CULTURETALK-ID: A Multi-Task Dialogue Benchmark for Cultural Commonsense in Indonesian Local Languages

Muhammad Dehan Al Kautsar **Salsabila Zahirah Pranida**
Bilal Elbouardi **Fajri Koto**
Mohamed bin Zayed University of Artificial Intelligence
{muhammad.dehan,fajri.koto}@mbzuai.ac.ae, salsabila.pranida@alumni.mbzuai.ac.ae

Abstract

Culture shapes how people interpret and navigate social interactions, yet existing Indonesian cultural commonsense benchmarks evaluate LLMs on short and isolated prompts, stripping away the dialogic context in which cultural nuances actually surface. We introduce **CULTURETALK-ID**, the first dialogue-based benchmark for cultural commonsense in Indonesian and its local languages, comprising 4,496 culturally grounded dialogues across 11 languages and 13 culturally salient topics, curated through a multi-stage human pipeline with native speakers to ensure authenticity. **CULTURETALK-ID** introduces three complementary tasks, namely dialogue-based multiple-choice cultural commonsense reasoning, culturally faithful machine translation, and language steering, which jointly probe whether LLMs can understand, transfer, and generate culturally grounded language. Our results show that open-source models still underperform, particularly in the generation task and local language settings, highlighting a gap in their ability to capture and generate culturally grounded knowledge in conversational contexts.[1]

1 Introduction

Culture encompasses patterns of thinking, communication, and behavior, along with material artifacts, that shape a community's way of life and are transmitted across generations (Liu et al., 2025; Nguyen et al., 2022). This transmission is inherently dynamic, as it is continually negotiated and renewed through everyday interactions between individuals and groups, emerging in conversations across languages and communities (Sharma and Liu, 2024). As LLMs are increasingly deployed as conversational agents for culturally diverse users, their ability to recognize and reason about culture within dialogue becomes critical, not only for producing contextually appropriate responses (Wu et al., 2025), but also for capturing the implicit, community-specific norms and expectations that guide interaction, often referred to as cultural commonsense.

However, despite this inherently interactive view of culture, prior work has largely evaluated cultural commonsense within simplified notions of commonsense reasoning, typically using short, minimally contextualized tasks (Sap et al., 2019; Bisk et al., 2020; Mahendra et al., 2021; Putri et al., 2024; Myung et al., 2024; Wibowo et al., 2024; Guo et al., 2025; Permadi et al., 2026). For example, Wibowo et al. (2024) relies on isolated prompts paired with single-turn multiple-choice questions. While useful for controlled benchmarking, such setups treat cultural knowledge as static, detached from the interactions through which it is expressed and negotiated. This mismatch becomes especially pronounced in culturally dense and multilingual environments, where practices vary across communities, as seen in Indonesia (Koto et al., 2024).

Indonesia is a highly diverse nation, with over 700 languages, numerous ethnic groups, and distinct regionally grounded cultural norms (Aji et al., 2022). It is also the second most linguistically diverse country in the world (Winata et al., 2023). In such a setting, cultural commonsense is deeply localized, shaped by both national context and community-specific practices. Consequently, evaluating LLMs solely on Indonesian-language benchmarks (Ponti et al., 2020; Koto et al., 2022; Wibowo et al., 2024; Koto et al., 2024) provides only a partial view of their cultural understanding. A more faithful assessment requires context-rich approaches that incorporate both Indonesian and local languages, capturing how cultural knowledge varies and unfolds across communities.

To address these challenges, we introduce **CULTURETALK-ID**, a new benchmark for culturally grounded dialogue in Indonesia, developed

[1]Dataset: 🤗 CultureTalk-ID

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2888–2912
November 6-10, 2026 ©2026 Association for Computational Linguistics

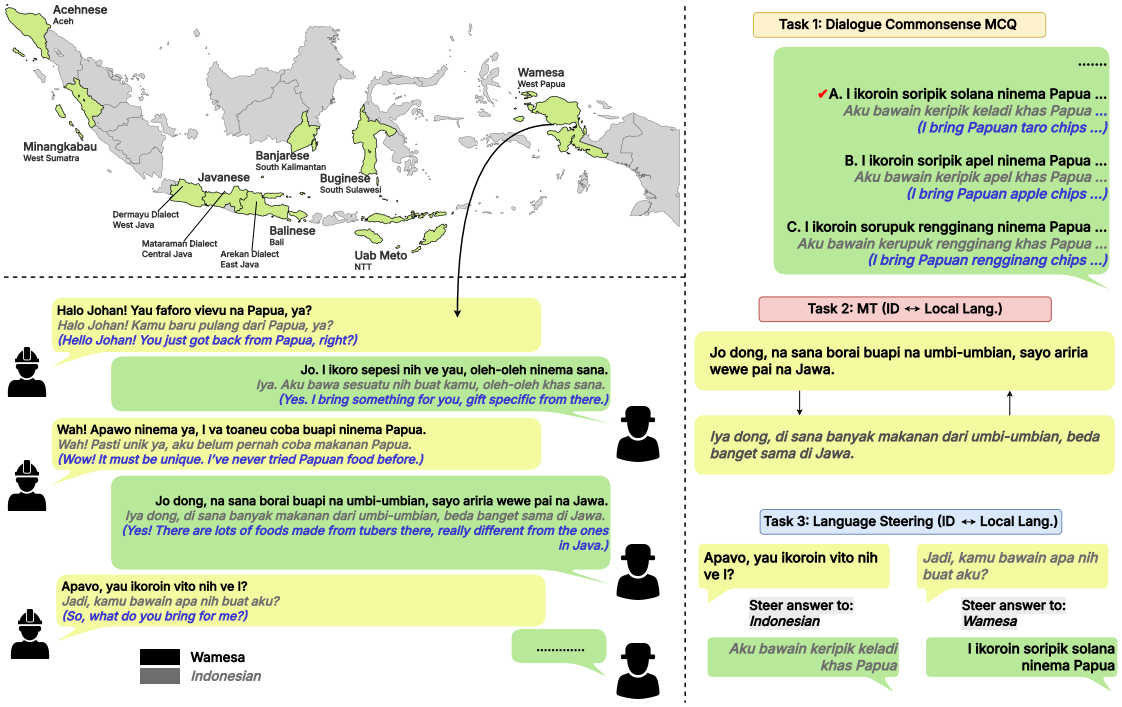


Figure 1: Overview of the **CULTURETALK-ID** benchmark. **CULTURETALK-ID** supports three tasks: (1) dialogue-based multiple-choice questions (MCQs) for cultural commonsense reasoning, (2) bidirectional machine translation between Indonesian and local languages, and (3) language steering.

through a multi-stage human pipeline involving native speakers to ensure cultural authenticity. Compared with recent benchmarks (Permadi et al., 2026; Aji and Cohn, 2025), **CULTURETALK-ID** incorporates extensive human revision, annotation, and translation across the dataset creation process. While prior work either focuses on multihop reasoning in Indonesian only (Permadi et al., 2026) or lacks a conversational setting (Aji and Cohn, 2025), our benchmark emphasizes multi-turn dialogue, enabling more natural and context-rich cultural reasoning. It covers both general Indonesian culture and the cultures of 10 provinces, including Aceh, West Sumatra, West Java, Central Java, East Java, Bali, Nusa Tenggara Timur (NTT), South Kalimantan, South Sulawesi, and West Papua, spanning Indonesian and 10 local languages across 13 cultural topics.

As illustrated in Figure 1, each dialogue is constructed in both Indonesian and the corresponding local language spoken in the region (e.g., Wamesa in West Papua, languages with limited digital presence and considered vulnerable (Ritchie et al., 2024)). We intentionally omit the final utterance in each dialogue, and the model selects the most appropriate continuation from multiple candidates. Furthermore, we introduce machine translation and language steering tasks to further evaluate LLMs' capabilities in Indonesian and regional languages, leveraging our parallel datasets for each region.

We summarize our contributions as follows:

- We introduce **CULTURETALK-ID**, a dialogue-based benchmark for culturally grounded interactions in Indonesia, developed through a multi-stage human pipeline involving native speakers. It covers Indonesian and 10 local languages across 13 cultural topics.
- We propose three evaluation tasks: (1) dialogue-based multiple-choice questions for cultural commonsense reasoning, (2) bidirectional machine translation between Indonesian and local languages, and (3) language steering with cultural grounding. These tasks assess LLMs' ability to reason about culturally appropriate responses in dialogue, handle multilingual variation, and generate contextually grounded outputs.
- We evaluate a diverse set of LLMs, including proprietary, general multilingual, and Southeast Asian (SEA)-centric models. Our results show that open-source models lag behind, revealing a substantial gap in capturing and generating culturally grounded knowledge in Indonesian conversational settings.

2 Related Works

Dialogue Benchmarks in Multilingual Settings Most existing dialogue benchmarks focus on high-resource languages, particularly English (e.g., DailyDialog (Li et al., 2017), PersonaChat (Zhang et al., 2018), and MultiWOZ (Budzianowski et al., 2018)), and lack explicit cultural grounding. Recent efforts extend these benchmarks to multilingual settings, such as XDailyDialog (Liu et al., 2023) and mDIA (Zhang et al., 2022), which expand coverage to up to 46 languages. However, these resources provide limited support for Indonesian and its diverse local languages and cultures. As a result, they do not capture multi-turn dialogue, cultural commonsense reasoning, and localized linguistic variation in this context, which **CULTURETALK-ID** is designed to address.

Conversational-based NLP task in Indonesian Context Most Indonesian conversational datasets are developed as subsets of broader multilingual dialogue benchmarks. For instance, COD (Majewska et al., 2023) is a multilingual task-oriented dialogue (TOD) benchmark covering four languages: Arabic, Indonesian, Russian, and Kiswahili. The benchmark includes tasks such as intent detection, slot labeling, and end-to-end dialogue modeling. Kautsar et al. (2023) focuses on task-oriented dialogue in a parallel Indonesian–English setting, where the system needs to understand the intent of the user and succeed in doing the user's task. More recently, Al Kautsar et al. (2026) introduced a culturally grounded multilingual dialogue benchmark for Southeast Asian languages. However, the benchmark has limited language coverage and focuses on evaluating dialogue generation quality, with dialogues automatically constructed using LLMs. In contrast, **CULTURETALK-ID** emphasizes culturally grounded, multi-turn dialogue in the Indonesian context, combining dialogue-based cultural MCQ, machine translation, and language steering across Indonesian and local languages.

Commonsense Reasoning in Indonesian and Local Languages Commonsense reasoning has been widely studied since the introduction of the Winograd Schema Challenge (Levesque et al., 2012). This line of work has since evolved, and such approaches have been extended to Indonesian through datasets such as XStoryCloze (Lin et al., 2022), IndoCloze (Koto et al., 2022), XCOPA (Ponti et al., 2020), and COPAL (Wibowo et al., 2024). However, these datasets remain limited in capturing cultural nuances and localized aspects of Indonesian contexts. Koto et al. (2024) introduce geographically grounded cultural commonsense reasoning across eleven Indonesian provinces, while Permadi et al. (2026) focus on multi-hop reasoning for Indonesian cultural understanding. However, these benchmarks rely on single-turn formats and do not incorporate dialogue-based interactions or coverage of Indonesian local languages. In parallel, Aji and Cohn (2025) introduce LoraxBench, a benchmark for evaluating cultural understanding in Indonesian and local languages through the translation of existing datasets. While expanding language coverage, it does not model conversational interactions or incorporate province-specific scenarios, limiting its ability to capture localized and context-dependent cultural communication. In contrast, **CULTURETALK-ID** explicitly incorporates multi-turn dialogue, as illustrated in Figure 1, to capture how cultural knowledge is expressed, contextualized, and negotiated in real-world interactions.

3 Dataset Creation

We construct the **CULTURETALK-ID** dataset using the pipeline outlined in Figure 2, transforming the IndoCulture (Koto et al., 2024) and COPAL-ID (Wibowo et al., 2024) datasets with the help of human annotators into Indonesian and various Indonesian local languages. It covers 4,496 dialogues, 10 regions across Indonesia, 11 languages, and 13 topics, closely related to Indonesian culture.

3.1 Phase 1: Dialogue Generation

To initiate the pipeline, we select IndoCulture (Koto et al., 2024) and COPAL-ID (Wibowo et al., 2024) as pivot datasets, due to their culturally grounded multiple-choice formats. We repurpose their premises as dialogue topics and treat the answer options as candidate utterances to facilitate dialogue construction and downstream human annotation. Given each instance, we prompt GPT-5 to generate a dialogue in Indonesian, intentionally omitting the final utterance. The original answer options are then adapted as candidate endings, with only one being contextually appropriate. We generate the dialogues in Indonesian because local Indonesian languages are generally low-resource and less reliably handled by current LLMs, making human involvement necessary for subsequent

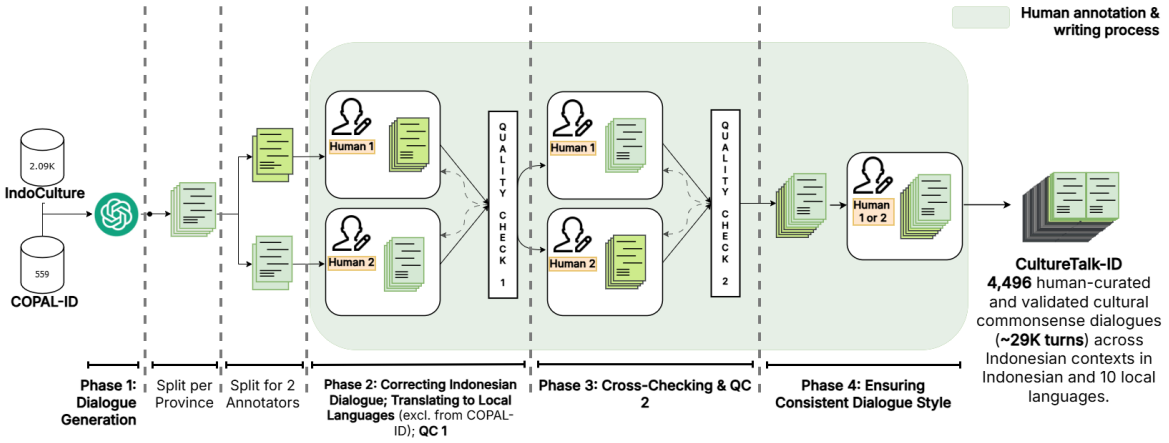


Figure 2: The CULTURETALK-ID dataset creation pipeline. It consists of four phases: dialogue generation; correction of Indonesian dialogues, translation into local languages, and initial quality control (QC1); cross-checking and secondary quality control (QC2); and finally, ensuring consistency in dialogue style.

annotation and translation. The resulting dialogues are grouped by province and prepared for human annotation and translation. The detailed process of this phase is provided in Appendix A.

3.2 Phase 2: Correction of Indonesian Dialogues, Translation to Local, and QC1

The dialogues generated by GPT-5 were subsequently verified through human evaluation. For this process, we recruited two annotators for each cultural group, covering 10 provinces and one general Indonesian category (22 annotators in total). All annotators were required to be fluent in both Indonesian and their respective local languages, and to have lived in the corresponding province for more than 10 years.

During the recruitment process, candidates were provided with the task description and annotation guidelines, and were asked to complete a pilot annotation task to assess their understanding of the objectives. We then selected the most suitable annotators based on the quality of their pilot annotation outputs. Then, all annotators participated in a preliminary briefing session to ensure a consistent understanding of the annotation guidelines. During the validation stage, annotators were instructed to verify cultural appropriateness and answer correctness, and ensure that incorrect answer choices were culturally inappropriate or implausible.

Once the Indonesian dialogues had been validated, annotators translated them into their respective local languages. We allow flexible, non-literal translation to preserve fluency and naturalness in the target language, while keeping the same meaning in each utterance. After the annotators have completed the task, as a final quality-assurance step, we conduct additional sampling-based reviews of the annotated dialogues. Dialogues that did not satisfy the annotation standards were returned to annotators for revision.

3.3 Phase 3: Cross-checking and QC2

Following the first phase, we conducted a cross-check between annotators, followed by a second quality control stage (QC2). Since each language was assigned two annotators, each annotator was asked to review their partner's work for cultural correctness and translation quality.

To assess annotator attentiveness during this process, we deliberately inserted identifiable incorrect entries into approximately 5% of the instances, following the methodology introduced by Muszyński (2023). If an annotator failed to detect and correct these deliberately incorrect entries, they were asked to re-evaluate the entire set of dialogues until all injected errors had been identified and fixed. Additionally, we asked annotators to answer the MCQ instances themselves to verify that the correct answers were consistently identifiable. Instances answered incorrectly by the annotators were subsequently revised. These processes served as our second-stage quality control mechanism. The transformation process from the original dataset until this phase is illustrated in Figure 4 (Appendix A).

3.4 Phase 4: Consistency in Dialogue Style

To ensure that the benchmark is not trivial, we identified two common issues arising from the

#dialogues	4,496
#region-specific dialogue	2,980
#turns	29,420
avg. words per dialogue	60.23
avg. utterance per dialogue	6.54
avg. words per utterance	8.28
#words	270,793
#unique words	27,391
train : test split	1:9
#unique words (Indonesian ∧ Local Indonesian languages)	9,385 (39.51% of local languages' unique words)

Table 1: Statistics of the **CULTURETALK-ID** dataset, combining Indonesian and local Indonesian language dialogues across all cultural regions.

LLM-generated dialogues in the early stage of our pipeline. The first issue is the non-uniform answer choices. In some cases, the correct option was written in a confident and natural manner, while the incorrect options contained obviously unnatural or doubtful expressions, making the correct answer easy to infer. For example, several incorrect options included phrases such as *"malah"* ("instead/contrarily") or *"anehnya"* ("weirdly"), which implicitly signaled implausibility. The second issue concerns the explicit presence of cultural entities within the answer choices. This may introduce shortcut cues within the utterance options, as illustrated in Appendix A.

To address these issues, we asked one annotator from each province to directly revise the answer options in both Indonesian and their corresponding local-language versions. The refined datasets were then merged into the final **CULTURETALK-ID** benchmark.

3.5 CULTURETALK-ID Benchmark

Each instance in **CULTURETALK-ID** consists of a culturally grounded dialogue with the final utterance omitted, together with three candidate responses, only one of which is culturally and contextually appropriate. Every dialogue is available in both Indonesian and its corresponding local Indonesian language, forming a parallel dialogue dataset spanning 10 Indonesian cultural regions. Table 1 presents the dataset statistics, where it contains 4,496 dialogues (2,980 region-specific dialogues), requiring knowledge tied to particular Indonesian cultural contexts. The more detailed statistics can be found in Appendix B.

To the best of our knowledge, **CULTURETALK-ID** is the first dialogue-based benchmark with extensive human curation for evaluating Indonesian cultural commonsense across both Indonesian and multiple local languages. In addition to dialogue-based cultural commonsense reasoning, its parallel structure also supports machine translation and language steering tasks.

4 Experimental Setup

Using **CULTURETALK-ID**, we evaluate dialogue-based cultural competence across Indonesian and local Indonesian languages. We consider three groups of models: (1) proprietary models, (2) multilingual LLMs, and (3) Southeast Asian (SEA)-centric LLMs. For proprietary models, we evaluate GPT-5.1, Gemini-2.5-flash, and Cohere-Command-A. The multilingual models include Qwen3-8B (Yang et al., 2025), Llama-3.1-8B Instruct (Grattafiori et al., 2024), and Gemma-2-9B Instruct (Team et al., 2024). Furthermore, the SEA-centric models include Sailor2-8B Chat (Dou et al., 2025), SEA-LION-v3.5-8B (Ng et al., 2025), and Sahabat-AI-v1 Instruct[2]. All open-weight models are selected at a comparable parameter scale, around 8–9B parameters, while proprietary models are included to represent the current state of the art in large language models.

Building on the parallel dialogue dataset described in the previous section, we evaluate models on three complementary tasks. Prompts are provided in Appendix C.

MCQ Evaluation. For the MCQ evaluation task, each model is given a dialogue context with the final utterance omitted, together with three candidate responses, only one of which is culturally and contextually appropriate. For open-source models, rather than using generation-based evaluation, we adopt a likelihood-based setup in which models score each candidate response and select the option with the highest likelihood as the prediction. We evaluate both in Indonesian and local-language dialogues, and we report average accuracy as the main evaluation metric.

Machine Translation. The machine translation task evaluates a model's ability to translate dialogue utterances between Indonesian and local Indonesian languages (as listed in Appendix D) in both directions. Translation quality is evaluated using BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and an LLM-as-a-judge framework

[2]Sahabat-AI HuggingFace

Dialogue-based MCQ	Context: None			Context: Prov. + Lang.		
Model	**Avg**	**PS**	**~PS**	**Avg**	**PS**	**~PS**
			Proprietary Models			
GPT-5.1	82.31 (↓6.92)	79.07 (↓7.37)	88.67 (↓6.05)	82.70 (↓7.25)	79.26 (↓8.12)	89.45 (↓5.54)
Gemini-2.5-flash	**87.77** (↓3.69)	**85.23** (↓4.43)	**92.74** (↓2.25)	**87.72** (↓3.96)	**85.23** (↓4.76)	**92.61** (↓2.38)
Cohere-Command-A	75.09 (↓9.65)	70.07 (↓10.67)	84.96 (↓7.65)	74.64 (↓10.37)	70.00 (↓11.41)	83.77 (↓8.31)
average proprietary	*81.72* (↓6.76)	*78.12* (↓7.49)	*88.79* (↓5.32)	*81.69* (↓7.19)	*78.16* (↓8.10)	*88.61* (↓5.41)
			Multilingual Models			
Qwen3-8B	52.89 (↓16.06)	48.86 (↓17.78)	60.82 (↓13.02)	54.72 (↓15.56)	50.74 (↓17.25)	62.53 (↓12.27)
Llama3.1-8B Instruct	51.87 (↓10.76)	49.60 (↓10.53)	56.33 (↓11.22)	53.87 (↓11.57)	51.95 (↓11.94)	57.65 (↓11.09)
Gemma2-9B Instruct	**65.17** (↓11.61)	**59.80** (↓13.02)	**75.73** (↓8.83)	**65.88** (↓11.83)	**60.87** (↓13.22)	**75.73** (↓9.10)
average multilingual	*56.64* (↓12.81)	*52.75* (↓13.78)	*64.29* (↓11.02)	*58.15* (↓12.99)	*54.52* (↓14.14)	*65.30* (↓10.82)
			SEA-Centric Models			
Sailor2-8B Chat	**70.15** (↓9.92)	**64.50** (↓11.94)	**81.27** (↓5.93)	**72.02** (↓9.30)	**66.85** (↓10.94)	**82.19** (↓6.07)
SEA-LION-v3.5-8B	36.57 (0.00)	36.85 (0.00)	36.02 (0.00)	36.57 (0.00)	36.85 (0.00)	36.02 (0.00)
Sahabat-AI-v1 Instruct	69.08 (↓11.79)	63.62 (↓13.63)	79.82 (↓8.17)	69.93 (↓11.61)	65.17 (↓13.35)	79.29 (↓8.18)
*average SEA-centric**	*69.62* (↓10.86)	*63.56* (↓12.79)	*80.55* (↓7.05)	*70.98* (↓10.46)	*66.01* (↓12.15)	*80.71* (↓7.13)

Table 2: **Dialogue-based MCQ** performance on local Indonesian language dialogues. Values in parentheses denote the performance difference compared to Indonesian dialogue settings, where ↓ indicates performance degradation in local language settings. **Bold** scores indicate the best score within each model category. *The SEA-LION-v3.5-8B-R scores are excluded.

on a 1-5 scale. We evaluate models in zero-shot settings and additionally report supervised fine-tuning (SFT) results for selected open-weight models. For open-weight models, inference is repeated three times, and the mean and standard deviation are reported.

Language Steering. The language steering task evaluates whether models can generate responses in a requested linguistic variety. This is important, as Indonesians are bilingual, able to speak in Indonesian and their own local language. Given a dialogue context, models are instructed to produce the next utterance either in Indonesian or in a specified local Indonesian language. We evaluate generated responses using an LLM-as-a-judge score, Glot-LID (Kargaran et al., 2023) language identification accuracy, BLEU, and chrF.

5 Results & Analysis

5.1 Dialogue-based Cultural Commonsense MCQ Evaluation

The evaluation results for dialogue-based cultural commonsense are presented in Table 2. Overall, proprietary models achieve the best performance on this benchmark, delivering consistent results across both settings, with Gemini-2.5-flash performing the best among the proprietary models. However, generic multilingual models lag substantially behind proprietary models, where in contrast, SEA-centric models consistently outperform generic multilingual baselines, with Sailor2-8B-Chat achieving the strongest performance on local Indonesian dialogues. However, SEA-LION-v3.5-8B-R exhibits degenerate behavior, consistently selecting the same answer choice (B) across questions.

We also find that performance drops substantially in local Indonesian language settings. Under the no-context setting (i.e., without province and language information), performance decreases by 6.76, 12.81, and 10.86 points for proprietary, generic multilingual, and SEA-centric models, respectively. This underscores the challenges of understanding low-resource languages, particularly Indonesian local languages, despite their large number of speakers.

We also observe that providing additional context improves performance for open-source model groups, particularly multilingual and SEA-centric models, while the gains are relatively limited for proprietary models. Province-specific (PS) dialogues benefit the most from this additional context. Moreover, as more contextual information is introduced, the performance gap between PS and ~PS dialogues narrows, suggesting that additional context helps models achieve better cultural grounding and capture localized cultural nuances. The more detailed analysis can be found in Appendix E.1.

Is further pretraining an LLM in a cultural context helpful? We investigate whether additional pretraining in an Indonesian cultural context im-

Machine Translation	BLEU Scores	BERTScore	LLM as a Judge (1–5)				
Model	B4	F1	adequacy	fluency	register	terminology	overall
Indonesian → local Indonesian languages							
Proprietary Models (0-shot)							
GPT-5.1	27.53	84.02	4.35	3.80	3.75	4.33	3.87 (2)
Gemini-2.5-flash	**28.36**	**84.33**	**4.48**	**3.91**	**3.91**	**4.48**	**4.00** (1)
Cohere-Command-A	21.45	81.40	3.35	2.82	2.69	3.29	2.93 (5)
Multilingual Models (0-shot)							
Qwen3-8B	19.78 (±0.1)	78.74 (±0.1)	2.82	2.11	1.90	2.73	2.22 (11)
Llama-3.1-8B Instruct	17.62 (±0.1)	78.71 (±0.1)	2.79	2.27	2.21	2.81	2.49 (10)
Gemma-2-9b Instruct	**20.54** (±0.1)	**80.15** (±0.1)	**3.20**	**2.49**	**2.40**	**3.11**	**2.73** (7)
SEA-Centric Models (0-shot)							
Sailor2-8B Chat	15.44 (±0.1)	77.58 (±0.1)	3.38	**2.85**	**2.65**	**3.42**	**2.94** (4)
SEA-LION-v3.5-8B	19.34 (±0.1)	78.98 (±0.1)	3.18	2.61	2.44	3.11	2.72 (9)
Sahabat-AI-v1 Instruct	**21.70** (±0.0)	**80.51** (±0.0)	**3.57**	2.63	2.46	3.21	2.80 (6)
SFT Models							
Gemma-2-9b Instruct	⇑ 22.60 (±0.0)	⇓ 79.39 (±0.0)	⇓ 3.16	⇑ 2.54	⇑ 2.46	⇑ 3.14	≈ 2.73 (7)
Sahabat-AI-v1 Instruct	⇑ **24.56** (±0.0)	⇓ **80.30** (±0.0)	⇓ **3.53**	⇑ **3.03**	⇑ **2.97**	⇑ **3.55**	⇑ **3.14** (3)

Table 3: **Machine translation** results for Indonesian → local languages under the Province + Language context. We report BLEU-4, BERTScore F1, and LLM-as-a-judge scores for adequacy, fluency, register, terminology, and overall quality. The $_{\text{scores}}$ in BLEU and BERTScore represent standard deviation, while the $_{\text{numbers}}$ in the overall score represent the ranking of each model. **Bold** scores indicate the best score within each model category.

proves model performance, and to what extent it benefits downstream tasks. To this end, we compare Gemma-2-9B Instruct with Sahabat-AI-v1 Instruct, where the latter is further adapted from the former using Indonesian-centric training data. As shown in Figure 10 (Appendix E.2), performance improves consistently across all provinces, where somehow the largest improvement is in South Kalimantan, despite the additional pretraining primarily introducing only Indonesian, Javanese, and Sundanese contexts. Overall, these results indicate that culturally grounded pretraining provides consistent and transferable benefits, supporting the conclusion that such pretraining is indeed effective.

5.2 Machine Translation Evaluation

We present machine translation results in Table 3 for Indonesian to local Indonesian languages[3], with even the best model achieving moderate performance, while most models fall below 3.5. Proprietary models, such as Gemini-2.5-flash and GPT-5.1, achieve the strongest results across evaluation metrics. Generic multilingual models lag significantly, while SEA-centric models perform better, which narrows the gap with proprietary systems. Across evaluation dimensions, fluency and register consistently receive lower scores than adequacy and terminology, suggesting that generating natural and contextually appropriate local expressions remains a key challenge.

Supervised fine-tuning on the **CULTURETALK-ID** training set yields consistent improvements. Both Gemma-2-9B Instruct and Sahabat-AI-v1 Instruct outperform their base versions. Additionally, Sahabat-AI-v1 Instruct also surpasses Cohere-Command-A in this setting. Substantial gains are also observed in fluency and register, indicating that supervised fine-tuning, even with a relatively small train set (448 dialogues), effectively enhances naturalness and stylistic appropriateness. The machine translation experiment details further can be found in Appendix F.

Correlation between the LLM-as-a-judge and human evaluation. Table 3 shows a general consistency between computation-based metrics and LLM-as-a-judge scores. To further assess the reliability of LLM-as-a-judge in culturally grounded translation tasks, we analyze its correlation with human judgments. Specifically, we ask fluent Javanese and Minangkabau speakers to evaluate the adequacy and fluency of 60 generated samples from the best-performing models in each category. As shown in Section 6, the mean absolute distance (MAD) is generally below 1, and the accuracy-at-one (Acc@1) average reaches 0.83, indicating that LLM-as-a-judge is a reliable proxy in this setting.

[3]We conduct statistical significance testing for all metrics with available standard deviations.

Language Steering	Indonesian → Local languages		
Model	LLM	Glot-LID	BLEU
Proprietary Models			
GPT-5.1	4.42	73.9	0.64
Gemini-2.5-flash	**4.55**	**92.7**	**0.94**
Cohere-Command-A	3.70	31.6	0.62
Multilingual Models			
Qwen3-8B	**3.11** (±0.0)	1.5 (±0.2)	0.42 (±0.1)
Llama-3.1-8B Inst.	2.91 (±0.0)	**9.2** (±0.4)	0.32 (±0.2)
Gemma-2-9b Inst.	3.00 (±0.0)	7.8 (±0.6)	**0.44** (±0.0)
SEA-Centric Models			
Sailor2-8B	3.19 (±0.0)	23.7 (±0.4)	0.23 (±0.0)
SEA-LION-v3.5-8B	3.01 (±0.0)	**27.5** (±0.7)	0.26 (±0.0)
Sahabat-AI-v1 Inst.	**3.32** (±0.0)	13.7 (±1.0)	**0.64** (±0.2)
SFT Models			
Qwen3-8B	⇓ 3.06 (±0.1)	⇑ 4.4 (±5.1)	⇑ 0.50 (±0.1)
Sahabat-AI-v1 Inst.	⇑ **3.33** (±0.1)	⇑ **18.2** (±8.9)	⇑ **0.77** (±0.0)

Table 4: **Language steering** to local languages results. We report LLM-as-a-judge (denoted as LLM), Glot-LID, and BLEU scores. The $_{\text{scores}}$ represent the standard deviation. **Bold** scores indicate the best score within each model category.

We also observe that stronger models, typically proprietary models, exhibit closer alignment with human judgments.

5.3 Language Steering Evaluation

We present the language steering results from Indonesian to local Indonesian languages in Table 4. Overall, model performance drops substantially when generating local languages, in sharp contrast to the considerably stronger performance observed when generating Indonesian responses (see Appendix G). Only Gemini-2.5-flash and GPT-5.1 achieve moderate success, while Cohere-Command-A frequently fails to produce outputs in the correct language (31.6%), even before considering the correctness of the utterance content itself. Open-source models struggle considerably. The best-performing open-source model, SEA-LION-v3.5-8B, generates only 27.5% of utterances in the correct local language, while Qwen3-8B achieves just 1.5%. Interestingly, supervised fine-tuned models tend to perform better when generating local-language responses, although this improvement is often accompanied by a slight degradation in their generation semantic quality.

Upon further inspection, we find that many open-source models default to Indonesian, even when a local language is required. As illustrated in Table 5, both Qwen3-8B and Sahabat-AI-v1 Instruct (despite being trained on Javanese) mostly produce

Input	… Dewi: "Nasi jagung kan agak kering ya, biasanya kamu gimana biar makannya enak?" *(Dewi: "Corn rice can be a bit dry. What do you usually do to make it more enjoyable to eat?")* Syavira: …
Gold	"Aku mangan sego jagung diseloni kuah ben ora keselek." *from Javanese: ("I eat corn rice with soup so that I will not choke.")*
Gemini-2.5-flash	"Iyo bener, mangkane biasane tak tambahi jangan lodeh sing akeh kuahe karo iwak asin plus sambel bajak, wis jan mantep pol!" ✓ *from Javanese: ("Correct, that is why I usually add a lot of lodeh soup with salted fish and sambel bajak, it is really good!")*
Qwen3-8B	"Aku biasa rebus sedikit santan biar lembut dan lebih enak." ✗ *from Indonesian: ("I usually boil a bit of coconut milk to make it softer and tastier.")*
Sahabat-AI-v1 Instruct	"Aku biasanya tambahin sambal biar pedes, Dewi." ✗ *from Indonesian: ("I usually add sambal to make it spicy, Dewi.")*

Table 5: Example of language steering for Javanese. Open-source models tend to default to Indonesian and produce contextually incorrect responses.

output in Indonesian rather than Javanese. Moreover, the generated content is incorrect: the context requires adding more soup to prevent the food from drying out, but these models instead suggest adding *"santan"* (coconut milk) or even more sambal, which is not appropriate for the given situation. We present the more detailed language steering analysis in Appendix G.

6 Discussion

LLM-as-a-judge – Human Evaluation Correlation. The correlation between LLM-as-a-judge and human evaluation is presented in Table 19 for adequacy and Table 20 for fluency. Across both translation directions, proprietary models consistently achieve mean absolute distance (MAD) scores below 1 while maintaining Acc@1 scores above 0.84, indicating strong agreement with human judgments. Open-source models also demonstrate reasonably good correlations, although weaker than proprietary systems. Gemma-2-9B Instruct achieves average MAD scores of 0.98 and 0.80 for adequacy and fluency, respectively, while Sailor2-8B Chat obtains 0.90 and 0.87. Their Acc@1 scores remain relatively high at 0.78 for Gemma-2-9B Instruct and 0.775 for Sailor2-8B Chat, suggesting that LLM-as-a-judge remains a

reliable proxy for human evaluation even in open-source settings.

Implications for Future Model Development. Given the results in Section 5, we envision this benchmark not only as a means of comparing LLM performance, but also as a diagnostic tool for guiding the development of Indonesian and regional language models. Our experiments identify several aspects across three complementary tasks, including large performance gaps in local language generation, improvements from culturally grounded pretraining, and gains from supervised fine-tuning. Moving forward, the benchmark can therefore serve as a testbed for measuring whether such developed models have better cultural and linguistic alignment in Indonesian contexts, particularly in conversational settings where cultural exchange naturally occurs.

7 Conclusion

In this work, we present **CULTURETALK-ID**, a human-curated, culturally grounded dialogue dataset covering Indonesian and 10 province-specific cultural contexts. The dataset consists of 4,496 dialogues and is designed to benchmark language models' ability to understand and generate culturally nuanced interactions across regions. We ensure data quality through a multi-stage pipeline and incorporate human annotators. We benchmark a range of proprietary and open-source models on three tasks: dialogue-MCQ cultural commonsense, machine translation, and language steering.

Our results show that while proprietary models consistently achieve strong performance, open-source models still struggle, particularly in generation tasks such as machine translation and language steering. These results highlight a gap in culturally grounded understanding and generation capabilities in dialogue, suggesting the need for future work in incorporating richer Indonesian cultural context and broader local language coverage into model pretraining and adaptation.

Limitations

While our dataset is the first to cover cultural contexts in Indonesian and local Indonesian language dialogues, it still does not represent all provinces in Indonesia. Indonesia consists of 38 provinces with highly diverse cultural backgrounds, and several regions and cultural variations remain underrepresented in our data [4]. In addition, the topic distribution is not entirely balanced. Some domains, such as 'agriculture', 'farm', and 'game', are less represented compared to other topics, although the imbalance is relatively moderate. The difficulty level of dialogues across provinces may also be imbalanced, as some provinces may require less complex cultural reasoning to answer correctly, which could influence the resulting scores. Future work can expand both the topical coverage and the regional diversity represented in the benchmark.

For the human evaluation correlation in the MT experiments, we were only able to conduct detailed correlation analyses between LLM-as-a-judge and the human evaluation for Javanese and Minangkabau. Although our dataset creation effort involved 22 annotators from 11 provinces, recruiting qualified annotators for many local languages remains challenging. This is partly due to the differences in local language standardization, as discussed in Novitasari et al. (2020), and the speaker availability. Nevertheless, the results for Javanese and Minangkabau show reasonably strong correlations, particularly for the proprietary models.

Ethical Consideration

This dataset is released under the CC BY-NC-SA 4.0 license. The dataset is intended for benchmarking large language model performance on Indonesian cultural understanding in dialogue contexts. **We allow the dataset to be used as training data, provided that only the designated training split is used and that the usage is strictly non-commercial.** Users should also be aware that models trained on this dataset may inherit biases toward the languages and cultural groups represented in the data, as our benchmark does not cover the full diversity of Indonesian cultures, as discussed in the Limitations section.

Additionally, all 22 annotators were compensated based on agreements established before the annotation process. The compensation was considered appropriate within the Indonesian economic context and proportional to the workload involved. The annotators were informed that the dataset would be publicly released. Furthermore, the dataset does not contain personally identifiable information, as the annotations were conducted

[4] We nevertheless believe it is preferable to include these regions, even with limited data, rather than omit them entirely, especially since evaluation resources for languages such as Uab Meto in NTT are particularly scarce.

on third-person point-of-view conversations rather than personal user interactions.

Acknowledgements

We thank all reviewers for their valuable feedback during the rebuttal phase and the annotators for their help with the annotation process and their collaboration throughout the dataset creation. This research was supported by a grant from Mohamed bin Zayed University of Artificial Intelligence (MBZUAI).

References

Alham Fikri Aji and Trevor Cohn. 2025. LORAXBENCH: A multitask, multilingual benchmark suite for 20 Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17421–17446, Suzhou, China. Association for Computational Linguistics.

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.

Muhammad Dehan Al Kautsar, Aswin Candra, Muhammad Alif Al Hakim, Maxalmina Satria Kahfi, Fajri Koto, Alham Fikri Aji, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Genta Indra Winata. 2026. SeaDialogues: A Multilingual Culturally Grounded Multi-turn Dialogue Dataset on Southeast Asian Languages. In *WOCHAT2026: Workshop on Chatbots and Agentic Technologies*, pages 94–122.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, and 22 others. 2025. Sailor2: Sailing in south-east asia with inclusive multilingual llms. *Preprint*, arXiv:2502.12982.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32866–32895, Suzhou, China. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Muhammad Kautsar, Rahmah Nurdini, Samuel Cahyawijaya, Genta Winata, and Ayu Purwarianti. 2023. IndoToD: A multi-domain Indonesian benchmark for end-to-end task-oriented dialogue systems. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 85–99, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

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.

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

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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, and 2 others. 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 Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023. XDailyDialog: A multilingual parallel dialogue corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253, Toronto, Canada. Association for Computational Linguistics.

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.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Marek Muszyński. 2023. Attention checks and how to use them: Review and practical recommendations. *Ask: Research and Methods*.

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, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. Sea-lion: Southeast asian languages in one network. *Preprint*, arXiv:2504.05747.

Minh-Hoang Nguyen, Viet-Phuong La, Tam-Tri Le, and Quan-Hoang Vuong. 2022. Introduction to bayesian mindsponge framework analytics: An innovative method for social and psychological research. *MethodsX*, 9:101808.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138, Marseille, France. European Language Resources association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vynska Amalia Permadi, Xingwei Tan, Nafise Sadat Moosavi, and Nikolaos Aletras. 2026. No shortcuts to culture: Indonesian multi-hop question answering for complex cultural understanding. *Transactions of the Association for Computational Linguistics*, 14:1243–1265.

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

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified metadata for thousands of languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

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.

Suraj Sharma and Leigh Anne Liu. 2024. Culture and communication. In Michele J. Gelfand and Miriam Erez, editors, *The Oxford Handbook of Cross-Cultural Organizational Behavior*. Oxford University Press.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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.

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.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. SocialCC: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022. Mdia: A benchmark for multilingual dialogue generation in 46 languages. *Preprint*, arXiv:2208.13078.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Dataset Creation Additional Details

As mentioned, we repurpose the IndoCulture (Koto et al., 2024) and COPAL-ID (Wibowo et al., 2024) as both are grounded in Indonesian cultural contexts and share a similar structure: multiple-choice questions based on culturally situated scenarios, where each instance consists of one premise and three answer options with a single correct answer. This structure makes them well-suited for transformation into a unified dialogue dataset. Notably, IndoCulture contains both province-specific and non-province-specific instances, whereas COPAL-ID focuses specifically on Jakarta-related cultural contexts. Using GPT-5, we generate dialogues by treating the premise as the dialogue context and the three answer options as candidate final utterances, which are then passed to annotators for validation and refinement. The prompt used for dialogue generation is shown in Figure 3, while the transformation process is illustrated in Figure 4.

> You are a dialogue creation agent, and you have the task of creating a daily-life dialogue in Indonesian between two people, with a maximum length of 6 sentences. The dialogue must contain cultural elements from the region of {province}, and relate to the topic of {topic}. Within the dialogue, there must be an implied {context}, and it should end with one 'correct utterance' that matches the correct {answer} and two 'incorrect utterances' from all of the options in {option} except the answer. Do not mention the answer nor state it explicitly within the dialogue, as it must only occur on the one 'correct utterance'! It is also important that the relationship between the context and the answer reflects commonsense knowledge that is typical of the region.
>
> Please create a dialogue based on these variables using markdown format, so it can be extracted easily in the postprocess.
>
> ==FEW SHOT EXAMPLES==
>
> {3_fewshot_examples}
>
> ==DIALOGUE==
>
> {dialogue}

Figure 3: Prompt for the dialogue generation.

B Detailed Dataset Statistics

Table 6 and Table 7 present detailed dataset statistics by region and topic, respectively. The dataset covers 10 provinces and one general Indonesian cultural category, with the largest number of dialogues originating from the Minangkabau culture in West Sumatra and the fewest from NTT. Across topics, food-related dialogues are the most common, while agriculture-related dialogues are the least represented. We also observe that local-language dialogues generally contain higher word counts and lexical diversity across both regions and topics.

C CULTURETALK-ID Tasks Prompt

We present the prompts used for evaluation in Figure 5, 6, 7, 8, and 9. To ensure fairness, we use the same system and user prompts across all models for each task. For the MCQ task, province and language information can be optionally included to control the amount of contextual information provided, resulting in the different settings reported in Table 2. For open-source models, we adopt a likelihood-based evaluation setup by computing the likelihood of each answer option (A, B, or C) and selecting the highest-probability choice. This approach provides a more stable evaluation than generation-based prompting, which can be highly sensitive to prompt variations, especially for smaller models.

For the machine translation and language steering tasks, the same prompt templates are used across all model categories, with only the language and province variables adjusted according to the target setting.

D Language Covered by CULTURETALK-ID

We provide details of the covered provinces, languages, and estimated speaker populations in Table 8. Our benchmark covers 10 provinces and one general Indonesian cultural category. Javanese is further divided into three dialects: Dermayu, primarily spoken in eastern West Java; Mataraman, spoken in Central Java; and Arekan, spoken in Surabaya (East Java) and surrounding regions.

The estimated speaker populations highlight the importance of evaluating cultural understanding in LLMs for these languages, many of which have millions of speakers. Despite this, language technology support for several local languages remains limited, and some languages, such as Acehnese, Minangkabau, and Wamesa, are considered endangered languages.

E MCQ Evaluation Details

E.1 MCQ Scores

The more detailed dialogue-based MCQ cultural commonsense results are presented in Table 9, while scores across provinces and topics are in

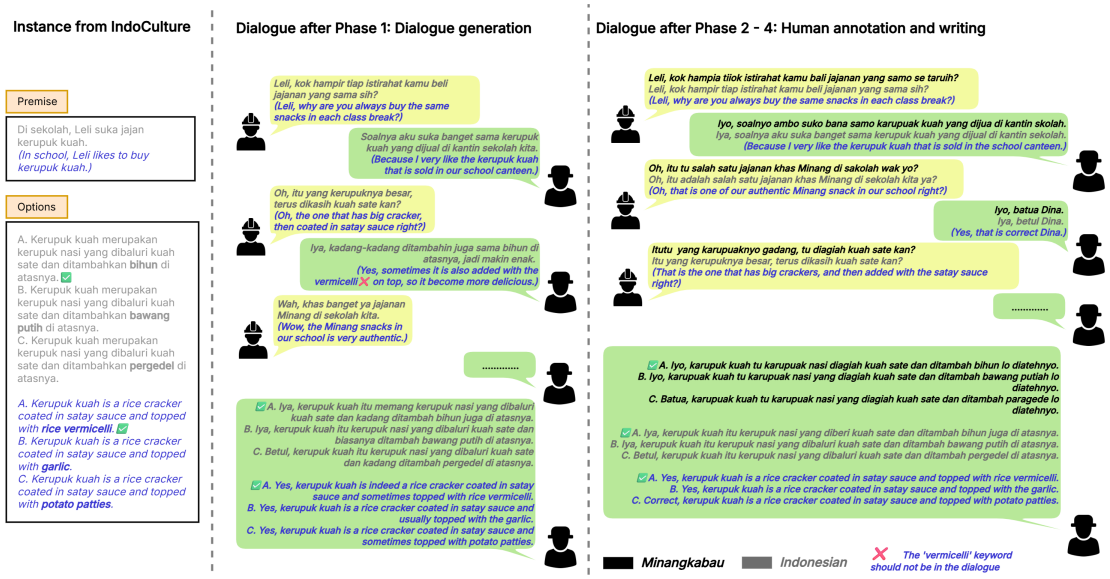


Figure 4: An example of the dataset transformation process from an original IndoCulture instance into **CULTURETALK-ID** across each phase. This example uses the Minangkabau cultural setting.

Metric	Indonesia	Aceh	Bali	W. Java	C. Java	E. Java	S. Kalimantan	NTT	W. Papua	S. Sulawesi	W. Sumatera
General Dialogue Statistics											
#dialogues	500	492	482	462	466	342	466	34	506	148	598
#region specific dialogues	0	348	368	268	324	228	390	28	446	134	446
Indonesian Data											
avg. words per dialogue	52.27	59.80	60.21	60.04	59.86	59.68	61.19	60.06	62.04	60.26	59.08
avg. utterances per dialogue	6.02	6.96	6.22	7.73	6.61	6.81	6.40	6.88	6.45	6.23	6.63
avg. words per utterance	7.72	7.73	8.72	7.00	8.15	7.90	8.63	7.85	8.70	8.71	8.03
#word counts	13,068	14,710	14,511	13,870	13,947	10,205	14,257	1,021	15,695	4,459	17,666
#unique word counts	2,895	2,936	2,974	3,157	2,867	2,563	2,958	481	3,063	1,383	3,636
Local Language Data											
avg. words per dialogue	52.27	63.52	65.12	61.21	60.85	60.69	62.68	67.94	62.53	65.85	59.70
avg. utterances per dialogue	6.02	6.96	6.22	7.73	6.61	6.81	6.40	6.88	6.45	6.23	6.63
avg. words per utterance	7.72	8.27	10.06	7.15	8.38	8.17	8.86	9.76	8.79	9.68	8.75
#word counts	13,068	15,626	15,695	14,140	14,177	10,378	14,604	1,155	15,819	4,873	17,849
#unique word counts	2,895	3,388	3,905	3,400	3,469	3,052	3,329	499	3,156	1,664	3,868

Table 6: Detailed dataset statistics, split by region. W, C, E, and S denote west, central, east, and south, respectively.

Table 10, 11, 12, and 13. Overall, Gemini-2.5-flash achieves the best performance on this benchmark, delivering consistent results across both settings, followed closely by other proprietary models, such as GPT-5.1 and Cohere-Command-A, which also demonstrate robust performance across all settings, including both province-specific and non-province-specific dialogue metrics.

Generic multilingual models lag substantially behind proprietary models, achieving average scores of 69.45 on Indonesian dialogues and 56.64 on local-language dialogues. In contrast, SEA-centric models consistently outperform generic multilingual baselines. Among them, Sahabat-AI-v1 Instruct achieves the strongest performance on Indonesian dialogues, trailing only the lowest-performing proprietary model (Cohere Command-A) by an average of 3.72 points.

We also observe that providing context improves performance across open-source model groups. For example, in the Indonesian dialogue setting, the average PS score for generic multilingual models increases by 2.17 with province context, whereas the ~PS score improves by only 0.89.

E.2 Effect of Pretraining in SEA Contexts

We further examine the effect of further pretraining the multilingual model in Indonesian-related context in Indonesian (see Figure 10, Table 10, and Table 12) and local language settings (see Table 11 and Table 13) by comparing the best-performing generic multilingual and SEA-centric models. The SEA-centric model shows improvements across all provinces and topics, even for languages not

Metric	Food	Wedding	Family	Birth	Death	Religion	Agriculture	Farm	Art	Game	Daily	Socio-religious
General Dialogue Statistics (By Topic)												
#dialogues	646	522	344	424	258	300	154	170	424	158	246	350
#country specific dialogues	544	426	246	264	162	186	106	126	398	134	146	242
Indonesian Data												
avg. words per dialogue	58.31	61.07	59.87	61.31	61.19	61.99	59.48	60.92	61.07	58.14	57.93	60.91
avg. utterances per dialogue	6.47	6.66	6.66	6.98	7.13	6.67	7.47	7.45	6.29	6.03	6.78	6.63
avg. words per utterance	8.09	8.29	8.10	7.93	7.75	8.40	7.16	7.38	8.76	8.67	7.67	8.29
#word counts	18,834	15,938	10,298	12,997	7,893	9,298	4,580	5,178	12,947	4,593	7,125	10,660
#unique word counts	3,183	2,875	2,434	2,592	1,725	2,021	1,308	1,570	2,690	1,316	1,992	2,732
Local Language Data												
avg. words per dialogue	61.40	63.78	62.16	63.10	62.77	63.41	61.43	62.34	62.30	59.18	59.41	62.54
avg. utterances per dialogue	6.35	6.50	6.53	6.97	7.12	6.57	7.38	7.32	6.21	5.92	6.69	6.60
avg. words per utterance	8.89	8.99	8.68	8.25	8.02	8.80	7.56	7.74	9.10	9.04	8.04	8.62
#word counts	19,831	16,646	10,691	13,378	8,097	9,511	4,730	5,299	13,207	4,675	7,307	10,944
#unique word counts	6,361	5,385	4,181	4,697	3,092	3,457	2,151	2,486	4,381	2,142	3,288	4,562

Table 7: Detailed dataset statistics by topic.

You are an Indonesian professional that able to reason in the Indonesian culture.

Rules:
1. Output only the OPTIONS [A/B/C].
2. Do not add explanations, comments, or quotation marks. Only the option label [A/B/C].

You are tasked with selecting the most culturally appropriate option based on the context provided below.

Location: {province}
Language: {language}

Conversation:
{dialogue}

Give the option label only [A/B/C].

Options:
{choices_str}

Answer:

Figure 5: Task 1 - Dialogue-based cultural commonsense MCQ Evaluation's prompt.

Province	Language Covered (code)	Estimated Speaker
West Java	Javanese - Dermayu dialect (*jav*)	
Central Java	Javanese - Mataraman dialect (*jav*)	91,000,000 *
East Java	Javanese - Arekan dialect (*jav*)	
Aceh	Acehnese (*ace*)	3,700,000 ***
West Sumatra	Minangkabau (*min*)	8,000,000 ***
Bali	Balinese (*ban*)	4,800,000 **
South Kalimantan	Banjarese (*bjn*)	4,000,000 *
South Sulawesi	Buginese (*bug*)	4,300,000 *
NTT	Uab Meto (*aoz*)	720,000 **
West Papua	Wamesa (*wad*)	5,000 ***
-	Indonesia (ind)	171,000,000 *

Table 8: Provinces, languages (along with their ISO-639 codes), estimated speaker populations, and language status covered in **CULTURETALK-ID**, based on (Ritchie et al., 2024). *Institutional language; **Stable language; ***Endangered language according to https://www.ethnologue.com/language/, accessed on 11 May 2026.

You are a professional Indonesian translation system, specialized in Indonesian and Indonesian local languages.

Rules:
1. Output only the translated text in Latin (both Indonesian and other local languages).
2. Do not add explanations, comments, labels, or quotation marks.
3. Preserve meaning, tone, and level of formality.
4. Do not change names, numbers, or entities.
5. If the input already matches the target variety, return it unchanged.

Extra Rules:
- Output ONLY the translation
- No explanation
- No reasoning (DO NOT EXPLAIN YOUR REASONING)
- No extra text

Figure 6: Task 2 - Machine translation's unified system prompt.

Translate the following Indonesian text into the {language}, the local language of {province}. Do not include any explanation, just return the {language} translation.

Indonesian Text: {indonesian_dialogue}

{language} Translation:

Figure 7: Task 2 - Machine translation's (Indonesian → local Indonesian) user prompt.

Translate the following {language} text from the {province} into Indonesian. Do not include any explanation, just return the Indonesian translation.

{language} Text: {local_dialogue}

Indonesian Translation:

Figure 8: Task 2 - Machine translation's (local Indonesian → Indonesian) user prompt.

You are a helpful assistant who writes in {target_name}{code_suffix}. Continue the dialogue with a single natural utterance in that variety.

Rules: output ONLY the continuation text (no analysis, no translation, no English explanation, no meta commentary, no speaker labels).

Do NOT output <think> blocks or any reasoning traces.

Figure 9: Task 3 - Language steering's user prompt.

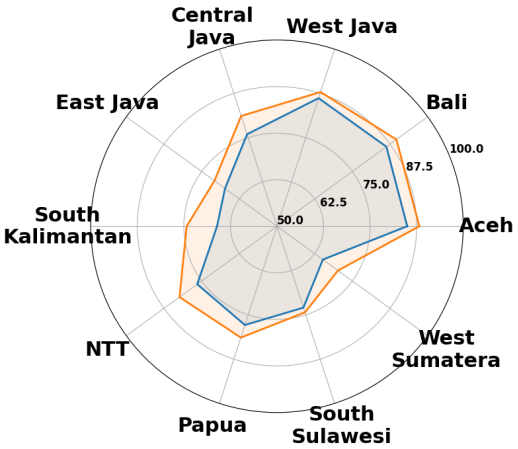


Figure 10: MCQ cultural commonsense evaluation score differences by province between *Gemma2-9B Instruct* (generic-multilingual model) and *Sahabat-AI-v1 Instruct* (SEA-centric model).

explicitly included in its training data.

F MT Evaluation Details

We present the full machine translation results in Table 14. Overall, models perform substantially better when translating from local Indonesian languages into Indonesian than in the reverse direction. This likely reflects the stronger representation and preservation of Indonesian in current language models, while many local Indonesian languages remain under-resourced. In the local-to-Indonesian setting, the best-performing model (GPT-5.1) achieves an overall score of 4.69 and a BLEU score of 50.54.

Consistent with the main results, generic multilingual models perform considerably worse than proprietary and SEA-centric models. For example, the strongest multilingual baseline, Gemma-2-9B Instruct, achieves only 20.54 BLEU and a 2.73 overall score in the Indonesian-to-local setting. SEA-centric models generally perform better, with Sailor2-8B Chat approaching Cohere-Command-A in the Indonesian-to-local direction, while Sahabat-AI-v1 Instruct achieves competitive performance in the local-to-Indonesian setting with an overall score of 3.98.

Supervised fine-tuning further improves performance for both Gemma-2-9B Instruct and Sahabat-AI-v1 Instruct across both translation directions, with especially large gains in local-to-Indonesian translation. We additionally report province-level translation results in Table 15, 16, 17, and 18. Across most models, Javanese tends to achieve stronger performance than other local languages, likely due to its relatively larger resource availability and speaker population.

G Language Steering Evaluation Details

As shown in Table 21, models generally perform well when generating Indonesian utterances. Even the lowest-performing model (Gemma-2-9B Instruct) produces Indonesian outputs correctly in 91.3% of cases, as measured by the Glot-LID metric. GPT-5.1 achieves the best performance, with an LLM-as-a-judge score of 4.65 on a 5-point Likert scale.

Consistent with the main results, model performance drops substantially when generating Indonesian local languages. For example, Sahabat-AI-v1 Instruct, which achieves an LLM-as-a-judge score of 4.15 and a 95.0% correct language generation rate in Indonesian, drops to 3.32 and 13.7%, respectively, in local language generation. We also observe that SFT improves the ability of open-source models to generate the correct local languages, although this comes with a slight degradation in Indonesian generation performance.

We further provide province-level results in Table 22. Overall, even strong proprietary models such as Gemini-3-Flash still struggle with languages spoken in Eastern Indonesia, including those from NTT, South Sulawesi, and Papua. In a good sign, models achieve relatively strong LLM-as-a-judge scores despite generating responses primarily in Indonesian rather than the target local language. It is also worth noting that the GLOT-LID score for Papua (Wamesa, '*wad*') is reported as '*NA*' because Wamesa is not covered in the GLOT-LID label set.

Model	Context: None			Context: Prov.			Context: Prov. + Lang.		
	Avg	PS	~PS	Avg	PS	~PS	Avg	PS	~PS
Indonesian Dialogue									
Proprietary Models									
GPT-5.1	89.23$_{(2)}$	86.44	94.72	89.86$_{(2)}$	87.18	95.12	89.95$_{(2)}$	87.38	94.99
Gemini-2.5-flash	**91.46**$_{(1)}$	**89.66**	**94.99**	**91.55**$_{(1)}$	**89.80**	**94.99**	**91.68**$_{(1)}$	**89.99**	**94.99**
Cohere-Command-A	84.74$_{(3)}$	80.74	92.61	85.54$_{(3)}$	82.28	91.95	85.01$_{(3)}$	81.41	92.08
average proprietary	*88.48*	*85.61*	*94.11*	*88.98*	*86.42*	*94.02*	*88.88*	*86.26*	*94.02*
Multilingual Models									
Qwen3-8B	68.95$_{(7)}$	66.64	73.84	70.82$_{(7)}$	68.52	75.33	70.28$_{(7)}$	67.99	74.80
Llama3.1-8B Instruct	62.63$_{(8)}$	60.13	67.55	65.44$_{(8)}$	63.62	69.00	65.44$_{(8)}$	63.89	68.74
Gemma2-9B Instruct	**76.78**$_{(6)}$	**72.82**	**84.56**	**77.45**$_{(6)}$	**73.96**	**84.30**	**77.71**$_{(6)}$	**74.09**	**84.83**
average multilingual	*69.45*	*66.53*	*75.32*	*71.24*	*68.70*	*76.21*	*71.14*	*68.66*	*76.12*
SEA-Centric Models									
Sailor2-8B Chat	80.07$_{(5)}$	76.44	87.20	81.01$_{(5)}$	77.45	**87.99**	81.32$_{(5)}$	77.79	**88.26**
SEA-LION-v3.5-8B	36.57$_{(9)}$	36.85	36.02	36.57$_{(9)}$	36.85	36.02	36.57$_{(9)}$	36.85	36.02
Sahabat-AI-v1 Instruct	**80.87**$_{(4)}$	**77.25**	**87.99**	**81.72**$_{(4)}$	**78.99**	87.07	**81.54**$_{(4)}$	**78.52**	87.47
*average SEA-centric**	*80.47*	*76.85*	*87.60*	*81.37*	*78.22*	*87.53*	*81.43*	*78.16*	*87.87*
Local Indonesian Dialogue									
Proprietary Models									
GPT-5.1	82.31$_{(2)}$	79.07	88.67	82.52$_{(2)}$	78.86	89.71	82.70$_{(2)}$	79.26	89.45
Gemini-2.5-flash	**87.77**$_{(1)}$	**85.23**	**92.74**	**87.90**$_{(1)}$	**85.70**	**92.22**	**87.72**$_{(1)}$	**85.23**	**92.61**
Cohere-Command-A	75.09$_{(3)}$	70.07	84.96	75.67$_{(3)}$	71.21	84.43	74.64$_{(3)}$	70.00	83.77
average proprietary	*81.72*	*78.12*	*88.79*	*82.03*	*78.59*	*88.79*	*81.69*	*78.16*	*88.61*
Multilingual Models									
Qwen3-8B	52.89$_{(7)}$	48.86	60.82	55.29$_{(7)}$	51.28	63.19	54.72$_{(7)}$	50.74	62.53
Llama3.1-8B Instruct	51.87$_{(8)}$	49.60	56.33	54.09$_{(8)}$	52.55	57.12	53.87$_{(8)}$	51.95	57.65
Gemma2-9B Instruct	**65.17**$_{(6)}$	**59.80**	**75.73**	**66.01**$_{(6)}$	**61.21**	**75.46**	**65.88**$_{(6)}$	**60.87**	**75.73**
average multilingual	*56.64*	*52.75*	*64.29*	*58.46*	*55.01*	*65.26*	*58.15*	*54.52*	*65.30*
SEA-Centric Models									
Sailor2-8B Chat	**70.15**$_{(4)}$	**64.50**	**81.27**	**71.44**$_{(4)}$	**66.24**	**81.66**	**72.02**$_{(4)}$	**66.85**	**82.19**
SEA-LION-v3.5-8B-R	36.57$_{(9)}$	36.85	36.02	36.57$_{(9)}$	36.85	36.02	36.57$_{(9)}$	36.85	36.02
Sahabat-AI-v1 Instruct	69.08$_{(5)}$	63.62	79.82	69.80$_{(5)}$	64.77	79.68	69.93$_{(5)}$	65.17	79.29
*average SEA-centric**	*69.62*	*63.56*	*80.55*	*70.62*	*65.51*	*80.67*	*70.98*	*66.01*	*80.71*

Table 9: Model performance in **dialogue-based cultural commonsense MCQ** under three context settings: no context, province context, and province plus language context. Scores are reported as average (Avg), province-specific (PS), and non-province-specific (∼PS). (*) The SEA-LION-v3.5-8B-R scores are excluded.

Model	Province	Avg	CS	∼CS
Gemini-2.5 Flash	Aceh	92.68	90.80	97.22
	Bali	95.44	95.65	94.74
	W. Java	96.10	97.01	94.85
	C. Java	92.27	91.36	94.37
	E. Java	87.72	83.33	96.49
	S. Kalimantan	88.41	87.69	92.11
	NTT	100.00	100.00	100.00
	W. Papua	92.89	91.93	100.00
	S. Sulawesi	91.89	91.04	100.00
	W. Sumatera	81.94	79.82	88.16
	Indonesia	96.00	–	–
Gemma-2 9B Instruct	Aceh	84.96	81.61	93.06
	Bali	86.31	84.78	91.23
	W. Java	86.15	84.33	88.66
	C. Java	75.97	72.84	83.10
	E. Java	67.25	58.77	84.21
	S. Kalimantan	66.09	63.08	81.58
	NTT	76.47	78.57	66.67
	W. Papua	77.87	75.78	93.33
	S. Sulawesi	72.97	73.13	71.43
	W. Sumatera	65.22	61.43	76.32
	Indonesia	82.00	–	–
Sahabat-AI-v1 Instruct	Aceh	88.21	84.48	97.22
	Bali	89.63	88.04	94.74
	W. Java	87.88	88.81	86.60
	C. Java	81.12	78.40	87.32
	E. Java	70.76	59.65	92.98
	S. Kalimantan	74.25	72.31	84.21
	NTT	82.35	78.57	100.00
	W. Papua	81.42	80.27	90.00
	S. Sulawesi	74.32	74.63	71.43
	W. Sumatera	70.23	65.92	82.89
	Indonesia	85.60	–	–

Table 10: Performance of dialogue-based cultural commonsense MCQ in Indonesian by model across provinces. Scores reported as average (Avg), country-specific (CS), and non-country-specific (∼CS). W, C, E, and S denote west, central, east, and south, respectively.

Model	Province	Avg	CS	∼CS
Gemini-2.5 Flash	Aceh	91.46	90.80	93.06
	Bali	95.44	96.20	92.98
	W. Java	93.94	94.78	92.78
	C. Java	94.85	93.83	97.18
	E. Java	87.13	83.33	94.74
	S. Kalimantan	85.41	83.59	94.74
	NTT	35.29	35.71	33.33
	W. Papua	68.77	69.06	66.67
	S. Sulawesi	83.78	85.07	71.43
	W. Sumatera	83.28	81.61	88.16
	Indonesia	96.00	–	–
Gemma-2 9B Instruct	Aceh	67.89	64.37	76.39
	Bali	72.61	69.02	84.21
	W. Java	79.65	78.36	81.44
	C. Java	71.67	74.07	66.20
	E. Java	59.65	52.63	73.68
	S. Kalimantan	61.37	58.46	76.32
	NTT	29.41	28.57	33.33
	W. Papua	50.99	50.22	56.67
	S. Sulawesi	37.84	38.81	28.57
	W. Sumatera	53.51	49.78	64.47
	Indonesia	82.00	–	–
Sailor2-8B Chat	Aceh	69.51	64.94	80.56
	Bali	79.67	80.43	77.19
	W. Java	83.12	82.84	83.51
	C. Java	76.39	73.46	83.10
	E. Java	71.35	63.16	87.72
	S. Kalimantan	64.81	61.54	81.58
	NTT	58.82	64.29	33.33
	W. Papua	51.38	50.22	60.00
	S. Sulawesi	41.89	41.79	42.86
	W. Sumatera	61.20	57.85	71.05
	Indonesia	86.80	–	–

Table 11: Performance of dialogue-based cultural commonsense MCQ in Indonesian local languages by model across provinces. Scores reported as average (Avg), country-specific (CS), and non-country-specific (∼CS). W, C, E, and S denote west, central, east, and south, respectively.

Model	Topic	Avg	CS	∼CS
	Food	91.95	92.28	90.20
	Wedding	86.97	85.92	91.67
	Family	87.79	86.99	89.80
	Birth	90.09	87.12	95.00
	Death	89.15	86.42	93.75
Gemini-2.5	Religion	97.33	96.77	98.25
Flash	Agriculture	94.81	94.34	95.83
	Farm	88.24	84.13	100.00
	Art	91.98	91.46	100.00
	Game	92.41	92.54	91.67
	Daily	93.50	93.15	94.00
	Socio-religious	90.29	86.78	98.15
	Food	73.07	72.79	74.51
	Wedding	70.88	67.61	85.42
	Family	76.16	72.36	85.71
	Birth	78.77	72.73	88.75
	Death	76.74	70.37	87.50
Gemma-2 9B	Religion	84.00	87.10	78.95
Instruct	Agriculture	88.31	84.91	95.83
	Farm	70.59	65.08	86.36
	Art	74.53	73.37	92.31
	Game	73.42	73.13	75.00
	Daily	80.49	72.60	92.00
	Socio-religious	76.57	71.07	88.89
	Food	78.64	77.94	82.35
	Wedding	76.63	74.18	87.50
	Family	81.98	78.86	89.80
	Birth	80.19	74.24	90.00
	Death	80.62	76.54	87.50
Sahabat-AI-v1	Religion	86.00	86.02	85.96
Instruct	Agriculture	88.31	84.91	95.83
	Farm	76.47	71.43	90.91
	Art	78.30	77.39	92.31
	Game	73.42	73.13	75.00
	Daily	84.55	78.08	94.00
	Socio-religious	82.86	77.69	94.44

Table 12: Performance of dialogue-based cultural commonsense MCQ in Indonesian by model across topics. Scores reported as average (Avg), country-specific (CS), and non-country-specific (∼CS).

Model	Topic	Avg	CS	∼CS
	Food	84.21	84.19	84.31
	Wedding	82.76	82.16	85.42
	Family	84.88	82.11	91.84
	Birth	87.26	83.33	93.75
	Death	85.27	83.95	87.50
Gemini-2.5	Religion	92.67	92.47	92.98
Flash	Agriculture	90.91	90.57	91.67
	Farm	83.53	80.95	90.91
	Art	92.45	91.96	100.00
	Game	86.08	85.07	91.67
	Daily	89.43	87.67	92.00
	Socio-religious	85.14	80.99	94.44
	Food	52.32	51.47	56.86
	Wedding	64.75	61.97	77.08
	Family	60.47	54.47	75.51
	Birth	62.74	56.06	73.75
	Death	67.44	62.96	75.00
Gemma-2 9B	Religion	74.00	74.19	73.68
Instruct	Agriculture	72.73	75.47	66.67
	Farm	57.65	50.79	77.27
	Art	64.62	62.81	92.31
	Game	55.70	55.22	58.33
	Daily	67.48	63.01	74.00
	Socio-religious	67.43	64.46	74.07
	Food	63.16	62.87	64.71
	Wedding	64.37	61.97	75.00
	Family	66.28	60.98	79.59
	Birth	66.98	58.33	81.25
	Death	66.67	65.43	68.75
Sailor2-8B	Religion	77.33	72.04	85.96
Chat	Agriculture	80.52	79.25	83.33
	Farm	67.06	61.90	81.82
	Art	74.06	73.37	84.62
	Game	67.09	64.18	83.33
	Daily	66.67	58.90	78.00
	Socio-religious	68.00	60.33	85.19

Table 13: Performance of dialogue-based cultural commonsense MCQ in Indonesian local languages by model across topics. Scores reported as average (Avg), country-specific (CS), and non-country-specific (∼CS).

Model	BLEU Scores	BERTScore	LLM as a Judge (1–5)				
	B4	F1	adequacy	fluency	register	terminology	overall
Indonesian → local Indonesian languages							
Proprietary Models (0-shot)							
GPT-5.1	27.53	84.02	4.35	3.80	3.75	4.33	3.87 (2)
Gemini-2.5-flash	**28.36**	**84.33**	**4.48**	**3.91**	**3.91**	**4.48**	**4.00** (1)
Cohere-Command-A	21.45	81.40	3.35	2.82	2.69	3.29	2.93 (5)
Multilingual Models (0-shot)							
Qwen3-8B	19.78 (±0.1)	78.74 (±0.1)	2.82	2.11	1.90	2.73	2.22 (11)
Llama-3.1-8B Instruct	17.62 (±0.1)	78.71 (±0.1)	2.79	2.27	2.21	2.81	2.49 (10)
Gemma-2-9b Instruct	**20.54** (±0.1)	**80.15** (±0.1)	**3.20**	**2.49**	**2.40**	**3.11**	**2.73** (7)
SEA-Centric Models (0-shot)							
Sailor2-8B Chat	15.44 (±0.1)	77.58 (±0.1)	3.38	**2.85**	**2.65**	**3.42**	**2.94** (4)
SEA-LION-v3.5-8B	19.34 (±0.1)	78.98 (±0.1)	3.18	2.61	2.44	3.11	2.72 (9)
Sahabat-AI-v1 Instruct	**21.70** (±0.0)	**80.51** (±0.0)	**3.57**	2.63	2.46	3.21	2.80 (6)
SFT Models							
Gemma-2-9b Instruct	⇑ 22.60 (±0.0)	⇓ 79.39 (±0.0)	⇓ 3.16	⇑ 2.54	⇑ 2.46	⇑ 3.14	≈ 2.73 (7)
Sahabat-AI-v1 Instruct	⇑ **24.56** (±0.0)	⇓ **80.30** (±0.0)	⇓ **3.53**	⇑ **3.03**	⇑ **2.97**	⇑ **3.55**	⇑ **3.14** (3)
local Indonesian → Indonesian languages							
Proprietary Models (0-shot)							
GPT-5.1	50.54	90.78	**4.64**	**4.96**	**4.88**	**4.70**	**4.69** (1)
Gemini-2.5-flash	**52.74**	**91.46**	4.60	**4.96**	4.85	4.68	4.67 (2)
Cohere-Command-A	42.00	88.41	3.96	4.85	4.52	4.17	4.17 (4)
Multilingual Models (0-shot)							
Qwen3-8B	24.75 (±0.3)	80.25 (±0.2)	2.53	2.80	2.70	2.65	2.54 (11)
Llama-3.1-8B Instruct	26.60 (±0.2)	83.71 (±0.1)	2.66	3.61	3.19	2.93	2.89 (10)
Gemma-2-9b Instruct	**30.41** (±0.1)	**84.46** (±0.1)	**3.30**	**4.14**	**3.74**	**3.49**	**3.47** (7)
SEA-Centric Models (0-shot)							
Sailor2-8B Chat	22.84 (±0.1)	80.89 (±0.1)	3.25	4.14	3.68	3.41	3.40 (8)
SEA-LION-v3.5-8B	31.66 (±0.0)	84.87 (±0.0)	3.13	4.00	3.58	3.31	3.34 (9)
Sahabat-AI-v1 Instruct	**39.20** (±0.0)	**87.41** (±0.0)	**3.80**	**4.61**	**4.28**	**3.96**	**3.98** (5)
SFT Models							
Gemma-2-9b Instruct	⇑ 45.79 (±0.0)	⇑ 89.45 (±0.1)	⇑ 3.69	⇑ 4.81	⇑ 4.31	⇑ 3.87	⇑ 3.93 (6)
Sahabat-AI-v1 Instruct	⇑ **49.81** (±0.0)	⇑ **90.39** (±0.0)	⇑ **3.99**	⇑ **4.93**	⇑ **4.51**	⇑ **4.14**	⇑ **4.20** (3)

Table 14: Machine translation results for **Indonesian → local** (top) and **local → Indonesian** (bottom) under the **Province + Language** context. We report BLEU-4, BERTScore F1, and LLM-as-a-judge scores for adequacy, fluency, register, terminology, and overall quality.

Province	BLEU Scores				BERTScore			LLM as a judge (1–5)				
	B1	B2	B3	B4	P	R	F1	adequacy	fluency	register	terminology	overall
Gemini-3-Flash (reasoning none)												
Aceh	0.39	0.25	0.16	0.09	0.78	0.80	0.79	4.67	4.10	4.26	4.68	4.26
Bali	0.32	0.19	0.11	0.07	0.80	0.81	0.80	4.61	3.99	3.99	4.67	4.02
West Java	0.51	0.38	0.28	0.19	0.84	0.85	0.84	4.78	3.97	3.70	4.75	3.99
Central Java	0.59	0.47	0.38	0.29	0.87	0.87	0.87	**4.86**	**4.29**	4.11	**4.83**	**4.30**
East Java	0.57	0.44	0.33	0.25	0.86	0.86	0.86	4.80	4.13	4.03	4.77	4.15
South Kalimantan	0.62	0.49	0.39	0.30	0.87	0.88	0.87	4.74	4.09	4.13	4.76	4.15
NTT	0.20	0.09	0.04	0.02	0.74	0.74	0.74	4.00	3.00	3.40	3.80	3.47
Papua	0.14	0.07	0.03	0.01	0.69	0.69	0.69	2.32	1.87	1.96	2.32	2.18
South Sulawesi	0.27	0.16	0.08	0.04	0.78	0.77	0.78	4.25	3.61	3.73	4.21	3.76
West Sumatera	**0.68**	**0.57**	**0.48**	**0.40**	**0.89**	**0.90**	**0.90**	4.80	4.07	**4.14**	4.80	4.14
Gemma-2-9B Instruct												
Aceh	0.33	0.21	0.12	0.07	0.77	0.76	0.77	2.62	1.89	1.89	2.47	2.30
Bali	0.21	0.11	0.06	0.03	0.75	0.74	0.75	2.52	1.87	1.90	2.55	2.33
West Java	**0.50**	**0.37**	**0.26**	**0.18**	**0.84**	**0.83**	**0.84**	3.49	2.53	2.41	3.26	2.85
Central Java	0.48	0.35	0.24	0.16	0.83	0.81	0.82	3.47	2.61	2.53	3.26	2.89
East Java	0.45	0.31	0.21	0.13	0.82	0.81	0.82	3.62	2.68	2.58	3.49	3.03
South Kalimantan	0.48	0.34	0.24	0.16	0.82	0.80	0.81	3.49	2.07	1.66	3.20	2.30
NTT	0.09	0.04	0.02	0.01	0.65	0.64	0.64	1.27	1.07	1.07	1.40	1.13
Papua	0.15	0.07	0.03	0.02	0.66	0.65	0.65	1.39	1.12	1.07	1.37	1.18
South Sulawesi	0.16	0.09	0.04	0.02	0.69	0.66	0.68	2.13	1.61	1.58	2.12	1.87
West Sumatera	0.46	0.32	0.22	0.14	0.83	0.82	0.82	**3.76**	**2.93**	**2.88**	**3.81**	**3.14**
Sahabat-AI-v1 Instruct												
Aceh	0.34	0.22	0.14	0.08	0.77	0.75	0.76	2.80	1.26	1.04	1.88	1.55
Bali	0.22	0.12	0.07	0.03	0.75	0.73	0.74	3.18	1.83	1.66	2.60	2.33
West Java	0.53	**0.40**	**0.30**	0.21	**0.85**	**0.84**	**0.85**	4.57	3.69	3.52	4.49	3.73
Central Java	**0.54**	**0.40**	**0.30**	**0.22**	**0.85**	0.83	0.84	4.49	3.67	3.50	4.40	3.70
East Java	0.50	0.36	0.26	0.18	0.84	0.83	0.83	**4.58**	**3.71**	**3.57**	**4.51**	**3.75**
South Kalimantan	0.45	0.31	0.21	0.14	0.82	0.81	0.81	3.49	1.64	1.28	2.37	1.99
NTT	0.08	0.05	0.02	0.01	0.66	0.65	0.66	1.67	1.20	1.00	1.60	1.27
Papua	0.17	0.09	0.04	0.02	0.68	0.65	0.66	1.29	1.09	1.01	1.18	1.09
South Sulawesi	0.17	0.10	0.04	0.02	0.69	0.65	0.67	2.00	1.24	1.09	1.67	1.33
West Sumatera	0.47	0.33	0.23	0.16	0.83	0.82	0.82	3.79	2.66	2.45	3.52	2.92

Table 15: Translation results for **Indonesian** → **local** by provinces. We compare the best proprietary model (Gemini-2.5-flash), best multilingual model (Gemma-2-9b-it), and best SEA-centric model (Sahabat-AI-v1 Instruct) with zero shot using BLEU, BERTScore, and LLM-as-a-judge metrics.

Province	BLEU Scores				BERTScore			LLM as a judge (1–5)				
	B1	**B2**	**B3**	**B4**	**P**	**R**	**F1**	**adequacy**	**fluency**	**register**	**terminology**	**overall**
Gemma-2-9B Instruct (SFT)												
Aceh	0.34	0.22	0.14	0.09	0.75	0.75	0.75	2.32	1.64	1.74	2.38	2.09
Bali	0.24	0.15	0.09	0.05	0.73	0.72	0.72	2.64	1.89	1.85	2.62	2.26
West Java	0.51	0.39	0.29	0.20	0.84	0.83	0.83	3.23	2.48	2.37	3.12	2.71
Central Java	0.50	0.37	0.27	0.19	0.83	0.82	0.82	3.25	2.53	2.44	3.22	2.80
East Java	0.46	0.32	0.22	0.15	0.82	0.82	0.82	3.25	2.54	2.44	3.31	2.75
South Kalimantan	0.53	0.39	0.28	0.20	0.83	0.81	0.82	3.75	2.64	2.26	3.58	2.70
NTT	0.04	0.02	0.01	0.01	0.57	0.57	0.57	1.40	1.07	1.07	1.40	1.27
Papua	0.11	0.06	0.03	0.02	0.62	0.61	0.62	1.31	1.11	1.03	1.24	1.12
South Sulawesi	0.05	0.03	0.01	0.01	0.58	0.57	0.57	1.57	1.22	1.22	1.52	1.37
West Sumatera	**0.55**	**0.41**	**0.30**	**0.22**	**0.85**	**0.84**	**0.84**	**4.23**	**3.45**	**3.47**	**4.26**	**3.54**
Sahabat-AI-v1 Instruct (SFT)												
Aceh	0.37	0.24	0.16	0.10	0.78	0.76	0.77	2.37	1.76	1.86	2.43	2.14
Bali	0.27	0.16	0.10	0.06	0.74	0.73	0.73	2.60	1.98	1.99	2.64	2.35
West Java	0.52	0.38	0.28	0.19	0.85	0.84	0.85	4.42	3.76	3.69	4.42	3.80
Central Java	**0.57**	**0.45**	**0.35**	**0.27**	**0.86**	0.84	0.85	**4.53**	3.90	3.80	4.51	3.92
East Java	0.55	0.42	0.31	0.23	0.85	0.84	0.85	4.49	**4.03**	**3.92**	**4.56**	**4.03**
South Kalimantan	0.54	0.40	0.30	0.21	0.84	0.82	0.83	4.13	3.28	2.99	4.08	3.23
NTT	0.04	0.02	0.01	0.00	0.59	0.60	0.59	1.33	1.13	1.13	1.40	1.27
Papua	0.09	0.05	0.02	0.01	0.59	0.60	0.59	1.46	1.22	1.21	1.40	1.30
South Sulawesi	0.05	0.03	0.01	0.01	0.57	0.57	0.57	1.51	1.19	1.21	1.51	1.30
West Sumatera	**0.57**	0.44	0.34	0.25	**0.86**	**0.85**	**0.86**	3.92	3.38	3.42	4.02	3.49

Table 16: Translation results for **Indonesian** → **local** under **SFT models**, broken down by provinces. We compare the best multilingual model (Gemma-2-9b-it) and SEA-centric model (Sahabat-AI-v1 Instruct) using BLEU, BERTScore, and LLM-as-a-judge metrics.

Province	BLEU Scores				BERTScore			LLM as a judge (1–5)				
	B1	B2	B3	B4	P	R	F1	adequacy	fluency	register	terminology	overall
Gemini-3-Flash (reasoning none)												
Aceh	0.76	0.67	0.59	0.51	0.93	0.92	0.92	4.50	4.96	4.74	4.66	4.58
Bali	0.69	0.58	0.50	0.42	0.91	0.91	0.91	4.41	4.98	4.80	4.58	4.48
West Java	**0.82**	**0.75**	**0.69**	**0.63**	**0.94**	**0.93**	**0.94**	4.83	4.97	4.88	4.91	4.87
Central Java	0.78	0.70	0.63	0.57	0.93	**0.93**	0.93	4.92	4.97	4.91	**4.94**	**4.93**
East Java	0.73	0.63	0.54	0.46	0.91	0.91	0.91	**4.93**	4.95	4.89	**4.94**	4.90
South Kalimantan	0.80	0.72	0.65	0.59	0.93	0.92	0.93	4.80	4.97	**4.92**	4.85	4.84
NTT	0.39	0.25	0.16	0.10	0.83	0.82	0.82	3.53	4.53	4.27	3.53	3.53
Papua	0.35	0.21	0.13	0.09	0.80	0.79	0.79	3.55	4.93	4.68	3.82	3.86
South Sulawesi	0.61	0.48	0.38	0.30	0.87	0.87	0.87	3.88	4.96	4.78	3.88	3.93
West Sumatera	0.78	0.71	0.64	0.58	0.93	0.92	0.92	4.89	**4.99**	4.91	4.86	4.91
Gemma-2-9B Instruct												
Aceh	0.50	0.37	0.27	0.20	0.85	0.83	0.84	2.87	4.13	3.60	3.18	3.23
Bali	0.46	0.33	0.24	0.17	0.83	0.83	0.83	2.65	3.69	3.14	2.81	2.88
West Java	0.68	0.57	0.48	0.40	0.89	0.88	0.89	3.40	4.19	3.85	3.67	3.56
Central Java	**0.71**	**0.61**	**0.52**	**0.45**	**0.91**	**0.90**	**0.90**	**4.18**	**4.64**	**4.41**	**4.43**	**4.26**
East Java	0.66	0.54	0.44	0.36	0.89	0.88	0.88	3.89	4.42	4.08	4.10	3.92
South Kalimantan	0.67	0.56	0.47	0.40	0.89	0.88	0.88	3.71	4.37	4.11	3.89	3.81
NTT	0.20	0.10	0.04	0.02	0.76	0.73	0.74	2.00	3.60	3.07	2.20	2.53
Papua	0.25	0.13	0.07	0.04	0.76	0.75	0.76	2.41	4.12	3.57	2.61	2.89
South Sulawesi	0.28	0.16	0.09	0.05	0.76	0.76	0.76	1.46	3.37	2.52	1.72	2.00
West Sumatera	0.70	0.60	**0.52**	**0.45**	0.90	0.88	0.89	3.84	4.43	4.22	3.98	3.91
Sahabat-AI-v1 Instruct												
Aceh	0.53	0.39	0.30	0.22	0.85	0.84	0.85	3.23	4.51	4.03	3.48	3.51
Bali	0.48	0.34	0.25	0.18	0.84	0.83	0.83	2.86	4.09	3.49	3.03	3.17
West Java	0.74	0.65	0.57	0.49	0.91	0.90	0.91	4.17	4.64	4.44	4.31	4.25
Central Java	**0.75**	**0.66**	**0.58**	**0.51**	**0.92**	**0.91**	**0.92**	**4.65**	**4.90**	**4.76**	**4.74**	**4.67**
East Java	0.70	0.59	0.50	0.42	0.90	0.89	0.90	4.50	4.82	4.64	4.63	4.54
South Kalimantan	0.71	0.61	0.52	0.44	0.90	0.89	0.90	4.12	4.73	4.52	4.24	4.22
NTT	0.22	0.11	0.05	0.03	0.75	0.73	0.74	2.13	4.13	3.53	2.53	2.80
Papua	0.25	0.13	0.07	0.04	0.76	0.75	0.76	2.64	4.47	3.83	2.92	3.14
South Sulawesi	0.28	0.16	0.09	0.05	0.76	0.76	0.76	1.49	3.70	2.79	1.75	2.15
West Sumatera	0.73	0.63	0.55	0.48	0.91	0.90	0.90	4.25	4.79	4.57	4.37	4.34

Table 17: Translation results for **local** → **Indonesian** under the **0-shot setting**, broken down by provinces. We compare the best proprietary, multilingual, and SEA-centric models using BLEU, BERTScore, and LLM-as-a-judge metrics.

Province	BLEU Scores				BERTScore			LLM as a judge (1–5)				
	B1	B2	B3	B4	P	R	F1	adequacy	fluency	register	terminology	overall
Gemma-2-9B Instruct (SFT)												
Aceh	0.60	0.48	0.38	0.31	0.88	0.87	0.87	3.29	4.87	4.25	3.53	3.60
Bali	0.57	0.44	0.35	0.27	0.86	0.85	0.85	2.94	4.65	3.81	3.10	3.30
West Java	0.76	0.67	0.60	0.54	0.92	0.91	**0.92**	3.81	4.74	4.34	4.04	4.00
Central Java	**0.77**	**0.69**	**0.62**	**0.55**	**0.93**	**0.92**	**0.92**	**4.23**	4.81	**4.47**	**4.48**	**4.35**
East Java	0.72	0.61	0.52	0.44	0.91	0.90	0.91	3.94	4.75	4.33	4.14	4.07
South Kalimantan	0.76	0.67	0.60	0.53	0.92	0.91	**0.92**	4.02	**4.93**	4.46	4.20	4.18
NTT	0.31	0.17	0.08	0.04	0.79	0.77	0.78	2.13	4.73	3.80	2.27	2.87
Papua	0.33	0.19	0.11	0.07	0.80	0.78	0.79	2.61	4.79	4.00	2.88	3.18
South Sulawesi	0.35	0.21	0.11	0.06	0.80	0.78	0.79	1.70	4.55	3.51	1.97	2.61
West Sumatera	**0.77**	0.68	0.61	**0.55**	0.92	0.91	**0.92**	4.12	4.84	4.41	4.22	4.21
Sahabat-AI-v1 Instruct (SFT)												
Aceh	0.62	0.50	0.42	0.34	0.88	0.87	0.88	3.50	4.90	4.34	3.68	3.75
Bali	0.59	0.47	0.38	0.30	0.87	0.86	0.86	3.23	4.82	4.04	3.38	3.52
West Java	**0.82**	**0.75**	**0.69**	**0.63**	**0.94**	**0.93**	**0.94**	4.31	4.95	4.64	4.52	4.47
Central Java	0.81	0.73	0.67	0.61	**0.94**	**0.93**	**0.94**	**4.69**	**4.98**	**4.75**	**4.80**	**4.77**
East Java	0.76	0.66	0.58	0.50	0.92	0.92	0.92	4.61	4.91	4.71	4.70	4.67
South Kalimantan	0.80	0.71	0.65	0.59	0.93	0.92	0.93	4.28	4.96	4.65	4.47	4.41
NTT	0.29	0.17	0.10	0.06	0.78	0.76	0.77	2.00	4.80	3.60	2.40	2.87
Papua	0.34	0.21	0.12	0.07	0.80	0.79	0.79	2.75	4.89	4.15	2.98	3.30
South Sulawesi	0.39	0.25	0.15	0.09	0.80	0.79	0.80	2.03	4.76	3.78	2.30	2.82
West Sumatera	0.79	0.72	0.65	0.59	0.93	**0.93**	0.93	4.46	4.97	4.62	4.57	4.55

Table 18: Translation results for **local → Indonesian** under **SFT models**, broken down by provinces. We compare the best multilingual model (Gemma-2-9b-it) and SEA-centric model (Sahabat-AI-v1 Instruct) using BLEU, BERTScore, and LLM-as-a-judge metrics.

Indonesian → Local						
	Gemini-2.5-flash		Gemma-2-9B Instruct		Sailor2-8B Chat	
	MAD	Acc@1	MAD	Acc@1	MAD	Acc@1
Indonesian → Javanese	0.52	0.84	0.76	0.84	0.72	0.84
Indonesian → Minangkabau	0.28	0.96	1.08	0.71	1.00	0.72
Local → Indonesian						
	GPT-5.1		Gemma-2-9B Instruct		Sahabat-AI-v1 Instruct	
	MAD	Acc@1	MAD	Acc@1	MAD	Acc@1
Javanese → Indonesian	0.12	0.96	1.24	0.68	0.64	0.76
Minangkabau → Indonesian	0.04	1.00	0.83	0.79	1.24	0.76

Table 19: MAD and Acc@1 scores between LLM-as-a-judge and human evaluation in "adequacy" for Indonesian ↔ local translation across models.

Indonesian → Local						
	Gemini-2.5-flash		Gemma-2-9B Instruct		Sailor2-8B Chat	
	MAD	Acc@1	MAD	Acc@1	MAD	Acc@1
Indonesian → Javanese	0.80	0.92	0.80	0.84	1.12	0.76
Indonesian → Minangkabau	0.80	0.92	0.96	0.83	1.40	0.56
Local → Indonesian						
	GPT-5.1		Gemma-2-9B Instruct		Sahabat-AI-v1 Instruct	
	MAD	Acc@1	MAD	Acc@1	MAD	Acc@1
Javanese → Indonesian	0.40	0.96	0.92	0.76	0.60	0.88
Minangkabau → Indonesian	0.16	0.92	0.50	0.88	0.36	0.92

Table 20: MAD and Acc@1 scores between LLM-as-a-judge and human evaluation in "fluency" for Indonesian ↔ local translation across models.

Model	Local → Indonesian				Indonesian → Local			
	LLM	Glot-LID	BLEU	chrF	LLM	Glot-LID	BLEU	chrF
Proprietary Models								
GPT-5.1	**4.65**	**98.3**	1.80	**26.02**	4.42	73.9	0.64	20.57
Gemini-2.5-flash	4.62	97.8	2.25	25.28	**4.55**	**92.7**	**0.94**	**20.98**
Cohere-Command-A	4.37	97.9	**2.96**	25.23	3.70	31.6	0.62	18.35
Multilingual Models								
Qwen3-8B	**3.94** (±0.0)	98.7 (±0.1)	**1.75** (±0.0)	21.38 (±0.0)	**3.11** (±0.0)	1.5 (±0.2)	0.42 (±0.1)	15.79 (±0.0)
Llama-3.1-8B Instruct	3.80 (±0.0)	**99.4** (±0.1)	1.34 (±0.1)	**22.67** (±0.1)	2.91 (±0.0)	**9.2** (±0.4)	0.32 (±0.2)	**16.68** (±0.1)
Gemma-2-9b Instruct	3.78 (±0.0)	91.3 (±0.5)	1.62 (±0.1)	14.96 (±0.0)	3.00 (±0.0)	7.8 (±0.6)	**0.44** (±0.0)	10.92 (±0.1)
SEA-Centric Models								
Sailor2-8B Chat	3.93 (±0.0)	95.5 (±0.1)	0.67 (±0.0)	21.42 (±0.1)	3.19 (±0.0)	23.7 (±0.4)	0.23 (±0.0)	**17.43** (±0.1)
SEA-LION-v3.5-8B	3.82 (±0.0)	**99.1** (±0.2)	1.22 (±0.0)	**22.56** (±0.1)	3.01 (±0.0)	**27.5** (±0.7)	0.26 (±0.0)	16.64 (±0.1)
Sahabat-AI-v1 Instruct	**4.15** (±0.0)	95.0 (±0.5)	**3.13** (±0.1)	20.58 (±0.1)	**3.32** (±0.0)	13.7 (±1.0)	**0.64** (±0.2)	13.65 (±0.0)
SFT Models								
Qwen3-8B	⇓ 3.84 (±0.0)	⇓ **96.0** (±0.0)	⇑ **3.79** (±0.0)	⇑ **25.12** (±0.2)	⇓ 3.06 (±0.1)	⇑ 4.4 (±5.1)	⇑ 0.50 (±0.1)	⇑ **16.18** (±0.6)
Sahabat-AI-v1 Instruct	⇓ **4.12** (±0.0)	⇓ 94.6 (±0.0)	⇓ 3.08 (±0.1)	⇓ 20.47 (±0.1)	⇑ **3.33** (±0.1)	⇑ **18.2** (±8.9)	⇑ **0.77** (±0.0)	⇑ 14.74 (±1.8)

Table 21: Language steering results in **Indonesian** and **Local languages**. We report LLM-as-a-judge (denoted as LLM), Glot-LID, BLEU, and chrF scores.

Province	Gemini-3-Flash				Qwen3-8b				SEA-LION-v3.5-8B			
	LLM	Glot-LID	BLEU	chrF	LLM	Glot-LID	BLEU	chrF	LLM	Glot-LID	BLEU	chrF
Aceh	4.62	99.50	0.42	20.03	3.04	0.00	0.45	14.66	3.01	1.51	0.31	11.63
Bali	4.48	94.50	0.36	21.35	3.17	0.00	0.38	16.37	3.21	2.00	0.52	14.10
West Java	4.43	92.79	0.88	21.54	2.98	**6.60**	0.57	16.38	**3.54**	**40.10**	1.04	15.26
Central Java	4.67	99.00	1.29	21.01	3.10	1.30	0.57	15.85	3.48	21.90	0.67	13.87
East Java	4.61	98.70	1.00	23.07	3.16	2.60	**0.89**	16.52	3.53	23.59	**1.55**	15.54
South Kalimantan	4.60	99.00	0.53	24.48	3.12	1.10	0.47	**18.15**	3.27	6.35	0.90	**15.82**
NTT	4.45	0.00	0.37	14.53	**3.26**	0.00	0.68	12.72	3.26	0.00	0.89	10.95
Papua	4.42	*NA*	0.29	15.65	3.25	*NA*	0.29	13.28	3.26	*NA*	0.33	10.05
South Sulawesi	4.15	3.00	1.10	17.55	3.00	0.00	0.09	13.05	3.18	0.00	0.28	10.50
West Sumatera	**4.68**	**100.00**	**1.88**	**24.73**	3.11	0.50	0.60	17.74	3.38	10.78	1.05	16.4

Table 22: Language steering results across provinces comparing the best proprietary, multilingual, and SEA-centric models in this task. We report LLM-as-a-judge (denoted as LLM), Glot-LID, BLEU, and chrF scores.