IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces

Fajri Koto[1] **Rahmad Mahendra**[2,3] **Nurul Aisyah**[4] **Timothy Baldwin**[1,5]

[1]Department of Natural Language Processing, MBZUAI, UAE
[2]Universitas Indonesia, Indonesia   [3]RMIT University, Australia
[4]Quantic School of Business and Technology, USA   [5]The University of Melbourne, Australia
`fajri.koto@mbzuai.ac.ae, rahmad.mahendra@cs.ui.ac.id`

Abstract

Although commonsense reasoning is greatly shaped by cultural and geographical factors, previous studies have predominantly centered on cultures grounded in the English language, potentially resulting in an Anglocentric bias. In this paper, we introduce `IndoCulture`, aimed at understanding the influence of geographical factors on language model reasoning ability, with a specific emphasis on the diverse cultures found within eleven Indonesian provinces. In contrast to prior work that has relied on templates (Yin et al., 2022) and online scrapping (Fung et al., 2024), we create `IndoCulture` by asking local people to manually develop a cultural context and plausible options, across a set of predefined topics. Evaluation of 27 language models reveals several insights: (1) the open-weight Llama–3 is competitive with GPT–4, while other open-weight models struggle, with accuracies below 50%; (2) there is a general pattern of models generally performing better for some provinces, such as Bali and West Java, and less well for others; and (3) the inclusion of location context enhances performance, especially for larger models like GPT–4, emphasizing the significance of geographical context in commonsense reasoning.[1]

1 Introduction

The reasoning abilities of multilingual language models are frequently evaluated using English texts, potentially amplifying an Anglocentric bias toward culture grounded in the English language, and leading to less inclusive models (Thomas, 1983; Ponti et al., 2020). Cultures, however, vary significantly from one location to another and profoundly shape the way speakers of a language reason (Hershcovich et al., 2022). Recent

---

[1]`IndoCulture` is available from `https://huggingface.co/datasets/indolem/IndoCulture`.

evaluations of models' commonsense reasoning ability (OpenAI, 2023; Sengupta et al., 2023; Liu et al., 2023) have been conducted on English datasets such as Social IQA (Sap et al., 2019) and PIQA (Bisk et al., 2020), and thus often overlook geographical aspects, thereby risking cultural bias.

Culture is a multifaceted concept encompassing the way of life (Giddens and Sutton, 2021), including our thoughts and actions (Macionis, 2012). It includes tangible elements like food, art, and clothing, as well as intangible aspects such as ideas, values, attitudes, and norms. Culture is shaped by geographical location and ethnicity, influencing the commonsense reasoning of people within a region. For example, in Indonesia, it is culturally acceptable to eat rice with your hands but it is considered unusual to use chopsticks. Similarly, at traditional Indonesian weddings, it is common to sit on the floor while eating, whereas this practice is less common in Australia.

This work focuses on understanding the influence of geographical contexts in cultural commonsense reasoning, with the main focus on Indonesian culture. Indonesia is a highly multicultural country (Putra et al., 2019), home to over 1,300 recognized ethnic groups and more than 700 languages (Zarbaliyev, 2017; Aji et al., 2022). As the targest archipelagic country in the world, Indonesia has a population exceeding 270 million spread across 38 provinces, stretching from Aceh province in the west to Papua province in the east. Few prior studies on commonsense reasoning in Indonesian contexts (Mahendra et al., 2021; Wibowo et al., 2024; Putri et al., 2024) have explicitly addressed the geographical nuances and rich diversity of Indonesian cultures.

This paper introduces `IndoCulture`, a novel dataset to evaluate cultural reasoning in eleven Indonesian provinces, manually developed by local people in each province based on predefined

topics. In prior work, cultural reasoning has primarily relied on datasets constructed through templates (Yin et al., 2022), and online scraping (Nguyen et al., 2023; Fung et al., 2024). While these studies offer valuable insights, they may be susceptible to training data contamination when used to assess large language models (LLMs). For instance, Fung et al. (2024) reported a zero-shot accuracy of 92% when using ChatGPT (Ouyang et al., 2022) to evaluate low-resource data.

`IndoCulture` contains cultural commonsense knowledge data from eleven provinces in Indonesia (blue colored in Figure 1), namely, Aceh, North Sumatra, West Sumatra, West Java, Central Java, East Java, Bali, South Borneo, East Nusa Tenggara (NTT), South Sulawesi, and Papua. These provinces span breadth of Indonesia, each representing a major island in the country, with the addition of Bali and NTT. Figure 1 also shows three examples in `IndoCulture` for three provinces: Aceh, North Sumatra, and Papua.[2] The first example focuses on cultural artifact, specifically, the traditional wedding dress from Aceh. The second example examines family relationships while the third example focuses on cultural beliefs and norms regarding pregnancy in Papua.

Can large language models effectively reason based on the diverse cultures of Indonesia? To capture the rich diversity of Indonesian cultures, we predefined 12 fine-grained topics as guidelines for data construction. Figure 2 displays the topic distribution in `IndoCulture`, with the majority focusing on food, weddings, art, pregnancy and children, and family relationships. Additionally, we also pose the question: *Is there any influence of geographical location on the commonsense reasoning of language models?* We address these questions through comprehensive experiments across different language models, incorporating several levels of location granularity as additional context in the prompt.

Our contributions can be summarized as follows:

- We present `IndoCulture`, a high-quality cultural reasoning dataset in the Indonesian language, covering eleven provinces of Indonesia and twelve fine-grained cultural

Figure 1: `IndoCulture` covers eleven provinces spanning from eastern to western Indonesia. The highlighted regions in the map represent the provinces examined in `IndoCulture`. We present examples from Aceh, North Sumatra, and Papua, with three plausible options and correct answers indicated in bold. English translations are provided for illustrative purposes.

topics. Our dataset has 2,429 instances, and was developed by local people with rigorous quality controls in place.

- We assess 19 open-weight multilingual models, 6 open-weight Indonesian-centric models, and 2 closed-weight models. Although local individuals can answer all questions correctly (i.e., 100% accuracy), most open-weight models struggle to comprehend Indonesian cultures. Interestingly, we observed that Llama–3 (Dubey et al., 2024) is competitive with GPT–4 (OpenAI, 2023).

---

[2]Although Papua consists of six provinces, for the purpose of this study, we treat it as a single entity (referred to as Papua) due to the relatively recent establishment of most of these provinces.
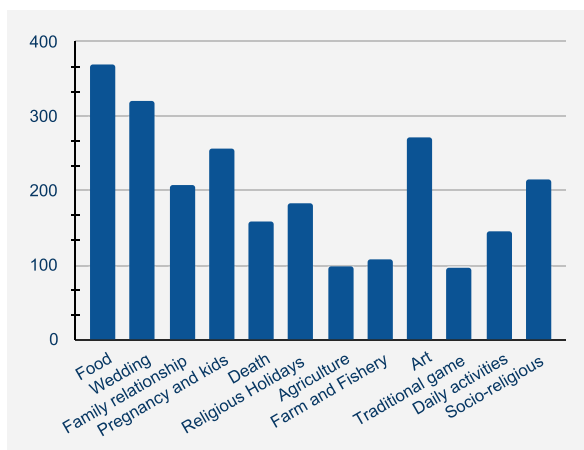
Figure 2: Topic distribution in `IndoCulture`.

- We conduct a thorough analysis over various dimensions: (1) model performance for each province and topic; (2) the influence of different granularities of location context (i.e., none, province, country); (3) model performance over English translations; and (4) analysis of model explanations for a given answer.

2 Related Work

Commonsense Reasoning in English Many studies have focused on commonsense reasoning in English, often overlooking considerations of culture and geographical location. Early work included the `Winograd Schema Challenge` (Levesque et al., 2012) and `WinoGrande` (Sakaguchi et al., 2021) for pronoun coreference resolution. Other research areas include reasoning based on cause-effect relationships (Roemmele et al., 2011), physical activities (Bisk et al., 2020), social interactions (Sap et al., 2019), cloze story completion (Mostafazadeh et al., 2016), sentence completion (Zellers et al., 2019), numerical reasoning (Lin et al., 2020), and temporal reasoning (Qin et al., 2021). Additionally, pretrained language models have been employed in other work to extract structured commonsense knowledge by providing seed words (Davison et al., 2019), and using *code* language models (Madaan et al., 2022).

Cultural Commonsense Reasoning with Geographical Contexts Previous studies have explored commonsense reasoning with geographical context. Shwartz (2022) investigated time perception (e.g., morning and night) across different locations, while Yin et al. (2022) examined

cultural knowledge of language models across five countries using datasets built from templates and translations. Other work has focused on automatically extracting cultural knowledge from various sources, including Wikipedia (Fung et al., 2024), conversations (Fung et al., 2023), and Common Crawl (Nguyen et al., 2023), incorporating location context with the assistance of LLMs. Relatedly, Ziems et al. (2023) created a knowledge bank for situational norms, using English-speaking Mechanical Turk annotators and incorporating a country taxonomy. Unlike this work, `IndoCulture` specifically concentrates on cultural reasoning across Indonesian provinces, developed and validated manually by local people (experts). Compared to the automatic method and English-speaking crowd workers for data construction, `IndoCulture` arguably contains less noise, and is free from the training data contamination of LLMs.

Commonsense Reasoning with Indonesian Contexts Table 1 shows a comparison of `IndoCulture` with other Indonesian datasets for cultural knowledge and reasoning evaluation. Commonsense reasoning in Indonesian language models has been studied using translated English–Indonesian datasets, such as `XCOPA` (Ponti et al., 2020) and `XStoryCloze` (Lin et al., 2022). However, these datasets potentially introduce a cultural bias toward culture grounded in the English language. `IndoCloze` (Koto et al., 2022) was the first commonsense reasoning dataset in Indonesian, developed by native Indonesian workers following the cloze story completion framework (Mostafazadeh et al., 2016). However, `IndoCloze` lacks local cultural nuances and fine-grained geographical context. Wibowo et al. (2024) followed the COPA framework (Roemmele et al., 2011) to build a dataset with contexts limited to Jakarta. In other work, Putri et al. (2024) studied the capability of LLMs in generating questions with cultural norms, for both general Indonesian and specific Sundanese contexts, while Liu et al. (2024) used proverbs and LLMs to generate conversational data. In contemporary work, Myung et al. (2024) released BLEnD, a large-scale cultural knowledge dataset, built using templates, translation, and human validations, covering the West Java province in Indonesia. BLEnD specifically focuses on short-answer questions, limiting its capacity for

Dataset	Size	Data Construction Method	Cultural?	Location?	#province	#topic
IndoCulture (**ours**)	2,429	Manually built and validated by native	✓	✓	11	66
COPAL-ID (Wibowo et al., 2024)	559	Manually built and validated by native	✓	–	–	–
MAPS (Liu et al., 2024)	371	LLM generation & human generation	✓	–	–	1
ID-CSQA (Putri et al., 2024)*	4,416	LLM generation & human generation	✓	–	–	5
BLEnD (Myung et al., 2024)	1,000	Template, translation, human validation	✓	✓	1	6
IndoCloze (Koto et al., 2022)	2,335	Manually built and validated by native	–	–	–	–
XCOPA (Ponti et al., 2020)	600	Translated from English data	–	–	–	–
XStoryCloze (Lin et al., 2022)	1,872	Translated from English data	–	–	–	–

Table 1: Comparison of IndoCulture with other cultural knowledge and reasoning datasets containing instances in Indonesian. The metadata includes **Size** (number of Indonesian instances), **Cultural?** (whether the data considers cultural nuances), **Location?** (whether the data includes fine-grained location information, such as provinces, as context), **#province** (number of Indonesian provinces covered), and **#topic** (number of fine-grained topics covered). * indicates the dataset involves question generation with less emphasis on reasoning.

reasoning evaluation. Unlike most other datasets that do not consider geographical factors, IndoCulture has broad coverage across eleven provinces, thereby providing greater inclusivity for local communities in Indonesia.

3 IndoCulture

As illustrated in Figure 1, IndoCulture is a sentence completion task in the Indonesian language featuring a one-sentence premise, three plausible options, and one correct option to evaluate reasoning ability and cultural knowledge across eleven Indonesian provinces. While sentence completion tasks are straightforward for humans, answering IndoCulture requires machines to engage in cultural reasoning to logically conclude which of the three options is logically consistent with the first sentence (Huang and Chang, 2023). The dataset includes a total of 2,429 instances.

3.1 Data Construction

IndoCulture was constructed manually by humans, and verified through a two-step process.

Worker Recruitment Culture generally arises from the shared experiences, traditions, and beliefs of a specific group over time, often closely intertwined with native populations. With this in mind, we engaged individuals from various provinces across Indonesia to assist in preparing data for the IndoCulture benchmark.

During recruitment, we presented a few examples of the intended IndoCulture data and requested each candidate to generate similar instances tailored to the context of their respective provinces. From a pool of 58 applicants, we carefully selected 22 expert workers representing 11 provinces (with 2 workers selected per province). These recruited expert workers are local residents and have resided in their respective provinces for a minimum of 10 years, thereby possessing a profound understanding of local customs and culture. The age range of our workforce spans from 21 to 35 years old, with educational backgrounds distributed as follows: 3 high school graduates, 14 bachelor's degree holders, 4 master's degree holders, and 1 PhD holder.

During data construction, each expert worker fulfilled the dual roles of instance writer and quality controller. Each worker was compensated above the monthly minimum wage in Indonesia.

Province Selection The provinces covered in this study represent the diversity of Indonesian cultures. The 11 provinces (in Figure 1) are spread across 6 islands of the Indonesian archipelago, which are inhabited by different ethnic groups who speak different regional languages and adhere to different religions.

Topic Taxonomy IndoCulture consists of 12 topics and 66 fine-grained subtopics, carefully constructed based on discussions and brainstorming with Indonesian natives. The selection of these topics and subtopics was guided based on several criteria and motivations: (1) relevance to Indonesian culture; (2) diversity and coverage; (3) regional representation (e.g., religious holidays); (4) practicality; and (5) expert consultation (i.e., native speaker feedback). Compared to the other Indonesian datasets in Table 1, IndoCulture includes a richer array of fine-grained topics. Below is a list of the topics along with their detailed subtopics. The numbers following each topic

indicate the total number of instances required to be written by one worker (with a total of 150 per worker).

1. **Food** (22): breakfast (2); lunch (3); dinner (2); snacks (2); food souvenirs (3); traditional foods and beverages (5); eating habits (1); cutlery (1); cooking ware (1), fruit (2).

2. **Wedding** (20): traditions before marriage (3); traditions when getting married (3); traditions after marriage (3); men's wedding clothes (2); women's wedding clothes (2); invited guests (2); wedding location (1); foods at a wedding (2); gifts brought to weddings (2).

3. **Family relationship** (13): relationships within the main family (3); relationships in the extended family (3); relations with society/neighbors (5); clan/descendant system (2).

4. **Pregnancy and kids** (16): traditions during pregnancy (4); traditions after birth (2); how to care for a newborn baby (2); how to care for toddlers (2); how to care for children (2); how to care for teenagers (2); parents and children interactions as adults (2).

5. **Death** (10): when death occurs (2); the process of dealing with a corpse (2); traditions after the body is buried (2); the clothes of the mourners (2); inheritance matters (2).

6. **Religious holiday** (12): traditions before religious holidays (2); traditions leading up to religious holidays (4); traditions during religious holidays (5); traditions after religious holidays (1).

7. **Agriculture** (6): what to plant (2); traditions when planting (2); harvest (2).

8. **Fisheries and trade** (7): traditions of taking care of livestock/fish (5); buying and selling traditions (2)

9. **Art** (16): musical instruments (3); folk songs (3); traditional dances (3); use of art at certain events (5); poetry or similar literature (2)

10. **Traditional games** (5): game types; (3), location played (2).

11. **Daily activities** (10): morning activities (1); afternoon activities (1); evening activities (1); leisure activities (3); house, household, and transportation (4).

12. **Socio-religious aspects of life** (13): regular religious activities (2); mystical things (2); traditional ceremonies (1); lifestyle (3); self care (1); traditional medicine (3); traditional sayings (1).

Instance Writing For each instance, workers were asked to craft two culturally relevant sentences that align with the predefined subtopic. The first sentence serves as the premise context, and the last sentence acts as the correct answer. Subsequently, the annotator generates two additional plausible sentences as distractors by modifying cultural objects or activities from the correct sentence. These distractors are designed to reflect local cultural contexts, ensuring they are challenging yet unambiguous, and could potentially serve as correct answers in other regional contexts. Workers were given a period of two months to complete the task.

Two Stages of Quality Control In stage 1, we implemented quality control by pairing two annotators from the same province. Each annotator was tasked with answering a set of questions prepared by the other annotator, and vice versa. During this phase, the annotator were presented with a premise sentence and three shuffled options. They were allowed to search for the answer from any source if they were unsure. Instances that were incorrectly answered by the second annotator were discarded, as we hypothesize that these instances may contain incorrect answers or possess a level of ambiguity. Additionally, annotators were required to identify whether the instance is province-specific (binary annotation: True/False), indicating that it is uniquely relevant in their province and not in others.

In stage 2 of quality control, the first two authors of this paper performed post-editing of data that passed the first stage of quality control. We first focused on correcting the linguistic aspects of the text, including checking for spelling errors. Although the text is written in Indonesian, some annotators may use dialects or be influenced by the structure or style of regional languages. In these cases, we corrected the text to adhere to Indonesian grammar.

To maintain the quality of IndoCulture, we rigorously filtered instances that contained: (1) poor writing, in the case that it was difficult to post-edit to enhance their quality; (2) obvious answer options, which allow for easy guessing of

Province	#	province specific (%)	μ(word)	μ(char)
Aceh	246	70.7	28.0	175.9
North Sumatra	234	83.8	36.8	246.0
West Sumatra	299	74.6	39.6	261.4
West Java	231	58.0	37.5	244.8
Central Java	171	66.7	39.3	260.5
East Java	233	69.5	46.0	310.4
Bali	241	76.3	33.3	216.1
NTT	103	72.8	31.8	203.6
South Borneo	233	83.7	33.3	226.0
South Sulawesi	185	90.3	33.6	227.8
Papua	253	88.1	37.3	245.0
All	2429	76.0	NA	NA

Table 2: Overall statistics of `IndoCulture` by province.

the correct choice without understanding the cultural context; and (3) ambiguous contexts, where all options are equally valid as the correct answer. For example, in a topic about breakfast, the three options might include one traditional food alongside two other very commonly consumed foods in Indonesia, and be considered too obvious.

Furthermore, we manually verified the province-specific annotations for each instance using the Google search engine. We annotated whether the instance pertains to national-level culture or not. If the example is specific to a province, it will be annotated as uncommon in national culture, and vice versa.

3.2 Data Statistics

After the instance writing process, we initially collected 3,162 instances out of a target of 3,300 instances (22 workers × 150 subtopics). Although we requested each annotator to produce 150 instances, not all were able to complete their allotted tasks within the given timeframe. Unfortunately, we were unable to find additional candidates from the same local province to address the data deficiencies (Winata et al., 2023).

In stage 1 of quality control, the initial pool of 3,162 instances was reduced to 2,801 instances, and stage 2 of quality control further reduced the sample to 2,429 high-quality samples. The data distribution of `IndoCulture` per province is presented in Table 2. Approximately three-quarters of `IndoCulture` instances contain province-specific content, with an average length of around 35 words. `IndoCulture` covers multiple topics, as illustrated in Figure 2.

Sentence completion prompt

Indonesian Prompt
Lokasi: [LOCATION]
[PREMISE] [OPTION-i]

English prompt
Location: [LOCATION]
[PREMISE] [OPTION-i]

Multiple-choice question prompt

Indonesian Prompt
Untuk konteks [LOCATION], sambungan yang tepat dari kalimat "[PREMISE]" adalah
[OPTIONS]
Jawaban:

English prompt
Given [LOCATION] as the location context, how would you continue the sentence "[PREMISE]"?
[OPTION]
Answer:

Figure 3: Templates for sentence completion and multiple-choice questions prompts.

4 Experiments

4.1 Set-Up

We evaluate 27 language models in zero-shot settings: (1) nineteen open-weight multilingual language models of varying sizes, namely, BLOOMZ (Muennighoff et al., 2023), mT0 (Muennighoff et al., 2023), Bactrian-X (Li et al., 2023), Llama–2 (Touvron et al., 2023), and Llama–3 (Dubey et al., 2024); (2) two South East Asian language models, namely, SeaLLM (Nguyen et al., 2024), and SeaLion (Singapore, 2023); (3) four Indonesian-centric language models, namely, IndoBART (Cahyawijaya et al., 2021), IndoGPT (Cahyawijaya et al., 2021), Merak (Ichsan, 2023), and Komodo (Owen et al., 2024); (3) two closed-weight models, namely, ChatGPT: `gpt-3.5-turbo` (Ouyang et al., 2022) and GPT–4: `gpt-4-0613` (OpenAI, 2023). Please refer to Appendix A for further details.

First, we evaluate the effectiveness of sentence completion and multiple-choice question strategies in predicting the correct options using the Indonesian and English prompt templates shown in Figure 3. In both scenarios, we conduct benchmarks across three distinct location contexts. Formally, given a premise s,

Model (#parameter)	Completion			MCQ		
	l = None	l = Ind	l = Prov	l = None	l = Ind	l = Prov
Human	–	–	100.0	–	–	100.0
Random	33.3	33.3	33.3	33.3	33.3	33.3
BLOOMZ (560M)	37.2	35.3	35.3	32.5	32.4	32.5
BLOOMZ (1.1B)	36.3	36.9	37.2	32.4	32.4	32.4
BLOOMZ (3B)	38.6	40.7	41.5	47.0	48.6	49.2
BLOOMZ (7B)	41.3	44.1	44.6	49.5	50.6	50.5
mT0$_{small}$ (300M)	28.3	28.1	28.3	34.1	33.1	32.6
mT0$_{base}$ (580M)	28.4	28.1	28.5	35.4	35.1	35.6
mT0$_{large}$ (1.2B)	29.6	29.5	30.1	35.6	35.7	35.8
mT0$_{xl}$ (3.7B)	31.9	31.0	31.2	49.8	50.5	50.7
mT0$_{xxl}$ (13B)	33.2	33.5	34.3	52.7	51.4	52.1
Bactrian-X$_{LLaMA}$ (7B)	33.8	34.2	34.2	38.0	38.6	38.9
Bactrian-X$_{LLaMA}$ (13B)	33.3	35.2	35.1	38.6	38.2	38.6
Llama–2 (7B)	37.2	37.5	37.7	40.5	39.9	38.8
Llama–2 chat (7B)	37.3	37.4	37.9	40.6	41.3	40.7
Llama–2 (13B)	39.6	40.2	40.2	47.6	47.6	47.3
Llama–2 chat (13B)	38.6	38.9	39.3	47.8	49.6	49.6
Llama–3 (8B)	41.0	42.2	43.4	54.4	54.4	55.1
Llama–3 Instruct (8B)	41.9	41.5	42.3	56.7	57.6	59.0
Llama–3 (70B)	**51.2**	**51.7**	**54.3**	**68.6**	**69.9**	72.7
Llama–3 Instruct (70B)	49.2	49.6	52.2	68.5	69.3	**73.3**
IndoBART (132M)	42.4	41.3	42.1	32.6	32.4	32.7
IndoGPT (117M)	42.6	41.9	42.4	33.7	33.8	34.7
Merak (7B)	41.0	41.5	43.5	51.9	**53.1**	**53.2**
SeaLLM (7B)	39.1	39.3	41.1	**52.2**	**53.1**	53.0
SEA-LION (7B)	38.8	38.9	39.7	33.8	33.0	33.3
Komodo (7B)	**45.1**	**45.4**	**46.1**	37.6	35.1	36.1
GPT–3.5 (NA)	–	–	–	59.8	60.9	62.7
GPT–4 (NA)	–	–	–	**69.1**	**71.8**	**75.9**

Table 3: Zero-shot accuracy across various models and settings. "MCQ" refers to the multiple-choice question method, and l denotes the location as additional context ("Ind" and "Prov" denote the country of Indonesia, and the corresponding province). The bold numbers highlight the highest score within each model group.

three candidate options c_1, c_2, c_3, and location $l \in \{\texttt{none}, \texttt{Indonesia}, province\}$, for sentence completion, we select the correct option based on:

$$\underset{c}{\arg\max} \log P(\text{concat}(s, c) | l)$$

Here, $\text{concat}(s, c)$ denotes the concatenation of premise s and candidate option c, separated by a space. In the case of multiple-choice questions, we devise a template for the prompt question and determine the answer by selecting the option with the highest probability among letters A, B, and C.

For GPT–3.5 and GPT–4, we exclude experiments with sentence completion because the closed-weight models do not provide an overall probability score. For multiple-choice questions,

we select the first generated token that corresponds to the letters A, B, or C using a regular expression.

4.2 Results

Overall Observation The results presented in Table 3 display the performance across various models and settings. The overall observation is that most open-weight models struggle to understand Indonesian culture, contrasting sharply with the 100% accuracy achieved by humans (i.e., natives of the given province). Among open-weight models, Llama–3 achieves the highest accuracy of 73.3%. Other open-weight models such as Merak and mT0$_{xxl}$ achieve accuracy of 52–53%, while closed-weight models, such as GPT–3.5 and GPT–4, achieve performances of 62.7% and

75.9%, respectively. These findings underscore the challenging nature of the `IndoCulture` dataset.

The Multiple-choice Question Method is Generally Better. Our findings suggest that the multiple-choice question method tends to outperform the sentence completion method, with exceptions noted for BLOOMZ (560M, 1.1B), IndoBART, IndoGPT, and Komodo. Interestingly, in the sentence completion task, the Indonesian-focused language model Komodo (7B) outperforms nearly all large multilingual models, with the exception of Llama–3 (70B). However, Komodo experiences a significant decline for the multiple-choice question method, with a notable margin of 10–12 points. This discrepancy could potentially be attributed to differences in the nature of language model training and instruction-tuning.

Impact of Location Context on Model Performance Our investigation reveals that incorporating various levels of location granularity has a noticeable effect on zero-shot performance, especially for models with larger parameter sizes. Detailed location context notably enhances the accuracy of BLOOMZ (7B), Llama–2 (13B), Llama–3 (70B), Merak (7B), SeaLLM (7B), Komodo (7B), GPT–3.5, and GPT–4. For instance, in GPT–4, the accuracy gap between $l = \texttt{none}$ and $l = \texttt{Indonesia}$ is 2.7, and this gap further increases by 7 points when $l = \texttt{Province}$ is assigned.

4.3 Analysis

Given the exceptional performance of models with large parameter sizes using the multiple-choice question method and location $l = \texttt{Province}$, we employ these configurations for our analysis. In this section, our main focus is on the top three performing models: Llama–3 Instruct (70B), Merak (7B), and GPT–4. These models represent a multilingual open-weight model, Indonesian-centric open-weight model, and closed-weight model, respectively.

Results by Province Table 4 highlights that the top 3 performing LLMs exhibit a nuanced understanding of culture within Indonesian provinces, particularly excelling in the cultures of West Java and Bali compared to other provinces. Llama–3

Province	Merak		Llama–3		GPT–4	
	¬PS	PS	¬PS	PS	¬PS	PS
Aceh	59.7	53.4	77.8	68.4	93.1	73.6
North Sumatra	58.3	46.9	69.4	67.5	75.0	73.2
West Sumatra	50.0	41.3	73.7	64.6	85.5	63.7
West Java	65.3	58.8	91.6	86.3	92.6	81.7
Cental Java	58.8	47.1	79.4	67.1	82.4	72.9
East Java	57.4	37.6	88.9	57.4	87.0	63.4
Bali	78.9	65.6	96.5	86.7	93.0	89.4
NTT	64.3	52.0	78.6	61.3	85.7	68.0
South Borneo	65.7	46.4	85.7	66.9	77.1	67.4
South Sulawesi	53.3	47.9	73.3	66.9	80.0	70.6
Papua	76.7	55.2	90.0	73.1	90.0	71.3

Table 4: Top-3 model accuracy by province. "PS" indicates instances containing province-specific context, while "¬PS" indicates otherwise. The green and red cells indicate the top three and bottom three scores, respectively.

and GPT–4, for instance, achieve the best accuracy of more than 90%, while Merak achieves the best performance in Bali, Papua, and West Java, with accuracies ranging between 55% and 79%. In other provinces like West Sumatra and South Borneo, the models typically exhibit poorer performance. Specifically, for Llama–3, the performance gap compared to Bali ranges from 10% to 30%. This highlights the presence of cultural biases and a lack of inclusivity in model reasoning abilities, likely stemming from the distribution of training data. The proximity of West Java to Jakarta (Indonesia's capital) and Bali's global status as a tourism destination may contribute to the abundance of textual data on these two cultures.

We also note a consistent disparity between non-province and province-specific contexts across all models, with models generally finding non-province contexts easier to comprehend. On average, this gap ranges from 12 to 13 points for the three models, highlighting the challenge posed by province-specific content and emphasizing the significant influence of location context on the reasoning ability of LLMs.

Results by Topic Table 5 shows the accuracy of the top 3 performing models across different topics. Similar to Table 4, the models perform worse in province-specific contexts for all topics, with the notable exception of food. For province-specific contexts, GPT–4 excels for the themes of food, religious holidays, and arts

Topic	Merak		Llama–3		GPT–4	
	¬PS	PS	¬PS	PS	¬PS	PS
Food	58.7	50.2	69.6	71.8	73.3	77.3
Wedding	60.0	50.2	80.0	66.9	82.5	66.9
Family relationship	55.1	39.6	75.5	66.9	85.2	63.0
Pregnancy and kids	57.1	44.3	83.5	69.0	84.7	64.6
Death	56.9	51.0	86.2	62.5	79.4	63.5
Religious holidays	66.7	64.3	79.7	78.6	91.5	78.6
Agriculture	65.5	54.5	100.0	80.3	93.8	72.7
Farm and fishery	67.9	42.9	85.7	62.3	80.0	72.7
Art	77.8	55.2	83.3	72.2	90.0	82.5
Traditional game	38.5	48.8	69.2	68.3	78.6	64.6
Daily activities	68.5	52.3	92.6	72.7	93.1	76.1
Socio-religious	70.5	51.0	88.5	71.8	86.4	75.8

Table 5: Top-3 model accuracy by topic. ''PS'' indicates instances containing province-specific contexts, while ''¬PS'' indicates otherwise. The green and red cells indicate the top three and bottom three scores, respectively.

while for non-specific contexts, Llama–3 achieves accuracy more than 90% for agriculture, daily activities, and religious holidays.

Results by Fine-grained Cultural Elements
We tasked two expert workers with annotating 200 random samples from IndoCulture based on six cultural elements, as derived from Axtell and Fornwald (1998); Williams (2014).[3,4] While these elements may not encompass every cultural aspect, we contend that they cover the most prominent or pivotal elements, including: (1) *symbols* (material or non-material objects representing meaning); (2) *artifacts* (material or non-material objects produced by society); (3) *values and beliefs* (principles, ideas, and concepts assumed to be ideal and correct in society); (4) *norms* (rules guiding values and beliefs); (5) *language*; (6) *rituals* (established procedures and ceremonies); and (7) *other*, for examples that do not fit into any of the defined elements. This annotation is a multi-label task, and the average Kappa score across the cultural elements is 0.56, with each ranging from 0.4 to 0.75. These scores indicate moderate to substantial agreement.

[3] https://www.languageeducatorsassemble.com/elements-of-culture.

[4] https://pressbooks.howardcc.edu/soci101/chapter/3-2-the-elements-of-culture/.

Cultural element (%)	Merak	Llama–3	GPT–4
Symbols (4)	50.0	50.0	70.8
Artifacts (42.5)	55.3	74.1	67.8
Values and Beliefs (10.5)	61.9	85.7	69.8
Norms (37.5)	38.7	62.7	73.6
Language (19.5)	38.5	66.7	72.0
Ritual (30)	53.3	65.0	70.7
Other (7.5)	66.7	66.7	69.2

Table 6: Accuracy comparison of Merak, Llama–3–Instruct (70B), and GPT–4 across 200 random samples, categorized by cultural elements. The numerical value following each cultural element indicates its proportion within the samples.

Table 6 displays the distribution of each cultural element in our dataset, along with the performance breakdown across Merak, Llama–3, and GPT–4. Among the 200 random samples, we observe that 42.5% of our data contains *artifacts*, 37.5% *norms*, and 30% *rituals*. Only 4% of the data pertains to *symbols*, while 7.5% belongs to the *other* category. Merak shows lower accuracy in *norms*, with a 24% decrease compared to *values and beliefs*. Conversely, Llama–3 performs best in *values and beliefs* with 85% accuracy, but accuracy drops by 23% for *norms*. GPT–4 maintains relatively stable accuracies across cultural elements, with differences averaging between 3% and 5%. Furthermore, *language* presents a challenge for Merak, achieving only 38% accuracy, whereas Llama–3 and GPT–4 achieve 66% and 72% accuracy, respectively.

Can the Model Provide a Reasonable Explanation to Support the Answer? We conduct a manual investigation of the text generation output for Merak (7B), Llama–3, and GPT–4 across 200 random samples. This involves manually examining the generated answer along with its explanation. To obtain the explanation, we modify the Indonesian prompt in Figure 3 by adding the string *Jelaskan jawabanmu!* ''Explain your answer!''. Our annotation process is binary, categorizing explanations as either True or False. We label an explanation as False if it is absent, contains hallucinations, or provides inaccurate information.[5]

[5] We use the Google search engine to verify the correctness of the explanation.
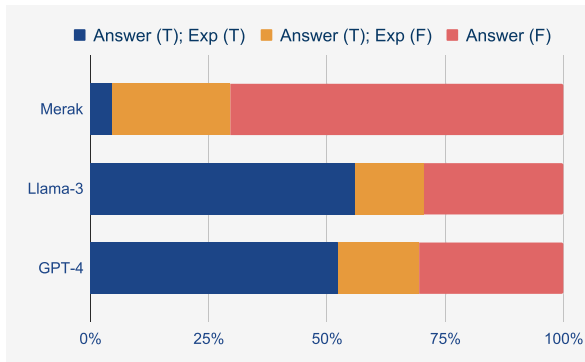
Figure 4: Performance comparison between Merak (7B), Llama–3 Instruct (70B), and GPT–4 based on text generation output. ``Answer (T)'' indicates that the generated answer is true, while ``Exp(F)'' denotes that the answer explanation is false.
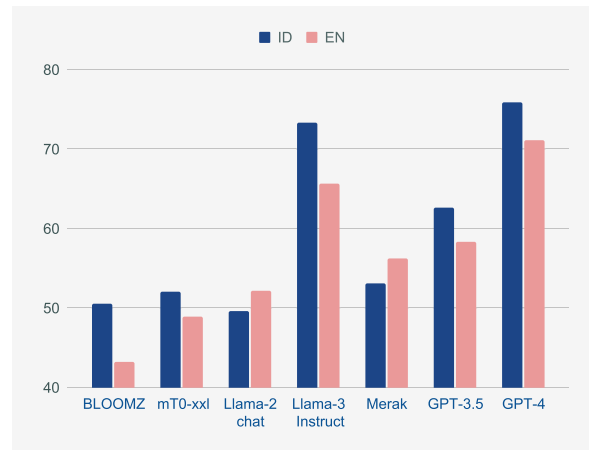

Figure 5: The accuracy of Indonesian and English translations across BLOOMZ (7B), mT0$_{xxl}$ (13B), Llama–2 chat (13B), Llama–3 Instruct (70B), Merak (7B), GPT–3.5, and GPT–4.

As anticipated, there is a substantial drop in accuracy for Merak (7B) from 53.2% (score reported in Table 3) to 29.5% (see Figure 4). This discrepancy underscores the limitations of relying solely on token probabilities to assess the true capability of a language model. Interestingly, only 4.5% of the samples are answered correctly with the appropriate explanation by Merak, despite it being the top performer among the Indonesian-centric language models. Larger models like Llama–3 and GPT–4 achieve more robust accuracies of 70.5% and 69.5% (shown in Figure 5), which are 3–5% lower than those indicated in Table 3. However, both models encounter challenges in generating appropriate explanations for correctly-answered samples, with Llama–3 and GPT–4 producing explanation errors in 29% and 34% of cases, respectively.

Does Language Affect Model Performance? We automatically translated `IndoCulture` to English using the Google Translate API[6] and used the English prompt in Figure 3 to evaluate the models. Specifically for this part, we include more models for comparison. All results over English text dropped, except for Llama–2 and Merak. This could be attributed to two reasons. First, Llama–2 is an English-centric model, and Merak is fine-tuned from Llama–2. Second, the performance drop for other models could be caused by translation errors. We further investigated this with 100 random samples and found that 81 samples had acceptable translations. We observed translation errors such as pronoun mismatches,

inaccurate proverb translations, and inaccurate translations of local terms, such as *pupuik* translated as ``fertilizer''.[7] To better understand the cultural gap in language models, we followed the approach of Liu et al. (2024) to manually correct the translations and reevaluate the models. We found that GPT–4's performance over the 100 random samples was 77.0 for the original Indonesian text, 68.0 for the English machine-translated text, and 72.0 for the English translation fixed by humans.

5 Discussion

A recent study (Wang et al., 2024) has demonstrated that evaluating language models using token probabilities in multiple-choice question types does not align well with the generated text. This discrepancy is reported to be more pronounced in models fine-tuned on conversational or safety data. In response to this issue, we conducted a manual evaluation of 200 random samples (in Section 4.3) and found that the performance in the generated text, especially for Merak—the best open-weight Indonesian-centric language model—deteriorates. However, this issue is less apparent in larger models such as Llama–3 and GPT–4. Conducting manual evaluation on all data and models is expensive, and we plan to address this issue in future works.

---

[6] Accessed in March 2024.

[7] *Pupuik* is a traditional musical instrument in West Sumatra. It is worth noting that the word *pupuik* closely resembles the word *pupuk* in Indonesian, which means fertilizer.

This work primarily focuses on introducing a novel dataset constructed for evaluating cultural commonsense reasoning within the Indonesian context, including preliminary evaluation results based on standard methods used in previous studies (OpenAI, 2023; Touvron et al., 2023; Koto et al., 2023; Li et al., 2024; Koto et al., 2024).

6 Conclusion

`IndoCulture` is a cultural commonsense reasoning dataset encompassing the diversity of Indonesian cultures, spanning from Aceh province in the west to Papua province in the east. Through collaboration with local individuals across eleven provinces and rigorous quality control measures, we introduce `IndoCulture` for the purpose of evaluating language models. Our findings reveal that large language models, whether Indonesian-centric or multilingual, demonstrate a limited understanding of Indonesian cultures. Notably, incorporating location as additional context significantly enhances model performance, particularly for GPT–3.5 and GPT–4.

Limitations

`IndoCulture` is specifically designed to explore the influence of geographical location on cultural commonsense reasoning, with a focus on the present time. It does not consider temporal aspects. Our dataset was created in the year 2023, and we recognize that cultures may evolve over time, as discussed by Mesoudi (2016).

Furthermore, as demonstrated in Section 4.3, a significant portion of `IndoCulture` comprises cultural elements such as artifacts, norms, and rituals. Symbols, values and beliefs, and language represent smaller proportions, ranging from 4% to 20%. We encourage future research to further explore these cultural elements and to expand the geographical coverage beyond the eleven provinces studied in this paper.

We also acknowledge that our dataset size is relatively small, to cover all the provinces. However, compared to existing Indonesian datasets that have been manually curated by natives (see COPAL–ID in Table 1), ours is significantly larger in terms of both size and regional coverage. Future work may extend the size and the coverage of `IndoCulture` to get a more holistic picture of Indonesian cultures.

Ethical Considerations

Acknowledgments

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.

Roger E. Axtell and Mike Fornwald. 1998. *Gestures: The do's and Taboos of Body Language Around the World*. Wiley.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on*

---

[8]https://creativecommons.org/licenses/by-nc-sa/4.0/.

Educational Advances in Artificial Intelligence, EAAI 2020, pages 7432–7439. AAAI Press. `https://doi.org/10.1609/aaai.v34i05.6239`

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898. `https://doi.org/10.18653/v1/2021.emnlp-main.699`

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D19-1109`

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.emnlp-main.941`

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Anthony Giddens and Philip W. Sutton. 2021. *Essential Concepts in Sociology*. John Wiley & Sons.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.482`

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.findings-acl.67`

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 (EMNLP)*. Singapore. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.emnlp-main.760`

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Cloze evaluation for deeper understanding of commonsense stories in Indonesian. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.csrr-1.2`

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting.

Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.findings-acl.334`

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11260–11285, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.findings-acl.671`

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.naacl-long.112`

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2023. LLM360: Towards fully transparent open-source LLMs. *arXiv preprint arXiv:2312.06550*.

John J. Macionis. 2012. *Sociology: Fourteenth Edition*. Pearson.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-main.90`

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.emnlp-main.821`

Alex Mesoudi. 2016. Cultural evolution: Integrating psychology, evolution and culture. *Current Opinion in Psychology*, 7:17–22. `https://doi.org/10.1016/j.copsyc.2015.07.001`

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation

for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1098

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.891

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917. https://doi.org/10.1145/3543507.3583535

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting*

of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 294–304, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-demos.28

OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into Indonesia's regional languages. *arXiv preprint arXiv:2403.09362*.

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. https://doi.org/10.18653/v1/2020.emnlp-main.185

Hadi Syah Putra, Rahmad Mahendra, and Fariz Darari. 2019. Budayakb: Extraction of cultural heritage entities from heterogeneous formats. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, WIMS* 2019, pages 6:1–6:9. ACM. https://doi.org/10.1145/3326467.3326487

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? Case study in Indonesian and Sundanese. *arXiv e-prints*, arXiv–2402.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui.

2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. 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*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106. `https://doi.org/10.1145/3474381`

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D19-1454`

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Vered Shwartz. 2022. Good night at 4 pm?! Time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853,

Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.findings-acl.224`

AI Singapore. 2023. SEA-LION (Southeast Asian languages in one network): A family of large language models for Southeast Asia. `https://github.com/aisingapore/sealion`

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied Linguistics*, 4(2):91–112. `https://doi.org/10.1093/applin/4.2.91`

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. ''My answer is C'': First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.findings-acl.441`

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.naacl-long.77`

Raymond Williams. 2014. *Keywords: A Vocabulary of Culture and Society*. Oxford University Press.

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. `https://doi.org/10.18653/v1/2023.eacl-main.57`

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-demos.6`

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-main.132`

Habib Zarbaliyev. 2017. Multiculturalism in globalization era: History and challenge for Indonesia. *Journal of Social Studies (JSS)*, 13(1):1–16. `https://doi.org/10.21831/jss.v13i1.16966`

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P19-1472`

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.acl-long.429`

A Model Details

Table 7 lists all model artifacts evaluated in our experiments.

Models (#parameters)	Source
BLOOMZ (560M)	bigscience/bloomz-560m
BLOOMZ (1.1B)	bigscience/bloomz-1b1
BLOOMZ (1.7B)	bigscience/bloomz-1b7
BLOOMZ (3B)	bigscience/bloomz-3b
BLOOMZ (7.1B)	bigscience/bloomz-7b1
mT0$_{small}$ (300M)	bigscience/mt0-small
mT0$_{base}$ (580M)	bigscience/mt0-base
mT0$_{large}$ (1.2B)	bigscience/mt0-large
mT0$_{xl}$ (3.7B)	bigscience/mt0-xl
mT0$_{xxl}$ (13B)	bigscience/mt0-xxl
Llama–2 (7B)	meta-llama/Llama-2-7b
Llama–2 chat (7B)	meta-llama/Llama-2-7b-chat
Llama–2 (13B)	meta-llama/Llama-2-13b
Llama–2 chat (13B)	meta-llama/Llama-2-13b-chat
Llama–3 (8B)	meta-llama/Meta-Llama-3-8B
Llama–3 Instruct (8B)	meta-llama/Meta-Llama-3-8B-Instruct
Llama–3 (70B)	meta-llama/Meta-Llama-3-70B
Llama–3-chat (70B)	meta-llama/Meta-Llama-3-70B-Instruct
Bactrian-X$_{LLaMa}$ (7B)	MBZUAI/bactrian-x-llama-7b-merged
Bactrian-X$_{LLaMa}$ (13B)	MBZUAI/bactrian-x-llama-13b-merged
IndoBART (132M)	indobenchmark/indobart-v2
IndoGPT (117M)	indobenchmark/indogpt
Merak (7B)	Ichsan2895/Merak-7B-v5-PROTOTYPE1
SeaLLM (7B)	SeaLLMs/SeaLLM-7B-v2
SEA-LION (7B)	aisingapore/sea-lion-7b
Komodo (7B)	Yellow-AI-NLP/komodo-7b-base

Table 7: With the exception of GPT–3.5 and GPT–4, all the models used in this study were sourced from HuggingFace (Wolf et al., 2020).