Culturally-Nuanced Story Generation for Reasoning in Low-Resource Languages: The Case of Javanese and Sundanese

Salsabila Zahirah Pranida[*]   **Rifo Ahmad Genadi**[*]   **Fajri Koto**

Department of Natural Language Processing, MBZUAI

{salsabila.pranida,rifo.genadi,fajri.koto}@mbzuai.ac.ae

Abstract

Culturally grounded commonsense reasoning is underexplored in low-resource languages due to scarce data and costly native annotation. We test whether large language models (LLMs) can generate culturally nuanced narratives for such settings. Focusing on Javanese and Sundanese, we compare three data creation strategies: (1) LLM-assisted stories prompted with cultural cues, (2) machine translation from Indonesian benchmarks, and (3) native-written stories. Human evaluation finds LLM stories match natives on cultural fidelity but lag in coherence and correctness. We fine-tune models on each dataset and evaluate on a human-authored test set for classification and generation. LLM-generated data yields higher downstream performance than machine-translated and Indonesian human-authored training data. We release a high-quality benchmark of culturally grounded commonsense stories in Javanese and Sundanese to support future work.[1]

1 Introduction

Reasoning, the ability to draw conclusions, make inferences, and relate concepts, is a core evaluation target in recent LLM work (Dubey et al., 2024; OpenAI et al., 2024; Hurst et al., 2024; Almazrouei et al., 2023). Yet widely used English benchmarks such as StoryCloze (Mostafazadeh et al., 2016, 2017), WinoGrande (Sakaguchi et al., 2021), and HellaSwag (Zellers et al., 2019) encode Western norms. Because reasoning is culturally shaped, relying on machine-translated English datasets (Ponti et al., 2020; Lin et al., 2022; Hershcovich et al., 2022) risks erasing local context.

Recent datasets for medium-resource languages (e.g., Indonesian (Koto et al., 2024) and Arabic (Sadallah et al., 2025)) add cultural grounding but

mainly target sentence-level classification. Story-level commonsense, how people interpret events across narratives, remains underexplored in low-resource languages due to limited speaker access, high annotation costs, and scarce culturally relevant materials.

We address story comprehension in two underrepresented languages, Javanese and Sundanese, spoken by roughly 80M and 32M people respectively (Badan Pusat Statistik, 2025; Eberhard et al., 2025). Beyond sheer scale, both carry rich sociolinguistic systems: Sundanese encodes politeness and hierarchy phonologically, while Javanese employs elaborate speech levels (Wolff and Poedjosoedarmo, 1982). We adopt a StoryCloze-style setup (Mostafazadeh et al., 2016, 2017): given a four-sentence story, models either generate a plausible fifth sentence (generation) or choose the correct continuation from two options (classification).

We compare three dataset creation strategies for culturally grounded story comprehension: (1) LLM-assisted generation with culturally informed prompts, (2) machine translation from Indonesian benchmarks, and (3) native-authored stories. Each has distinct benefits, scalability, resource reuse, and authenticity, respectively. Human evaluation shows LLM outputs are culturally plausible and comparable to native-written stories.

To assess cultural reasoning on narratives, we use the native-authored set as a zero-shot testbed. As in Figure 1, Indonesian/SEA-centric models, though covering the target languages, often produce culturally inappropriate endings, showing that language support alone is insufficient. We then fine-tune on synthetic data and find that LLM-generated training yields stronger results than machine-translated data for both classification and generation.

Our contributions are as follows:

- We release the **first benchmark** for culturally grounded commonsense reasoning in
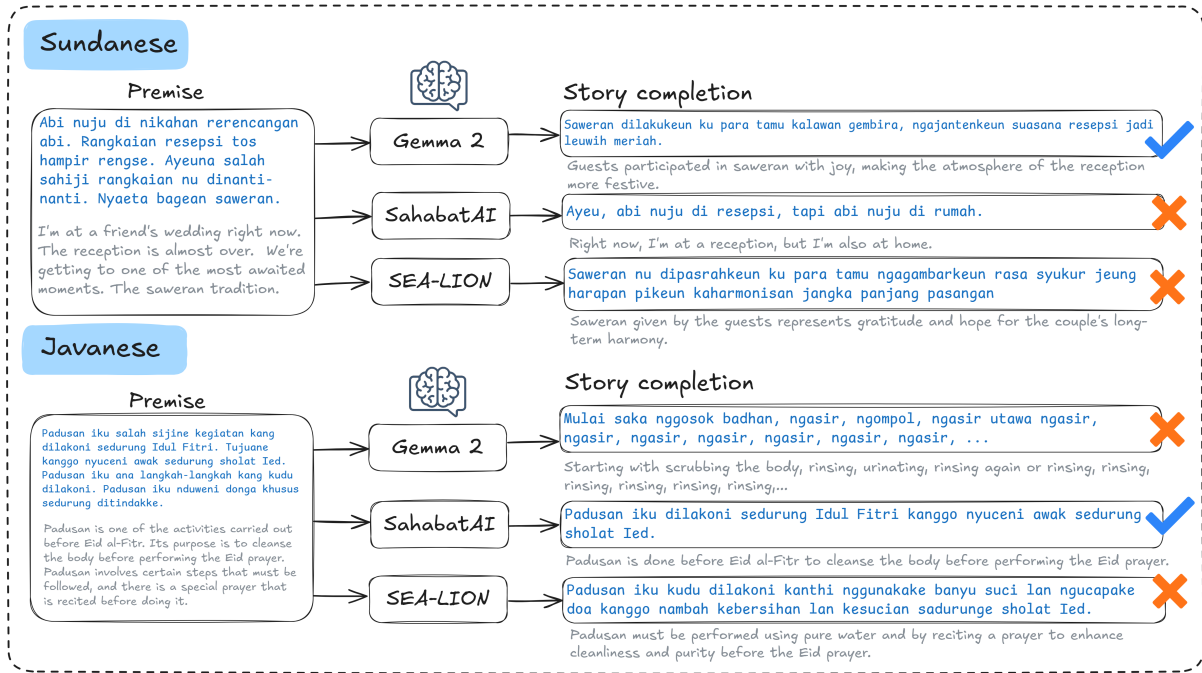

Figure 1: Examples of human-written stories in Sundanese and Javanese. English translations in gray color are provided for reference. A cross (✗) indicates a culturally irrelevant ending, generated by LLMs

Javanese and Sundanese, comprising 3.3K high-quality stories. This includes 1.12K human-written samples, 1K human-reviewed machine-translated texts, and 1.22K filtered LLM-generated samples.

- We conduct extensive human evaluation of multiple dataset creation strategies, including (i) LLM-assisted generation, (ii) direct machine translation, (iii) culturally localized translation, and (iv) native-authored stories.

- We evaluate model performance through zero-shot inference and supervised fine-tuning in both classification and generation settings to assess their cultural reasoning capabilities.

2 Related Works

2.1 Commonsense Reasoning in English Story Comprehension

Story comprehension in NLP involves reasoning over causal, temporal, and commonsense relations within narratives. The StoryCloze test, introduced by Mostafazadeh et al. (2016, 2017) is a landmark benchmark, requiring models to select the most plausible ending for a short four-sentence story. Many commonsense reasoning datasets, however, focus on sentence-level challenges include WinoGrande (Sakaguchi et al., 2021) for pronoun resolution, COPA (Gordon et al., 2012) for

causal reasoning, and HellaSwag (Zellers et al., 2019) for adversarial sentence completion. While effective for probing localized reasoning, these do not capture broader discourse coherence or character motivations.

Recent work has shifted toward narrative-level reasoning with longer contexts and richer event dynamics. NarrativeQA (Kočiský et al., 2018) covers full books and movie scripts, CosmosQA (Huang et al., 2019) infers implicit causes and intentions, and TellMeWhy (Lal et al., 2021) targets causal and motivational "why" questions. Yet these remain English-centric and question-answering-oriented. Our work instead addresses narrative completion in low-resource languages, particularly in Javanese and Sundanese, providing a culturally grounded alternative to high-resource, English-dominant benchmarks.

2.2 Commonsense Reasoning in Languages Beyond English

Early multilingual commonsense benchmarks often extended English datasets via translation. XCOPA (Ponti et al., 2020) translated COPA into 11 typologically diverse languages, including Indonesian, while X-CSQA (Lin et al., 2021) adapted CommonSenseQA across languages. Although useful for cross-lingual evaluation, such resources inherit Anglocentric biases, as progress in English

does not always transfer culturally or linguistically (Lin et al., 2022; Shwartz et al., 2020). Direct translations risk embedding English social contexts rather than local commonsense (Lin et al., 2021).

Story comprehension tasks like StoryCloze (Mostafazadeh et al., 2016) have been similarly extended. One such extension is XStoryCloze (Lin et al., 2022), by translating English narratives into multiple languages. Yet such approaches still struggle to capture culture-specific narrative norms.

For Indonesian, culturally grounded datasets such as COPAL-ID (Wibowo et al., 2024) and IndoCulture (Koto et al., 2024) model regional practices and norms across 11 provinces, advancing evaluation in a medium-resource language. However, they primarily focus on short-form, sentence-level reasoning such as multiple-choice or cloze-style questions, rather than full-narrative comprehension. Beyond Indonesia, CultureBank (Shi et al., 2024) compiles large-scale cultural knowledge from community narratives to support culturally aware language technologies, while CultureLLM (Li et al., 2025) incorporates cultural differences into LLMs via semantic data augmentation. However, these resources primarily focus on short-form, structured tasks rather than full-narrative comprehension. Our work fills this gap by introducing the first benchmark for **story-level commonsense reasoning** in low-resource languages, specifically Javanese and Sundanese, two of Indonesia's most widely spoken local languages.

2.3 LLM-Generated Data Creation

One possible solution to tackle data scarcity in NLP is applying data augmentation (Feng et al., 2021; Ding et al., 2020; Ahmed and Buys, 2024; Liu et al., 2024; Yong et al., 2024; Guo and Chen, 2024; Liu et al., 2022), with LLMs increasingly used to produce high-quality synthetic data that complements or substitutes manual annotation. Most prior work targets classification tasks. For example, WANLI (Liu et al., 2022) used GPT-3 (Brown et al., 2020) to generate synthetic English natural language inference data (Bowman et al., 2015) refined by humans, while Yong et al. (2024) generated English sentiment and topic classification data before translating it into low-resource languages using bilingual lexicons.

For low-resource languages, Putri et al. (2024) employed GPT-4 (OpenAI et al., 2024) to create question-answering datasets, showing LLM potential in under-resourced settings. However, such efforts often overlook cultural reasoning and narrative coherence. Our work instead focuses on culturally nuanced story generation in Javanese and Sundanese, targeting story-level commonsense reasoning. We compare multiple data creation strategies, including LLM-assisted generation with open- and closed-weight models, machine translation from Indonesian, and native-authored stories.

3 Dataset Construction

As in Figure 2, we build two parallel streams: training and test. For training, the IndoCloze (Koto et al., 2022) train split serves both as seeds for LLM-guided generation and as sources for machine translation into Javanese and Sundanese. For testing, we translate the IndoCloze test split and human-verify it for linguistic quality and cultural relevance, and we add a fully new set of native-authored stories from predefined topics. Each instance follows the StoryCloze format: a four-sentence premise with one correct and one incorrect ending.

Indonesian is chosen as the seed language for its national status and cultural proximity to Javanese and Sundanese. To ensure authenticity, native speakers authored and validated stories, embedding local names, places, foods, and customs. Following Mostafazadeh et al. (2016); Koto et al. (2022), the dataset targets everyday commonsense reasoning grounded in local culture

3.1 Training Data

We construct our training set using three strategies: (1) LLM-assisted data generation, (2) direct machine translation, and (3) machine translation followed by cultural localization using an LLM.

3.1.1 LLM-Assisted Data Generation

We synthesize training data with three open models: Gemma2-27B-it (Rivière et al., 2024), Llama3.1-70B (Dubey et al., 2024), Mixtral-8x7B-Instruct (Jiang et al., 2024), and three closed models: GPT-4o (Hurst et al., 2024; OpenAI et al., 2024), Cohere Command-R-Plus (Cohere, 2024), Claude-3-Opus (Anthropic, 2024).

For each language and LLM, we supply seed examples and topics. We manually translate 50 IndoCloze train samples into Javanese and Sundanese with cultural localization (e.g., foods like *Gudeg* and rituals like *Sawéran*). Topics are de-
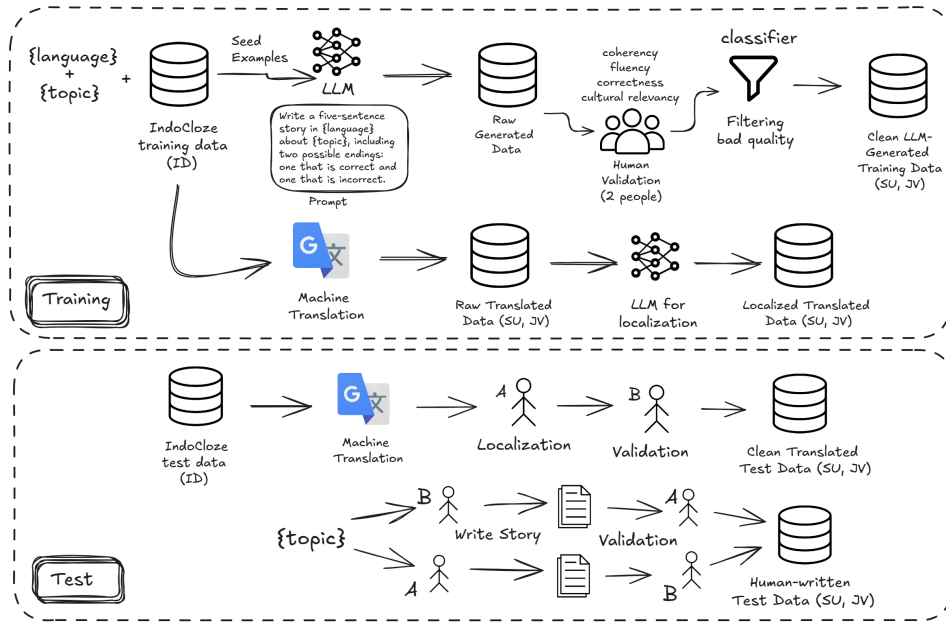
Figure 2: Overall pipeline of dataset creation.

rived from `IndoCulture` (Koto et al., 2024). The full prompt appears in Figure 4.

Each LLM produces 2,000 stories (1,000 per language), yielding six training sets. Per call, five random seeds guide generation at temperature 0.7; we repeat until each model reaches 1,000 valid samples per language ($\approx$ 166 per topic on average).

To filter quality, we train an XLM-R binary classifier on 600 human-rated outputs (Section 4.2)[2]. Applied to all 12,000 generations, it retains 1.2K high-quality stories (592 Sundanese, 628 Javanese). Roughly half of Claude and GPT-4o outputs pass, while far fewer from other LLMs, especially Mixtral, survive (Appendix B).

3.1.2 Machine Translation (MT$_{\text{train}}$)

We translate 1,000 `IndoCulture` training instances into Javanese and Sundanese (1,000 each) using Google Translate[3] (see Appendix F for quality). As a training resource, this set is *not* human-validated. We refer to it as **MT$_{\text{train}}$**.

3.1.3 Machine Translation + Localization (MT$_{\text{train}}$+GPT4o)

We prompt GPT-4o to culturally localize the MT outputs, following the same guidelines given to human annotators. The model adapts names,

events, foods, settings, and social norms to Javanese/Sundanese contexts. This probes whether LLMs can align literal translations with local cultural values to produce more authentic, context-appropriate narratives.

3.2 Test Set

We carefully construct the test set using two strategies: (1) machine translation with human verification, resulting in 500 Javanese and 500 Sundanese instances, and (2) manual writing by native speakers based on pre-defined topics, producing roughly 529 Javanese instances and 595 Sundanese instanecs. Each strategy undergoes rigorous quality control to ensure accuracy and reliability. In total, we create a high-quality test set of 2,124 instances across both languages.

To ensure the authenticity and quality of the dataset, we recruited 4 expert workers (2 per language) who are not only native Indonesian speakers but also fluent in Javanese and Sundanese. Each expert worker has a deep understanding of their respective language, culture, and customs. They have at least 10 years of experience speaking Javanese or Sundanese and possess strong linguistic and cultural expertise. The recruited workers, aged between 21 and 35 years, hold bachelor's degrees and were carefully selected for their proficiency in both language and cultural knowledge.

---

[2] 68.33% accuracy on the dev split (20% of train); setup in Appendix A.

[3] https://translate.google.com/, accessed September 2024.

Dataset	Sundanese				Javanese			
	Coherence	Fluency	Correctness	Cultural Rel.	Coherence	Fluency	Correctness	Cultural Rel.
Human Written (HW)	**5.0**	**5.0**	**100**	**96**	**5.0**	**5.0**	**100**	66
LLM Generated Data								
GPT-4o	4.7	4.2	80	**96**	4.9	4.5	97	91
Claude	4.7	4.4	86	92	4.9	4.3	96	**93**
Cohere	3.4	3.0	28	46	4.6	4.1	80	65
Llama3.1	3.7	3.4	56	70	4.5	4.2	65	50
Gemma2	3.9	3.1	42	78	4.8	3.6	83	81
Mixtral	2.0	1.5	0	4	2.0	1.9	3	22
Translated								
MT$_{train}$	3.24	4.36	80	20	4.36	4.46	98	12
Translated Data + Localization								
MT$_{train}$ + GPT4o	4.36	4.68	86	80	4.6	4.64	98	76
MT$_{test}$ + Human	**5.0**	**5.0**	**100**	14	**5.0**	**5.0**	**100**	18

Table 1: Quality analysis of models on Sundanese and Javanese. Higher scores indicate better performance in each category.

Dataset	Sundanese				Javanese			
	#data	#vocab	LW (%) ↓	MATTR ↑	#data	#vocab	LW (%) ↓	MATTR ↑
Human Written (HW)	594	4693	**0**	**0.84**	529	3497	**0**	**0.80**
LLM Generated Data								
GPT-4o	1000	3444	**0**	0.82	1000	3073	**0**	**0.80**
Claude	1000	3898	**0**	0.80	1000	3104	**0**	0.79
Cohere	1000	2654	3	0.65	1000	2254	3	0.68
Llama3.1	1000	3126	0	0.69	1000	2836	2	0.69
Gemma2	1000	3758	5	0.72	1000	3334	4	0.71
Mixtral	1000	4584	16	0.71	1000	4215	13	0.67
Translated Data								
MT$_{train}$	1000	5272	**0**	0.83	1000	5007	1	**0.81**
Translated Data + Localization								
MT$_{train}$ + GPT4o	1000	4985	1	0.82	1000	4692	1	**0.81**
MT$_{test}$ + Human	500	3877	**0**	0.82	500	3620	3	**0.81**

Table 2: Lexical diversity analysis of different models on Sundanese and Javanese. "LW" denotes the percentage of loanword.

3.2.1 Machine translation with human verification (MT$_{test}$+Human)

As shown in Figure 2, we translate 500 randomly selected samples from the IndoCloze test set into Javanese and Sundanese using Google Translate.[4] To ensure accuracy and naturalness of the machine-translation, we employ two native speakers for each language and implement a two-stage quality control process (Winata et al., 2023). In Stage 1, the first worker manually corrects translation errors and localizes content by replacing entities (e.g., names, buildings, food) with culturally relevant alternatives.[5] In Stage 2, a second worker validates the revised text and directly corrects any remaining errors from the first stage. From this point forward, we refer to this data as **MT$_{test}$+Human**.

3.2.2 Human-written Dataset (HW)

Each expert worker in Sundanese and Javanese is tasked with writing 600 short stories following the IndoCloze format: a four-sentence premise, a correct fifth sentence, and an incorrect fifth sentence. Stories are written based on 12 predefined topics, adhering to the same topic taxonomy used for training (see Section 3.2).See Appendix E for further details on the writing guidelines.

To ensure quality, each expert worker reviewed their peer's written stories. The reviewing worker was presented with a premise and two randomized alternate endings from another worker's story and was asked to identify the correct one. Instances incorrectly identified by the second worker were discarded, as they likely contained incorrect endings or exhibited ambiguity. After quality control, 529 Javanese and 595 Sundanese instances remained from the original 600 per language. From this point forward, we refer to this human-written dataset as **HW**.

---

[4]Google Translate was accessed in September 2024

[5]Note that while we applied cultural localization, not all examples could be fully adapted to Javanese or Sundanese contexts, as some stories reflect general Indonesian cultural elements that are not specific to either group.

4 Data Analysis

4.1 Overall Statistics

For LLM-assisted dataset creation, GPT-4o and Claude demonstrated the highest efficiency, generating nearly 1,000 clean samples with minimal discarded output, while Mixtral was the least efficient, requiring significantly more samples to reach the same threshold. The LLM-generated data does not have a uniform topic distribution due to variations in broken samples.

In total, the LLM-generated datasets contain 72K sentences and approximately 557K words. The word distribution across sentence positions remains consistent across the six LLMs, with word counts per position relatively uniform and a median sentence length ranging from 5 to 10 words. MT_{test}+Human (both with and without localization) and HW datasets exhibit a similar word distribution pattern to the LLM-generated datasets. MT contains around 6K sentences with 44,5K words, while HW has 6,7K sentences with 58,2K words. Despite the slight difference in word count, both datasets maintain a consistent distribution, with a median sentence length ranging from 4 to 11 words.

4.2 Quality Analysis based on Human Evaluation

We evaluate the quality of all the constructed dataset based on four key criteria: coherence, fluency, correctness, and cultural relevance. Specifically, we randomly select 50 samples (5–10% of the total dataset) for both Sundanese and Javanese and engage two native speakers for evaluation. Coherence and fluency are rated on a Likert scale from 0 to 5, while correctness and cultural relevance are assessed using binary annotation and reported as percentages. More details on the annotation guideline can be found in Appendix D. Table 1 summarizes the results of human evaluation, with scores for coherence and fluency being averaged between the annotators. Meanwhile, for correctness and cultural relevance, we count the percentage of data perceived as correct or culturally relevant by both annotators. Inter-annotator agreement scores ranges from 0.4 to 0.7, as shown in Appendix C.

We observe that among the evaluated LLMs, both GPT-4o and Claude consistently demonstrate strong performance across all metrics. Notably, their cultural relevance scores are comparable to those of human-written texts. Applying localiza-

tion to the machine-translated data using GPT-4o (MT_{train}+GPT-4o) improves coherence and cultural relevance. Finally, while human post-editing ensures near-perfect scores in coherence, fluency, and correctness, achieving full cultural localization remains challenging, as not all content can be naturally adapted without compromising narrative plausibility.

4.3 Lexical Diversity

We analyze the lexical diversity of LLM-generated data for Sundanese and Javanese using key metrics such as vocabulary size, moving-average type-token ratio (MATTR) (Covington and McFall, 2010), and the proportion of loanwords. To measure loanword presence, we manually review the top 100 most frequent words for each model.

As shown in Table 2, GPT-4o achieves a high MATTR score, making it highly comparable to human-written data. Among the LLM-generated datasets, Mixtral has the largest vocabulary size but also introduced the highest proportion of loanwords, with 16% in Sundanese and 13% in Javanese, suggesting a significant reliance on non-native terms. In contrast, GPT-4o and Claude generate text entirely in Sundanese and Javanese without incorporating foreign words, highlighting their ability to produce better datasets.

Upon manual inspection, we find that LLMs frequently generate common Javanese names such as *Ayu*, *Dwi*, *Bayu*, *Eko*, and *Sari*. For Sundanese, commonly produced names include *Lia*, *Budi*, *Dewi*, and *Rina*. While these names are widely used across Indonesia, the variation in honorific terms is limited across all models, only *Pak* and *Bu*, along with their formal variants *Bapak* and *Ibu*, appear consistently. This suggests that, although some models demonstrate surface-level lexical diversity, deeper sociolinguistic features such as honorific variation remain underrepresented.

5 Experiments and Analysis

5.1 Classification

5.1.1 Setup

We adopt the classification accuracy metric, as proposed in both Mostafazadeh et al. (2016) and Koto et al. (2022), defined as the ratio of correctly predicted instances to the total number of test cases. For our experiments, we fine-tune several models: Qwen 2.5 7B (Qwen et al., 2025), Llama 3.1 8B (Grattafiori et al., 2024), Gemma 2 9B (Rivière

Model	Javanese					Sundanese				
	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt
XLM-R	NA	69.00	67.67	80.72	**81.47**	NA	75.42	50.34	**75.93**	72.73
Qwen 2.5	76.94	85.44	84.31	84.88	**87.71**	64.48	74.75	72.22	**80.64**	79.63
Llama 3.1	75.24	90.17	86.96	91.68	**92.44**	48.32	79.97	71.55	**81.14**	78.62
Gemma 2	85.07	93.95	**95.46**	**95.46**	95.27	51.52	86.20	**89.23**	87.88	85.02
SahabatAI Llama	77.13	95.46	94.14	97.16	**97.73**	34.51	87.88	87.21	91.25	**92.42**
SEA-LION Llama	69.19	96.60	94.71	93.76	**95.27**	25.93	88.55	83.67	**82.15**	77.44

Table 3: Accuracy of models on the human-written test set in Javanese and Sundanese across training sets.

Model	Javanese					Sundanese				
	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt
XLM-R	NA	69.40	**70.20**	66.80	62.60	NA	56.60	**73.00**	67.60	66.00
Qwen 2.5	68.00	**80.00**	75.00	75.80	75.80	63.60	78.80	74.20	76.60	**77.60**
Llama 3.1	47.60	**86.00**	70.40	77.40	73.40	48.60	**83.40**	74.00	75.00	76.00
Gemma 2	63.00	91.80	**92.80**	83.60	83.40	57.60	90.00	**91.80**	80.80	83.40
SahabatAI Llama	34.80	90.00	87.60	84.20	**88.40**	30.80	87.40	86.60	83.60	**87.60**
SEA-LION Llama	32.20	**92.60**	87.40	76.20	72.20	29.80	92.20	88.80	77.60	71.20

Table 4: Accuracy of models on the machine-translated test set in Javanese and Sundanese across training sets.

et al., 2024), SahabatAI Llama 8B (GoToCompany, 2024), SEA-LION Llama 8B (Ng et al., 2025), and XLM-R (Conneau et al., 2019). For LLMs, we conduct instruction fine-tuning with a multiple-choice question framework using a LoRA adapter (Hu et al., 2022) on the combined Javanese and Sundanese training set. Full training details are provided in Appendix A. To ensure robustness, results are averaged over three runs. For comparison, we also report the zero-shot performance of each model prior to fine-tuning.

In addition to the zero-shot setting, we experiment with the following training data variations: (i) MT$_{train}$ (machine-translated data), (ii) MT$_{train}$+GPT-4o (machine-translated data localized using GPT-4o), (iii) All LLM (the full set of LLM-generated samples), and (iv) LLM_Filt (a filtered subset comprising the top 10% of All LLM based on human evaluation). We evaluate each model on two test sets: (i) a human-written set and (ii) a machine-translated set.

5.1.2 Overall Performance

Table 3 and Table 4 present the classification accuracies of all models trained on different training sets, evaluated on the human-written and machine-translated test sets, respectively. In the zero-shot setting, model performance ranges from approximately 34% to 70% across both Javanese and Sundanese, underscoring the inherent challenges of cultural commonsense reasoning in these languages. Fine-tuning consistently improves model performance over the zero-shot baseline, with LLMs showing particularly strong gains. Notably, XLM-R underperforms relative to the LLMs, suggesting that large generative models are more effective at

capturing cultural nuances.

We observe that models trained on LLM-generated data perform best when evaluated on the human-written test set (Table 3), while those trained on machine-translated data tend to perform better on the machine-translated test set (Table 4), indicating some sensitivity to data distribution alignment. Interestingly, localizing the machine-translated training data (MT$_{train}$+GPT-4o) does not consistently lead to improved model performance compared to using the original machine-translated data (MT$_{train}$).

5.1.3 Performance Across Different Topics

Figure 3 presents the topic-wise performance of SahabatAI on the human-written test set. The chart shows that models fine-tuned on both MT$_{train}$ and LLM$_{Filtered}$ consistently achieve higher accuracy across most categories compared to the zero-shot baseline. Interestingly, the zero-shot setting performs relatively well on the *Art* category. Fine-tuning on LLM$_{Filtered}$ yields notable improvements in culturally rich topics such as *Wedding*, *Pregnancy*, and *Art*, outperforming models trained on MT$_{train}$.

5.2 Generation

5.2.1 Setup

Using the same set of LLMs from the classification experiments, we fine-tune the models to perform story continuation: given a four-sentence premise, the model is trained to generate a coherent fifth sentence in either Javanese or Sundanese. We apply supervised fine-tuning using QLoRA (Dettmers et al., 2023), with training details provided in Appendix A. To evaluate generation quality, we use

Model	Javanese					Sundanese				
	0-shot	MT	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt
Qwen 2.5	72.5 / 21.8	71.7 / 19.2	72.0 / 19.5	**72.4 / 21.7**	71.9 / 19.5	69.2 / 15.4	70.3 / 15.6	**70.3 / 16.5**	**70.5 / 16.7**	69.2 / 14.6
Llama 3.1	62.4 / 8.3	72.7 / 21.5	72.4 / 21.1	**72.7 / 22.9**	72.7 / 22.2	59.4 / 5.6	70.0 / 16.1	70.3 / 16.23	**70.6 / 17.9**	68.9 / 15.5
Gemma 2	70.9 / 17.7	**73.2 / 24.0**	73.1 / 23.1	72.8 / 22.9	72.6 / 22.2	68.2 / 13.2	**70.3 / 18.1**	70.2 / 17.9	70.8 / 17.9	70.2 / 17.1
SahabatAI Llama	63.1 / 15.3	**71.8 / 20.1**	59.7 / 9.2	62.6 / 13.6	67.2 / 16.6	57.7 / 8.9	69.2 / 16.2	59.1 / 7.5	62.4 / 11.7	**70.6 / 17.6**
SEA-LION Llama	**72.7 / 23.0**	72.4 / 20.9	72.5 / 21.5	**72.8 / 23.9**	72.6 / 22.5	68.9 / 16.7	70.1 / 16.3	70.3 / 17.0	71.0 / 18.3	**71.3 / 18.6**

Table 5: BERTScore / ROUGE-L F1 of models on the human-written test set in Javanese and Sundanese across training sets.

Model	Javanese					Sundanese				
	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt	0-shot	MT$_{Train}$	MT$_{Train}$+GPT-4o	All_LLM	LLM_Filt
Qwen 2.5	70.5 / 15.6	70.7 / 18.3	71.0 / 18.1	**71.6 / 18.8**	70.6 / 15.4	69.9 / 14.0	70.9 / 17.8	**70.9 / 18.1**	71.3 / 16.7	70.2 / 13.7
Llama 3.1	59.7 / 4.4	**72.1 / 21.4**	72.0 / 20.7	71.8 / 20.4	71.3 / 18.8	59.3 / 4.2	**71.4 / 19.2**	71.2 / 18.6	71.3 / 17.7	68.8 / 15.7
Gemma 2	69.7 / 14.5	**72.3 / 22.5**	72.1 / 21.5	71.9 / 20.9	71.4 / 19.7	69.4 / 14.1	**72.4 / 21.3**	72.1 / 20.7	71.9 / 19.0	70.4 / 17.3
SahabatAI Llama	59.2 / 9.3	**72.1 / 22.2**	60.2 / 8.9	62.5 / 12.4	64.6 / 12.8	58.1 / 7.9	**71.5 / 20.5**	59.9 / 8.0	62.7 / 10.9	70.4 / 16.6
SEA-LION Llama	69.7 / 16.9	**72.4 / 21.3**	72.0 / 20.5	71.9 / 20.7	71.6 / 19.6	69.7 / 15.4	**72.1 / 19.5**	71.5 / 18.0	71.5 / 18.0	71.1 / 18.5

Table 6: BERTScore / ROUGE-L F1 of models on the machine-translated test set in Javanese and Sundanese across training sets.

Language	Model	Coherence	Fluency	Correctness	Cultural Rel.
Javanese	Gemma 2	4.92	4.94	76	**92**
	SEA-LION Llama	**4.98**	**4.98**	**84**	**92**
Sundanese	Gemma 2	4.22	4.74	74	86
	SEA-LION Llama	**4.40**	**4.92**	**78**	**94**

Table 7: Human evaluation results for Javanese and Sundanese.

two automatic metrics: ROUGE-L (Lin, 2004) for lexical overlap and BERTScore (Zhang et al., 2020) for semantic similarity.

5.2.2 Overall Performance

Table 5 and Table 6 presents the automated evaluation metrics for different training sets across Sundanese and Javanese on both human-written and machine-translated test sets. Finetuning consistently improves performance over the 0-shot. Notably, on the human-written data that contains more cultural nuanced story, model fine-tuned with LLM-generated data gives higher improvement compared to others.

Among the models, SEA-LION Llama achieves the highest score across most test, then followed closely by Gemma 2. The LLM_Filt data often matches or outperforms the All_LLM settings that uses more samples. Similar to classification, in human-written test set, LLM-generated-training data tends to be better than in machine-translated test set.

5.2.3 Human Evaluation

We conducted a human evaluation to compare models between SEA-LION Llama and Gemma 2 fine-tuned on All LLM-generated data, the two models that showed strong performance in both classification and generation tasks (see Table 3, 4, 5, and 6).

Using the human-written test set, annotators were presented with a story premise and the generated ending sentence. They were asked to rate the outputs based on coherence, fluency, correctness, and cultural relevance, following the same guidelines as in the earlier manual evaluation. As shown in Table 7, both models perform well in Javanese and Sundanese, with SEA-LION Llama slightly outperforming Gemma 2 across most human evaluation criteria. However, this advantage is less evident when measured using automatic metrics.

6 Conclusion

We explored the potential and limitations of LLM-generated data for commonsense reasoning and story generation in Javanese and Sundanese, introducing the first cloze dataset for these languages with high-quality test sets. Our preliminary analysis in classification and generation settings shows that GPT-4o and Claude-3 Opus demonstrate strong capabilities in generating plausible short stories but face challenges in fluency and cultural accuracy. Despite these limitations, our findings suggest that LLM-assisted data generation is a practical and effective approach for constructing datasets in low-resource languages.

7 Limitations

This study acknowledges several limitations in terms of cultural nuance, language scope, and dialectal representation. While we carefully curated data across 12 cultural topics, our focus was limited to only two local languages—Javanese and Sundanese. Although these are the most widely spoken regional languages in Indonesia, they do not cap-
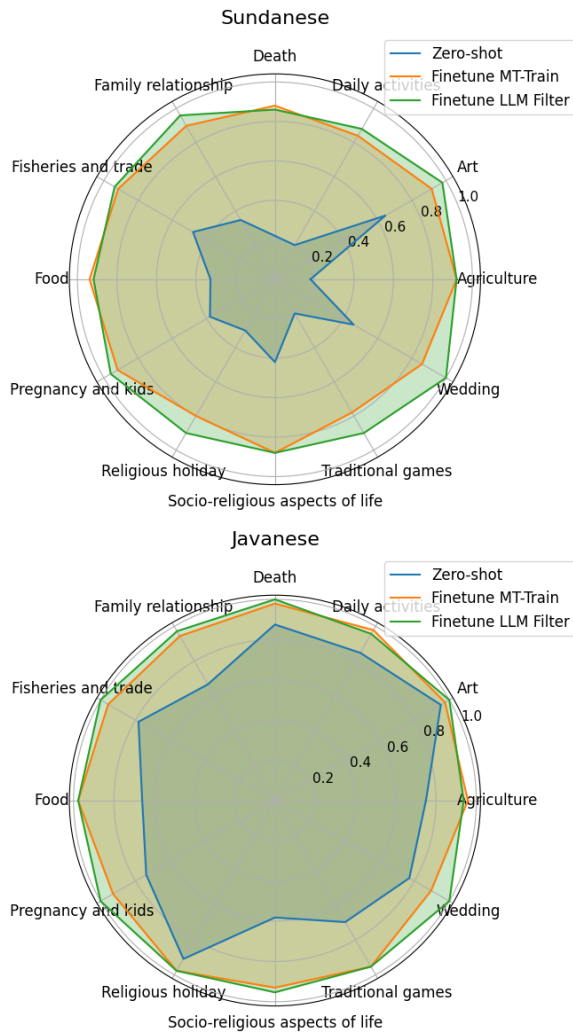
Figure 3: Topic-wise accuracy of SahabatAI (an Indonesian-centric LLM) on the human-written test set for Javanese and Sundanese, comparing zero-shot, and fine-tuning on MT_{train} fine-tuned, and $LLM_{filtered}$.

ture the full linguistic and cultural diversity present in the world. Moreover, our predefined topics and data sources may not comprehensively reflect the rich variation of cultural practices, dialects, and regional expressions within these languages.

Additionally, due to resource constraints and the primary focus on exploring the potential of LLM-generated data for commonsense reasoning and story generation, this research explored a limited range of LLM prompts and hyperparameter configurations. Future research could investigate a wider language scope, prompt variation, and other settings to identify configurations that maximize the performance of LLMs in generating culturally nuanced common sense reasoning in Javanese and Sundanese.

8   Ethical Considerations

All human-written datasets have been manually validated to ensure that harmful or offensive questions are not present in the dataset. We paid our expert workers fairly, based on the monthly minimum wage in Indonesia[6]. All workers were informed that their stories submitted would be used and distributed for research. Furthermore, no sensitive or personal information about the workers would be disclosed.

References

Khalid Ahmed and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158, Torino, Italia. ELRA and ICCL.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Al-shamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *CoRR*, abs/2311.16867.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Badan Pusat Statistik. 2025. *Statistik Indonesia 2025*, 1 edition. Badan Pusat Statistik (BPS), Jakarta, Indonesia. Nomor Katalog: 1101001, Nomor Publikasi: 03200.25004. Tanggal Rilis: 28 Februari 2025.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cohere. 2024. *Command R+ Documentation*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

---

[6]The average monthly minimum wage in Indonesia is approximately 3,000,000 IDR. The workload to complete all the tasks equates to roughly 8 days of full-time work. Each worker was paid 1,250,000 IDR accordingly

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and

et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas. Online version.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

GoToCompany. 2024. Gotocompany.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der

378

Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu,

379

Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xu Guo and Yiqiang Chen. 2024. Generative AI for synthetic data generation: Methods, challenges and the future. *CoRR*, abs/2403.04190.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson,

Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

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.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2025. Culturellm: incorporating cultural differences into large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

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

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

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, Brandon Ong, Zhi Hao Ong, Jann Railey Montalan, Adwin Chan, Sajeban Antonyrex, Ren Lee, Esther Choa, David Ong Tat-Wee, Bing Jie Darius Liu, William Chandra Tjhi, Erik Cambria, and Leslie Teo. 2025. Sea-lion: Southeast asian languages in one network. *Preprint*, arXiv:2504.05747.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-

der, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn,

Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in arab culture. *arXiv preprint arXiv:2502.12788*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

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 Romad-

hony, 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.

John U Wolff and Soepomo Poedjosoedarmo. 1982. Communicative codes in central java. Technical report, Southeast Asia Program, Dept. of Far Eastern Studies, Cornell University.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2024. LexC-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13990–14009, Miami, Florida, USA. Association for Computational Linguistics.

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

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 Training Configurations

For classification, we set the maximum token length for the pre-trained language model to 450 for the premise and 50 for the ending sentence. The model was trained over 20 epochs with early stopping (patience set to 5), using a batch size of 40, Adam optimizer, an initial learning rate of 5e-6 for XLM-R, and a warm-up phase comprising 10% of the total training steps.

For training the text generation models, we use 4-bit quantization with a LoRA rank of 64 and a LoRA alpha of 128. The models are trained with a batch size of 8, gradient accumulation of 8, a learning rate of 2e-4, and for a single epoch. We employ the Unsloth.ai framework for efficient fine-tuning (Daniel Han and team, 2023).

For training the classifier for data filtering, we fine-tune the XLM-R model with a maximum token length of 1024 for the premise and 128 for the ending sentence. The model was trained over 26 epochs, using a batch size of 16, Adam optimizer,

an initial learning rate of 1e-5 for XLM-R, and a warm-up phase comprising 5% of the total training steps.

Figure 4 shows the prompt template used for in-context learning, guiding the LLM to generate new Javanese and Sundanese.

B Distribution of Filtered LLM-Generated Training Data

Initially, training data was generated using six different LLMs, with each model contributing approximately 16.67% of the total 10K samples (around 2K samples per model). However, after filtering the bad examples, the final dataset composition shifted. The data consist of 1,220 samples, the distribution is as follows: Claude (37.7), GPT-4o (29.0%), LLama (8.6%), Cohere (4.3%), and Gemma-2 (1.7%).

Figure 4: Prompt template instructing LLM to generate a new example for Javanese and Sundanese.

C Agreement Score

We measure the fluency and coherency agreement scores using Pearson's correlation and computed the correctness and cultural relevance scores using Cohen's kappa. They are summarized in Table 8.

Metric	Javanese	Sundanese
Fluency	0.73	0.76
Coherency	0.77	0.71
Correctness	0.74	0.50
Cultural Relevance	0.75	0.49

Table 8: Agreement Scores for Javanese and Sundanese

D Workers Scoring Guidelines

- **Fluency (0–5):** Each sentence should be grammatically correct and fluent.
  - **5:** All sentences are grammatically correct and fluent.
  - **0:** Sentences are grammatically incorrect and lack fluency.

- **Coherency (0–5):** The story should be coherent, with all sentences logically connected.
    - **5:** Story is highly coherent, with clear and logical flow between sentences.
    - **0:** Sentences are disconnected and lack a logical sequence.

- **Correctness (Binary):** The correct story closure should be valid, while the incorrect closure should clearly be wrong.
    - **1:** The correct ending is indeed correct, and the incorrect ending is clearly wrong.
    - **0:** Either the correct ending is not valid, or the incorrect ending is not clearly wrong.

- **Cultural Relevance (Binary):** The story should reflect appropriate cultural norms, values, or symbols relevant to the language.
    - **1:** The story contains relevant cultural norms, values, symbols for the corresponding language.
    - **0:** The story lacks cultural relevance or includes irrelevant cultural aspects.

Scoring Guidelines:

- **Coherency:** Ranges from 0 to 5, where 5 means each sentence is strongly connected and flows well with the previous and next sentence.

- **Fluency:** Ranges from 0 to 5, where 5 indicates all sentences are grammatically sound and highly fluent.

- **Correctness:** A binary score of 0 or 1 to ensure that the correct ending is truly valid and the incorrect ending is clearly wrong.

- **Cultural Relevance:** A binary score of 0 or 1 to ensure the whole story contains appropriate and relevant cultural symbols and norms for the language being used.

E   Topic and Story-Writing Guidelines

We create a total of 300 native-authored stories as part of the Javanese and Sundanese Cloze Project. These stories are evenly distributed across 12 predefined topic categories, with 25 stories per topic. Each story must reflect traditional Javanese and Sundanese values and customs, with attention to

detail and coherence in the narrative. The topic categories and their subcategories follows IndoCulture (Koto et al., 2024). Each story must consist of 4 sentences and two endings: one correct and one incorrect. Ensure that all stories adhere to the topics and categories outlined in the taxonomy and reflect traditional values and cultural relevance.

F   Machine Translation Evaluation

	Javanese	Sundanese
Google Translate	0.11 / 0.41 / 45.2	0.09 / 0.35 / 42.8
NLLB-200-3.3B	0.18 / 0.30 / 53.1	0.12 / 0.33 / 43.8

Table 9: BLEU F1 / METEOR / ChrF comparison of Google Translate and NLLB Dense Transformers Translation Model (Costa-Jussà et al., 2022) for Indonesian to Javanese / Sundanese translation on NusaX dataset (Winata et al., 2023)