COPAL-ID: Indonesian Language Reasoning with Local Culture and Nuances

Haryo Akbarianto Wibowo[1], **Erland Hilman Fuadi**[2], **Made Nindyatama Nityasya**[2], **Radityo Eko Prasojo**[2], **Alham Fikri Aji**[1]
[1]MBZUAI [2]Independent Researcher
haryo.wibowo@mbzuai.ac.ae, {erland.hilman366,made.nindyatama,radityoeko}@gmail.com
alham.fikri@mbzuai.ac.ae

Abstract

We present COPAL-ID, a novel, public Indonesian language common sense reasoning dataset. Unlike the previous Indonesian COPA dataset (XCOPA-ID), COPAL-ID incorporates Indonesian local and cultural nuances, and therefore, provides a more natural portrayal of day-to-day causal reasoning within the Indonesian cultural sphere. Professionally written by natives from scratch, COPAL-ID is more fluent and free from awkward phrases, unlike the translated XCOPA-ID. In addition, we present COPAL-ID in both standard Indonesian and in Jakartan Indonesian–a dialect commonly used in daily conversation. COPAL-ID poses a greater challenge for existing open-sourced and closed state-of-the-art multilingual language models, yet is trivially easy for humans. Our findings suggest that general multilingual models struggle to perform well, achieving 66.91% accuracy on COPAL-ID. South-East Asian-specific models achieve slightly better performance of 73.88% accuracy. Yet, this number still falls short of near-perfect human performance. This shows that these language models are still way behind in comprehending the local nuances of Indonesian.

1 Introduction

A predominant challenge in multilingual NLP is to capture the sociolinguistic nuances and contexts that vary from culture to culture (Kabra et al., 2023; Hershcovich et al., 2022). This is especially important in localized language reasoning tasks where knowledge of local context and culture is crucial. For example, while the fact that a "chanting crowd" logically follows "Super Bowl" might be obvious within the US cultural sphere, the same cannot be said in Japan, where "Summer Koshien" is the more locally appropriate context for a "chanting crowd". As another example, while "wearing suits" follows "attending a wedding", within the Indonesian culture "wearing batik" is probably the more appropriate consequence.

Existing multilingual language reasoning datasets such as XNLI (Conneau et al., 2018) and XCOPA (Ponti et al., 2020) do not capture such local nuances because of two reasons. First, they are largely sanitized from localized and cultural elements. General, common-sense instances such as "water flows" following "opening faucet" are the ones typically found in the datasets. Second, any cultural element appearing in the dataset is typically based on US/Western context. Even when translated to other languages, it retains the original context while ignoring the cultural mismatch between the context and the common culture of the people speaking the target language. As a consequence, the current language reasoning benchmarks for multilingual models are lacking the crucial aspect of local and cultural context.

To provide a better benchmark for multilingual models that also capture local nuances for Indonesian, we introduce **COPAL-ID**.[1] It follows COPA's (Roemmele et al., 2011) commonsense causal reasoning format. COPAL-ID is handcrafted by native long-term Jakarta residents to capture Indonesian cultural and local nuances, especially in Jakarta. Specifically, we define three categories of locality: **Culture**, which captures local customs or norm; **Local Terminology**, terms commonly known by locals, yet not for outsiders; **Language**, which tests the nuance of the language, including uses of homonymy and non-compositionality. Each category contains data that can be considered uniquely Indonesian and can be understood as common customs or general knowledge by locals. Additionally, the dataset comes in pairs of two forms: standard Indonesian and colloquial Indonesian (Wibowo et al., 2021, 2020), with the latter being the go-to form in day-to-day contexts.

We find that the COPAL-ID dataset is trivially easy for native Jakartans, with our human scorers

[1]Data: https://huggingface.co/datasets/haryoaw/COPAL, Code: https://github.com/haryoa/COPAL-ID

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

Category	Premise	Correct Option	Incorrect Option	Note
T	Pria itu memperbaharui KK miliknya *(The man updated his KK)*	Ia baru saja menikah *(He just got married)*	Ia baru saja lulus kuliah *(He had just graduated from college)*	KK is a legal document that lists all the family members in a household.
L	Rumah tetangga saya baru saja dibobol maling *(My neighbor's house was just broken into by thieves)*	Dia cuma bisa gigit jari *(He can only bite his fingers)*	Dia menggigit jarinya *(He bit his finger)*	*gigit jari* is a figure of speech to express helplessness. The 2nd option is more literal.
T + C	Anak itu diterima masuk UI *(That kid was accepted into UI)*	Sekeluarga makan nasi kuning *(The whole family eats yellow rice)*	Sekeluarga makan nasi uduk *(The whole family eats uduk rice)*	UI is one of the top universities in Indonesia. *Nasi kuning* is often served for celebrations.

Table 1: COPAL-ID examples. T, L, and C denote Local Terminology, Language, and Culture respectively. Some samples may contain multiple categories at once.

achieving near-perfect accuracy. However, multilingual NLP models,[2] both fine-tuned and zero-shot prompted struggle, with many models showing performance close to random chances. In contrast, these models achieve better results in XCOPA-ID. This confirms that although the model might understand the Indonesian language, it struggles to comprehend the cultural aspect that comes with it.

Amongst the models, we find that ChatGPT and GPT-4 perform well, though it is difficult to conclude why given their proprietary nature. Nevertheless, just like the open models, the test on XCOPA-ID yields better results than on our COPAL-ID, denoting the cultural understanding gap.

2 Related Work

Multilingual Datasets. **XCOPA** (Ponti et al., 2020), an 11-language dataset translated from the English COPA (**C**hoice **o**f **P**lausible **A**lternatives) (Roemmele et al., 2011), is a commonsense reasoning (CSR) dataset. Each question in the dataset is composed of a premise and two causal alternatives, with the task being to choose the more plausible alternative with respect to the premise. **XStoryCloze** (Lin et al., 2022), a multilingual version of StoryCloze (Mostafazadeh et al., 2016), is another CSR dataset that introduces five-sentence stories capturing causal and temporal relations between everyday events in 10 languages. XCOPA and XStoryCloze each contain a version in Indonesian through translation, though without much care towards localized nuances and the more natural choices of phrases that locals commonly use. Others have built benchmark datasets in Indonesian from the ground up (Mahendra et al., 2021; Leong et al., 2023), and some have already focused on some cultural aspects, such as proverbs (Liu et al., 2023a; Kabra et al., 2023) and the cultural nuance within public school exams (Koto et al., 2023). COPAL-ID differs in its novel focus on cultural commonsense reasoning, which is challenging because it simultaneously tests the cultural knowledge and the logical inference capability of the models.

Cultural Aspect. Some previous work measured the inherent cultural knowledge of LLMs. Ramezani and Xu, 2023 analyzed whether language models understand cultural norms by evaluating them on two public datasets on morality. Meanwhile, Dwivedi et al., 2023 probed LLMs for their knowledge of etiquette norms in five regions. Lin et al., 2021 addressed the challenge of advancing commonsense reasoning beyond English, which is crucial for bridging the gap between different cultures and eliminating language barriers. Liu et al., 2021 analyzed and created a dataset consisting of image and caption pairs in five languages, including Indonesian, to introduce more languages and cultures to multimodal models. Similar to these studies, our dataset also enables measurement of LLMs' cultural knowledge, specifically on their cultural commonsense reasoning in Indonesian.

3 COPAL-ID

3.1 Data Language and Demography

The main purpose of COPAL-ID is to facilitate the benchmarking of NLP models based on their understanding and reasoning capabilities related to Indonesian local and cultural nuances. Relying on existing multilingual benchmarks (e.g., XCOPA (Ponti et al., 2020), XNLI (Conneau et al., 2018)) is inadequate as those data are just translations from English datasets.

Indonesia itself is rich in culture, consisting of many islands, provinces, and languages. Cultural

[2] Section 5 explains the models that we use for experiments.

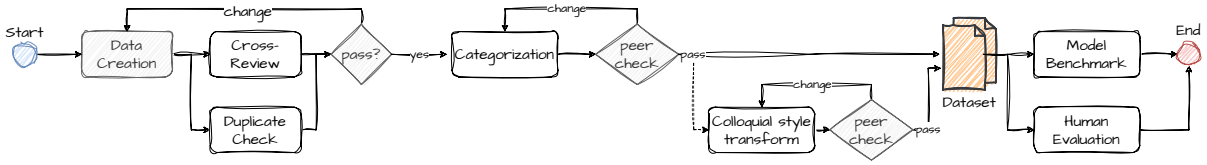

Figure 1: COPAL-ID creation and evaluation.

norms vary in different regions of Indonesia. Therefore, to limit the scope, we focus on the local and cultural aspects of Jakarta, the capital and the most populous city of Indonesia. While the standard Indonesian language is the official language, it's important to note that the Jakartan-Indonesian dialect (often referred to as colloquial Indonesian) is more prevalent in daily conversations in Jakarta than standard Indonesian. Hence, our dataset covers both standard Indonesian and Jakartan-Indonesian dialects, allowing us to evaluate NLP models' proficiency in understanding dialectal variations.

3.2 Task

We follow the COPA dataset, where we provide a premise and two plausible alternatives with one that is more likely to happen than the other.

3.3 Capturing Local Nuances

We break down local nuances into three categories: culture, local terminology, and languages. Every dataset entry should capture at least one category.

The culture category captures local customs or norms in Indonesia, especially for Jakartan. Local terminology captures long-tail terms or entities that are common and well-known for locals yet alien or long-tail for outsiders. This includes common local foods, animals, places, famous Indonesian public figures, ceremonies, common local abbreviations, and so on. Lastly, language category captures nuances in the language itself, for example, dealing with a figure of speech or word ambiguity. Some examples are shown in Table 1.

4 Data Creation

COPAL-ID is created through several steps. As a preliminary step, we formulate the end-to-end plan shown in Figure 1. In the data creation step, each of the five data creators (native and raised in Indonesia, accustomed to Jakarta culture) is tasked to create data following the definition in Section 3.3, i.e., to write the premise and both alternatives, in standard Indonesian. Each person is set a target of 110 data (but more is welcome). The resulting data then goes through a check-and-review process. This involves (1) an automated duplicate checker using TF-IDF vectors and (2) a double-blind cross-review process where each created data instance is assigned to two other creators for review. The assignment is distributed uniformly, meaning that there is no bias in the creator-reviewer pairings. The reviewers see the data with the alternatives shuffled, so they do not know the correct answer.

4.1 Cross Review

During cross-review, each reviewer performs two tasks for each data instance: (1) to pick the alternative that they deem to be more plausible and (2) to provide a qualitative analysis for the data instance. The analysis concerns several aspects.

- **Appropriateness**, whether the provided data falls within the definition of a cultural COPA described in Section 3.3. Common concerns here include non-cultural data and out-of-scope/obscure culture.
- **Difficulty**, or lack thereof: the improbable alternative is too obvious.
- **Correctness**, whether the logic, idea, or concept that is relied upon by the data is correct by the common cultural wisdom.
- **Ambiguity**, whether multiple common interpretations can lead to an ambiguity as to which alternative is more plausible.
- **Ethics**, whether the data contains sensitive or discriminating messages.
- **Clarity and format**, whether the provided data has issues with phrasing or spelling.
- **Duplicate**, whether the data reuses the same concept/idea as some other that the reviewer has come across. This is used to supplement the TF-IDF duplicate checker (Section 4.2).

Data that are answered correctly by both reviewers and have received no qualitative concerns are immediately passed. Meanwhile, data that (1) are answered incorrectly by at least one reviewer or (2) have received qualitative concerns are then decided via a discussion by all creators whether to be accepted, rephrased, or sent back for a change. Changed data go through the same process until at

Reasoning Category	#Sample	
	Cause	Effect
Terminology	186	181
Culture	136	146
Language	49	57
Total	279	280

Table 2: COPAL-ID statistic overview

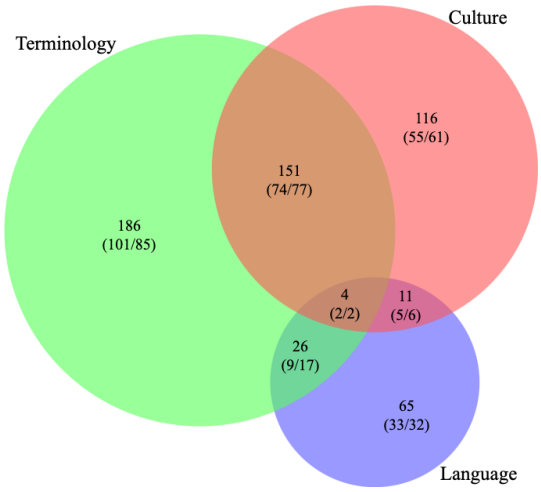

Figure 2: COPAL-ID statistic breakdown per-category. The number shows total for (Cause/Effect)

least 550 data are accepted. Statistics and examples of rejected data can be seen in Appendix B.

4.2 Duplicate Check

We perform a simple duplicate check on the dataset to have more variations of topics or concepts in the dataset. This duplicate checking was done semi-automatically. First, we automatically identified similar data and grouped them into a cluster. Then, we manually went through each group and checked whether we keep the data or replace it.

We use a TF-IDF algorithm to get a similarity score between each data. We use stopwords to reduce similarity caused by common words and use both unigram and bigram for the evaluated tokens. We eliminate dissimilar pairs by setting a threshold and then get the final pair pools.

For the grouping, we first pick a random pair from the pair pools. Then, we check in the pair pools if one of the data has another related pair and then pick it. This process will be done iteratively until we cannot find other associated data in the pair pools. All the related data will be merged into the same group to be evaluated together. From this process, we got 71 groups where each group contains data ranging from 2 to 5.

The final process is doing a manual check for each group to decide whether to accept or reject the data. For every rejected data, we ask the creator to replace it with a new one. An important thing to note is that for every group, it is possible to accept all the data even though it mentions the same topic, as long as the context is different. Some examples of these can be seen in Appendix B, Table 8.

4.3 Categorization

We categorize our data to local nuance according to three categories described in Section 3.3. First, the original data creator is asked to annotate the categories of their own data. Then, a different reviewer is asked to validate the label as a peer check. This second reviewer is requested to raise any category label that they deem incorrect and resolve it with the original data creator.

We note that category labeling is very ambiguous for some data. Therefore, one final check is done through a third reviewer. This last reviewer is asked to validate the category label across all COPAL-ID to ensure consistency. Table 2 and Figure 2 show the statistics of COPAL-ID categories.

4.4 Paraphrase to Colloquial Indonesian

We paraphrase all the datasets into colloquial/Jakartan-dialect. The paraphrase is done by the original data creator while making sure that the meaning of the premise and both alternatives are kept, therefore preserving both the plausibility label and the nuance categorization.

Then, a peer check review is executed to confirm that the newly constructed colloquial dataset still maintains the same meaning, while also making sure that the colloquial text is natural (i.e., commonly used by Jakartan). The data creator will change data entries that do not pass the requirement by the peer reviewer until both are in agreement.

To measure lexical similarity, we compute the BLEU score between the final colloquial dataset and the original standard Indonesian dataset. We obtain a low BLEU score of 3.98, which indicates a high lexical distinction between the two datasets while still keeping the semantics preserved.

4.5 Human Evaluation

Lastly, we perform a human evaluation on both standard Indonesian and colloquial Indonesian of COPAL-ID. This evaluation is performed by a completely different group of annotators. We provide the annotators with the premise and two alternatives and ask them to pick the more plausible alter-

Scenario		Mono (Std.)	Mono (Colloq.)	Cross	Translated
Finetune	Training set	ID GEN-X	ID GEN-X	EN COPA	EN COPA
	Validation set	XCOPA-ID	XCOPA-ID	EN COPA	EN COPA
	Test set	COPAL-ID	COPAL-ID-C	COPAL-ID	COPAL-ID-T
In-context	5-shot examples	COPAL-ID	COPAL-ID-C	EN COPA	EN COPA
	Test set	COPAL-ID	COPAL-ID-C	COPAL-ID	COPAL-ID-T

Table 3: Data used in finetuning and in-context learning.

native. Both standard and colloquial Indonesian of COPAL-ID are annotated by two annotators independently. In both datasets, we achieve consistently high accuracy of approximately 95%, therefore confirming that our dataset is trivial for humans accustomed to Jakartan culture.

5 Experiment Setup

Our data evaluation involves three different setups: monolingual, cross-lingual, translate-test, and colloquial test. Each setup is tailored to a specific scenario. For **monolingual** setup, we use the same language (Indonesian and Colloquial Indonesian) for training and testing. **Zero-shot cross-lingual** utilizes the English COPA dataset as the few-shot examples or training data for zero-shot classification and uses the COPAL-ID dataset for testing. In contrast, the **translate-test** employs an English-translated version of the COPAL-ID for testing instead. We use Seamless-M4T (Seamless Communication, 2023) for translation.

The evaluation under the monolingual setup is performed twice, one using COPAL-ID with standard Indonesian and the other with the colloquial dataset instead. We forego performing the colloquial testing on the two other setups, noting that the high lexical distinction (Section 4.4) would hamper the translation quality on the translate-test setup and that prior work has noted big degradation in cross-lingual performances of Indonesian local languages even for those with high lexical similarity with standard Indonesian (Winata et al., 2023).

Detailed data and model setup can be seen in Table 3 and will be elaborated further in this section.

5.1 Finetuning

We select four pre-trained models to finetune: MBERT (Libovický et al., 2019), IndoBERT (Koto et al., 2020), XLM-R Base, and XLM-R Large (Conneau et al., 2020). MBERT (Devlin et al., 2019), XLM-R Base, and XLM-R Large are multilingual models that can handle various languages, including Indonesian language, while IndoBERT is a model that is specifically fine-tuned for Indonesian Language. In this work, we fine-tune our models in monolingual, zero-shot cross-lingual, and translate-test defined above. Due to the unavailability of the training set in XCOPA-ID and COPAL-ID, we use Indonesian COPA data from GEN-X (Whitehouse et al., 2023), a multilingual augmented commonsense reasoning dataset produced from GPT-4, instead. Additionally, We use XCOPA-ID as the validation set. Meanwhile, we use English COPA dataset as training and validation set for both the zero-shot cross-lingual and translate-test, and COPAL-ID and COPAL-ID-T as test sets for each scenario, respectively. For each data point, we use two templates to represent the input, which can be found in Appendix A.

The models are trained with Huggingface's `transformers` (Wolf et al., 2019). We do grid-search hyperparameter tuning with batch size of {8, 16, 32} and learning rate of $\{5e^{-6}, 10e^{-6}\}$. We used a weight decay of 0.01 while the rest of the parameters were set to their default values. We pick the best hyperparameter based on model performance on the validation set. We finetuned each model for 5000 steps, with an early stopping patience of 500. This was done on a 24 GB GPU and completed in under 20 hours in total.

5.2 Prompting with Language Models

We test our dataset with in-context learning, which tests the data directly without explicit fine-tuning (Brown et al., 2020), in zero-shot and few-shot settings. For the few-shot setting, for each scenario defined above, we follow the experiment setup defined in MEGA (Ahuja et al., 2023). We benchmark BLOOMZ (Muennighoff et al., 2022), Bactrian-X (Li et al., 2023), Llama-2 (Touvron et al., 2023), and PolyLM (Wei et al., 2023) to represent the open-source (multilingual) prompting models. We also include SeaLLM (Nguyen et al., 2023) and Sailor (Dou et al., 2024), as recent models designed specifically for South-East Asia

Test Data	XCOPA-ID	Monolingual COPAL-ID (standard)	Monolingual COPAL-ID (colloquial)	Translate-test COPAL-ID (translated)	Cross-lingual COPAL-ID (standard)
Finetuned models					
XLMR-Base	66.80	55.99	57.42	52.24	53.67
XLMR-Large	**79.40**	64.22	62.97	52.06	52.95
mBERT	59.60	55.64	56.35	56.53	55.10
IndoBERT	67.60	61.00	60.64	-	-
Open models prompting					
Bactrian-X-7B (5-shot)	71.20	63.51	59.39	56.35	63.69
Llama-2-7B (5-shot)	64.20	57.96	54.29	55.81	58.86
BLOOMZ-7B (5-shot)	76.20	66.91	58.68	56.53	65.65
PolyLM-13B (5-shot)	71.20	63.15	58.50	56.89	63.33
Regional open models prompting					
SeaLLM-7B (5-shot)	77.80	73.21	64.11	59.11	68.39
Sailor-7B (5-shot)	76.80	**73.88**	**66.07**	**59.57**	**73.93**
Closed models prompting					
ChatGPT (5-shot)	90.80	76.74	76.57	-	-
GPT-4 (5-shot)	97.20	92.13	91.06	-	-
Human	-	95.00	95.62	-	-

Table 4: Comparison of accuracy score of finetuned, prompting with open and closed models. **Bold** indicates best results for open models in each column. For IndoBERT, we did not train on COPA-EN as it is inherently not a multilingual model. The score results of **open models prompting** are the maximum score among different prompt templates and scenarios (few-shots vs zero-shots).

(SEA) regions. Lastly, we also include ChatGPT[3] and GPT-4 (OpenAI, 2023)[4] for the proprietary or closed prompting models.

We use five examples for the few-shots scenario with respect to the test data type. In monolingual settings (COPAL-ID and COPAL-ID-C), we use five new in-context examples that we create outside of data produced from section 4. We use the English COPA dataset for both cross-lingual and test-translate and test them to COPAL-ID and COPAL-ID-T, respectively. The setup is in Table 3.

To prompt, we benchmark multiple templates since it is widely known that the performance of an LLM depends on its prompt (Liu et al., 2023b). We use templates from (Ahuja et al., 2023) (MEGA), BLOOMZ, and LM-Harness.[5] To check the effect of choosing the language in prompting, we used Indonesian and English templates for each template.[6] After that, we predict the chosen class by computing the logits by applying the template to each choice, comparing each logits, and taking the highest one as the chosen choice. For closed-source models, we use an instruction to make these models output in `<ANSWER>` format. Then, we extract the predicted answer and match it with the gold label (either 1 or 2, which represents the choice that it predicts). Indecisiveness to pick an answer is classified as incorrect instead. The prompting templates can be found in Appendix A. Prompting on open model is approximately 5 minutes per prompt template in a single A100 GPU.

6 Experiment Results

Our experiment results can be seen in Table 4. By comparing XCOPA-ID and COPAL-ID , it is clear that the latter's accuracy is consistently lower across different models. The best performing open model on XCOPA-ID and COPAL-ID, XLMR-Large and BLOOMZ-7B (respectively), have a performance drop of 15% and 9% (respectively). ChatGPT and GPT-4 also have a performance drop of 14% and 5% from XCOPA-ID to COPAL-ID. Smaller performance degradation is seen in SEA-focused models such as SeaLLM and Sailor, signifying that these models have a smaller 'knowledge gap' between general reasoning and locally nuanced reasoning. Yet, Based on this evidence, COPAL-ID can be considered more challenging than XCOPA-ID, which does not incorporate local nuance.

Among open-source models, XLMR-Large exhibits the best performance among the finetuned models, surpassing IndoBERT, which is pre-trained using the Indonesian dataset. On the other hand, for in-context learning, SEA-specific models gen-

[3] https://openai.com/blog/chatgpt

[4] Both GPT models were accessed in Sep 2023, 4th week

[5] github.com/EleutherAI/lm-evaluation-harness

[6] Template is manually translated to ID if it is only in EN.

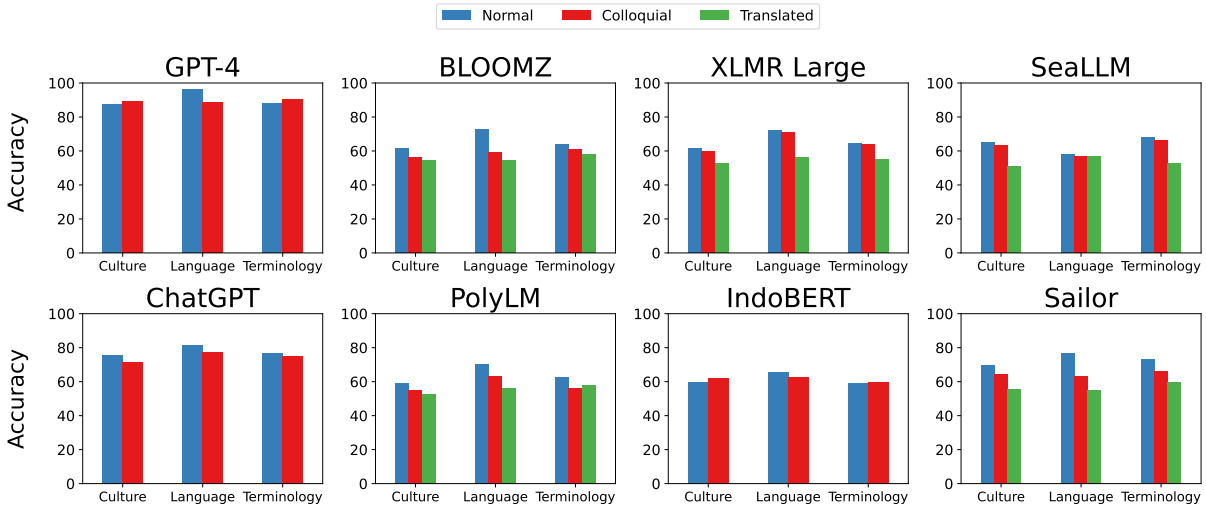

Figure 3: Models' accuracy score based on Culture, Language, and Terminology category.

erally achieve the best performance. For general models, BLOOMZ outperforms others, followed by PolyLM. PolyLM does not surpass BLOOMZ despite having more parameters, showing that it takes other factors, not just a bigger size, for this task. Regarding the closed-source models, GPT-4 outperforms ChatGPT by a significant margin yet cannot beat human performance.

Does colloquial Indonesian pose an additional difficulty? COPAL-ID-C tests the impact of colloquial data on models' performance. It is evident that human performance has comparable results for both colloquial and non-colloquial forms of COPAL-ID, unaffected by it. On the other hand, all of the open models used for prompting performance dropped by about 3-9 percent. Surprisingly, the differences are negligible for each fine-tuned model, demonstrating that finetuning them with the GEN-X dataset is quite robust to colloquial texts. This also indicates a lack of representation of colloquial Indonesian within the pretraining data of these open models. On the other hand, both ChatGPT and GPT-4 models exhibit minimal degradation when evaluated under colloquial Indonesian. However, due to their proprietary nature, it is impossible to pinpoint the reason.

For COPAL-ID, does using zero-shot cross-lingual instead of a translate-test improve the score? Performing translate-test on our dataset tends to degrade the model's performance notably. For instance, XLMR-Large exhibits a significant decline from its standard one by approximately 12%. We hypothesize that this drop is due to translation errors caused by long-tail words or local terminologies. In this scenario, we recommend using a local-nuance-aware machine translation model to obtain optimal performance, though admittedly such a model (or an open parallel data to build it) currently does not exist.

For finetuned models, the cross-lingual approach (train on EN COPA, then evaluate on COPAL-ID) yields comparably bad performance similar to the translate-test, aside from Sailor. Therefore in this scenario, English data is not really helpful, whether we use it for cross-lingual or translate tests. Interestingly, cross-lingual prompting is visibly better than translate-test, noting that language models can perform in-context learning from a different language.

Do the in-context learning setting and prompt template matter for COPAL-ID? In Table 5, it is evident that all models benefit from the few-shot scenario on COPAL-ID.[7] Moreover, comparing six different templates demonstrates that the best results are yielded with the LM-Harness template in either English or Indonesian languages with comparable scores. This suggests that few-shot scenarios contribute to improving the model performance and underline the requirement of a suitable prompt template for a model to achieve better. As a result, this indicates that those models are sensitive to the choice of prompt template.

How is the models' performance across different categories? Based on Figure 3, the Language category of COPAL-ID achieves the best accuracy across all models, while the Culture category performs the worst. This underscores the challenges posed in this category. The Terminology category

[7]Also applies on XCOPA with one exception: BLOOMZ, where 0-shot yields best results. (Appendix D, Table 14).

Model	Shots	Prompt Template							
		M-ID	M-EN	BL-ID	BL-EN	LH-ID	LH-EN	LLH-ID	LLH-EN
Bactrian-X	5	51.16	52.24	50.45	50.27	61.90	62.43	**63.51**	61.72
	0	49.19	48.84	46.51	48.84	62.08	61.18	**63.51**	60.11
LLAMA2	5	51.16	50.63	52.59	52.24	**57.96**	**57.96**	56.35	57.42
	0	49.02	48.30	50.09	51.52	49.19	53.85	57.60	56.53
BLOOMZ	5	54.74	55.10	60.82	57.25	65.47	65.12	65.65	**66.91**
	0	56.89	58.68	61.90	63.15	62.79	63.86	64.22	66.55
PolyLM (13B)	5	54.03	55.10	54.20	53.31	62.08	62.79	62.43	**63.15**
	0	52.77	50.45	52.06	51.16	61.90	60.47	60.64	**63.15**
SeaLLM	5	69.29	67.32	70.89	**73.21**	64.82	64.82	65.83	62.25
	0	58.04	51.96	56.61	60.54	61.79	61.96	61.54	59.21
Sailor	5	62.68	66.43	62.50	62.86	73.57	73.57	**73.88**	**73.88**
	0	51.61	52.14	52.86	51.96	73.57	72.68	72.81	69.95

Table 5: Indonesian monolingual scenario comparison for every template defined with the addition of juxtaposition between 5-shots and 0-shots. **Bold** indicates the best results for each model. **M**=MEGA, **BL**=BLOOMZ, **LH**=LM Harness, **LLH**=Local LM Harness, while **-ID** and **-EN** refer to the respective languages.

has a slightly higher overall score than the Culture category. We posit that since the dataset incorporates local nuances that require prior knowledge to answer, it might either conflict with other regions' cultural knowledge or be nonexistent within the embedded knowledge of the pre-training models.

Using colloquial test data also tends to hinder the model performance, especially in the Language category, where the open-source prompting model is more affected. In contrast, fine-tuned models are relatively robust to colloquial text, with some categories demonstrating comparable scores between the original and the colloquial.

Would Explicitly Instructing the Models to Reason from a Local Point-of-View Help? We ran our experiment using previously existing prompts. However, these prompts are in English, and none are explicitly designed for Indonesian causal reasoning. The closest indication for these LLMs to reason based on local aspects is when we translate the prompt into Indonesian.

In Table 6, we observe that when we explicitly prompt the model to reason from the point of view of an "Indonesian accustomed to Jakartan culture", there is a slight increase in overall performance. However, besides GPT-4, the improvement is minimal, and generally, the models still perform poorly, highlighting the challenge of Indonesian cultural reasoning. Prompt details and more comprehensive results are shown in Appendices A and D.

Prompt Template	Evaluation			
	Standard		Colloquial	
	0-shot	5-shot	0-shot	5-shot
LM Harness ID	58.39	60.64	55.89	56.99
Local LM Harness ID	60.23	60.59	54.96	56.48
LM Harness EN	59.03	60.82	55.56	56.31
Local LM Harness EN	60.06	60.67	54.60	56.30
ChatGPT	67.26	76.74	62.25	74.42
Local ChatGPT	69.94	76.74	62.79	76.57
GPT-4	83.00	89.80	82.28	89.80
Local GPT-4	86.58	92.13	85.86	91.06

Table 6: Model performance across prompts with and without local indicator. LM Harness's results are each an average of open model performances (Bactrian-X, Llama-2, BLOOMZ, PolyLM, SeaLLM, Sailor)

7 Qualitative Analysis

Upon analyzing the output of the models, we discover some interesting findings where some instances are predicted correctly by at most one model yet answered correctly by all humans. These are displayed in Appendix C, Table 9.

For *nasi uduk* and *nasi kucing* (lit: cat rice), both are dish names with a cultural context where one is often consumed as breakfast while the latter is for dinner. The second example is more intricate. The term *jaga lilin* (translated: keeping the candle), a common black magic practice widely known in Indonesia, requires the model to possess the right cultural knowledge. Other examples show some activities that are not commonly practiced outside Indonesia, such as kissing our parents' hands.

Although the LLMs are trained on multilingual data, some long-tail terms and local nuances are bound to be missed. Without imposing their context explicitly on the models, it would remain challenging for them to infer the correct reasoning.

8 Conclusion

We release COPAL-ID to benchmark common sense reasoning with local nuance in Indonesian that consists of culture, terminology, and language in COPA format. We maintain our data quality by having several cross-reviews and automatic duplicate checks. We then benchmark COPAL-ID by evaluating it through in-context learning and finetuning. Our experiment shows that COPAL-ID proves to be challenging for current open-source models across different experiment setups, yet easy for natives. We also provide additional insight that in-context learning is sensitive to the template, and using few-shot examples helps improve the accuracy of the models.

We believe that each region has a different culture, which results in different common sense on some specific things, such as norms and values. Hence, we hope that publishing this dataset encourages the NLP community to build new datasets and models that incorporate diverse local nuances, including Indonesia. We leave out how to build a local nuance-aware model for future work.

9 Limitations

Our data is categorized into three categories, which is not granular enough, making more detailed analysis not possible. Nevertheless, making thoroughly fine-grained categorization is challenging not only in itself but also because we need to increase the size of the dataset as a consequence. After all, the resulting categories would only be useful if an enough number of data instances fall into each of them and therefore are sufficiently statistically representative for analysis purposes.

Additionally, we scope the region only for Jakartan, even though Indonesia, the origin of the Indonesian language, has multiple regions with diverse cultures. Although Jakarta has multi-ethnic citizens and is arguably portrayed the most in mass and social media, not all culture or local nuances are present in Jakarta. Therefore, COPAL-ID has not captured common local nuances in all different Indonesian regions yet. We leave scaling up our dataset to other Indonesia's regions for future work.

Ethical Considerations

COPAL-ID has been carefully crafted and reviewed to avoid sensitive and discriminating content. The annotators who are hired for human evaluation are fairly compensated above the minimum wage stipulated by the law in Jakarta.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

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.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2310.04928*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models.

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.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How Language-Neutral is Multilingual BERT?

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning.

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 Language Models.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories.

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.

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.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. 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*.

Yu-An Chung Mariano Coria Meglioli David Dale Ning Dong Mark Duppenthaler Paul-Ambroise Duquenne Brian Ellis Hady Elsahar Justin Haaheim John Hoffman Min-Jae Hwang Hirofumi Inaguma Christopher Klaiber Ilia Kulikov Pengwei Li Daniel Licht Jean Maillard Ruslan Mavlyutov Alice Rakotoarison Kaushik Ram Sadagopan Abinesh Ramakrishnan Tuan Tran Guillaume Wenzek Yilin Yang Ethan Ye Ivan Evtimov Pierre Fernandez Cynthia Gao Prangthip Hansanti Elahe Kalbassi Amanda Kallet Artyom Kozhevnikov Gabriel Mejia Robin San Roman Christophe Touret Corinne Wong Carleigh Wood Bokai Yu Pierre Andrews Can Balioglu Peng-Jen Chen Marta R. Costa-jussà Maha Elbayad Hongyu Gong Francisco Guzmán Kevin Heffernan Somya Jain Justine Kao Ann Lee Xutai

Ma Alex Mourachko Benjamin Peloquin Juan Pino Sravya Popuri Christophe Ropers Safiyyah Saleem Holger Schwenk Anna Sun Paden Tomasello Changhan Wang Jeff Wang Skyler Wang Mary Williamson Seamless Communication, Loïc Barrault. 2023. Seamless: Multilingual expressive and streaming speech translation.

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

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered Data Augmentation for Enhanced Crosslingual Performance.

Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasojo, and Derry Tanti Wijaya. 2021. IndoCollex: A Testbed for Morphological Transformation of Indonesian Colloquial Words. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3170–3183, Online. Association for Computational Linguistics.

Haryo Akbarianto Wibowo, Tatag Aziz Prawiro, Muhammad Ihsan, Alham Fikri Aji, Radityo Eko Prasojo, Rahmad Mahendra, and Suci Fitriany. 2020. Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation.

Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, and Pascale Fung. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Prompt Template

This section provides all the templates that we use for prompting and fine-tuning. We first explain the individual variables in each template[8] as follows.

- `{p}`: the premise of the data,
- `{c1}`: the first choice of the data,
- `{c2}`: the second choice of the data,
- `{r}`: the text of the correct choice,
- `{rl}`: lowercased `{r}`

`{c1}` and `{c2}` end with a dot (.). No case change is applied unless indicated otherwise.

A.1 MEGA-EN

Prompt (Cause): `{p}`. This happened because...\nHelp me pick the more plausible option: - choice1: `{c1}`, choice2: `{c2}\n\n{r}`

Prompt (Effect): `{p}`. As a consequence...\nHelp me pick the more plausible option: - choice1: `{c1}`, choice2: `{c2}\n\n{r}`

A.2 MEGA-ID

Prompt (Cause): `{p}`. Ini terjadi karena...\nBantu saya memilih opsi yang paling mungkin: - opsi1: `{c1}`, opsi2: `{c2}\n\n{r}`

Prompt (Effect): `{p}`. Konsekuensinya...\nBantu saya memilih opsi yang paling mungkin: - opsi1: `{c1}`, opsi2: `{c2}\n\n{r}`

A.3 Bloomz-EN

Prompt (Cause): `{p}`.\n\nselect the most plausible cause:\n - `{c1}`\n - `{c2}`\n\n`{r}`

Prompt (Effect): `{p}`.\n\nselect the most plausible effect:\n - `{c1}`\n - `{c2}`\n\n`{r}`

A.4 Bloomz-ID

Prompt (Cause): `{p}`.\n\npilih penyebab yang paling mungkin:\n - `{c1}`\n - `{c2}`\n\n`{r}`

Prompt (Effect): `{p}`.\n\npilih efek yang paling mungkin:\n - `{c1}`\n - `{c2}`\n\n`{r}`

A.5 Fine-tuning

Prompt (Cause): `{p}`.What was the cause?

Prompt (Effect): `{p}`.What happened as a result?

A.6 LM Harness-EN

Prompt (Cause): `{p}` because `{rl}`

Prompt (Effect): `{p}` therefore `{rl}`

A.7 LM Harness-ID

Prompt (Cause): `{p}` karena `{rl}`

Prompt (Effect): `{p}` maka `{rl}`

A.8 Local LM Harness-EN

Instruction : Please answer the following question about commonsense causal reasoning from the perspective of someone accustomed to Jakartan culture in Indonesia.

Prompt (Cause): `{p}` because `{rl}`

Prompt (Effect): `{p}` therefore `{rl}`

A.9 Local LM Harness-ID

Instruction : Jawablah pertanyaan berikut mengenai penalaran umum sebab akibat dari sudut pandang seseorang yang terbiasa dengan budaya Jakarta di Indonesia.

Prompt (Cause): `{p}` karena `{rl}`

Prompt (Effect): `{p}` maka `{rl}`

A.10 ChatGPT and GPT-4

Instruction: You are an AI assistant whose purpose is to perform open-domain commonsense causal reasoning. You will be provided a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Answer as concisely as possible in the same format as the examples below:[9]

Prompt (Cause): `{p}`.\nThis happened because...\nHelp me pick the more plausible option and give me the option index without the text between angle brackets at the end of your answer like this <index>. Be concise. No talk; just go:\n - `{c1}`\n- `{c2}`\n\n### Response:

Prompt (Effect): `{p}`.\nAs a consequence...\nHelp me pick the more plausible option and give me the option index without the text between angle brackets at the end of your answer like this <index>. Be concise. No talk; just go:\n - `{c1}`\n- `{c2}`\n\n### Response:

[8] Please note that for the fine-tuning, we treat it as a classification task. Therefore, there is no response in the template.

[9] For zero-shot, this last sentence in the instruction is simply "Answer as concisely as possible."

A.11 Local ChatGPT and GPT-4

Instruction: Please answer the following questions about commonsense causal reasoning from the perspective of someone accustomed to Jakartan culture in Indonesia. You will be provided a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Answer as concisely as possible in the same format as the examples below:[10]

Prompt (Cause): `{p}.\nThis happened because...\nHelp me pick the more plausible option and give me the option index without the text between angle brackets at the end of your answer like this <index>. Be concise. No talk; just go:\n - {c1}\n- {c2}\n\n### Response:`

Prompt (Effect): `{p}.\nAs a consequence...\nHelp me pick the more plausible option and give me the option index without the text between angle brackets at the end of your answer like this <index>. Be concise. No talk; just go:\n - {c1}\n- {c2}\n\n### Response:`

[10] See Footnote 9.

B Data Creation & Review

Tables 7 and 8 show examples of review decisions. Figure 4 shows statistics of the initial review.

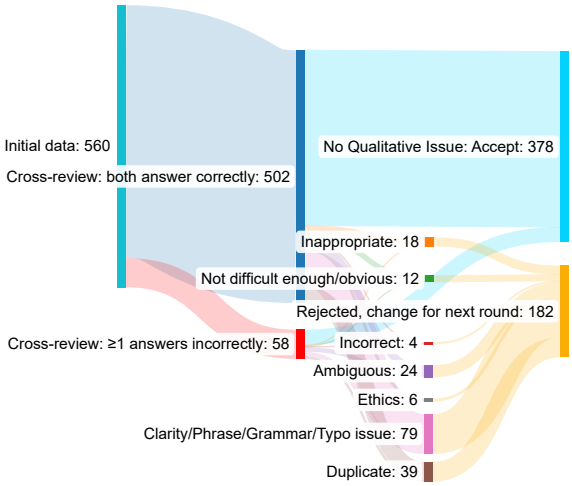

Figure 4: Initial data creation statistics

Premise	Correct Option	Incorrect Option	Verdict	Note
Ia ingin memberbaharui SIM-C miliknya *(He/she wants to renew his/her SIM-C)*	Ia datang ke kantor polisi *(He/she goes to police station)*	Ia datang ke toko ponsel *(He/she goes to phone shop)*	U	**Correctness**. We need to go to SAMSAT (not the police station) to obtain SIM-C (a driving license)
Hari ini ada gangguan sinyal *(Today there is a signalling failure)*	Perjalanan kereta terlambat *(Trains are delayed)*	Perjalanan pesawat terlambat *(Flights are delayed)*	R	**Appropriateness**. Signalling failures are not uniquely cultural to Indonesia.
Dia hendak beribadah pada hari Minggu *(He/she is going to pray on Sunday)*	Dia pergi ke gereja *(He/she goes to the church)*	Dia pergi ke wihara *(He/she goes to the vihara)*	R	**Ambiguity**. Churches and Viharas both provide religious services on Sunday.
Saat lebaran, suasana kota Jakarta sangat sepi *(During Eid, Jakarta becomes quiet.)*	Masyarakatnya sedang pulang kampung *(The citizens are going to their hometown)*	Ada pawai di Jakarta *(There is a parade in Jakarta)*	R	**Difficulty**. Parade is never quiet, making this data too easy.
Ayah tidak kunjung tiba karena pesawatnya terlambat berjam-jam *(Father has not arrived yet because his airline is delayed for hours)*	Ayah naik <Maskapai A> *(Father flies with <Airline A>)*	Ayah naik <Maskapai B>) *(Father flies with <Airline B>)*	R	**Ethics**. The original data directly mentions certain airline brands, perpetuating certain stereotypes about them.
Anak itu ketakutan *(That child is terrified)*	Ia mimpi melihat pocong *(He/she dreams of pocong)*	Ia mimpi melihat porli *(He/she dreams of porli)*	U	**Clarity**. Pocong is a local ghost, while porli is a typo of Polri (abbreviation of Indonesian national police)

Table 7: Candidate of COPAL-ID examples from the first iteration that are rejected or need revision. U means "needs small update/rephrase", while R means rejected and needs to be replaced with new data.

Premise	Correct Option	Incorrect Option	Verdict	Note
Saya mau berangkat sekolah *(I am going to school)*	Saya mencium tangan orang tua *(I kiss my parent's hand)*	Saya menunggu bis sekolah *(I wait for the school bus)*	A	These two data have the same topic of 'kissing parent's hand' and both phrasings are almost identical to each other. In this case, we only accept one of them.
Anak itu akan berangkat sekolah *(That child is going to school)*	Anak itu cium tangan orang tuanya *(That child kisses his/her parent's hand)*	Anak itu cium dahi orang tuanya *(That child kisses his/her parent's forehead)*	R	
Ada kasus demam berdarah di komplek itu *(There's a dengue fever case in that area)*	Mereka mengadakan penyemprotan *(They spray the area)*	Mereka memanggil detektif *(They call a detective)*	A	Although these two data mention the same topic of "dengue fever", the context on how the topic is used is quite different. Thus, we accept both data.
Di musim penghujan, kampungku mengadakan fogging *(In rainy season, my village conducts a fogging)*	Pemerintah ingin membasmi demam berdarah *(The government wants to prevent dengue fever)*	Pemerintah ingin membasmi malaria *(The government wants to prevent malaria)*	A	

Table 8: Candidate of COPAL-ID examples that are marked as duplicate or similar to each other. A means accepted, R means rejected and needs to be replaced with new data.

Category	Premise	Correct Option	Incorrect Option	Note
T+C	Malam-malam begini saya kelaparan *(I feel hungry at night)*	Saya keluar berburu nasi kucing angkringan *(I went out to buy nasi kucing angkringan)*	Saya keluar berburu nasi uduk warteg *(I went out to buy nasi uduk warteg)*	nasi uduk and nasi kucing are dish names, but the former is usually consumed for breakfast while the latter is for late dinner.
C+L	Kemarin malam, ia baru selesai jaga lilin *(Yesterday night, he/she just finished jaga lilin)*	Ia percaya dengan ilmu hitam *(He/she believed in black magic)*	Ia adalah orang yang taat beribadah *(He/she is a religious person)*	jaga lilin (keeping the candle alight) is one of the tasks performed in a black magic practice/rituals to get rich.
L	Kakek suka sambil menyelam minum air *(Grandfather likes to drink water while diving)*	Pekerjaan kakek seringkali cepat selesai semuanya *(Grandfather often finishes his work quickly)*	Kakek sering tersedak air *(Grandfather often chokes on water)*	sambil menyelam minum air is a popular idiom that expresses multitasking ability
C	Adik yang masih TK akan berangkat ke sekolah *(Little brother/sister is going to school)*	Adik mencium tangan orang tua *(Little brother/sister kisses his/her parent's hand)*	Adik menunggu bis sekolah *(Little brother/sister waits for the school bus)*	Kissing the back of parent's hand is a common culture to show respect and love. School buses are rare in Indonesia.

Table 9: COPAL-ID examples that is hard for model but trivial for human. T, C, and L depict Terminology, Culture, and Language, respectively.

C Qualitative Analysis

Table 9 provides examples of qualitative result analysis that showcase how our tested models struggle to understand the cultural nuance in COPAL-ID.

D Detailed Experiment Results

Tables 10, 11, 13, and 14 provide prompting results for cross monolingual & translate-test, colloquial, cross-lingual, and XCOPA respectively.

E Answer Rejection From Closed LMs

Table 12 shows the number of instances for which closed-source LMs reject to answer.

Normal	Colloquial
15	11

Table 12: Number of instances GPT-4 refused to answer. ChatGPT answered all questions.

Model	Shots	Translate-test							
		M-ID	M	BL-ID	BL	LH-ID	LH-EN	LLH-ID	LLH-EN
Bactrian-X	5-shots	50.81	50.98	47.23	47.94	54.38	53.85	56.35	55.09
	0-shot	50.63	47.94	50.81	49.37	54.92	53.31	55.99	54.74
Llama-2	5-shots	50.27	53.49	50.98	52.59	55.81	54.03	55.81	54.91
	0-shot	50.81	49.37	49.19	51.34	55.46	54.20	56.35	55.46
Llama-2-Chat	5-shots	53.31	54.56	50.81	52.06	55.28	54.20	54.92	55.10
	0-shot	50.45	48.66	52.95	50.98	53.31	54.03	55.10	55.64
BLOOMZ	5-shots	53.31	53.67	54.38	53.85	55.81	54.74	56.53	54.74
	0-shot	52.95	55.46	55.10	55.81	54.92	54.20	56.17	56.17
PolyLM (13B)	5-shots	48.12	51.34	47.23	47.58	55.99	54.20	56.89	56.53
	0-shot	50.27	49.37	49.55	51.52	55.28	54.74	55.64	56.53
SeaLLM	5-shots	57.5	56.96	59.11	58.93	53.04	53.75	55.1	55.64
	0-shot	54.29	55.00	51.07	57.68	56.61	55.89	55.64	55.99
Sailor	5-shots	51.07	55.36	52.86	52.50	57.14	56.07	57.25	56.89
	0-shot	50.71	51.07	50.36	53.21	58.57	57.86	59.39	59.57

Table 10: Prompting experiment results for translate-test with respect to model and few shot choices. **M**=MEGA, **BL**=BLOOMZ, **LH**=LM Harness, **LLH**=Local LM Harness, while **-ID** and **-EN** refer to the respective languages.

Model	Shots	Colloquial							
		M-ID	M	BL-ID	BL	LH-ID	LH-EN	LLH-ID	LLH-EN
Bactrian-X	5-shots	50.27	51.88	51.52	50.27	57.96	58.04	59.39	59.21
	0-shot	48.84	51.16	49.19	48.66	56.53	57.42	59.21	56.71
Llama-2	5-shots	50.81	52.77	50.45	49.91	54.29	53.85	53.49	53.31
	0-shot	51.34	50.27	51.16	51.70	52.24	52.06	52.06	52.42
Llama-2-Chat	5-shots	51.70	51.16	49.91	49.19	53.67	52.59	55.28	53.49
	0-shot	49.55	47.05	49.55	49.91	52.77	51.70	52.24	51.88
BLOOMZ	5-shots	55.28	54.92	58.50	57.96	58.14	58.68	58.32	58.50
	0-shot	53.31	57.78	59.39	59.57	56.17	55.10	56.89	58.50
PolyLM (13B)	5-shots	50.45	51.70	53.31	53.49	58.32	58.32	58.50	57.07
	0-shot	51.16	50.09	52.42	51.52	57.07	56.71	59.03	58.32
SeaLLM	5-shots	60.36	62.14	58.39	64.11	56.61	58.04	58.14	57.42
	0-shot	51.79	50.54	50.00	52.68	56.07	56.25	56.17	54.56
Sailor	5-shots	53.21	54.11	55.89	55.36	65.00	63.93	65.12	64.58
	0-shot	48.75	49.11	51.25	51.25	66.07	64.64	63.51	62.79

Table 11: Colloquial experiment results. **M**=MEGA, **BL**=BLOOMZ, **LH**=LM Harness, **LLH**=Local LM Harness, while **-ID** and **-EN** refer to the respective languages.

Model	M-ID	M	BL-ID	BL	LH-ID	LH-EN	LLH-ID	LLH-EN
Bactrian-X	51.52	49.73	50.45	50.81	60.29	59.93	63.69	60.64
Llama-2	51.70	51.34	50.45	50.45	56.89	55.81	58.86	56.71
Llama-2-Chat	53.13	52.59	50.09	49.91	55.99	55.28	55.46	56.35
BLOOMZ	55.10	54.56	55.64	57.60	62.97	62.97	65.65	65.29
PolyLM (13B)	53.67	51.16	55.64	51.52	63.33	61.90	62.97	62.97
SeaLLM	68.39	65.71	68.04	68.04	61.79	62.68	63.51	62.61
Sailor	62.50	67.68	60.54	59.64	73.93	71.43	71.43	72.09

Table 13: Cross-lingual experiment results (5-shot).

Model	Shots	XCOPA					
		M-ID	M	BL-ID	BL	LH-ID	LH-EN
Bactrian-X	5-shots	52.60	51.60	55.00	51.00	69.60	71.20
	0-shot	54.20	53.00	53.00	54.80	67.20	67.60
Llama-2	5-shots	56.00	56.40	56.80	58.80	63.20	64.20
	0-shot	51.00	50.00	51.60	51.80	61.00	60.60
Llama-2-Chat	5-shots	54.80	59.60	57.00	60.80	60.20	61.80
	0-shot	51.60	51.00	51.00	49.20	58.80	57.40
BLOOMZ	5-shots	61.60	62.00	66.60	71.20	71.80	69.60
	0-shot	59.60	71.00	66.80	76.20	60.60	59.80
PolyLM (13B)	5-shots	48.60	49.00	48.80	50.00	71.20	68.60
	0-shot	47.00	49.60	48.60	50.20	70.20	68.00
SeaLLM	5-shots	75.80	74.40	69.60	77.80	72.00	69.60
	0-shot	59.80	53.20	51.20	60.60	69.80	66.0
Sailor	5-shots	60.80	65.00	58.40	55.60	75.40	76.60
	0-shot	54.00	52.80	55.20	53.20	76.60	73.40

Table 14: XCOPA experiments results

F Data Creation & Review Instruction

This section provides the guidelines for data creation and review.

F.1 Data Creation

You are tasked to create at least 110 instances of commonsense causal reasoning data in the COPA format. Two examples of COPA instances:

Type: Cause.
Premise: Water flows from faucet.
Correct Choice: He turns the faucet open.
Incorrect Choice: The faucet is broken.

Type: Effect.
Premise: Two cars crash into each other.
Correct Choice: I am stuck in traffic.
Incorrect Choice: There are 1000 victims.

As you can see, we have a premise followed by two choices, with one being more plausible (hence correct) than the other. Note the cause and effect types. In the former, the choice causes the premise, and vice versa. As mentioned above, in this task, you should create data following the above format but injected with Indonesian local or cultural information. Two examples:

(Cause) I got a mild stomachache. **I just ate seblak**. I just ate nagasari.

(Effect) That kid's got circumcised. **He received a lot of money**. His family booked a holiday trip.

(Cause) A guy shouted "You're a dog!" **His friend is annoying**. His friend is cute and sweet.

(**Bold** indicates the correct alternative.)

For the first example, "Seblak" is a name of typically spicy food that is more likely to cause stomachache, rather than "Nagasari" which is a sweet dessert. For the second example, it is a common practice to gift money to a young boy who has just got a circumcision. For the third example, "anjing" (EN: "dog") is a common swear word in Indonesia. As you can see, in these three examples, it is impossible for a person who does not know Indonesian culture to know the correct answer, and your task is to create at least 110 such instances, using the standard formal Indonesian language.

You can also see that these three examples each showcase a different category of local information. For the seblak vs. nagasari example, the name or

terminology of these foods is the source of locality. For the circumcision example, the **cultural context** surrounding circumcision itself is the source of locality. In the last example (dog), the **language** usage is the source of locality. Later, you will be tasked to acutely categorize your data into these three categories. But for now, to ensure the variety of your data, you should try to come up with roughly ~50 terminologies, ~50 cultural contexts, and ~10 languages in your data. You should also ensure that the ratio between causes and effects of your data instances is 50:50.

During data creation, you should also take into account the following criteria:

- **Appropriateness**, ensure that your data contains the appropriate local or cultural nuance well-known by Indonesians, especially native Jakartans.
- **Difficulty**, ensure that your data is not too easy, but also not too difficult or obscure. Especially when making the incorrect choice, put something that is obviously incorrect for natives, but difficult to guess for foreigners.
- **Correctness**, ensure that the logical reasoning contained in your data is correct.
- **Ambiguity**, ensure that there is no ambiguity in wordings and in the choices. Check again your incorrect option, it might be the case that it is still plausible given the premise.
- **Ethics**, ensure your data is not discriminatory towards any person or organization.
- **Clarity and format**, check your capitalization, grammar, and spelling.

F.1.1 Data Peer Review

You will be tasked to blindly review other creators' data. The two choices order will be randomly swapped so you cannot see the intended correct answer. Your task is to first try to pick one choice that you deem to be more plausible. Second, you will be asked to put a qualitative comment on each data if you feel that there is a problem with regard to appropriateness, difficulty, correctness, ambiguity, ethics, or clarity. You should also put a comment for any instance that you find to be duplicated, whether with your own data or with other data you have reviewed until now.

Once the peer review is done, all data creators will meet together to discuss data that has incorrect answers by the reviewers and data that has a qualitative review. The discussion will result in a decision on whether each data would be accepted, rejected, or slightly modified. Once this process is done, you should rework your own data again, make the required changes, and repeat these review processes all over again until at least 550 data are accepted.

F.2 Data Categorization

Once your data passes peer review, you are tasked to categorize your data. As briefly touched above, your data should be put into three categories.

1. **Language**. If your data uses non-literal words or phrases, then it falls into this category.
2. **Local terminology**. If your data uses any Indonesian entities, famous people, food names, location names, abbreviations, local concepts, etc., then it falls into this category. Note that this means entities that originate from outside Indonesia, such as Pizza Hut, KFC, NATO, are not considered local terminologies.
3. **Culture.** Any cultural context that does not pertain directly to the use of local terminologies or language goes into this category. If the cultural context arises from the local term or language *immediately*, then it does not fall into this category.

Note that a single data instance should always fall into one or more categories. You may find that categorizing your data is not easy, as some things can be ambiguous. Here are some tips.

First, to ensure whether a local term is really local, you should try to google it first. If there is no direct one-to-one translation to your term, then you can be sure that it is indeed local. For instance "kobokan" is local because its closest English term, "finger bowl" does not mean exactly the same thing.

Second, you might find that, rather confusingly, some language usage can also be a local term. For example, "polisi tidur" (EN: "speedbump", literally "sleeping police") is a non-literal phrase, but it is also a local term because there is no one-to-one literal replacement in Indonesian for it. On the other hand, "datang bulan" is a language, but is not a local term because "haid" and "menstruasi" (EN: "menstruation") are appropriate literal replacements.

Third, you may find your data that uses local term is also cultural at the same time. For example, you may want to connect "Lebaran Haji" with animal slaughter. However, this should not be labeled as a culture because animal slaughter is implied

immediately by “Lebaran Haji”. On the other hand, you may want to connect “Lebaran” with the emptiness of Jakarta streets. This is OK to be labeled as culture because the cultural impact is not immediately implied.

F.2.1 Categorization Peer Review

In this step, you are tasked to review another data creator’s categorization. You are asked to highlight any categorization that you are unsure/disagree with. Meanwhile, another creator will review your work in the same way. Any disagreement should be resolved together with your peer until a final categorization is achieved.

F.3 Colloquial Form Translation

Next, you are tasked to translate all your data which is in standard Indonesia, into their colloquial forms. To do this, you should imagine your data being spoken/talked in a day-to-day Jakartan conversational context, then within that context, transform your data in a natural way. For example, from a standard sentence “Saya sedang dalam perjalanan ke sana.” (En: “I am on my way there.”), you can imagine a context where you are texting your close friends that you are on your way in a colloquial manner: “Gw otw”, which is natural. You should not translate it to “I sedang going ke there.”, which is non-standard but is highly unnatural.

To preserve variety in the colloquial forms, we will not be providing a detailed guideline here. You should use your own knowledge and experience of colloquial Indonesian and not seek outside influence too much.

F.3.1 Colloquial Peer Review

In this step, you are tasked to review another creator’s colloquial translation. You are asked to highlight any translation that you deem unnatural or inaccurate with respect to the original standard data.

Similarly, another creator will also review your work. Any disagreement should be resolved together with your peer until a final colloquial translation for each data is achieved.

G Human Scoring Guideline

The following contains the elaborated guidelines for human scoring of COPAL-ID. The original guideline is in Indonesian but we translate it to English for this appendix section.

G.1 What is COPA?

COPA stands for Choice of Plausible Alternatives. As the name suggests, COPA is a test that consists of a set of multiple-choice questions. Each question contains a premise or situation that serves as the basis for the question. The premise is followed by two alternative options. The participant’s task is to choose the option that is most plausible among the two. Below are some example of COPA questions.

Example 1
Premise: My little brother wakes up late.
Option 1: Mother is angry because it’s Monday.
Option 2: Mother is angry because it’s Sunday.

The correct answer is Option 1. Although it is possible that the mother scolds the little brother for waking up late on Sunday, this situation is less likely compared to Option 1, which is more common because Monday is a school day.

Example 2
Premise: The national team wins the Thomas Cup.
Option 1: My brother buys beer at the minimarket to celebrate.
Option 2: My brother buys pizza hut to celebrate.

The correct answer in the context of Indonesian society is Option 2. While it is possible that the brother buys beer for the celebration, this is less common, and most minimarkets in Indonesia are prohibited from selling beer.

Example 3
Premise: My little brother is wearing a uniform.
Option 1: He is going to school.
Option 2: He is going to play.

The correct answer is Option 1.

G.2 Effect vs Cause

Participants will encounter two types of questions in COPA: Effect and Cause. Examples 1 and 2 are Effect-type questions, while Example 3 is a Cause-type question. Participants can connect the premise and options using phrases like "as a result" or "because" to confirm.

G.3 Choosing the Most Plausible Option

If participants find a question where both options seem equally plausible, they are still asked to choose the option that is more likely or more plausible. Factors that can be used as a basis for choosing include:

1. **Choose the most likely option in the context of Indonesia and Jakarta.** Both options may be plausible in an international context, but participants are instructed to choose the one that is more likely in the national context of Indonesia and the city of Jakarta.
2. **Facts and statistics.** Consider factual information and statistics. For example, in Example 2, the fact that most minimarkets statistically do not sell beer and that, statistically, Indonesians do not consume alcohol should guide the decision.
3. **Sensible stereotypes.** Consider stereotypes that make sense. For instance, the stereotype that graduates from Islamic schools are usually more proficient in reciting religious texts than those from public schools may be a sensible guideline.
4. **Cause and effect relationships.** Consider cause-and-effect relationships that make sense. For example, it is more likely that eating spicy food causes stomachache rather than causing a headache.
5. **Personal knowledge/common sense.** Utilize personal knowledge or common sense, as well as insights from individuals within the participant's family or social circle. This can provide additional context and perspectives that may aid in making a more informed choice.

G.4 Technical Instructions

Participants are required to provide their Gmail addresses to the specified contact. The questions will be sent via Google Sheets, where each row contains one question (1 premise and 2 options). Participants are instructed to choose the most plausible option by checking the checkbox provided next to each option.

The total number of questions for each participant is 559. Participants are requested to complete all 559 questions on their own and refrain from consulting with third parties, including search engines and AIs such as ChatGPT.

G.5 Submission and Honorarium

Participants are required to complete all questions by Sunday, September 17, 2023. Once finished, participants can notify the specified contact that the task is complete. After that, edit access for the participant on the provided Google Sheet will be revoked.

Participants will receive a basic honorarium of IDR 250,000 plus IDR 180 for each correct answer. For example, if all answers submitted are incorrect, the honorarium is IDR 250,000. If all answers are correct, the honorarium is IDR 250,000 + (180 x 559) = IDR 350,620.[11]

The honorarium will be given to participants at most one day after the calculations are completed and after the participant's bank information is provided to the specified contact (whichever occurs last).

[11] Jakarta minimum wage is, as of December 2023, slightly below IDR 5,000,000 per month. Assuming 20 working days and 8 working hours per day, this translates to IDR 31,250 per hour. Each participant requires 2-4 hours to complete all questions, which means that we have paid at least double the minimum wage stipulated by the government.