Data
(eng)
Concept: snow
Question: What is it called when snow falls fast enough to obscure vision?
Options:
A. shovelling
B. avalanches
C. blizzard
D. christmas
E. blackout
Data Samples
Human
Annotators
Category: culinary
Concepts: sour, duck, lime leaves, …
Category: place
Concepts: 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.

2 Related Work

Commonsense Dataset Datasets and benchmarks 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 question-answering, 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 a concept knowledge from ConceptNet (Speer et al., 2017), built through crowdsourcing. One question corresponds to one target concept, and other related concepts become distractors, acting as parts of the answer choices. However, since the dataset is crowdsourced, the crowdworkers bring their cultural background as part of their common knowledge, 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. However, this dataset is relatively small, comprising only 184 questions and focusing solely on US, Indian, and Chinese customs. In the context of Indonesian, 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 an Indonesian COPA-style dataset written by native speakers, thus incorporating more Indonesian cultural nuances compared to X-COPA. COPAL-ID is composed of approximately 300 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

The Cambridge Dictionary defines commonsense as *"the basic level of practical knowledge and judgment that we all need to help us live in a reasonable and safe way."* van Holthoon and Olson (1987) describe commonsense as *"knowledge, judgment, and taste which is more or less universal and which is held more or less without reflection or argument."* Both definitions highlight commonsense as the practical knowledge in daily life. Talmor et al. (2019) did not explicitly define the scope of commonsense in their QA dataset. Nevertheless, they outlined various skills required to answer the questions in their dataset correctly, such as spatial, cause & effect, social, activity, and many more. To build our dataset, we follow similar skills but focus on adding common or cultural knowledge specific to Indonesian and Sundanese contexts to reduce bias towards Western culture. Generally, one data point is a triple of *concept*, *question*, and *options (with one correct answer)*. The data samples are shown in Figure 1.

3.2 Languages in Indonesia

Indonesia is one of the most culturally and linguistically diverse countries in the world, with more than 700 languages spoken across the country (Aji et al., 2022; Eberhard et al., 2021). Of the many languages in Indonesia, Indonesian is a unifying language used nationally. It utilizes the Latin script and was developed from literary "Classical Malay" of the Riau-Johor sultanate (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. Sundanese is primarily spoken in West Java, as well as Banten, Jakarta, Central Java, and more.[2] 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 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.

4 Data Generation Methods

To investigate whether LLMs can generate culturally relevant commonsense QA data in Indonesian and Sundanese, we construct a dataset using various methods with LLMs as data generators. In addition, we employ humans to generate data for comparison. As illustrated in Figure 1, we apply three dataset generation methods: (1) *Automatic Data Adaptation*, where we leverage LLMs to automatically adapt English CommonsenseQA data to our target languages; (2) *Manual Data Generation*, where we ask native-speaker human annotators to manually construct the dataset; and (3) *Automatic Data Generation*, where we utilize LLMs to generate data based on the list of categories and concepts used in method (2).

4.1 Automatic Data Adaptation

We build the first LLM-generated data by adapting the CommonsenseQA dataset (Talmor et al., 2019) to make it culturally relevant to Indonesian and Sundanese. We employ a multi-stage approach to adapt the data. First, we remove data that contains ambiguous options and offensive keywords. Next step involves assessing three elements: *concept*, *name*, and *location*, in order to sample the data that will be adapted. We take the question concepts from the existing CommonsenseQA data. As for the name and location extraction, we use Stanford's Stanza NLP library (Qi et al., 2020) and ConceptNet (Speer et al., 2017) API. We use GPT-3.5 Turbo results ensembled from five different prompts to determine the relevance of the data based on the three elements. Data that are considered irrelevant in at least one of these three elements are selected to be adapted.

The next step is to transform the selected irrelevant data. First, we prompt GPT-4 Turbo[3] to rephrase the sampled questions and options to align with Indonesian cultures. Subsequently, for data flagged with Western-centric names, a second rephrasing is conducted using the GPT-3.5 Turbo to replace all person names. The fully rephrased and adapted data are then translated from Indonesian to Sundanese using Google Translation API. Finally, we filter the data by examining the appearance of concepts in the questions and checking Sundanese-Indonesian back-translation with a similarity threshold of 90%.

4.2 Manual Data Generation

We employ annotators with different ethnicities from diverse regions covering Java and Bali island to ensure the variability of the knowledge in the human-generated dataset. The data collection process involves two main phases: (1) creating commonsense question-answer pairs and (2) answering commonsense questions. We perform quality control with additional human annotators to ensure the

---

[2] `https://petabahasa.kemdikbud.go.id/infobahasa2.php?idb=56&idp=Jawa%20Barat`

[3] 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.

data quality.

Creating Question-Answer Pairs We first instruct the annotators to create commonsense question-answer pairs based on the given category and question concept. We also ask them to rely on their existing knowledge when making questions and avoid using internet search or LLM as much as possible. The categories and question concepts are manually curated to ensure wide domain coverage and to enrich the diversity of commonsense knowledge in the data. In total, we cover five categories: (1) *culinary*, (2) *place*, (3) *culture*, (4) *history*, and (5) *activity*, with 150 predefined question concepts for each category. Two annotators use the same 50 question concepts from each category to generate the questions, options, and answers. This process produces 1,500 QA pairs for each language in total.

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 questions each (excluding the set of data they have made in the first phase). This phase aims to evaluate 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) to ensure data correctness through manual human validation. 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 have a better understanding and do not repeat similar errors. 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. To maintain dataset quality and uniqueness, we filter out duplicate entries and questions that do not explicitly contain the question concepts.

Dataset Version	**Indonesian**			**Sundanese**		
	Train	**Valid**	**Test**	**Train**	**Valid**	**Test**
LLM_ADAPT	1,506	191	158	1,506	191	158
HUMAN_GEN	-	-	1,498	-	-	1,499
LLM_GEN	-	-	1,063	-	-	1,183
Total	1,506	191	2,719	1,506	191	2,840

Table 1: Statistics of our generated Indonesian and Sundanese CommonsenseQA dataset.

4.4 Resulting Data

Our final CommonsenseQA dataset contains 4,416 QA pairs for Indonesian and 4,537 QA pairs for Sundanese. Table 1 provides detailed statistics for both languages. We also have three data variations: (1) LLM_ADAPT: LLM-generated dataset constructed through data adaptation method (§4.1); (2) HUMAN_GEN: Human-generated dataset, as explained in §4.2; (3) LLM_GEN: LLM-generated dataset given the same set of categories and concepts in HUMAN_GEN (§4.3).

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. Based on the evaluation, we then calculate the accuracy of the generated concepts, questions, and options. Furthermore, we evaluate 300 randomly selected samples from the LLM_GEN using the same procedure.

5.1.1 Concept Analysis

Concept Quality In the English (`eng`) to Indonesian (`ind`) concept adaptation, 95.57% are correctly adapted, including 16 out of 19 concepts needing

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: Samples of English (`eng`) to Indonesian (`ind`) concept adaptation.

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: Samples of Indonesian (`ind`) to Sundanese (`sun`) concept adaptation.

major adaptation for cultural relevance, such as 'snow' to '*hujan abu vulkanik*' (volcanic ashfall). Most adaptations are minor, primarily involving direct translations of `eng` to `ind`, as shown in Table 2. For Indonesian to Sundanese, the correct adaptation rate drops to 77.22%, reflecting weaker machine translation performance for Sundanese. The samples are shown in Table 3.

Concept Variation Despite the high accuracy of concept adaptation, as shown in Figure 2, we observe that the adapted concepts are skewed towards '*komodo*'. This indicates that the model is biased toward a specific entity within a category, which could arise from its limited knowledge or the absence of a direct equivalent for the original English concept in Indonesian, leading it to default to one standard concept. This finding highlights the need to manually develop concepts across various categories to improve the dataset's quality in terms of concept diversity and coverage.

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 show 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. This suggests that synthetic data adaptation and generation is not useful for low-resource languages.

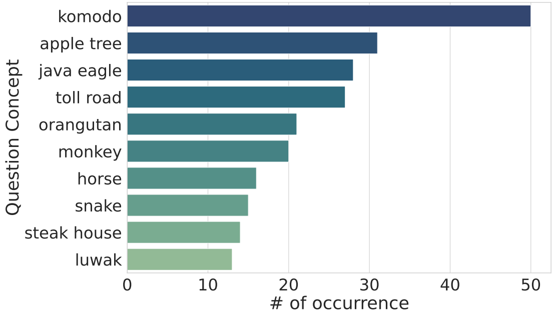

Figure 2: Top-10 adapted question concepts taken from train, validation, and test set of LLM_ADAPT data.

Common Mistakes To further analyze the quality of the questions, we delve deeper into the mistakes that are commonly made, which are listed in Table 5. We observe that most of the errors in LLM_ADAPT come from translation errors. For LLM_GEN, even though the accuracy of question generation is lower than LLM_ADAPT, 54.26% of the errors in LLM_GEN are actually minor typos or mechanics errors. In the case of Sundanese, most errors in LLM_GEN are due to sentence fluency, indicating that the model is still struggling to generate fluent sentences in Sundanese.

5.1.3 Options and Answer Analysis

Options Quality We also evaluate the options in a similar manner to the evaluation of the question. From Table 6, we observe that LLM_GEN generates higher quality options compared to LLM_ADAPT. However, there is still a significant gap in performance between Indonesian and Sundanese. This once again highlights the performance discrepancy between medium- and lower-resource languages.

Common Mistakes We conduct a detailed analysis to identify the common errors found in each dataset, shown in Table 7. For LLM_ADAPT, the most common issues are typos and mechanical errors, particularly with capitalization (e.g., "indonesia" instead of "Indonesia"), along with translation errors. In the Sundanese dataset, the errors are

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

Table 4: Question generation accuracy of LLM-generated datasets. The accuracy reflects the percentage of a 'clean' question without any major or minor errors.

Error Type	% of error type			
	LLM_ADAPT		LLM_GEN	
	`ind`	`sun`	`ind`	`sun`
Translation	**33.33%**	**48.51%**	0.00%	13.61%
Sent. structure	12.82%	11.94%	0.00%	0.00%
Sent. fluency	28.21%	27.61%	36.17%	**36.73%**
Sent. context	7.69%	2.24%	9.57%	16.33%
Subjectivity	2.56%	0.75%	0.00%	0.00%
Typo/mechanics	15.38%	8.96%	**54.26%**	33.33%

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

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

Table 6: Choices generation accuracy of LLM-generated datasets. The accuracy reflects the percentage of 'clean' choices.

Error Type	% of error type			
	LLM_ADAPT		LLM_GEN	
	`ind`	`sun`	`ind`	`sun`
Translation	10.17%	**77.32%**	0.00%	**76.61%**
Sent. fluency	1.69%	1.03%	9.52%	3.23%
Sent. context	1.69%	1.03%	0.00%	0.00%
Invalid options	3.39%	2.06%	**80.95%**	20.16%
Typo/mechanics	**83.05%**	18.56%	9.52%	0.00%

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

mainly due to major translation issues. As for LLM_GEN, although errors constitute less than 10% of the total evaluated samples, the primary issue is the presence of invalid options, particularly when no correct answers are among the options. For Sundanese, alongside translation errors, the model also produces some invalid options.

5.2 LLM vs. Human-Generated Data

To compare the data generated by LLM with human-generated data, we analyze the proportion of shared unigram and bigram tokens between the LLM_GEN and HUMAN_GEN datasets. The overlap percentage is calculated by dividing the number of shared tokens by the total number of 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. 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*.[4]

Upon sample-level analysis, we also find token variance disparities, particularly related to cultural context. For instance, given *'kerupuk'* (crackers) concept, human annotators can formulate questions tied to their cultural background, such as asking about *'kerupuk rambak'* (rambak crackers). In contrast, LLM tends to ask more general questions, like *"What is the common primary ingredient of crackers?"* Also, although LLM can generate some culturally related entities such as *'tari kecak'*, a traditional dance from Bali, it is more popular compared to *'tari pendet'*, which is generated by the human annotator. More examples are outlined in Table 11.

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), an LLM trained on a multilingual dataset, predominantly in English and Chinese; BLOOMZ-7B (Muennighoff et al., 2022), an LLM fine-tuned on the xP3 dataset; and SeaLLM-7B (Nguyen et al., 2023), an LLM with a focus on Southeast Asian languages.

Monolingual LLMs We include Merak-v4 (Ichsan, 2023), an Indonesian LLM fine-tuned based on MistralOrca-7B. We also include

---

[4] 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.

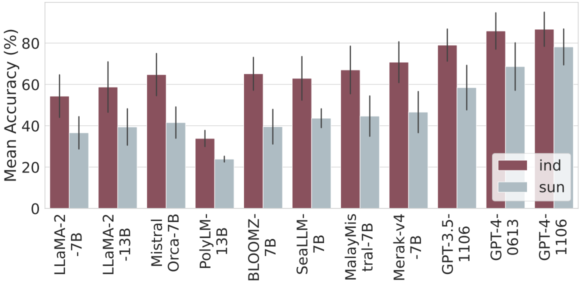

Figure 3: Overall performance of LLMs on our combined test set. We run three different prompts for each question in a zero-shot setting.

MalaysianMistral-7B (Zolkepli et al., 2024), a Malaysian LLM extended from Mistral-7B.

Proprietary LLMs We include GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo, (OpenAI, 2023), which are accessed via OpenAI API.

We use three prompt variations for all models (see Appendix E). 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 accuracy of ∼80%. Among open models, Indonesian LLM, Merak-v4, scores highest but cannot surpass GPT-3.5. Interestingly, the score difference between Merak-v4 and MalaysianMistral is small, possibly due to the significant lexical similarities between Indonesian and Standard Malay. This could also be attributed to the fact that some of the instruction data used to train MalaysianMistral is generated using GPT-4, which tends to produce Indonesian rather than standard Malay. To address this issue, Zolkepli et al. (2024) translated the Indonesian texts to standard Malay using MT, but some cultural nuances may still be related to Indonesia. On the other hand, the performance gap between Indonesian and Sundanese is notable, with accuracy dropping by 10–20%. This suggests that current LLMs struggle to answer commonsense questions in Sundanese, even in multiple-choice settings.

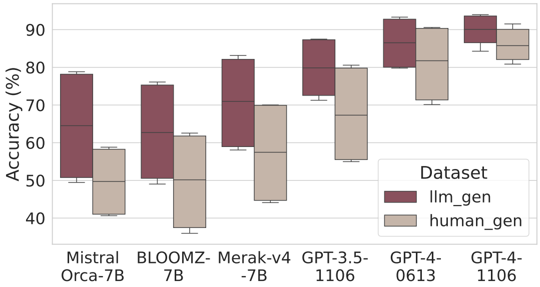

Figure 4: LLMs' performance on LLM_GEN vs. HUMAN_GEN.

6.3 LLM vs. Human-Generated Data

To assess LLMs' ability to answer both LLM and human-generated data, we compare their performances on our LLM_GEN and HUMAN_GEN datasets. Our result, shown in Figure 4, reveals that LLMs perform higher on LLM_GEN, especially on MistalOrca, which is trained on English-centric data. This indicates that the model may possess some knowledge to answer LLM-generated questions, even though it was not specifically trained on Indonesian or Sundanese data. However, the model still struggles to answer new knowledge present in HUMAN_GEN data. The performance gap is also observed in Indonesian LLM, Merak-v4. The accuracy of this model in Sundanese is also low, indicating that training LLM solely on Indonesian texts is not enough to make it perform well in local languages such as Sundanese.

6.4 Performance by Question Category

Figure 5 shows the LLMs' performance across different question categories. It reveals that LLMs perform better in the *activity* and *place* questions, but struggle in *culinary*. A notable performance gap exists between LLM_GEN and HUMAN_GEN dataset, with weaker results on the latter. Moreover, open LLMs exhibit a significant performance gap compared to proprietary models, especially in HUMAN_GEN, indicating their shortcomings in capturing commonsense knowledge. Interestingly, a significant improvement is seen from GPT-4 (`gpt-4-0613`) to GPT-4 Turbo (`gpt-4-1106`), especially in Sundanese HUMAN_GEN. This suggests that the model is acquiring more knowledge, one reason for which possibly comes from the interactions on the ChatGPT web. Despite these advancements, some categories still score below 80%, showing there is still room for improvement.

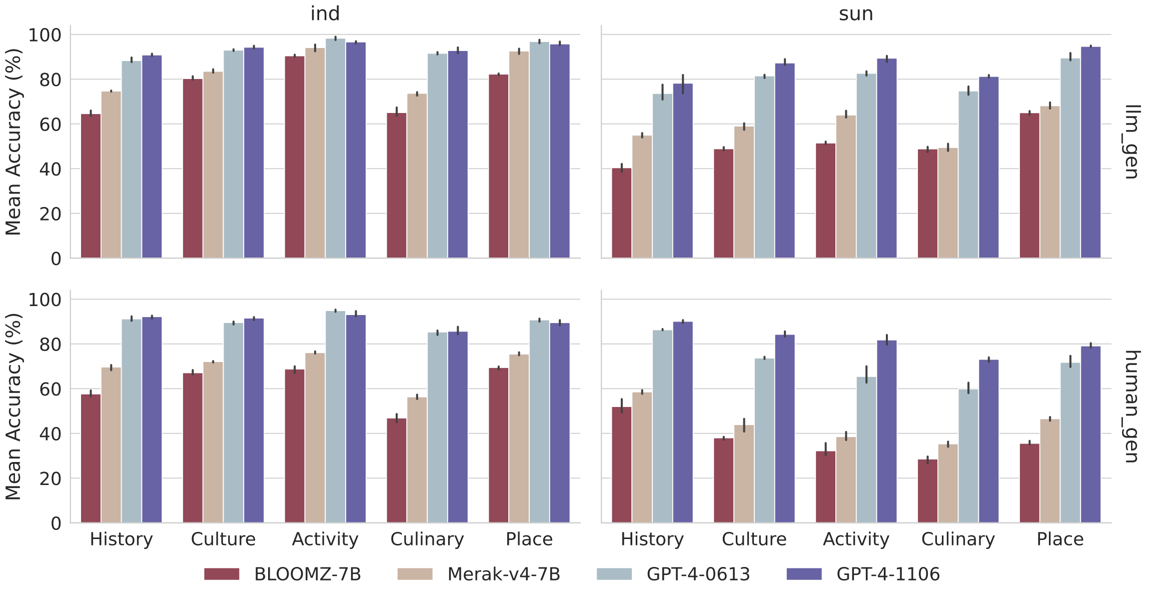

Figure 5: LLMs performance by question category in LLM_GEN and HUMAN_GEN for Indonesian and Sundanese.

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' genuine 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 a 77% accuracy rate, highlighting the model's struggle with open-ended questions and confirming our dataset's importance. Among the wrong answers, 18 cases stemmed from overly general answers not specific to Indonesia. For example, when asked *"Where can we find many villas?"* the model answers *"Bali"* but also listed other cities in mostly Western countries. In contrast, the annotator writes *"Bandung"* as the correct answer. As for the other wrong cases, the model struggles the most in answering questions in the culinary category. More failure examples can be seen in Table 12.

7.2 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 good-quality synthetic data could be particularly beneficial for extremely low-resource languages, such as Buginese or Toba Batak, given the difficulty of finding native speakers. Yet, our case study in Sundanese indicates that LLMs might struggle even more with understanding cultural nuances in such languages.

8 Conclusion

In this study, we created Indonesian and Sundanese CommonsenseQA datasets using various dataset collection methods, including automatic dataset generation utilizing LLMs and manual dataset generation employing human annotators from diverse regions. Our thorough analysis reveals a substantial gap in the performance of LLMs between Indonesian and Sundanese, both in the dataset generation and in answering the questions in the dataset itself. Thus, one must be cautious when using existing LLMs in generating datasets, particularly for lower-resource languages, since their culturally relevant commonsense knowledge may be limited. Our findings highlight the urgent need to create high-quality resources for these underrepresented languages to prevent them from being left behind.

Limitations

Our work has some limitations that should be acknowledged. Firstly, 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 questions.

Additionally, 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 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.

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 the Institutional Review Board (IRB) in our institution. All recruited annotators were paid above the minimum wage. Our datasets will be publicly available under the Creative Commons Non-Commercial (CC BY-NC 4.0) license.

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.

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`.

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.

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.

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*.

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*.

F.L. van Holthoon and D.R. Olson. 1987. *Common Sense: The Foundations for Social Science*. Common Sense. University Press of America.

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.

A LLM-Generated Data

A.1 Prompts Details

A.1.1 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}
###
```

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).

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.1.2 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:
```

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

Table 8 shows the win rate accuracy of concept, question, and choices 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, the model 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?).

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.3 Additional Analysis of the 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 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.

Categories and Concepts Five selected categories are used in the data creation process. Each category has 150 concepts that need to be used and developed to become commonsense questions. The following is a list of the categories.

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.

General Rules There are several rules in the data creation process. The following is the explanation.

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*___unjung___*i*, *di*___kunjung___*i*).
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.

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*).

C Annotators Demographics

In accordance with the aim of constructing the dataset, we involved all Indonesian native-speaker annotators in the manual data generation process. 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. Each dataset development team consists of 6 annotators. The detailed annotators' demographics are shown in Table 10.

D Manual Data Generation Findings

During the manual data generation process, we discovered several interesting findings from the data created by the annotators.

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 the limitations of this data generation were the Indonesian and Sundanese contexts (specifically for the Sundanese dataset). For instance, *"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.

Choices The findings related to choices 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 Choices: 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.

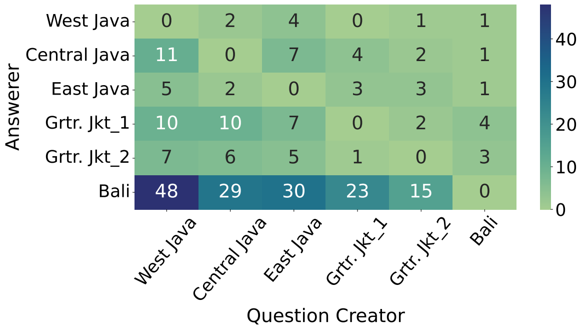

Figure 6: Answer conflict across Indonesian annotators.

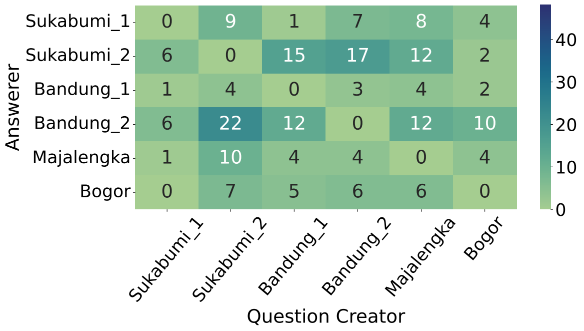

Figure 7: Answer conflict across Sundanese annotators.

Knowledge Variations between Annotators As we employ annotators from different regions, it enables us to explore 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 the 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.

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. This finding points to the fact that differences in annotators' knowledge levels may arise from various factors, one of which could be the regional culture to which they are most exposed.

E 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:
```

F 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.

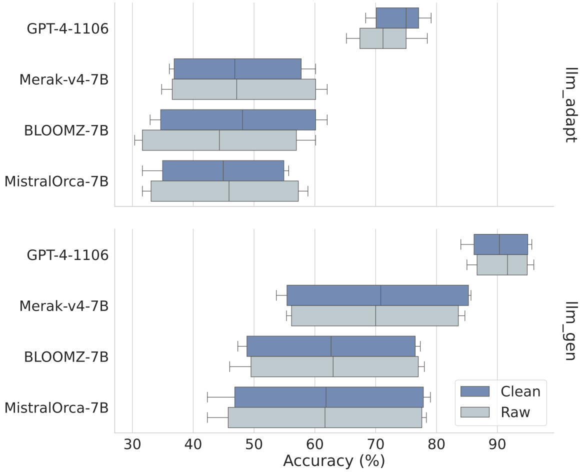

Figure 8: Performance comparison of raw vs. cleaned version of LLM-generated data.

G LLM vs. Human-Generated Data Examples

We provide some GPT-4 Turbo generation samples compared with human-generated data in Table 11.

H 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 12.

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 11: 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 12: Examples of incorrect responses by GPT-4 Turbo (`gpt-4-1106-preview`) in 'free' or open-ended generation settings.