SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages

Holy Lovenia[★,1,2] **Rahmad Mahendra**[★,3,2] **Salsabil Maulana Akbar**[★,2] **Lester James V. Miranda**[★,4] **Jennifer Santoso**[★,5] **Elyanah Aco**[★,6] **Akhdan Fadhilah**[★,7] **Jonibek Mansurov**[★,8] **Joseph Marvin Imperial**[★,9,10] **Onno P. Kampman**[★,11] **Joel Ruben Antony Moniz**[★,6] **Muhammad Ravi Shulthan Habibi**[★,3,2] **Frederikus Hudi**[★,12,13] **Railey Montalan**[★,1] **Ryan Ignatius**[6] **Joanito Agili Lopo**[14] **William Nixon**[15] **Börje F. Karlsson**[16] **James Jaya**[6] **Ryandito Diandaru**[6] **Yuze Gao**[6] **Patrick Amadeus**[15] **Bin Wang**[6] **Jan Christian Blaise Cruz**[8,17] **Chenxi Whitehouse**[18] **Ivan Halim Parmonangan**[19] **Maria Khelli**[15] **Wenyu Zhang**[6] **Lucky Susanto**[20] **Reynard Adha Ryanda**[21] **Sonny Lazuardi Hermawan**[22] **Dan John Velasco**[17] **Muhammad Dehan Al Kautsar**[15] **Willy Fitra Hendria**[6] **Yasmin Moslem**[23] **Noah Flynn**[24] **Muhammad Farid Adilazuarda**[8] **Haochen Li**[6] **Johanes Lee**[15] **R. Damanhuri**[25] **Shuo Sun**[6] **Muhammad Reza Qorib**[26] **Amirbek Djanibekov**[8] **Wei Qi Leong**[1] **Quyet V. Do**[27] **Niklas Muennighoff**[28] **Tanrada Pansuwan**[18] **Ilham Firdausi Putra**[6] **Yan Xu**[29,27] **Ngee Chia Tai**[1] **Ayu Purwarianti**[6,30] **Sebastian Ruder**[31] **William Tjhi**[1] **Peerat Limkonchotiwat**[★,32] **Alham Fikri Aji**[★,8] **Sedrick Keh**[★,33] **Genta Indra Winata**[★,35,2] **Ruochen Zhang**[★,34] **Fajri Koto**[★,8,2] **Zheng-Xin Yong**[★,34] **Samuel Cahyawijaya**[★,31,27,2]

[1]AI Singapore [2]IndoNLP [3]Universitas Indonesia [4]Allen Institute for Artificial Intelligence [5]RevComm, Inc. [6]Independent Researcher [7]Tohoku University [8]MBZUAI [9]University of Bath [10]National University Philippines [11]MOH Office for Healthcare Transformation (MOHT) [12]NAIST [13]Works Applications Lab [14]Universitas Gadjah Mada [15]Institut Teknologi Bandung [16]Beijing Academy of Artificial Intelligence (BAAI) [17]Samsung Research Philippines [18]University of Cambridge [19]Queensland University of Technology [20]Monash University Indonesia [21]Imperial College London [22]Independent Design Engineer [23]Bering Lab [24]Amazon [25]Universitas Diponegoro [26]NUS [27]HKUST [28]Contextual AI [29]Huawei Noah's Ark Lab [30]Prosa.ai [31]Cohere [32]VISTEC [33]Toyota Research Institute [34]Brown University [35]Capital One

[★]Major contributors

Abstract

Southeast Asia (SEA) is a region characterized by rich linguistic diversity and cultural variety, with over 1,300 indigenous languages and a population of 671 million people. However, the performance of contemporary AI models for SEA languages is compromised by a significant lack of representation of texts, images, and auditory datasets from SEA. Evaluating models for SEA languages is challenging due to the scarcity of high-quality datasets, compounded by the predominance of English training data, which raises concerns regarding potential cultural misrepresentation. To address these challenges, we introduce SEACrowd, a collaborative initiative that consolidates a comprehensive resource hub[1] to bridge the resource gap by providing standardized corpora and benchmarks[2] in nearly 1,000 SEA languages across three modalities. We assess the performance of AI models on 36 indigenous languages across 13 tasks included in SEACrowd, offering valuable insights into the current AI landscape in SEA. Furthermore, we propose strategies to facilitate greater AI advancements, maximizing potential utility and resource equity for the future of AI in Southeast Asia.

1 Introduction

Despite Southeast Asia (SEA) being home to 1,300 indigenous languages (18% of the world's languages) and 671 million people (8.75% of the world's population), the representation of texts, images, and audio datasets from this region is significantly lacking in machine learning models. This deficiency adversely affects the model quality for SEA languages. The language coverage of SEA languages in two common pre-training resources, Common Crawl[3] and C4 (Xue et al., 2021), is extremely limited, with only 2.36% (in 11 languages) and 10.62% (in 11 languages), respectively. In modalities beyond text, the representation is even more limited. For instance, Common Voice, one of the largest multilingual speech corpora, includes six SEA indigenous languages (Conneau et al., 2021; Ardila et al., 2020), and LAION-5B, one of the largest multilingual vision-language

[1] https://seacrowd.github.io/seacrowd-catalogue/

[2] https://github.com/SEACrowd/seacrowd-datahub/

[3] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5155–5203
November 12-16, 2024

(VL) corpora, includes 12 SEA indigenous languages (Schuhmann et al., 2022). Datasets for other SEA indigenous languages exist, but are often scattered, insufficiently documented, or varied in quality and formatting, thereby making access and usage challenging (Cahyawijaya et al., 2023a; Joshi et al., 2020; Aji et al., 2023).

In terms of evaluation, the sparse availability of high-quality test sets for these languages also complicates evaluating models for SEA languages. Despite there being 1,300+ languages in the SEA region, prior works (Winata et al., 2023; Cahyawijaya et al., 2021; Koto and Koto, 2020; Zhang et al., 2024; Wang et al., 2024; Nguyen et al., 2023; Leong et al., 2023; Yong et al., 2023) have only evaluated fewer than 10 SEA languages collectively. The actual performance of current models on most SEA languages remains largely unknown.

Moreover, the dominance of Anglocentric training data may result in cultural bias when generating texts, images, or audio in underrepresented SEA languages (Søgaard, 2022; Talat et al., 2022). Further, Durmus et al. (2023); AlKhamissi et al. (2024); Cahyawijaya et al. (2024a) have shown that the learned representations in large language models (LLMs) often fail to reflect local cultural values in SEA (Koto et al., 2024; Liu et al., 2024; Adilazuarda et al., 2024). This raises concerns about the ability of current LLMs to generate natural, high-quality texts for this region. In addition, the discrepancy in language support creates language barriers in technological access and risks marginalizing minority groups who do not speak the dominant language.

In this work, we investigate the current AI progress for SEA languages by addressing the challenges of resources, evaluation, and generation quality. Our contributions are three-fold:

- We bridge the resource gap by centralizing and standardizing ∼500 corpora in nearly 1,000 SEA languages in SEACrowd, a comprehensive and standardized resource center, across three modalities: text, image, and audio.
- We close the evaluation gap in SEA languages with the SEACrowd Benchmarks, which cover 38 SEA indigenous languages on 13 tasks across 3 modalities, providing insights into the performance of a diverse spectrum of AI models. Further, our study reveals that the generative outputs of existing LLMs exhibit a closer resemblance to "translationese" rather than natural data in nine SEA languages.
- We offer insights and strategies for the future development of AI in SEA.

2 SEACrowd

SEACrowd represents the first comprehensive AI dataset collection initiative for SEA, developed through a collaborative effort among researchers and engineers primarily based in the SEA region. As addressed in §1, resource scarcity and the scattered nature of the data are crucial challenges in SEA. SEACrowd addresses these issues through two primary contributions: 1) **consolidating datasheets** to enhance data discoverability; and 2) **standardizing dataloaders** for easier use, especially in multiple dataset loading. We also follow data provenance practices (Longpre et al., 2023) to preserve the proprietary rights of dataset owners.

Consolidating datasheets We invited contributors to submit datasheet forms (Gebru et al., 2021) for publicly available datasets across all modalities including text, audio, and image in SEA languages and/or cultures. These datasheets include detailed information about each dataset, such as data subset(s), description, task, language, license, URL access, annotation method(s), annotation validation, relevant publications, publication venue, and data splits. For each submission, we manually verify and correct it as necessary to ensure datasheet accuracy.

Standardizing dataloaders For each approved datasheet, we created a standardized dataloader wrapper to facilitate ready-to-use data access since only 38.4% of the consolidated data sources were originally hosted on Hugging Face[4]. To support diverse task types, we carefully designed the standardized `seacrowd` schema to support different data structures and modalities (see Appendix F). We also adhere to data provenance practices (Longpre et al., 2023) and document the relevant metadata (e.g., license) in the dataloaders. Furthermore, we engaged with data owners and successfully converted three private datasets into public ones.

These efforts have culminated in 498 datasheets in SEACrowd Catalogue and 399 dataloaders in SEACrowd Data Hub (§2.1). Notably, our centralized data repository covers ∼1,000 SEA languages, underscoring the extensive linguistic diversity captured by SEACrowd. We elaborate on the

[4] https://huggingface.co/

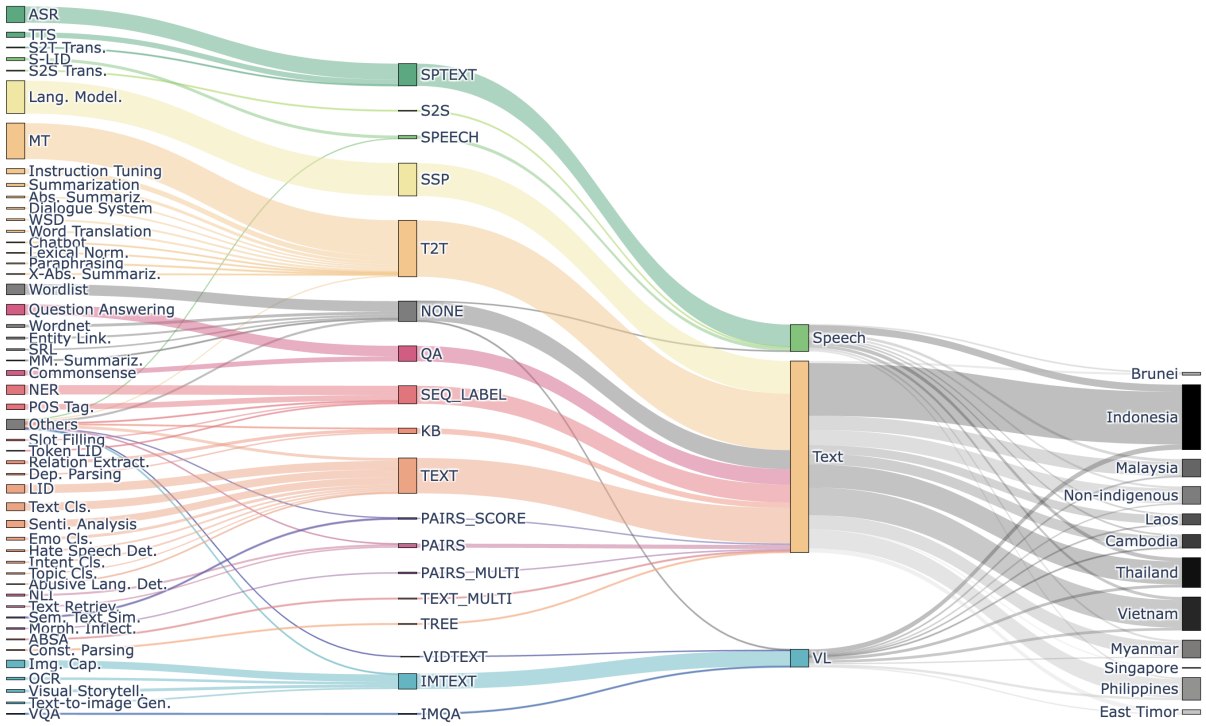


Figure 1: Mapping between tasks, schemas, modalities, and language regions across 498 datasheets in SEACrowd.

SEACrowd dataset statistics in §2.2. SEACrowd's contribution guidelines, progression details, and reviewing procedure are in Appendix C, D, and E.

2.1 SEACrowd Catalogue & Data Hub

SEACrowd comprises two interconnected platforms: SEACrowd Catalogue[5] and SEACrowd Data Hub. These platforms work in tandem to consolidate the datasheet submissions and provide a standardized pipeline for SEACrowd. Specifically, Catalogue houses the datasheets (metadata), while Data Hub stores the standardized dataloaders and the `seacrowd` library[6] for the schemas and configurations (Appendix F). These systems share information on the datasheets and dataloaders, allowing users to seamlessly explore and utilize them.

2.2 Datasets in SEACrowd

SEACrowd consolidates 498 datasheets with diverse tasks in SEA languages and provides standardized access through dataloaders to 399 of them. As shown in Figure 1, approximately 81% of the datasets in SEACrowd are textual data, with the remaining ∼8% and ∼11% being VL and speech, respectively. The complete list of SEA indigenous languages covered by SEACrowd and their mapping to the relevant SEA regions are provided in Appendix K. Around ∼53% of the datasets have a commercially permissive license.

A total of 83 tasks are provided in SEACrowd with a breakdown of 66 in NLP (e.g., abusive language detection, intent classification, instruction tuning, named entity recognition, etc.), 10 in VL (image-to-text generation, sign language recognition, video captioning, etc.), and 7 in speech (e.g., automatic speech recognition, text-to-speech, speech emotion recognition, and others). These tasks are then standardized into 20 dataloader schemas described in Appendix F. Further discussion regarding resources in SEACrowd is in §5.1.

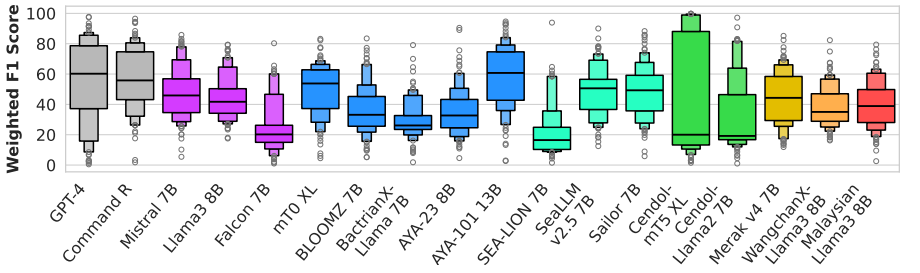
3 SEACrowd Benchmarks

To understand the capability of state-of-the-art models, we conduct comprehensive evaluations of existing LLMs, VLMs, and speech models from various architectures and training approaches. To construct a benchmark suite[7], we select a subset of the dataset that has been manually annotated and/or validated from the data presented in §2.2. More details regarding the data subsets, baselines, and prompts used for the evaluations are given in Appendix G.1, G.2, and G.3.

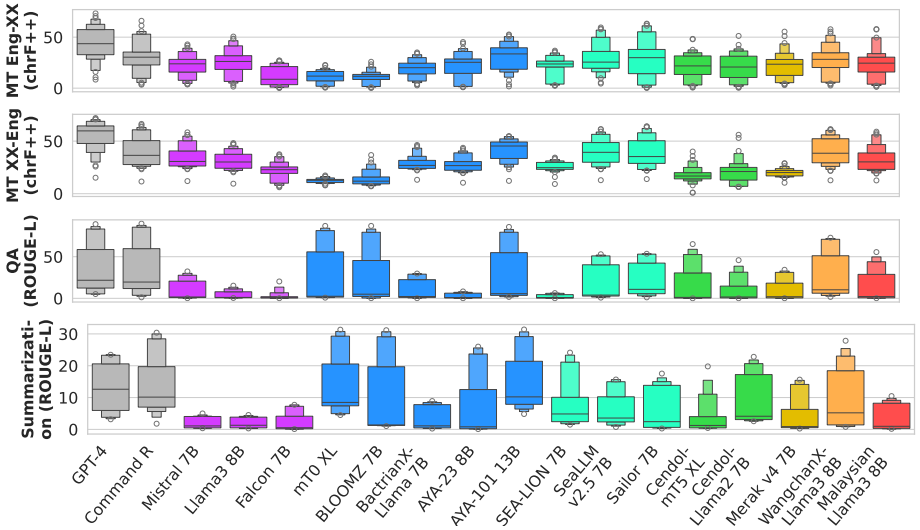
[5] SEACrowd Catalogue is also present in `csv` format.

[6] All codes are available under Apache License 2.0.

[7] https://github.com/SEACrowd/seacrowd-experiments



(a) NLU evaluation



(b) NLG evaluation

Figure 2: Zero-shot model performance across NLU and NLG tasks in SEA languages.

Model	Gini ↓
Commercial	
GPT-4	0.155
Command-R	0.184
English	
Mistral	0.159
Llama3	0.131
Falcon	0.238
Multilingual	
mT0	0.131
BLOOMZ	0.228
BactrianX-Llama	0.163
AYA-23	0.183
AYA-101	**0.095**
SEA regional	
SEA-LION	0.204
SeaLLM v2.5	0.116
Sailor	0.145
SEA country	
Cendol-mT5	0.378
Cendol-Llama2	0.267
Merak v4	0.199
WangchanX-Llama3	0.153
Malaysian Llama3	0.179

Table 1: Language equity across baselines based on Gini coefficient weighted by population ($\tau = 0.5$).

3.1 Datasets

NLP Our natural language understanding (NLU) benchmark consists of 131 data subsets and 7 tasks: sentiment analysis, topic classification, natural language inference (NLI), commonsense reasoning, exam-style multiple-choice question answering (QA), culture understanding, and reading comprehension. It covers English (ENG) and 33 SEA indigenous languages.

We utilize 100 data subsets for the natural language generation (NLG) benchmark, which covers machine translation (MT) between English and SEA languages from both directions, summarization, as well as extractive or abstractive question answering, covering 27 SEA indigenous languages.

Speech We employ 19 automatic speech recognition (ASR) data subsets to evaluate the capability of speech models in 15 SEA indigenous languages.

VL We assess the models on image captioning using four data subsets in 4 SEA indigenous languages, i.e., Filipino (FIL), Indonesian (IND), Thai (THA), and Vietnamese (VIE). This disparity in the evaluation scale is due to the fact that only a few datasets in SEACrowd are VL datasets, and even fewer are annotated by humans.

3.2 Baselines

Complete details regarding the model architectures, model sizes, seen languages, corresponding publications, and other aspects are in Appendix G.2.

NLP To evaluate the zero-shot performance of instruction-tuned LLMs on SEA languages, we benchmark two commercial, i.e., GPT-4 (OpenAI et al., 2024) and Command-R[8], and 17 open-source baselines, the majority of which are ∼7B-13B parameters. We categorize the open-source baselines according to the language(s) coverage in pre-training and/or instruction tuning, i.e., 1) **English**: Llama3 (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Falcon (Almazrouei et al., 2023); 2) **Multilingual**: AYA-101, AYA-23 (Üstün et al., 2024), mT0, BLOOMZ (Muennighoff et al., 2022), and BactrianX-Llama (Li et al., 2023a); 3) **SEA regional**: SEA-LION (Singapore, 2023), Sailor (Dou et al., 2024), and SeaLLM (Nguyen et al., 2023); and 4) **SEA country-specific**:

[8] https://docs.cohere.com/docs/command-r

Cendol-mT5, Cendol-Llama2 (Cahyawijaya et al., 2024b), and Merak (Ichsan, 2023) from Indonesia, WangchanX-Llama3 (Phatthiyaphaibun et al., 2024) from Thailand, and Malaysian-Llama3[9] from Malaysia.

Speech We evaluate the zero-shot performance of state-of-the-art **multilingual pre-trained** speech models in transcribing speech in SEA languages. Specifically, we consider Whisper v3 (Radford et al., 2023), MMS 1B (Pratap et al., 2024), and Seamless M4T v2 (Communication et al., 2023), which have shown proficiency in accurately transcribing multiple languages without fine-tuning. Additionally, we include models that are **fine-tuned on specific language(s)**, SEA or English, based on 1) Wav2Vec2 XLSR (Conneau et al., 2021) and 2) XLS-R (Babu et al., 2021), known for their cross-lingual speech representation learning by pre-training on raw speech waveforms across diverse languages, with XLS-R offering broader language coverage, and 3) Whisper, which leverages weakly supervised pre-training on spectrograms of speech in diverse languages. The specific fine-tuned models are evaluated: XLSR on IND, JAV, SUN; XLSR and Whisper on Indonesian (IND); XLSR and Whisper on Thai (THA); XLS-R on Tagalog (TGL); XLS-R on Burmese (MYA); XLS-R and Whisper on Khmer (KHM); and XLSR on English (ENG). See Appendix G.2 for details.

VL We consider state-of-the-art VLMs primarily trained on **English** pre-training and instruction-following data: LLaVA (Liu et al., 2023b,a), InstructBLIP (Dai et al., 2024), and Idefics2 (Laurençon et al., 2024), and VLMs trained in a **multilingual** manner: mBLIP (Geigle et al., 2023) and PaliGemma (Gemma Team et al., 2024), to assess their image captioning ability in SEA languages.

3.3 Experimental Settings

We conduct all evaluations in a zero-shot fashion. We employ 3 prompt templates in English for each NLU task and 1 for each NLG task. We utilize the weighted F1 score to measure the model performance on NLU tasks and n-gram reference-based metrics, i.e., chrF++ (Popović, 2015, 2017) and ROUGE-L (Lin, 2004), on NLG tasks. As for VL, aside from a prompt template in English, we also use a prompt template in the respective SEA indigenous language per data subset. We report

[9] https://huggingface.co/mesolitica/malaysian-llama-3-8b-instruct-16k

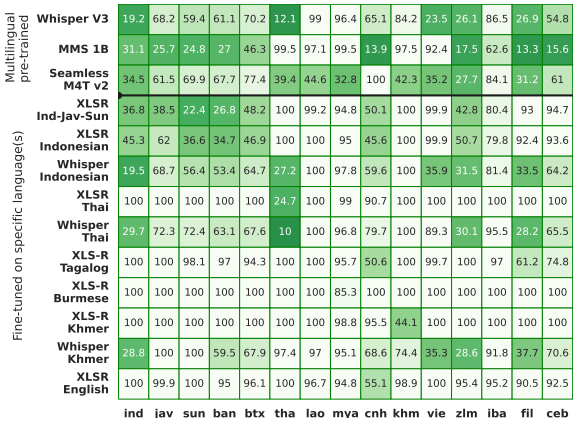


Figure 3: Speech model error rate (%↓) across existing ASR tasks in SEA languages.

CIDEr (Vedantam et al., 2015) for the image captioning task. For ASR, we use word error rate (WER) for languages with Latin script and character error rate (CER) for those with non-Latin script.

4 Result & Analysis

4.1 State-of-the-Art Models on SEA languages

LLMs Figure 2a and 2b illustrate the overall model performance of the LLM baselines in SEA languages for both NLU tasks and NLG tasks. In our NLU evaluation, AYA-101, a large multilingual instruction-tuned language model covering 101 languages, demonstrates the best zero-shot performance. It is followed by the commercial baselines, which achieve a median of ∼0.6 weighted F1-score. Sailor and SeaLLM, models specifically trained with SEA languages, also display competitive performance. Similarly, mT0 exhibits strong generalization abilities due to its exposure to ∼100 languages in pre-training, including those from the SEA region (Muennighoff et al., 2022). In contrast, most English and SEA country-specific baselines perform less effectively, likely due to their narrow focus on English or a limited set of SEA languages, such as Indonesian languages for Cendol and Thai for WangchanX-Llama3. Similar and consistent trends are observed on MT task, while the baselines' poorer scores on abstractive/extractive QA and summarization indicate their ineffectiveness in producing acceptable outputs in SEA languages for these tasks, which is especially pronounced in the open-source baselines. Appendix G.4 describes the performance of LLMs per language.

To analyze the equality in model performance across SEA languages, following Khanuja et al. (2023), we utilize the Gini coefficient—originally

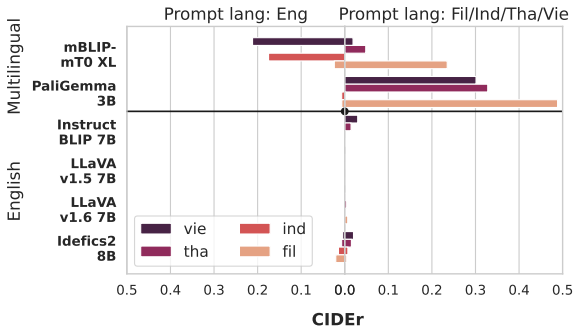


Figure 4: Existing VLMs produce subpar image captions in SEA languages. We report CIDEr (Vedantam et al., 2015).

Model	Natural outputs
SEA-LION	**58.57%**
AYA-23	43.57%
Sailor	37.86%
Cendol-Llama2	37.37%
Malaysian Llama3	36.90%
WangchanX-Llama3	30.24%
Falcon	29.52%
BactrianX-Llama	28.10%
SeaLLM	27.38%
Merak	26.19%
BLOOMZ	25.00%
Cendol-MT5	24.05%
Command-R	20.95%
mT0-XL	19.76%
Mistral	19.52%
GPT-4	16.67%
Llama3	14.05%
AYA-101	8.33%

(a) Avg. by models

Language	Natural outputs
Indonesian (IND)	**41.58%**
Vietnamese (VIE)	37.31%
Thai (THA)	34.21%
Khmer (KHM)	29.21%
Lao (LAO)	28.42%
Malay (ZLM)	22.24%
Burmese (MYA)	19.47%
Filipino (FIL)	12.22%
English (ENG)†	8.95%

(b) Avg. by languages

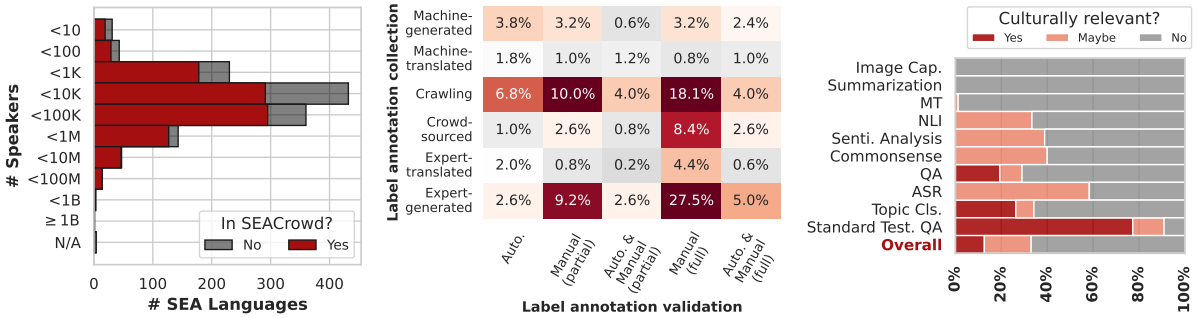
Table 2: Current LLMs are still incapable of generating natural texts in SEA languages. †As spoken in SEA regions, not worldwide.

used to observe income equality (Dorfman, 1979)—weighted by demand and parameterized by τ. Here, $\tau = 1$ corresponds to a demographic notion of demand, considering language population size, while $\tau = 0$ does not take population size into account (Blasi et al., 2022). Table 1 shows that models trained on more SEA languages, such as multilingual and SEA regional baselines, generally exhibit greater language equity. For instance, although Command-R and GPT-4 are competitive performance-wise against AYA-101 and mT0, AYA-101 and mT0 demonstrate higher equality across all SEA languages under study. This trend is consistent across different τ (see Appendix G.5).

Speech models Figure 3 presents the off-the-shelf speech model performance on ASR across languages in SEA, measured by the error rate percentage. 9 of the 15 SEA languages in our speech evaluation belong to the Austronesian language family. The other 6 are KHM and VIE, which belong to Austro-Asiatic, CNH and MYA belong to Sino-Tibetan, and THA and VIE belong to the Kra-Dai language family. The multilingual pre-trained baselines have a competitive generalization capability across languages, although it varies by language. For instance, Whisper v3 demonstrates significantly higher effectiveness for national languages such as IND, ZLM, FIL, THA, and VIE, while performing less optimally for other indigenous languages. Conversely, Seamless M4T v2 shows a more balanced performance across the languages. Regarding fine-tuned baselines, error rates decrease for their seen languages. The fine-tuned Whisper models, however, manage to better optimize for the target language while retaining their original capabilities in other SEA languages compared to their Wav2Vec2 XLSR and XLS-R counterparts, despite both having been pre-trained in a multilingual manner. This observation aligns with the findings of Rouditchenko et al. (2023), who find that the number of hours seen per language and language family during pre-training is predictive of how the models compare, in which Whisper's pre-training data duration for these four language families exceeds that of XLSR.

VLMs Figure 4 depicts the zero-shot performance of off-the-shelf VLMs on image captioning in SEA indigenous languages. Despite the capability of LLMs for zero-shot cross-lingual generalization (Huang et al., 2021; Täckström et al., 2012; Neubig and Hu, 2018; Artetxe et al., 2020), VLMs trained only in English (i.e., InstructBLIP, LLaVA, and Idefics2) fail to exhibit this capability, struggling to generate adequate image captions in SEA languages. Multilingual VL pre-training is crucial to achieving aligned multilingual representations (Burns et al., 2020; Li et al., 2023b; Huang et al., 2021). For instance, PaliGemma and mBLIP generate better image captions in THA and FIL when prompted in the relevant SEA languages.

However, when prompted in ENG, the performance of these multilingual baselines varies notably. PaliGemma's performance collapses completely, while mBLIP's performance shows both increases and decreases across different SEA languages. This raises the question of whether the multilingual VLMs can maintain consistent performance across different languages used in the instructions and the tasks. It highlights the need



(a) Language coverage (b) Annotation quality (c) Cultural relevance

Figure 5: The resource gap in SEA in terms of language coverage, annotation quality, and cultural relevance.

for further research into the mechanisms that drive these variations and how to achieve robust multilingual performance in VLMs across diverse linguistic contexts. Understanding these dynamics is crucial for improving VLMs' generalization capabilities and ensuring equitable performance across all languages, despite most related works focusing on monolingual visual instruction tuning (Liu et al., 2023b; Gong et al., 2023; Zhu et al., 2024).

4.2 Generation Quality in SEA Languages: Translationese vs. Natural Language

Classifying Translationese in SEA Languages To analyze the generation quality of LLMs in SEA languages, we build a text classifier to discriminate between translationese and natural texts (Riley et al., 2020). We construct a translationese classification training and testing dataset using 49 and 62 data subsets, respectively, covering approximately 39.9k and 51.5k sentences across English (ENG) and 8 SEA languages: Indonesian (IND), Khmer (KHM), Lao (LAO), Burmese (MYA), Filipino (FIL), Thai (THA), Vietnamese (VIE), and Malay (ZLM). The training and test data are detailed in Appendix H.1.

We fine-tune a classifier from mDeBERTaV3 (He et al., 2020, 2022)[10] using these data and achieve 79.08% accuracy on the test set in predicting translationese across these 9 languages. The detailed results and ablation studies of our translationese classifier experiments are provided in Appendix H.2. This classifier enables us to assess the generation quality of LLMs by distinguishing between translationese and naturally occurring text, providing insights into the models' performance in producing authentic language output.

[10] https://huggingface.co/microsoft/mdeberta-v3-base

Generation Quality of LLMs We evaluate the generation quality of LLMs in 9 SEA languages by generating answers to natural, general, and safety questions from Sea-Bench (Nguyen et al., 2023). As shown in Table 2a, LLMs with extensive language coverage but less focus on SEA languages, e.g., AYA-101 (Üstün et al., 2024), GPT-4 (OpenAI et al., 2024), mT0 (Muennighoff et al., 2023; Xue et al., 2021), and Llama3 (AI@Meta, 2024), tend to produce natural sentences less than 20% of the time. In contrast, models with narrower language coverage but a greater focus on SEA languages, such as Cendol-Llama2 (Cahyawijaya et al., 2024b), Sailor (Dou et al., 2024), AYA-23 (Aryabumi et al., 2024), and SEA-LION (Singapore, 2023), generate natural sentences over 35% of the time.

However, even the LLM with the least translationese generation, SEA-LION, only produces natural SEA sentences 57.71% of the time, highlighting a significant quality gap in generating natural sentences in SEA languages. As displayed in Table 2b, the translationese issue varies across SEA languages. Languages such as Tagalog (TGL), Burmese (MYA), and Malay (ZLM) have more severe translationese problems, with existing LLMs producing natural sentences only 11.58%, 19.47%, and 22.24% of the time, respectively. This underscores the need for further improvements in LLMs to more effectively address the linguistic diversity and complexity of SEA languages.

5 Discussions

5.1 Resource Gaps in SEA

Coverage SEACrowd covers 980 out of the 1,308 languages spoken in SEA (74.9%). Despite this high coverage, language representation in SEACrowd exhibits a very long-tail distribution, with over 700 languages having only 1 or 2 datasets,

and only 23 languages having 20 datasets or more. These less represented languages typically exist only in the form of lexicons (Asgari et al., 2020; List et al., 2022) or unlabeled data (Leong et al., 2022; Kudugunta et al., 2024; Nguyen et al., 2024). Existing tasks in SEACrowd still cover only a small portion of languages. For instance, sentiment analysis data is available for only 22 languages, and named entity recognition (NER) data is available for just 17 languages. Furthermore, for modalities beyond text, SEA resources are extremely underrepresented. Approximately 90% of SEA indigenous languages lack both speech and VL datasets.

Quality 78.7% of the datasets in SEACrowd are published in peer-reviewed venues, and most of the data has undergone external validation. The overall quality of the datasets in SEACrowd is depicted in Figure 5b. We compile the reported data construction methods by the authors, considering both the data collection method (i.e., data source) and label annotation validation (i.e., quality control). Nearly 19% of the datasets in SEACrowd have machine-generated and machine-translated annotations, while more than 80% were obtained from online texts (e.g., web crawling) and expert generation. In terms of label annotation validation, 62.4% of the datasets have been fully manually checked, while the remaining portion is partially validated and automatically checked. Note that these statistics only provide an initial indication of dataset collection quality on the surface and do not necessarily reflect the exact quality. Only a few datasets (6%) in SEACrowd report their detailed quality metrics (e.g., inter-annotator agreement scores). A deeper investigation is required for future work.

Cultural Relevance The resource gap in SEA extends to the cultural aspect, where misrepresentation can lead to offensive behaviors, e.g., cultural appropriation and stereotyping (Evans et al., 2020; Glotov, 2023). As a proxy of the cultural relevance of SEA datasets, we manually curated 259 data subsets used in SEACrowd evaluation based on their data source. Specifically, we categorize them whether they are 1) translated from another language, 2) crawled from local sources, or 3) hand-crafted to capture cultural relevance. In Figure 5c, approximately 70% lack cultural relevance, as many are machine-translated from English sources. About 20% are taken from local news, social media, or other local outlets, which potentially contain some culturally relevant data. Only the remaining 10% are designed to consider cultural relevance, derived from studies highlighting serious deficiencies in cultural understanding by LLMs for underrepresented languages (Kabra et al., 2023; Koto et al., 2023a; Wibowo et al., 2023; Liu et al., 2024; Koto et al., 2024).

5.2 Conclusion & Future Work

Southeast Asia is home to highly diverse languages and cultures; the majority of its people do not use English as their primary language. The utility of English-first AI is limited for the majority of Southeast Asian users, especially in critical sectors like healthcare and education. Through SEACrowd, we have explored the AI landscape in SEA and bridged the gaps in resources, evaluation, and naturalness analysis of AI models in SEA languages. Further, our initiative has nurtured an open-source research community, which will actively continue to add and maintain datasheets and dataloaders, as well as drive AI research and developments in SEA.

Nonetheless, AI development in SEA requires concentrated efforts by a range of stakeholders, who may prioritize differently when it comes to incorporating the region's 1,300+ languages into AI models. Moving forward, our work suggests AI development in SEA should prioritize two key metrics: 1) potential utility and 2) resource equity.[11]

Potential utility Potential utility is defined as the gap between current utility and ideal utility, in which model capability acts as a proxy for utility. Based on potential utility, unsurprisingly the development of the national languages (except for English and Chinese used in Singapore), i.e., Indonesian (IND), Burmese (MYA), Vietnamese (VIE), Thai (THA), Filipino (FIL), Khmer (KHM), Malay (ZLM), and Lao (LAO) in Figure 6, will bring the biggest benefit. Among them, we identify notable gaps in the naturalness of Malay, Burmese, and Filipino AI-generated outputs (§4.2). Focused efforts in resource building for these languages may move the needle the most for utility. Beyond the national languages, growing local languages or dialects with large speaker bases, e.g., Javanese (JAV), Sundanese (SUN), and Hmong (HMN), is key.

Resource equity Resource equity is defined as the gap between existing and ideal resource availability (Figure 6). We found that many local languages or dialects still fall short of the expected

[11] https://github.com/SEACrowd/globalutility

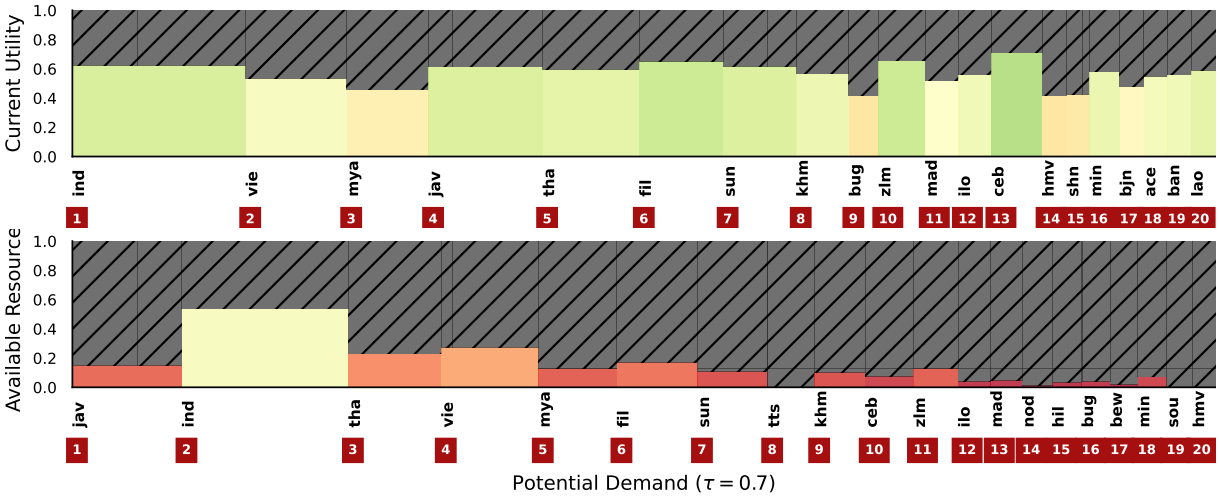


Figure 6: SEA languages prioritization based on **(top)** current utility and **(bottom)** resource availability. The languages are ranked based on the descending order of the area size of their missing potential .

level of resources. These include Northeastern Thai (TTS), Northern Thai (NOD), Hmong Do (HMV), Southern Thai (SOU), Cebuano (CEB), Ilocano (ILO), and others. Efforts to narrow these gaps would not only help preserve these languages but also ensure the continuation of the cultural heritage of the speakers of these languages. More details on SEA language prioritization for different weightings of demand can be found in Appendix I.

To improve these metrics, governments, and industry leaders in the region should invest in R&D activities to improve regional language capability for both the national languages and local dialects. This could include funding for open data collection and collaborations with local communities to address the resource gap in local languages. This also requires long-term sustainable strategies, such as catalyzing profitable use cases based on inclusive AI models, promoting fair and responsible compensation schemes for data workers, and orchestrating win-win exemplar collaborations between data owners, AI, and application developers.

Acknowledgments

We would like to thank our amazing contributors: Joshua Spergel, Tiezheng Yu, Parinthapat Pengpun, Ishan Jindal, Muhammad Satrio, Jipeng Zhang, Bhavish Pahwa, Haryo Akbarianto Wibowo, Hiroki Nomoto, Yohanes Sigit Purnomo W.P., Ahmad Fathan Hidayatullah, Bryan Wilie, Ruhiyah Faradishi Widiaputri, Rafif Rabbani, Fawwaz Mayda, Manoj Khatri, Supryadi Supryadi, Virach Sornlertlamvanich, Pavaris Ruangchutiphophan, Erland Hilman Fuadi, Mega Fransiska, Richardy Sapan, and Camilla Johnine Cosme, for their hard work in submitting datasheets and implementing dataloaders for SEACrowd.

This work is supported by the National Research Foundation, Singapore under its AI Singapore Programme; PhD Fellowship Award, the Hong Kong University of Science and Technology; and PF20-43679 Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong. JMI is funded by National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent AI [EP/S023437/1] of the University of Bath. In addition, we would like to express our gratitude to Cohere For AI for providing research grants that enabled us to conduct experiments using a commercial baseline, Command-R.

Limitations

While our work covers nearly 1,000 SEA languages, many dialects, which are considered as belonging to a parent language, are missing from our evaluation benchmark. For instance, for the Malay language, only Standard Malay (ZSM) is evaluated, but not other dialects such as Sarawak Malay (ZLM-SAR). Furthermore, the majority of our datasets also do not contain code-switched texts, which is a common linguistic phenomenon of SEA language usage (Aji et al., 2023). Moreover, the language coverage of different evaluation tasks varies significantly. For instance, NLP tasks cover 34 languages in total, whereas VL tasks only cover 4 languages. Tackling these limitations is essential to achieving a better representation of SEA, and we strongly encourage future works to prioritise these aspects.

Ethics Statement

In developing an evaluation benchmark for SEA languages, we have taken several steps to ensure ethical considerations are addressed comprehensively. First, the data used for this benchmark is sourced from publicly available resources, ensuring compliance with legal and ethical standards regarding data privacy. Where applicable, explicit consent was obtained from data contributors. Furthermore, all the datasets and resources utilized in this benchmark are used in accordance with their respective licenses. Second, our benchmark aims to be inclusive, representing a wide range of SEA languages, including those that are underrepresented in current linguistic resources. Lastly, our research process, including data collection, benchmark development, and evaluation methodologies, is entirely open-sourced and is documented transparently to enable reproducibility and accountability.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu

Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, and Ayu Purwarianti. 2023. The obscure limitation of modular multilingual language models. *ICLR Tiny Papers 2023*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

AI@Meta. 2024. Llama 3 model card.

Alham Fikri Aji, Jessica Zosa Forde, Alyssa Marie Loo, Lintang Sutawika, Skyler Wang, Genta Indra Winata, Zheng-Xin Yong, Ruochen Zhang, A. Seza Doğruöz, Yin Lin Tan, and Jan Christian Blaise Cruz. 2023. Current status of NLP in south East Asia with insights from multilingualism and language diversity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pages 8–13, Nusa Dua, Bali. 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.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *Preprint*, arXiv:2402.13231.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila B Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking large language models for cross-lingual few-shot transfer. *Preprint*, arXiv:2305.14857.

Ehsaneddin Asgari, Fabienne Braune, Benjamin Roth, Christoph Ringlstetter, and Mohammad Mofrad. 2020. UniSent: Universal adaptable sentiment lexica for 1000+ languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4113–4120, Marseille, France. European Language Resources Association.

Laksmita Widya Astuti, Yunita Sari, and Suprapto. 2023. Code-mixed sentiment analysis using transformer for twitter social media data. *International Journal of Advanced Computer Science and Applications*, 14(10).

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 197–213. Springer.

Samuel Cahyawijaya, Alham Fikri Aji, Holy Lovenia, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Fajri Koto, David Moeljadi, Karissa Vincentio, Ade Romadhony, and Ayu Purwarianti. 2022. Nusacrowd: A call for open and reproducible nlp research in indonesian languages. *Preprint*, arXiv:2207.10524.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024a. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, Dea Annisayanti Putri, Bryan Wilie, Genta Indra Winata, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2024b. Cendol: Open instruction-tuned generative large language models for indonesian languages. *Preprint*, arXiv:2404.06138.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jasper Kyle Catapang and Moses Visperas. 2023. Emotion-based morality in Tagalog and English scenarios (EMoTES-3K): A parallel corpus for explaining (im)morality of actions. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan van Esch, Vera Axelrod, Simran Khanuja, Jonathan Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason Riesa, and Melvin Johnson. 2022. XTREME-S: Evaluating Cross-lingual Speech Representations. In *Proc. Interspeech 2022*, pages 3248–3252.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and N. L. L. B. Team. 2024. Scaling neural machine translation to 200 languages. *Nature*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Robert Dorfman. 1979. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *Preprint*, arXiv:2404.03608.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL International.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Alexander Elias. 2018. Lio and the central flores languages. *Leiden: Leiden University Master thesis*.

Leanne M Evans, Crystasany R Turner, and Kelly R Allen. 2020. " good teachers" with" good intentions": Misappropriations of culturally responsive pedagogy. *Journal of Urban Learning, Teaching, and Research*, 15(1):51–73.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv*, abs/2307.06930.

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, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya,

Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Sergei Glotov. 2023. Intercultural film literacy education against cultural misrepresentation: Finnish visual art teachers' perspectives. *Journal of Media Literacy Education*, 15(1):31–43.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.0. leipzig: Max planck institute for evolutionary anthropology.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online. Association for Computational Linguistics.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Muhammad Ichsan. 2023. Merak-7b: The llm for bahasa indonesia. *Hugging Face Repository*.

Joseph Marvin Imperial, Jeyrome Orosco, Shiela Mae Mazo, and Lany Maceda. 2019. Sentiment analysis of typhoon related tweets using standard and bidirectional recurrent neural networks. *arXiv preprint arXiv:1908.01765*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Shengyi Jiang, Sihui Fu, Nankai Lin, and Yingwen Fu. 2022. Pretrained models and evaluation data for the khmer language. *Tsinghua Science and Technology*, 27(4):709–718.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab. 2015. Using resources from a closely-related language to develop asr for a very under-resourced language: A case study for iban. In *Proceedings of INTERSPEECH*, Dresden, Germany.

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

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian

languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Ichwanul Muslim Karo Karo, Mohd Farhan Md Fudzee, Shahreen Kasim, and Azizul Azhar Ramli. 2022. Sentiment analysis in karonese tweet using machine learning. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 10(1):219–231.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023a. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023b. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

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 and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. 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. *Preprint*, arXiv:2404.01854.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Thang Le and Anh Luu. 2023. A parallel corpus for Vietnamese central-northern dialect text transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13839–13855, Singapore. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

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

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. 2023b. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958, Toronto, Canada. 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.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *Preprint*, arXiv:2309.08591.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

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.

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

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Aad Muzad and Faisal Rahutomo. 2016. Korpus berita daring bahasa indonesia dengan depth first focused crawling. *Prosiding Sentrinov (Seminar Nasional Terapan Riset Inovatif)*, 2(1):11–20.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Xuan-Phi Nguyen, Wenxuan Zhang, Li Xin, Mahani Aljunied, Weiwen Xu, Hou Pong Chan, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms - large language models for southeast asia. *Preprint*, arXiv:arXiv:2312.00738.

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 Ryder, 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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Chester Palen-Michel and Constantine Lignos. 2023. LR-sum: Summarization for less-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6829–6844, Toronto, Canada. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai natural language processing in python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36, Singapore. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Surapon Nonesung, Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Chompakorn Chaksangchaichot, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Wangchanlion and wangchanx mrc eval. *Preprint*, arXiv:2403.16127.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

The Joshua Project. 2024. The joshua project.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Ayu Purwarianti, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2007. A machine learning approach for Indonesian question answering system. In *Artificial Intelligence and Applications*, pages 573–578.

I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2024. Snli indo: A recognizing textual entailment dataset in indonesian derived from the stanford natural language inference dataset. *Data in Brief*, 52:109998.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Riccosan and Karen Etania Saputra. 2023. Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review. *Data in Brief*, 50:109576.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Muhammad Razif Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. 2023. Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6. IEEE.

Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. 2023. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In *Proc. INTERSPEECH 2023*, pages 2268–2272.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Auliya Sani, Sakriani Sakti, Graham Neubig, Tomoki Toda, Adi Mulyanto, and Satoshi Nakamura. 2012. Towards language preservation: Preliminary collection and vowel analysis of indonesian ethnic speech data. In *2012 International Conference on Speech Database and Assessments*, pages 118–122.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Ken Nabila Setya and Rahmad Mahendra. 2018. Semi-supervised textual entailment on indonesian wikipedia data. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 416–427. Springer.

AI Singapore. 2023. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. https://github.com/aisingapore/sealion.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rhio Sutoyo, Said Achmad, Andry Chowanda, Esther Widhi Andangsari, and Sani M. Isa. 2022.

Prdect-id: Indonesian product reviews dataset for emotions classification tasks. *Data in Brief*, 44:108554.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Khanh Quoc Tran, Phap Ngoc Trinh, Khoa Nguyen-Anh Tran, An Tran-Hoai Le, Luan Van Ha, and Kiet Van Nguyen. 2021. An empirical investigation of online news classification on an open-domain, large-scale and high-quality dataset in vietnamese. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 367–379. IOS Press.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. New vietnamese corpus for machine reading comprehension of health news articles. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *NAACL*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. 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.

Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. 2024. Miners: Multilingual language models as semantic retrievers. *arXiv preprint arXiv:2406.07424*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023a. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 5484–5505. Curran Associates, Inc.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ICLR*.

A Key Takeaways of SEACrowd

Key findings include:

Model Performance.

- **LLMs:** SEA-specific models, such as AYA-101 and mT0, show strong performance on zero-shot tasks, outperforming English or country-specific models in the region. However, tasks like abstractive QA and summarization reveal limitations in existing models' ability to handle SEA languages effectively.
- **Speech:** Off-the-shelf models like Whisper v3 show competitive ASR performance for major SEA languages but struggle with indigenous languages. In contrast, Seamless M4T v2 offers more balanced results across SEA languages.
- **VLMs:** Current VLMs fail to generate high-quality image captions in SEA languages, highlighting the need for more effective multilingual pre-training.

LLM Generation Quality. SEA language outputs by LLMs are often plagued by translationese, with models like SEA-LION v1 producing natural sentences only 57.71% of the time. Languages like Tagalog, Burmese, and Malay suffer from unnatural generation.

Resource Gaps. SEACrowd covers 74.9% of SEA languages but reveals a long-tail distribution, where most languages lack comprehensive datasets. SEA languages also face cultural misrepresentation, with 70% of datasets being translations rather than culturally relevant sources.

Prioritizing Development. Focus should be placed on SEA national languages with significant gaps in naturalness (e.g., Malay, Burmese, Filipino), as well as under-resourced local languages like Javanese and Cebuano.

Collaboration. Governments, industries, and local communities must invest in R&D, data collection, and open collaboration to address resource equity and improve SEA AI development.

B Related Work

SEA data resources LLM research efforts for SEA languages are limited by the lack of available datasets and benchmarks. Up to this day, resources for SEA NLP tasks are concentrated on relatively higher-resource SEA indigenous languages,

Benchmark	# Languages	# Indigenous SEA Languages	# Datasets	# Tasks
SEACrowd (**ours**)†	39	38	254	13 (11 text, 1 speech, 1 vision)
NusaCrowd† (Cahyawijaya et al., 2023a)	19	19	137	12 (11 text, 1 speech)
BUFFET (Asai et al., 2023)	54	N/A	15	8 (8 text)
XTREME-UP (Ruder et al., 2023)	88	11	269	9 (7 text, 1 speech, 1 vision)

Table 3: Benchmark comparison. †The numbers in SEACrowd and NusaCrowd are the numbers of datasets included in the evaluation.

such as Indonesian (Mahendra et al., 2021; Wilie et al., 2020; Cahyawijaya et al., 2021, 2023a) and Vietnamese (Nguyen et al., 2020; Huynh et al., 2022; Le and Luu, 2023; Van Nguyen et al., 2022). NusaCrowd (Cahyawijaya et al., 2023a) introduce the first multimodal benchmark for Indonesian languages, including text and speech. Ruder et al. (2023) introduce a multimodal benchmark encompassing 11 indigenous languages from SEA, spanning a wide array of languages totaling 88.

Additionally, Asai et al. (2023) present an LLM benchmark for cross-lingual few-shot transfer, comprising 15 distinct tasks and 54 languages sourced from varied multilingual datasets. Furthermore, Dou et al. (2024) find that publicly available pretraining data for SEA languages suffer from quality issues such as textual duplicates and excessive occurrences of Unicode escapes. On the other hand, pre-trained LLMs specifically for SEA languages suffer from limited language coverage; for instance, Cendol (Cahyawijaya et al., 2024b), Sailor (Dou et al., 2024), SEA-LION (Singapore, 2023), and SeaLLMs (Nguyen et al., 2023) have only covered up to 11 different SEA languages, including English and Chinese.

Open-source Community Initiatives in NLP Open-source and open-science communities play a crucial role in engaging native speakers to curate large-scale multilingual NLP resources. In the past, collaborative efforts have been organized to collect data and train multilingual language models either on a global scale (Workshop et al., 2022; Singh et al., 2024; Üstün et al., 2024) or on a regional level, e.g., Masakhane for African languages (Adelani et al., 2021, 2022b,a, 2023), AI4Bharat for Indian languages (Kakwani et al., 2020; Kumar et al., 2022; Dabre et al., 2022, inter alia), and Americas-NLP for Latin American languages (Mager et al., 2021; Ebrahimi et al., 2022).

In the SEA region, there have been community-based initiatives, e.g., IndoNLP, PyThaiNLP, and RojakNLP, to study NLP on Indonesian languages (Aji et al., 2022; Wilie et al., 2020; Cahyawijaya et al., 2021, 2023a), Thai language (Phatthiyaphaibun et al., 2023), and the code-switching phenomenon in SEA (Aji et al., 2023; Yong et al., 2023; Winata et al., 2024), respectively.

Submission	Points	Max points
Public datasheet	2+bonus	6
Dataloader	3	6 if difficult
Private datasheet	1	-
Access to private data	4+bonus	10 if high-quality
Datasheet review	1	1
Dataloader review	2	4 if difficult
Private datasheet review	0.5	-
Private data contact	1	5 if succeeds

Table 4: Amount of points obtained for contributions related to datasheet, dataloader, and private data.

C Contributing to SEACrowd

C.1 Open Contributions

We identify four tasks for open contribution in SEACrowd.[12] These tasks and the workflow of SEACrowd are heavily influenced by and extended upon NusaCrowd (Cahyawijaya et al., 2023a, 2022), a collaborative effort to pool data resources for Indonesian NLP.

- **Submitting Metadata for Existing Public Datasets.** Contributors can submit detailed datasheets for existing datasets through this form.[13] Contributors must provide important information such as data license, size, language and dialect, annotation method, and so on. The approved datasheets, as well as under review datasheets, will show up and be indexed in a monitor spreadsheet and the SEACrowd Catalogue (Figure 7).
- **Building a Dataloader.** From the approved datasheets from the previous task, contributors can further contribute by building a HuggingFace dataset loader to ensure that all datasets

[12]Landing page: https://github.com/SEACrowd.

[13]Public datasheet form: https://form.jotform.com/team/232952680898069/seacrowd-sea-datasets.

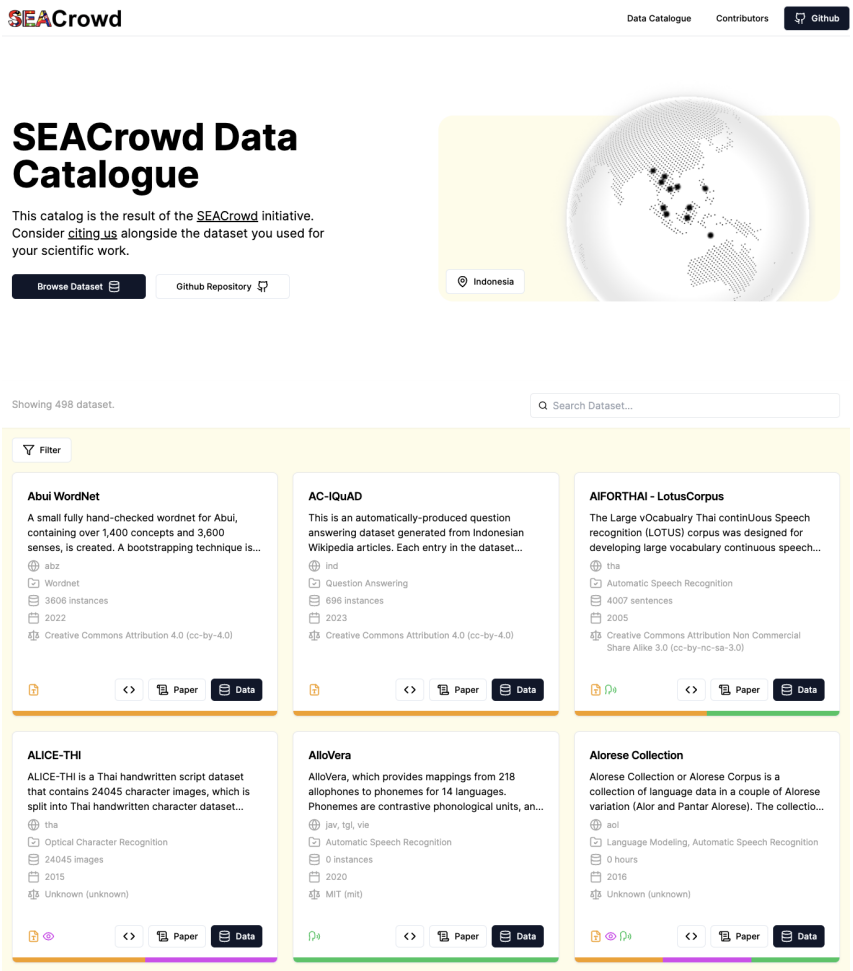


Figure 7: A glimpse of SEACrowd Catalogue.

in SEACrowd are standardized in terms of formatting and usage. Contributors can follow a dataloader guide and examples available[14] in the SEACrowd Data Hub. Dataloader maintainers and reviewers also monitor the self-assigned dataloader issues after 2 weeks of inactivity and ping contributors in case of a blocking impediment.

- **Identifying Private AI Datasets for SEA Languages, Cultures, and/or Regions.** Unfortunately, a number of prior works involving SEA languages are still not publicly available. These may be due to several different reasons, including (but not limited to): non-release contracts related to funding, inclusion of private and personally identifiable data, and the use of explicitly private data such as those used by for-profit companies.
  In this task, contributors can search for works that contain private data and fill out a corresponding record form.[15] The SEACrowd team then attempts to contact the original data owners and negotiate the open-sourcing of their resources.
- **Opening a Private AI Dataset of SEA.** If a contributor has previous work with closed data (or has been contacted by the SEACrowd team regarding closed-source data), they can decide to release their resources and register them in the collection via the public datasheet form. The resource will still be owned by the original contributor and is still tied to the contributor's previous work, as SEACrowd simply catalogs it and records its now open-source license.

[14] Dataloader guide: https://github.com/SEACrowd/seacrowd-datahub/blob/master/DATALOADER.md.

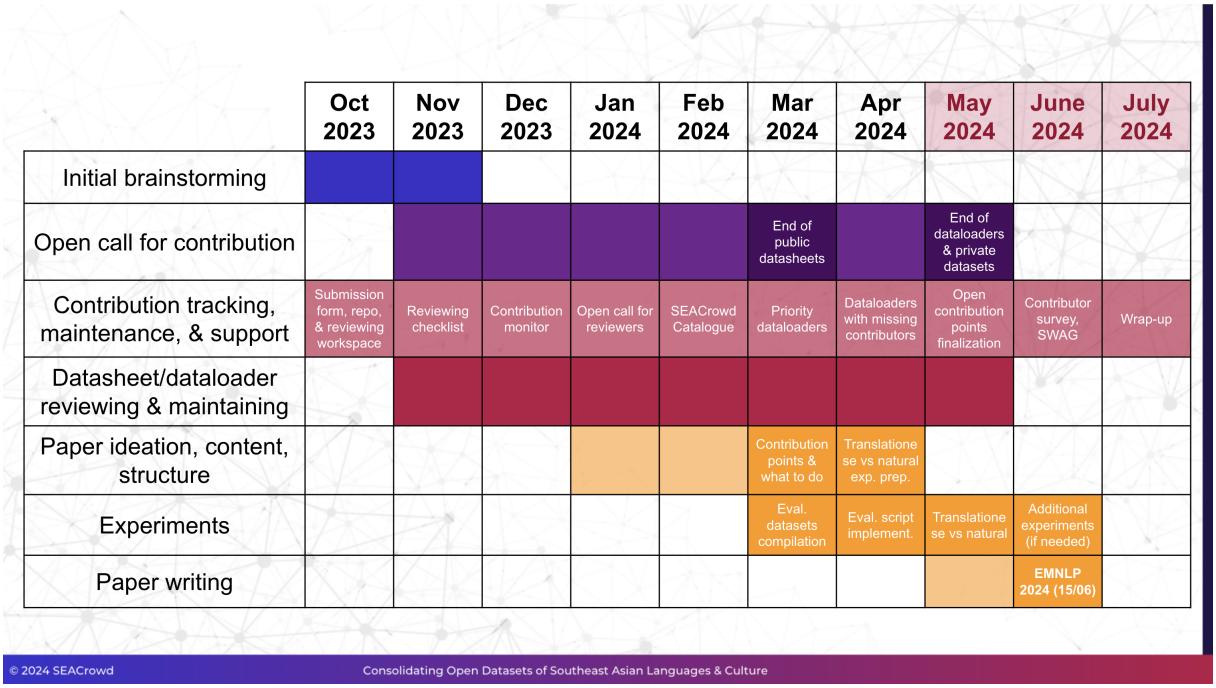


Figure 8: The timeline of SEACrowd's entire run.

C.2 Measuring Contributions

To be considered as a co-author, 20 contribution points are required.[16] To monitor how many points the contributors have obtained, the contribution point tracking is provided and updated regularly. The purpose of the point system is not to barrier collaboration but to reward rare and high-quality dataset entries. Table 4 describes the contribution points.[17] A bonus of 1 point is given if the dataset modality is speech or vision. We also provide a bonus based on the language rarity in terms of available resources as defined by Joshi et al. (2020)[18], consisting of 1 point for languages in level 1 and 2, and 2 points for languages in level 0 or absent from the list. For other contributions not mentioned in Table 4 (e.g., maintenance, design, experiment, paper writing, etc.), the amount of contribution points is adjusted to the bulk and the complexity of the relevant work.

D Progression of SEACrowd

D.1 Timeline

SEACrowd released the open call for contributions on 1 November 2023. This lasted until 31 March 2024, for datasheet submissions, and until 15 May 2024 for both dataloaders and private dataset submissions. SEACrowd contributors have a biweekly discussion regarding the challenges they face while contributing, the next steps they should take to proceed, and/or experiment and research ideas for the paper. The detailed timeline can be seen in Figure 8.

D.2 Contribution Progress

Figure 9 shows the number of submissions for public datasheets, dataloader pull requests, and papers with private datasets in SEACrowd.

E Reviewing SEACrowd's Submissions

We provide the complete reviewing guidelines in our Data Hub.[19]

E.1 Datasheet Reviewing

The datasheet reviewing standard operating procedure (SOP) ensures the integrity and completeness of datasets submitted to SEACrowd. It outlines procedures for verifying dataset availability,

[15] Papers with private dataset form: https://form.jotform.com/team/232952680898069/seacrowd-paper-with-private-dataset.

[16] Submissions past the deadlines (see Appendix D.1) are still recorded, but contribution points are no longer given.

[17] Contribution point guidelines: https://github.com/SEACrowd/seacrowd-datahub/blob/master/POINTS.md.

[18] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

[19] Reviewer SOP: https://github.com/SEACrowd/seacrowd-datahub/blob/master/REVIEWING.md

avoiding duplicates, and ensuring correctness and relevance to the SEA region. The SOP includes FAQs addressing common issues such as dataset duplicates and incorrect information, along with an approval checklist covering aspects like data availability, dataset splits, and licensing. Reviewers are instructed on how to handle various scenarios, including correcting errors and determining points allocation for multiple contributors. For instance, if the datasheet submitted has incorrect or missing information, the reviewer can either ask the contributor to fix it (with some guidance) or fix it themself. Upon completion of the review, reviewers update the status, add notes and points, and await the generation of a GitHub issue for the approved datasheet.

E.2 Dataloader Reviewing

The dataloader reviewing SOP governs the review process for dataloaders in SEACrowd, ensuring adherence to the data structure and `seacrowd` schema and config standards. It specifies checks for metadata correctness, subset implementation, test script passing, and adherence to coding conventions. Additionally, it outlines dataloader config rules based on dataset types and provides guidelines for multilingual datasets. The SOP emphasizes the importance of reviewer collaboration, with each dataloader requiring two reviewers per submitted pull request, and outlines the approval and reviewer assignment process, either by allocation or by self-assignment based on availability and promptness.

F Schemas in SEACrowd

Schemas define and format the attributes of the dataset returned by a dataloader. For each dataloader, we implement 2 schema types: the source schema and the `seacrowd` schema. The source schema presents the dataset in a format similar to its original structure, while the `seacrowd` schema standardizes the data structure across similar tasks.

The following subsections define the `seacrowd` schemas in NLP (F.1), speech (F.2), and VL (F.3).

F.1 NLP

- **Unlabeled text (`SSP`)**. This schema could be used for language modeling in self-supervised pre-training. It consists of `(id, text)`, where `id` denotes a unique row identifier of the dataset and `text` denotes an input text.
- **Single-label text classification (`TEXT`)**. This schema could be used for sentiment analysis, emotion classification, legal classification, and others. It consists of `(id, text, label)`, where `id` denotes a unique row identifier of the dataset, `text` denotes an input text, and `label` denotes a deterministic target variable.
- **Multi-label text classification (`TEXT MULTI`)**. This schema could be used for hate speech detection and aspect-based sentiment analysis. It consists of `(id, text, labels)`, where `id` denotes a unique row identifier of the dataset, `text` denotes an input text, and `labels` denotes a list of deterministic target variables.
- **Text-to-text (`T2T`)**. This schema could be used for machine translation, summarization, and paraphrasing. It consists of `(id, text_1, text_2, text_1_name, text_2_name)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote

Subset ID	Language	Region	# Samples
Sentiment Analysis → `*_seacrowd_text`			
lazada_review_filipino	FIL	Philippines	1001
gklmip_sentiment	MYA	Myanmar	716
indolem_sentiment	IND	Indonesia	1011
id_sentiment_analysis	IND	Indonesia	10806
karonese_sentiment	BTX	Indonesia	1000
wisesight_thai_sentiment	THA	Thailand	2671
wongnai_reviews	THA	Thailand	6203
typhoon_yolanda_tweets	FIL	Philippines	153
smsa	IND	Indonesia	500
prdect_id_sentiment	IND	Indonesia	5400
id_sent_emo_mobile_apps_sentiment	IND	Indonesia	21696
shopee_reviews_tagalog	FIL	Philippines	2250
nusatranslation_senti_abs	ABS	Indonesia	500
nusatranslation_senti_btk	BTX	Indonesia	1200
nusatranslation_senti_bew	BEW	Indonesia	1200
nusatranslation_senti_bhp	BHP	Indonesia	500
nusatranslation_senti_jav	JAV	Indonesia	1200
nusatranslation_senti_mad	MAD	Indonesia	1200
nusatranslation_senti_mak	MAK	Indonesia	1200
nusatranslation_senti_min	MIN	Indonesia	1200
nusatranslation_senti_mui	MUI	Indonesia	500
nusatranslation_senti_rej	REJ	Indonesia	500
nusatranslation_senti_sun	SUN	Indonesia	1200
nusax_senti_ind	IND	Indonesia	400
nusax_senti_ace	ACE	Indonesia	400
nusax_senti_jav	JAV	Indonesia	400
nusax_senti_sun	SUN	Indonesia	400
nusax_senti_min	MIN	Indonesia	400
nusax_senti_bug	BUG	Indonesia	400
nusax_senti_bbc	BBC	Indonesia	400
nusax_senti_ban	BAN	Indonesia	400
nusax_senti_nij	NIJ	Indonesia	400
nusax_senti_mad	MAD	Indonesia	400
nusax_senti_bjn	BJN	Indonesia	400
nusax_senti_eng	ENG	Non-indigenous	400
indonglish	IND	Indonesia	1011

Table 5: Sentiment analysis data subsets used in SEACrowd NLU evaluation.

Subset ID	Language	Region	# Samples
NLI → `*_seacrowd_pairs`			
indonli	IND	Indonesia	5183
wrete	IND	Indonesia	100
snli_indo	IND	Indonesia	9823
myxnli	MYA	Myanmar	5010
xnli.tha	THA	Thailand	5010
xnli.vie	VIE	Vietnam	5010

Table 6: NLI data subsets used in SEACrowd NLU evaluation.

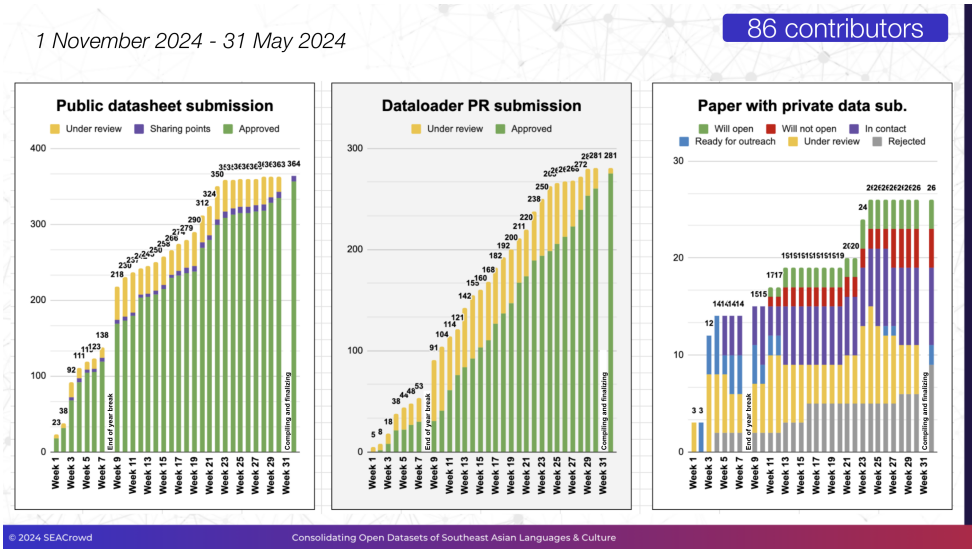


Figure 9: Weekly status update of the cumulative number of submissions in SEACrowd.

an input text pair, and `text_1_name` and `text_2_name` denote the names of the input text pair (e.g., `ind` and `jav` for translation input text pairs, or `document` and `summary` for summarization input text pairs).

- **Sequence labeling (`SEQ LABEL`)**. This schema could be used for named entity recognition (NER), POS tagging, and others. It consists of `(id, tokens, labels)`, where `id` denotes a unique row identifier of the dataset, `tokens` denotes a list of tokens of an input text, and `labels` denotes a list of targets for the tokens.
- **Question answering (`QA`)**. This schema could be used for extractive QA, multiple-choice QA, and others. It consists of `(id, question_id, document_id, question, type, choices, context, answer)`, where `id` denotes a unique row identifier of the dataset, `question_id` denotes a unique identifier of the question, `document_id` denotes a unique identifier of the context document, `question` denotes an input question to be answered, `type` denotes the type of the QA task (e.g., extractive, multiple-choice, open-generative, closed-generative, etc.), `choices` denotes a list of answer choices (if required), `context` denotes a passage that serves as the background information of the question (if required), and `answer` denotes the gold answer to the question (if required).
- **Single-label text pair classification (`PAIRS`)**. This could be used for textual entailment and next-sentence prediction. It consists of `(id, text_1, text_2, label)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `label` denotes the target variable.
- **Single-label text pair classification with continuous values or regression (`PAIRS SCORE`)**. This could be used for answer grading and semantic textual similarity. It consists of `(id, text_1, text_2, label)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `label` denotes a target variable as a continuous value.
- **Multi-label text pair classification (`PAIRS MULTI`)**. This could be used for morphological inflection. It consists of `(id, text_1, text_2, labels)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `labels` denotes a list of target variables.
- **Knowledge base (`KB`)**. This schema could be used for constituency parsing, dependency parsing, coreference resolution, dialogue systems, and other tasks with complex structures. It consists of `(id, passages, entities, events, coreferences, relations)`. Considering its intricate structure, we encourage readers to take a look at the implementation of the knowledge base schema.
- **Tree (`TREE`)**. This schema could be used for constituency parsing, this schema assumes a document with subnode elements and a tree hierarchy. It consists of `(id, passage,`

Subset ID	Language	Region	# Samples
Topic Classification → *_seacrowd_text			
gklmip_newsclass	KHM	Cambodia	1436
indonesian_news_dataset	IND	Indonesia	2627
uit_vion	VIE	Vietnam	26000
sib_200_ace_Arab	ACE	Indonesia	204
sib_200_ace_Latn	ACE	Indonesia	204
sib_200_ban_Latn	BAN	Indonesia	204
sib_200_bjn_Arab	BJN	Indonesia	204
sib_200_bjn_Latn	BJN	Indonesia	204
sib_200_bug_Latn	BUG	Indonesia	204
sib_200_ceb_Latn	CEB	Philippines	204
sib_200_ilo_Latn	ILO	Philippines	204
sib_200_ind_Latn	IND	Indonesia	204
sib_200_jav_Latn	JAV	Indonesia	204
sib_200_kac_Latn	KAC	Myanmar	204
sib_200_khm_Khmr	KHM	Cambodia	204
sib_200_lao_Laoo	LAO	Laos	204
sib_200_lus_Latn	LUS	Myanmar	204
sib_200_min_Arab	MIN	Indonesia	204
sib_200_min_Latn	MIN	Indonesia	204
sib_200_mya_Mymr	MYA	Myanmar	204
sib_200_pag_Latn	PAG	Philippines	204
sib_200_shn_Mymr	SHN	Myanmar	204
sib_200_sun_Latn	SUN	Indonesia	204
sib_200_tgl_Latn	FIL	Philippines	204
sib_200_tha_Thai	THA	Thailand	204
sib_200_vie_Latn	VIE	Non-indigenous	204
sib_200_war_Latn	WAR	Philippines	204
sib_200_zsm_Latn	ZSM	Malaysia	204
nusaparagraph_topic_btk	BTX	Indonesia	500
nusaparagraph_topic_bew	BEW	Indonesia	800
nusaparagraph_topic_bug	BUG	Indonesia	300
nusaparagraph_topic_jav	JAV	Indonesia	800
nusaparagraph_topic_mad	MAD	Indonesia	700
nusaparagraph_topic_mak	MAK	Indonesia	700
nusaparagraph_topic_min	MIN	Indonesia	800
nusaparagraph_topic_mui	MUI	Indonesia	400
nusaparagraph_topic_rej	REJ	Indonesia	350
nusaparagraph_topic_sun	SUN	Indonesia	900

Table 7: Topic classification data subsets used in SEACrowd NLU evaluation.

Subset ID	Language	Region	# Samples
Commonsense Reasoning → *_seacrowd_text/qa			
emotes_3k_tgl	FIL	Philippines	2905
emotes_3k_eng	ENG	Non-indigenous	2905
indo_story_cloze	IND	Indonesia	1135
xstorycloze_id	IND	Indonesia	1511
xstorycloze_my	MYA	Myanmar	1511

Table 8: Commonsense reasoning data subsets used in SEACrowd NLU evaluation.

Subset ID	Language	Region	# Samples
Standard Testing QA → *_seacrowd_qa			
indommlu_ind	IND	Indonesia	14979
indommlu_ban	BAN	Indonesia	14979
indommlu_mad	MAD	Indonesia	14979
indommlu_mak	MAK	Indonesia	14979
indommlu_sun	SUN	Indonesia	14979
indommlu_jav	JAV	Indonesia	14979
indommlu_bjn	BJN	Indonesia	14979
indommlu_abl	ABL	Indonesia	14979
indommlu_nij	NIJ	Indonesia	14979
seaeval_cross_mmlu_ind	IND	Indonesia	150
seaeval_cross_mmlu_vie	VIE	Vietnam	150
seaeval_cross_mmlu_zlm	ZSM	Malaysia	150
seaeval_cross_mmlu_fil	FIL	Philippines	150
seaeval_cross_logiqa_ind	IND	Indonesia	176
seaeval_cross_logiqa_vie	VIE	Vietnam	176
seaeval_cross_logiqa_zlm	ZSM	Malaysia	176
seaeval_cross_logiqa_fil	FIL	Philippines	176
m3exam_jav	JAV	Indonesia	371
m3exam_tha	THA	Thailand	2168
m3exam_vie	VIE	Vietnam	1789
okapi_m_arc_ind	IND	Indonesia	1170
okapi_m_arc_vie	VIE	Vietnam	1170
Cultural QA → *_seacrowd_qa			
copal_colloquial	IND	Indonesia	559
xcopa_tha	THA	Thailand	500
xcopa_vie	VIE	Vietnam	500
xcopa_ind	IND	Indonesia	500
seaeval_sg_eval_eng	ENG	Non-indigenous	103
seaeval_ph_eval_eng	ENG	Non-indigenous	100
mabl_ind	IND	Indonesia	1140
mabl_jav	JAV	Indonesia	600
mabl_sun	SUN	Indonesia	600
Reading Comprehension QA → *_seacrowd_qa			
belebele_ceb_latn	CEB	Philippines	900
belebele_ilo_latn	ILO	Philippines	900
belebele_ind_latn	IND	Indonesia	900
belebele_jav_latn	JAV	Indonesia	900
belebele_kac_latn	KAC	Myanmar	900
belebele_khm_khmr	KHM	Cambodia	900
belebele_lao_laoo	LAO	Laos	900
belebele_mya_mymr	MYA	Myanmar	900
belebele_shn_mymr	SHN	Myanmar	900
belebele_sun_latn	SUN	Indonesia	900
belebele_tgl_latn	FIL	Philippines	900
belebele_tha_thai	THA	Thailand	900
belebele_vie_latn	VIE	Vietnam	900
belebele_war_latn	WAR	Philippines	900
belebele_zsm_latn	ZSM	Malaysia	900

Table 9: Multiple-choice QA data subsets used in SEACrowd NLU evaluation.

nodes), where id denotes a unique row identifier of the dataset, passage denotes the passage to that particular id; this passage consist of (id, type, text, offsets), nodes denotes the nodes to that particular id; this nodes consists of (id, type, text, offsets, subnodes).

- **Conversational Chat (CHAT)**. This schema could be used for conversational chat and/or multi-turn conversation. It consists of (id, input, output, meta), where id denotes a unique row identifier of the dataset, input denotes a sequence that consists of content and role as an input prompt and the role of the entity inputting the prompt, output denotes an answer from that input prompt, and meta denotes relevant details to allow some flexibility of the schema (if required).
- **End-to-end Task Oriented Dialogue (TOD)**. This schema could be used for end-to-end task-oriented dialogue. It consists of (dialogue_idx, dialogue), where dialogue_idx denotes a unique row identifier of the dialogue, dialogue denotes some core details such as turn label, system utterance, turn idx, belief state (consist of slots and act), user utterance, and system acts.

Subset ID	Language	Region	# Samples
Extractive & Abstractive QA → *_seacrowd_qa			
facqa	IND	Indonesia	311
iapp_squad	THA	Thailand	739
qasina	IND	Indonesia	500
mkqa_khm	KHM	Cambodia	10000
mkqa_zsm	ZSM	Malaysia	10000
mkqa_tha	THA	Thailand	10000
mkqa_vie	VIE	Vietnam	10000

Table 10: Extractive and abstractive QA subsets used in SEACrowd NLG evaluation.

Subset ID	Language	Region	# Samples
Summarization → *_seacrowd_t2t			
lr_sum_ind	IND	Indonesia	500
lr_sum_vie	VIE	Vietnam	1460
lr_sum_lao	LAO	Laos	1496
lr_sum_tha	THA	Thailand	500
lr_sum_khm	KHM	Cambodia	486
lr_sum_mya	MYA	Myanmar	990
xl_sum_mya	MYA	Myanmar	570
xl_sum_ind	IND	Indonesia	4780
xl_sum_tha	THA	Thailand	826
xl_sum_vie	VIE	Vietnam	4013

Table 11: Summarization data subsets used in SEACrowd NLG evaluation.

Subset ID	Language	Region	# Samples
Image Captioning → *_seacrowd_imtext			
xm3600_fil	FIL	Philippines	2760
xm3600_id	IND	Indonesia	2775
xm3600_th	THA	Thailand	2798
xm3600_vi	VIE	Vietnam	2855

Table 12: Image captioning data subsets used in SEACrowd VL evaluation.

F.2 Speech

- **Speech-text (SPTEXT)**. This could be used for speech recognition, text-to-speech (TTS) or speech synthesis, and speech-to-text translation. It consists of (id, path, audio, text, speaker_id, metadata), where id denotes a unique row identifier of the dataset, path denotes the file path to an input audio source, audio denotes the audio data loaded from the corresponding path, text denotes an input text, speaker_id denotes a unique identifier of the speaker, metadata denotes relevant details such as the age and gender of the speaker (if required).
- **Speech-to-speech (S2S)**. This could be used for speech-to-speech translation. It consists of (id, path_1, audio_1, text_1, metadata_1, path_2, audio_2, text_2, metadata_2), where id denotes a unique row identifier of the dataset, path_1 and path_2 denote the file path to a respective input audio source, audio_1 and audio_2 denote the audio data loaded from the corresponding path, text_1 and text_2 denote input texts, and metadata_1 and metadata_2 denote relevant details such as the age of the speaker and their gender (if required).
- **Speech Classification (SPEECH)**. This schema could be used for speech classification, speech-language identification, and speech-emotion recognition for single-label use only. It consists of (id, path, audio, speaker_id, labels, metadata), where id denotes a unique row identifier of the dataset, path denotes the file path to an input audio source, audio denotes the audio data loaded from the corresponding path, speaker_id denotes a unique identifier of the speaker, labels denotes the label of that particular speech (only can be single-label), metadata denotes relevant details such as the age and gender of the speaker (if required).
- **Speech Classification for Multilabel (SPEECH MULTILABEL)**. This schema could be used for speech classification, speech-language identification, and speech-emotion recognition for multi-label use only. It consists of (id, path, audio, speaker_id, labels, metadata), where id denotes a unique row identifier of the dataset, path denotes the file path to an input audio source, audio denotes the audio data loaded from the corresponding path, speaker_id denotes a unique identifier of the speaker, labels denotes the sequence of labels of that particular speech (only can be multi-label), metadata denotes relevant details such as the age and gender of the speaker (if required).

F.3 VL

- **Image-text (IMTEXT)**. This schema could be used for image captioning, text-to-image generation, and vision-language pre-training. It consists of (id, text, image_paths, metadata), where id denotes a unique row identifier of the dataset, text denotes an input text, image_paths denotes a list of paths to the input image sources, and metadata denotes relevant details such as visual concepts and labels (if required).
- **General Image Classification (IMAGE)**. This schema could be used for image classifica-

Subset ID	Language	Region	# Samples
ASR → *_seacrowd_sptext			
asr_ibsc	IBA	Brunei	473
commonvoice_120_ind	IND	Indonesia	3647
commonvoice_120_tha	THA	Thailand	10964
commonvoice_120_cnh	CNH	Myanmar	763
commonvoice_120_vie	VIE	Vietnam	1302
fleurs_ind	IND	Indonesia	687
fleurs_jav	JAV	Indonesia	728
fleurs_tha	THA	Thailand	1021
fleurs_lao	LAO	Laos	405
fleurs_mya	MYA	Myanmar	880
fleurs_khm	KHM	Cambodia	771
fleurs_vie	VIE	Vietnam	857
fleurs_zlm	ZLM	Malaysia	749
fleurs_fil	FIL	Philippines	964
fleurs_ceb	CEB	Philippines	541
indspeech_newstra_ethnicsr_nooverlap_jav	JAV	Indonesia	1000
indspeech_newstra_ethnicsr_nooverlap_sun	SUN	Indonesia	1000
indspeech_newstra_ethnicsr_nooverlap_ban	BAN	Indonesia	1000
indspeech_newstra_ethnicsr_nooverlap_btk	BTX	Indonesia	1000

Table 13: ASR data subsets used in SEACrowd speech evaluation.

tion both single-label and multi-label. It consists of (`id`, `labels`, `image_path`, `metadata`), where `id` denotes a unique row identifier of the dataset, `labels` denotes the label of that particular image (can be single-label and multi-label), `image_path` denotes a list of paths to the input image sources, and `metadata` denotes relevant details such as visual concepts and labels (if required).

- **Image Question Answering (`IMQA`)**. This schema could be used for image/visual question answering. It consists of (`id`, `question_id`, `document_id`, `questions`, `type`, `choices`, `context`, `answer`, `image_paths`, `meta`), where `id` denotes a unique row identifier of the dataset, `question_id` denotes a unique identifier of the question, `document_id` denotes a unique identifier of the context document, `question` denotes an input question to be answered, `type` denotes the type of the QA task (e.g., extractive, multiple-choice, open-generative, closed-generative, etc.), `choices` denotes a list of answer choices (if required), `context` denotes a passage that serves as the background information of the question (if required), and `answer` denotes the gold answer to the question (if required), `image_path` denotes a list of paths to the input image sources, and `metadata` denotes relevant details to allow some flexibility of the schema (if required).
- **General Video-to-Text (`VIDEO`)**. This schema could be used for video-to-text retrieval and video captioning. It consists of (`id`, `video_path`, `text`, `metadata`), where `id` denotes a unique row identifier of the dataset, `video_path` denotes the file path to an input video source, `text` denotes the text associated with that particular frame/video, `metadata` denotes relevant details such as the resolution, duration, and FPS of the video (if required).

G Supplementary Details for SEA Evaluation

G.1 Datasets

Table 5, 6, 7, 8, and 9 provide the details of data subsets used in the NLU evaluation. Sentiment analysis dataset is originally from NusaX (Winata et al., 2023), NusaTranslation (Cahyawijaya et al., 2023b), SentiTaglish[20], SmSA (Purwarianti and Crisdayanti, 2019), PRDECT-ID (Sutoyo et al., 2022), code-mixed Indonesian-English sentiment (Astuti et al., 2023), Karonese tweet sentiment (Karo et al., 2022), Typhoon Yolanda sentiment (Imperial et al., 2019), GKLMIP Khmer sentiment (Jiang et al., 2022), Wisesight sentiment corpus[21], Filipino-Tagalog product reviews Sentiment[22], and multilabel sentiment of Indonesian mobile apps review (Riccosan and Saputra, 2023).

Topic classification dataset is originally from NusaParagraph (Cahyawijaya et al., 2023b), UIT-ViON (Tran et al., 2021), SIB-200 (Adelani et al., 2024), GKLMIP Khmer news (Jiang et al., 2022), and Indonesian news (Muzad and Rahutomo, 2016). Natural Language Inference dataset is originally from IndoNLI (Mahendra et al., 2021), WreTe (Setya and Mahendra, 2018), SNLI Indo (Putra et al., 2024), MyXNLI[23], and XNLI (Conneau et al., 2018). Commonsense reasoning dataset is originally from XStoryCloze (Lin et al., 2022), IndoCloze (Koto et al., 2022), and EMoTES-3K (Catapang and Visperas, 2023).

Open domain QA dataset is originally from IndoMMLU (Koto et al., 2023b), SeaEval (Wang et al., 2023), M3Exam (Zhang et al., 2023b), and Okapi (Dac Lai et al., 2023). Cultural QA dataset is originally from COPAL-ID (Wibowo et al., 2023), XCOPA (Ponti et al., 2020), SeaEval (Wang et al., 2023), and Multilingual Fig-QA (Kabra et al.,

[20] https://huggingface.co/datasets/ccosme/SentiTaglishProductsAndServices

[21] https://github.com/PyThaiNLP/wisesight-sentiment

[22] https://github.com/EricEchemane/Filipino-Tagalog-Product-Reviews-Sentiment-Analysis

[23] https://huggingface.co/datasets/akhtet/myXNLI

Subset ID		Language	Region	# Samples
Eng → XX	XX → Eng			
MT (Eng ⇔ XX) → *_seacrowd_t2t				
lio_and_central_flores_eng_ljl	lio_and_central_flores_ljl_eng	LJL	Indonesia	1658
flores200_eng_Latn_ace_Latn	flores200_ace_Latn_eng_Latn	ACE	Indonesia	1012
flores200_eng_Latn_ban_Latn	flores200_ban_Latn_eng_Latn	BAN	Indonesia	1012
flores200_eng_Latn_bjn_Latn	flores200_bjn_Latn_eng_Latn	BJN	Indonesia	1012
flores200_eng_Latn_bug_Latn	flores200_bug_Latn_eng_Latn	BUG	Indonesia	1012
flores200_eng_Latn_ceb_Latn	flores200_ceb_Latn_eng_Latn	CEB	Philippines	1012
flores200_eng_Latn_ilo_Latn	flores200_ilo_Latn_eng_Latn	ILO	Philippines	1012
flores200_eng_Latn_ind_Latn	flores200_ind_Latn_eng_Latn	IND	Indonesia	1012
flores200_eng_Latn_jav_Latn	flores200_jav_Latn_eng_Latn	JAV	Indonesia	1012
flores200_eng_Latn_kac_Latn	flores200_kac_Latn_eng_Latn	KAC	Myanmar	1012
flores200_eng_Latn_khm_Khmr	flores200_khm_Khmr_eng_Latn	KHM	Cambodia	1012
flores200_eng_Latn_lao_Laoo	flores200_lao_Laoo_eng_Latn	LAO	Laos	1012
flores200_eng_Latn_lus_Latn	flores200_lus_Latn_eng_Latn	LUS	Myanmar	1012
flores200_eng_Latn_min_Latn	flores200_min_Latn_eng_Latn	MIN	Indonesia	1012
flores200_eng_Latn_mya_Mymr	flores200_mya_Mymr_eng_Latn	MYA	Myanmar	1012
flores200_eng_Latn_pag_Latn	flores200_pag_Latn_eng_Latn	PAG	Philippines	1012
flores200_eng_Latn_shn_Mymr	flores200_shn_Mymr_eng_Latn	SHN	Myanmar	1012
flores200_eng_Latn_sun_Latn	flores200_sun_Latn_eng_Latn	SUN	Indonesia	1012
flores200_eng_Latn_tha_Thai	flores200_tha_Thai_eng_Latn	THA	Thailand	1012
flores200_eng_Latn_vie_Latn	flores200_vie_Latn_eng_Latn	VIE	Vietnam	1012
flores200_eng_Latn_war_Latn	flores200_war_Latn_eng_Latn	WAR	Philippines	1012
flores200_eng_Latn_zsm_Latn	flores200_zsm_Latn_eng_Latn	ZSM	Malaysia	1012
ntrex_128_eng-US_ind	ntrex_128_ind_eng-US	IND	Indonesia	1997
ntrex_128_eng-US_mya	ntrex_128_mya_eng-US	MYA	Myanmar	1997
ntrex_128_eng-US_fil	ntrex_128_fil_eng-US	FIL	Philippines	1997
ntrex_128_eng-US_khm	ntrex_128_khm_eng-US	KHM	Cambodia	1997
ntrex_128_eng-US_lao	ntrex_128_lao_eng-US	LAO	Laos	1997
ntrex_128_eng-US_zlm	ntrex_128_zlm_eng-US	ZSM	Malaysia	1997
ntrex_128_eng-US_tha	ntrex_128_tha_eng-US	THA	Thailand	1997
ntrex_128_eng-US_vie	ntrex_128_vie_eng-US	VIE	Vietnam	1997
ntrex_128_eng-US_hmv	ntrex_128_hmv_eng-US	HMV	Vietnam	1997
nusax_mt_eng_ind	-	IND	Indonesia	400
nusax_mt_eng_ace	nusax_mt_ace_eng	ACE	Indonesia	400
nusax_mt_eng_jav	nusax_mt_jav_eng	JAV	Indonesia	400
nusax_mt_eng_sun	nusax_mt_sun_eng	SUN	Indonesia	400
nusax_mt_eng_min	nusax_mt_min_eng	MIN	Indonesia	400
nusax_mt_eng_bug	nusax_mt_bug_eng	BUG	Indonesia	400
nusax_mt_eng_bbc	nusax_mt_bbc_eng	BBC	Indonesia	400
nusax_mt_eng_ban	nusax_mt_ban_eng	BAN	Indonesia	400
nusax_mt_eng_nij	nusax_mt_nij_eng	NIJ	Indonesia	400
nusax_mt_eng_mad	nusax_mt_mad_eng	MAD	Indonesia	400
nusax_mt_eng_bjn	nusax_mt_bjn_eng	BJN	Indonesia	400

Table 14: MT between English and SEA languages data subsets used in SEACrowd NLG evaluation.

2023). The reading comprehension dataset is originally from Belebele (Bandarkar et al., 2023).

Table 10, 11, and 14 provide the details of data subsets used in the NLG evaluation. The summarization dataset is originally from LR-Sum (Palen-Michel and Lignos, 2023) and XL-Sum (Hasan et al., 2021). The machine translation dataset is originally from Lio and the Central Flores corpus (Elias, 2018), Flores-200 (Costa-jussà et al., 2024) and NTREX-128 (Federmann et al., 2022). Question answering dataset is originally from FacQA (Purwarianti et al., 2007), QASiNa (Rizqullah et al., 2023), MKQA (Longpre et al., 2021), and Open Thai Wikipedia QA dataset[24].

Table 12 and 13 provide the details of data subsets used in the VL and speech evaluation. The image captioning dataset is originally from XM3600 (Thapliyal et al., 2022). Speech recognition dataset is originally from INDspeech NEWSTRA Ethnic collection (Sani et al., 2012), ASR Iban (Juan et al., 2015), FLEURS (Conneau et al., 2022), and Common Voice (Ardila et al., 2020).

G.2 Baselines

Table 20, 21, and 22 report the details of baseline models used in SEACrowd evaluation (§3). For each baseline model, we provide information regarding the model size, origin base model, seen languages in the training corpora use, and the URL where the models can be downloaded. In principle, this work does not aim to acquire and fit all available SEA-trained LLMs over the Internet, as this is computationally expensive. Rather, we want

[24] https://zenodo.org/records/4539916

Model	$\tau = 0.01$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 1.0$
Commercial					
GPT-4	0.199	0.192	0.155	0.118	**0.066**
Command-R	0.201	0.198	0.185	0.168	0.126
English					
Mistral	0.161	0.160	0.159	0.162	0.150
Llama3	0.138	0.137	0.131	0.129	0.113
Falcon	0.274	0.272	0.238	0.250	0.211
Multilingual					
mT0	0.151	0.148	0.131	0.112	0.074
BLOOMZ	0.238	0.236	0.228	0.217	0.167
BactrianX-Llama	0.163	0.162	0.163	0.168	0.149
AYA-23	0.183	0.182	0.183	0.179	0.135
AYA-101	**0.112**	**0.109**	**0.095**	**0.085**	0.069
SEA regional					
SEA-LION	0.250	0.242	0.204	0.164	0.102
SeaLLM v2.5	0.137	0.133	0.116	0.097	0.069
Sailor	0.152	0.151	0.145	0.139	0.113
SEA country					
Cendol-mT5	0.407	0.404	0.378	0.328	0.200
Cendol-Llama2	0.294	0.290	0.267	0.232	0.149
Merak v4	0.209	0.207	0.199	0.190	0.155
WangchanX-Llama3	0.163	0.161	0.153	0.150	0.131
Malaysian Llama3	0.181	0.181	0.179	0.176	0.143

Table 15: Language equity across baselines based on Gini coefficient weighted by population with different τ values. Lower Gini means higher equity.

Model	Hyperparameter	Value
Logistic Regression	`max_iter`	`100`
	`C`	`np.linspace(0.001, 10, 100)`
Naive Bayes	`alpha`	`np.linspace(0.001, 1, 50)`
	`distribution`	`MultinomialNB`
SVM	`C`	`1`
	`kernel`	`["rbf", "linear"]`

Table 16: Hyper-parameters of classical models for Translationese prediction through grid search.

Model	3-label	HT vs. MT-Nat	MT vs. HT-Nat	Nat vs. HT-MT
LR (TF-IDF)	39.73	53.03	56.01	75.20
LR (BoW)	45.63	55.90	61.39	75.60
NB (TF-IDF)	33.43	49.53	50.55	73.05
NB (BoW)	33.70	49.10	50.64	71.26
SVM (TF-IDF)	39.55	52.63	55.10	76.40
SVM (BoW)	46.84	56.85	61.40	75.65
mDeBERTa	51.51	64.77	59.16	**79.08**

Table 17: Results of translationese classifier (accuracy) averaged across languages.

Country	Affiliation	Origin
Indonesia	16	31
Malaysia	0	1
Philippines	3	7
Singapore	13	2
Thailand	1	2
Vietnam	0	1
Australia	1	0
Brazil/Sweden	0	1
Canada	1	0
China	2	8
Egypt	0	1
Germany	0	2
Hong Kong	2	0
India	0	1
Ireland	1	0
Japan	3	0
The Netherlands	0	1
UAE	5	0
UK	4	0
USA	9	1
Uzbekistan	0	2

Table 18: The demographics of the authors based on affiliation country and origin country.

to initiate the exploration of select publicly available models to serve as baselines for the evaluation of foundational capabilities on SEA languages through benchmarking on NLU, NLG, speech, and vision tasks aggregated via SEACrowd.

Across the various models explored, as listed in the tables, we prioritized the diversity of model variation in terms of scale, openness, and coverage of SEA languages. In NLP tasks, we covered 5 LLM groups for the main experiments: English-only, multilingual, regional, and country-specific models. Instruction-tuned LLMs demonstrate the ability to generalize to unseen tasks (Wei et al., 2021; Sanh et al., 2021; Ouyang et al., 2022). Some of these LLMs are based on a multilingual foundation, hence their proficiency in generalizing across languages (Muennighoff et al., 2022; Adilazuarda et al., 2023; Zhang et al., 2023a). For NLU, we compute the weighted F1-score and obtain the answers via log-likelihood for open-source baselines or string matching for commercial baselines.

For the speech benchmark, only two model families are available: multilingual models and models fine-tuned on specific SEA languages. For vision tasks, we covered English-only and one multilingual model. These models utilize a visual backbone pre-trained on image-text alignment, e.g., CLIP (Radford et al., 2021), to project image features into the input space of an existing pre-trained LM. In summary, we mostly explored open models readily accessible on HuggingFace but also included commercial models such as GPT-4 and Whisper V3 for performance benchmarking, reproducibility, and extension by future works.

G.3 Prompts

Tables 23, 24, and 25 describe the handwritten prompt templates used in NLU, NLG, and VL evaluation (§3). For all tasks, we used a zero-shot prompting procedure to serve as the baseline setup. Due to the task complexity and distribution of workload from volunteer contributors with available computing resources, we limited the experiment procedure for some setups to ensure the acquisition of results in line with target release dates. For

NLU, we explored three prompt styles for each dataset from core tasks, including commonsense reasoning, question-answering, and NLI. For more challenging tasks requiring more intensive computing power such as NLG and VL, we used only one uniform prompt style, but we also explored prompts translated into SEA languages, i.e., Filipino, Indonesian, Thai, and Vietnamese for VL.

G.4 Evaluation Results

Table 26 and 27 describes the NLU and NLG results per language.

G.5 Language Equity Results

Table 15 presents the language equity of LLMs used in the evaluation across different weights of the number of language speakers in the Gini coefficient calculation.

H Supplementary Details for Translationese Classifier

H.1 Training & Evaluation Data

We manually select and validate the text collection method of each data subset for training and evaluating the translationese classifier, in Tables 28 and 29, respectively. This validation is done by checking the relevant publication, domain, and annotation method. If the texts in the data subsets are a product of machine or human translation, we regard them as translationese. We label data subsets with human-generated texts as natural data.

H.2 Experiments

We aim to assess the capability of ML models to differentiate between human-generated/natural samples (`Nat`), human-translated samples (`HT`), and machine-translated samples (`MT`). Our approach involves training classifiers using classical ML techniques and fine-tuning mDeBERTa models to enhance learning. Furthermore, we experiment by combining two label classes into one to evaluate the predictive difficulty of distinguishing between these labels. This analysis provides valuable insights into the relative similarity of the samples across these categories. The following section provides a comprehensive overview of our methodology for this study.

Classical ML We use three classical machine learning methods: 1) Logistic Regression (LR), 2) Naive Bayes (NB), and 3) Support Vector Machine (SVM) with two different features, including TF-IDF and Bag-of-words (BoW). We run hyper-parameter tuning with grid search to find the best hyper-parameters for each method on validation set, and report the results on test set in Table 16.

Encoder LM We explore fine-tuning encoder-only LM for developing a translationese classifier. We utilize mDeBERTa-v3$_{base}$ model[25] (He et al., 2020, 2022)–a multilingual encoder-only LM–as our backbone. We train the model with AdamW (Loshchilov and Hutter, 2019) optimizer using a learning rate of 1e-5, batch size of 256, and warming up steps of 500 for a maximum of 10 epochs. We apply an early stopping of 3 epochs based on the validation accuracy. We show the results in Table 17.

I Supplementary Details for SEA Language Prioritization

Based on the results of the global utility metric (Blasi et al., 2022), we provide the top-20 SEA indigenous languages to be prioritized based on their demand (i.e., the number of SEA language speakers) and current utility (Figure 10) or resource availability (Figure 11).[26] We use the performance scores of AYA-101 as one of the best-performing models on SEA languages for the current utility. While the current utility, also known as the model capability, is relative to the model performance on ENG, the resource availability is relative to 500, which is approximately the number of datasets in Korean language available in HuggingFace. The Korean language is chosen as the pivot because it is considered a higher-resource language than most by Joshi et al. (2020).

J Contributor Demographics

Table 18 describes the geographical distribution of the authors in SEACrowd.

K Languages Under Study

Table 30-48 present the list of SEA indigenous languages covered by SEACrowd. Information regarding the ISO 639-3 code, language name, region, and population is obtained from (Eberhard et al., 2021; Hammarström et al., 2024; Project, 2024; Dryer and Haspelmath, 2013) and Wikipedia[27].

[25] https://huggingface.co/microsoft/mdeberta-v3-base

[26] https://github.com/SEACrowd/globalutility

[27] https://www.wikipedia.org/

No.	Name	C. Points
1	Holy Lovenia	549
2	Samuel Cahyawijaya	480
3	Rahmad Mahendra	317
4	Salsabil Maulana Akbar	243
5	Lester James V. Miranda	234
6	Zheng-Xin Yong	164
7	Jennifer Santoso	164
8	Elyanah Aco	158
9	Akhdan Fadhilah	157
10	Jonibek Mansurov	132
11	Fajri Koto	121
12	Joseph Marvin Imperial	118
13	Ruochen Zhang	114
14	Genta Indra Winata	108
15	Onno P. Kampman	107
16	Joel Ruben Antony Moniz	93
17	Muhammad Ravi Shulthan Habibi	92
18	Frederikus Hudi	83
19	Sedrick Keh	81
20	Alham Fikri Aji	80
21	Railey Montalan	78
22	Peerat Limkonchotiwat	72
23	Ryan Ignatius	56
24	Joanito Agili Lopo	50
25	William Nixon	50
26	Börje F. Karlsson	49
27	James Jaya	48
28	Ryandito Diandaru	48
29	Yuze Gao	48
30	William Tjhi	46
31	Patrick Amadeus	46
32	Bin Wang	44
33	Jan Christian Blaise Cruz	43
34	Chenxi Whitehouse	36
35	Ivan Halim Parmonangan	36
36	Maria Khelli	36
37	Sebastian Ruder	35
38	Wenyu Zhang	34
39	Lucky Susanto	33
40	Reynard Adha Ryanda	32
41	Sonny Lazuardi Hermawan	30
42	Dan John Velasco	29
43	Muhammad Dehan Al Kautsar	29
44	Willy Fitra Hendria	29
45	Yasmin Moslem	29
46	Noah Flynn	28
47	Muhammad Farid Adilazuarda	27
48	Haochen Li	27
49	Johanes Lee	27
50	R. Damanhuri	27
51	Shuo Sun	27
52	Muhammad Reza Qorib	26
53	Amirbek Djanibekov	25
54	Wei Qi Leong	25
55	Quyet V. Do	24
56	Niklas Muennighoff	24
57	Tanrada Pansuwan	22
58	Ilham Firdausi Putra	21
59	Yan Xu	21
60	Ayu Purwarianti	20
61	Ngee Chia Tai	20

Table 19: Co-authors ordered by their amount of contribution points.

L Amount of Contributions by Co-Authors

Table 19 provides a list of co-authors sorted by their amount of contributions in SEACrowd. The full details of their contributions can be seen in our contribution tracking.

Model name	Model size	Backbone	Seen langs	URL
			Commercial	
GPT-4	N/A	GPT-4	N/A	https://openai.com/index/gpt-4/. We used `turbo-2024-04-09` for NLU and `gpt-4o-2024-05-13` for NLG.
Command-R	36B	Command-R	2 SEA langs (VIE, IND), 22 non-SEA langs	https://cohere.com/blog/command-r
			English	
Mistral	7B	Mistral	N/A	`mistralai/Mistral-7B-Instruct-v0.3`
Llama3	8B	Llama3	N/A	`meta-llama/Meta-Llama-3-8B-Instruct`
Falcon	7B	Falcon	0 SEA langs (mainly English)	`tiiuae/falcon-7b-instruct`
			Multilingual	
mT0	3B	mT5	2 SEA langs (VIE, IND), 43 non-SEA langs	`bigscience/mt0-xl`
BLOOMZ	7B	BLOOM	2 SEA langs (VIE, IND), 43 non-SEA langs	`bigscience/bloomz-3b`
BactrianX-Llama	7B	Llama	6 SEA langs (IND, VIE, KHM, MYA, THA, TGL, VIE), 46 non-SEA langs	`MBZUAI/bactrian-x-llama-7b-merged`
AYA-23	8B	Command	2 SEA langs (IND, VIE), 21 non-SEA langs	`CohereForAI/aya-23-8B`
AYA-101	13B	T5	9 SEA langs (IND, VIE, THA, ZSM, MYA, CEB, FIL, JAV, SUN), 92 non-SEA langs	`CohereForAI/aya-101`
			SEA regional	
SEA-LION	7B	MPT	8 SEA langs (IND, VIE, THA, TGL, ZSM, KHM, LAO, MYA), 3 non-SEA langs	`aisingapore/sea-lion-7b-instruct`
SeaLLM v2.5	7B	SeaLLM	8 SEA langs (IND, VIE, THA, TGL, ZSM, KHM, LAO, MYA)	`SeaLLMs/SeaLLM-7B-v2.5`
Sailor	7B	Qwen 1.5	5 SEA langs (IND, VIE, LAO, ZLM, THA), 2 non-SEA langs	`sail/Sailor-7B-Chat`
			SEA country	
Cendol-mT5	3B	mT5	1 SEA lang (IND), 18 local Indonesian langs	`indonlp/cendol-mt5-xl`
Cendol-Llama2	7B	Llama2	1 SEA lang (IND), 18 local Indonesian langs	`indonlp/cendol-llama2-7b`
Merak v4	7B	Llama2	1 SEA lang (IND)	`Ichsan2895/Merak-7B-v4`
WangchanX-Llama3	8B	Llama3	4 SEA langs (IND, VIE, THA, MYA) and 26 non-SEA langs	`airesearch/LLaMa3-8b-WangchanX-sft-Demo`
Malaysian Llama3	8B	Llama3	1 SEA lang (ZLM)	`mesolitica/malaysian-llama-3-8b-instruct-16k`

Table 20: LLMs used in SEACrowd NLU and NLG evaluation.

Model name	Model size	Backbone	Seen langs	URL
			Multilingual	
Whisper v3	1.54B	Whisper v3	89 non-SEA & 9 SEA (IND, JAV, LAO, ZLM, MYA, TGL, THA, SUN, VIE)	`openai/whisper-large-v3`
MMS 1B	1B	MMS	993 non-SEA & 205 SEA (ABP, ACE, ACN, AGN, AHK, AKB, ALJ, ALP, AMK, AOZ, ATB, ATQ, AYZ, BAN, BBC, BCL, BDG, BDQ, BEP, BGR, BHZ, BKD, BLT, BLX, BLZ, BNO, BPR, BPS, BRU, BTD, BTS, BTX, BVZ, BZI, CEB, CEK, CFM, CGC, CMR, CNH, CTD, DBJ, DNT, DNW, DTP, EIP, FRD, GBI, GOR, HAD, HAP, HIL, HLT, HNN, HVN, IBA, IFA, IFB, IFK, IFU, IFY, ILO, IND, ITV, JAV, JMD, KAC, KAK, KDT, KHG, KHM, KJE, KJG, KLW, KMD, KML, KNB, KNE, KPQ, KPS, KQE, KQR, KRJ, KRR, KVW, KXF, KXM, KYB, KYO, KYU, KZF, LAO, LAW, LBW, LCP, LEW, LEX, LHU, LIS, LJE, LJP, LLG, LND, LSI, MAD, MAK, MBB, MBT, MEJ, MHX, MHY, MIN, MKN, MNB, MNW, MNX, MOG, MQF, MQJ, MQN, MRW, MTD, MTJ, MVP, MWQ, MWV, MYA, MYL, NFA, NIA, NIJ, NLC, NLK, NOD, NPY, NST, OBO, PAG, PAM, PCE, PEZ, PLW, PMF, PPK, PRF, PRK, PRT, PSE, PTU, PWW, RAW, REJ, RGU, RHG, RIL, ROL, SAJ, SAS, SBL, SDA, SEA, SGB, SHN, SJM, SLU, SML, SNE, SUC, SUN, SXN, SYA, SZA, TBK, TBL, TBY, TCZ, TDJ, TES, TGL, THA, TIH, TLB, TNT, TOM, TVW, TWB, TWE, TWU, TXA, TXQ, UBL, URK, URY, VIE, WAR, WLO, XDY, XMM, XSB, XTE, YKA, YLI, YVA, ZLM, ZYP)	`facebook/mms-1b-all`
Seamless M4T v2	2.3B	Seamless	83 non-SEA & 9 SEA (IND, JAV, KHM, LAO, MYA, TGL, THA, VIE, ZLM)	`facebook/seamless-m4t-v2-large`
			Fine-tuned on specific language(s)	
XLSR English	300M	Wav2Vec2	46 non-SEA & 7 SEA (CEB, CNH, IND, LAO, TAM, TGL, VIE) & fine-tuning language(s)	`jonatasgrosman/wav2vec2-large-xlsr-53-english`
XLSR Ind-Jav-Sun				`indonesian-nlp/wav2vec2-indonesian-javanese-sundanese`
XLSR Indonesian				`Galuh/wav2vec2-large-xlsr-indonesian`
XLSR Thai				`wannaphong/wav2vec2-large-xlsr-53-th-cv8-newmm`
XLS-R Tagalog				`sil-ai/wav2vec2-bloom-speech-tgl`
XLS-R Burmese				`sil-ai/wav2vec2-bloom-speech-mya`
XLS-R Khmer				`vitouphy/wav2vec2-xls-r-300m-khmer`
Whisper Indonesian	1.54B	Whisper	89 non-SEA & 9 SEA (IND, JAV, LAO, MSA, MYA, TGL, THA, SUN, VIE)	`cahya/whisper-large-id`
Whisper Thai				`biodatlab/whisper-th-large-v3-combined`
Whisper Khmer				`ksoky/whisper-large-khmer-asr`

Table 21: Speech models used in SEACrowd speech evaluation.

Model name	Model size	Backbone	Pre-training images	URL
			English	
LLaVA 1.5	N/A	N/A	N/A	N/A
LLaVA 1.6	7B	Mistral-7B	N/A	`liuhaotian/llava-v1.6-mistral-7b`
Idefics2	8B	Mistral-7B-v0.1	1.5B	`HuggingFaceM4/idefics2-8b`
PaliGemma	2B	Gemma-2B	N/A	`google/paligemma-3b-pt-224`
			Multilingual	
mBLIP	N/A	blip2-flan-t5-xl	N/A	`Gregor/mblip-mt0-xl`

Table 22: VLMs used in SEACrowd VL evaluation.

No.	Prompt template
	Sentiment Analysis
1	`Classify the sentiment of the text below.\n[INPUT] => Sentiment ([OPTIONS]): [LABEL_CHOICE]`
2	`Predict the sentiment of the following text.\nText: [INPUT]\nAnswer with [OPTIONS]: [LABEL_CHOICE]`
3	`[INPUT]\nWhat would be the sentiment of the text above? [OPTIONS]? [LABEL_CHOICE]`
	Topic Classification
1	`Classify the topic of the text below.\n[INPUT] => Topic ([OPTIONS]): [LABEL_CHOICE]`
2	`Predict the topic of the following text.\nText: [INPUT]\nAnswer with [OPTIONS]: [LABEL_CHOICE]`
3	`[INPUT]\nWhat would be the topic of the text above? [OPTIONS]? [LABEL_CHOICE]`
	Commonsense Reasoning → `*_seacrowd_text`
1	`Classify the morality of the text below.\n[INPUT] => Morality ([OPTIONS]): [LABEL_CHOICE]`
2	`Predict the morality of the following text.\nText: [INPUT]\nAnswer with [OPTIONS]: [LABEL_CHOICE]`
3	`[INPUT]\nWhat would be the morality of the text above? [OPTIONS]? [LABEL_CHOICE]`
	Commonsense Reasoning → `*_seacrowd_qa`
1	`Question: [QUESTION]\nWhat reply makes more sense to answer this question?\nChoices: [ANSWER_CHOICES]\nAnswer: [LABEL_CHOICE]`
2	`Based on the the following question: "[QUESTION]" and choices: [ANSWER_CHOICE the correct answer is: [LABEL_CHOICE]`
3	`Question: [QUESTION]\nChoices: [ANSWER_CHOICES]\nThe correct answer to the given question is: [LABEL_CHOICE]`
	All QAs
1	`Refer to the passage below and answer the following question:\nPassage: [CONTEXT]\nQuestion: [QUESTION]\nChoices: [ANSWER_CHOICES]\nAnswer: [LABEL_CHOICE]`
2	`[CONTEXT]\nBased on the above text, [QUESTION]\nChoices: [ANSWER_CHOICES]\nAnswer: [LABEL_CHOICE]`
3	`[CONTEXT]\nQuestion: [QUESTION]\nChoices:[ANSWER_ CHOICES]\nReferring to the passage above, the correct answer to the given question is: [LABEL_CHOICE]`
	NLI
1	`Hypothesis: [INPUT_A]\nPremise: [INPUT_B]\nQuestion: What is the relation between the hypothesis and the premise? [OPTIONS]? [LABEL_CHOICE]`
2	`Given the following premise and hypothesis:\nHypothesis: [INPUT_A]\nPremise: [INPUT_B]\nDetermine the logical relationship (([OPTIONS])): [LABEL_CHOICE]`
3	`Choose the most appropriate relationship ([OPTIONS]) between the premise and hypothesis:\nRelationship between "[INPUT_B]" and "[INPUT_A]": [LABEL_CHOICE]`

Table 23: Prompt templates used for NLU tasks.

No.	Prompt template
	Machine Translation (MT)
1	`Translate the following text from [SOURCE] to [TARGET]. Give your translation directly.\nText: [INPUT]\nTranslation:`
	Summarization
1	`Write a summary from the following text.\nText: [INPUT]\nSummary:`
	Abstractive & Extractive QA
1	`Refer to the passage below and answer the following question:\nPassage: [CONTEXT]\nQuestion: [QUESTION]\nAnswer:`

Table 24: Prompt templates used for NLG tasks.

Lang.	Prompt template
	Image Captioning
ENG	`Caption the following image in [LANGUAGE].`
FIL	`Ilarawan ang sumusunod na larawan.`
IND	`Deskripsikan gambar berikut.`

Table 25: Prompt templates used for the image captioning task in VL evaluation.

	ABL	ABS	ACE	BAN	BBC	BEW	BHP	BJN	BTX	BUG	CEB	ENG	FIL	ILO	IND	JAV	KAC	KHM	LAO	LUS	MAD	MAK	MIN	MUI	MYA	NIJ	PAG	REJ	SHN	SUN	THA	VIE	WAR	ZSM	Overall
GPT-4	63.3	39.0	39.3	60.3	7.1	68.5	2.8	60.4	27.8	40.4	85.6	52.1	55.9	69.5	60.7	59.7	30.8	66.4	51.8	70.0	37.1	44.3	57.9	71.8	47.6	40.2	79.4	34.0	21.7	58.5	59.6	56.1	84.9	61.6	51.9
Command-R	50.1	80.8	57.6	62.8	47.4	81.8	58.2	57.1	57.3	57.9	66.7	69.4	51.1	56.8	58.3	61.2	36.5	41.5	33.8	63.9	61.9	58.4	66.4	81.7	34.8	53.3	75.6	69.6	35.4	63.2	42.7	55.9	67.6	55.7	58.0
Mistral	36.7	53.6	46.4	49.6	33.0	59.3	44.3	44.6	44.3	48.8	53.5	69.2	48.4	49.1	52.5	46.7	33.2	29.8	30.7	56.1	45.7	44.8	51.2	62.6	27.4	40.1	69.2	48.6	31.9	48.3	40.8	45.2	54.4	49.6	46.8
Llama3	37.3	40.3	43.2	48.9	34.8	44.5	32.6	42.2	38.5	42.9	51.2	59.5	45.2	46.7	49.2	44.4	28.5	34.6	30.3	46.8	39.0	38.0	43.6	49.2	35.2	39.6	60.5	38.5	31.1	45.2	43.8	45.5	50.3	49.0	42.6
Falcon	21.1	63.2	13.3	19.0	23.0	37.9	62.1	15.6	31.9	15.7	19.5	43.7	25.1	18.8	30.8	27.0	14.2	10.2	12.7	15.0	30.3	32.3	23.6	37.0	18.0	23.0	18.8	36.0	14.1	28.2	15.9	18.8	19.1	17.4	25.1
mT0	37.6	63.6	43.7	51.2	37.0	66.1	38.4	43.6	41.3	50.3	62.5	49.4	41.0	59.0	47.2	56.0	40.9	57.5	61.2	57.0	46.7	45.8	52.6	68.8	45.9	40.9	62.6	47.8	47.0	58.8	41.8	41.4	61.4	49.4	50.5
BLOOMZ	25.6	66.5	28.4	34.2	35.8	53.9	48.0	30.4	36.3	33.3	30.9	51.7	28.9	27.8	44.7	38.2	23.1	18.9	23.6	28.1	37.8	34.5	39.9	60.2	23.0	34.6	33.1	42.2	19.8	41.3	25.9	34.8	32.1	34.3	35.3
BactrianX-Llama	24.9	48.6	21.2	28.5	26.9	33.4	45.9	22.8	31.4	22.7	27.9	45.6	32.0	24.3	38.3	30.0	19.9	17.0	20.7	21.0	30.0	28.8	26.2	35.7	22.8	27.2	26.5	29.2	20.5	30.2	24.5	27.1	28.3	31.5	28.6
AYA-23	43.3	21.2	26.9	35.0	24.3	31.2	16.8	30.9	25.1	26.5	36.0	50.8	33.5	32.7	46.8	36.9	20.5	15.1	22.0	27.4	31.0	31.7	27.3	35.5	23.7	37.3	32.6	22.8	20.8	34.9	32.7	44.8	37.1	47.9	31.3
AYA-101	42.5	64.3	71.2	65.2	58.8	68.2	43.3	63.5	52.7	60.7	71.7	62.8	52.8	65.0	54.2	62.6	43.1	62.2	67.8	71.8	56.9	49.0	69.3	70.2	51.5	57.2	75.7	52.9	53.8	67.2	49.5	48.0	70.5	56.4	59.8
SEA-LION	10.3	62.3	13.5	16.5	21.3	35.3	60.3	13.4	31.8	15.2	13.6	26.6	20.6	10.2	27.6	21.4	8.7	16.8	15.2	12.5	26.8	28.3	22.8	34.6	23.0	16.0	14.4	34.1	9.7	23.4	16.3	14.7	14.2	13.3	21.9
SeaLLM v2.5	50.7	55.1	34.5	43.4	36.3	53.9	53.2	45.8	45.8	37.7	47.6	42.5	52.6	44.7	53.4	49.8	27.4	42.6	50.3	45.8	48.7	49.8	46.8	58.4	41.0	39.1	55.7	47.8	28.7	50.1	49.0	54.5	55.4	60.6	47.0
Sailor	50.4	59.2	43.8	55.5	44.1	61.5	43.9	50.5	44.8	45.7	45.6	63.0	40.2	45.0	51.3	53.1	29.9	32.7	53.9	53.9	47.6	46.5	52.8	63.9	28.1	52.7	59.3	42.2	26.7	54.0	46.3	47.7	49.2	52.1	48.1
Cendol-mT5	15.0	98.5	38.3	42.3	84.7	99.4	95.6	33.3	92.6	68.6	14.1	38.7	23.8	12.2	33.4	50.5	10.4	20.3	15.3	9.6	76.5	70.2	65.2	99.6	16.6	52.6	12.8	98.9	7.2	56.6	26.4	14.7	15.1	15.9	44.8
Cendol-Llama2	17.5	80.0	30.8	33.5	60.6	49.3	73.4	27.9	45.1	32.3	18.7	36.8	21.4	17.8	37.4	35.1	14.7	13.2	15.9	15.0	46.3	38.1	37.1	51.6	19.9	40.3	17.7	47.7	16.5	38.5	20.6	17.3	18.5	18.4	32.5
Merak	37.0	68.6	37.7	48.3	36.4	66.1	60.1	41.4	50.4	47.8	42.4	59.6	37.9	39.7	48.5	48.4	27.9	24.2	28.0	44.3	51.7	51.0	50.5	70.3	27.2	40.0	58.6	57.9	28.6	50.8	29.3	35.3	43.7	47.1	45.2
WangchanX-Llama3	38.4	59.3	26.8	35.2	35.0	43.3	56.9	31.6	38.3	31.2	32.3	57.6	36.6	29.3	45.0	38.7	23.7	24.3	25.1	26.6	40.4	41.4	34.8	43.6	31.6	37.0	31.2	42.9	23.5	39.8	36.5	38.4	31.3	37.0	36.6
Malaysian Llama3	38.9	62.3	38.1	41.9	39.2	46.9	58.3	39.5	40.5	35.9	37.8	55.5	34.5	33.1	48.6	42.6	24.7	18.9	20.4	33.6	42.1	41.0	42.5	48.5	22.2	39.6	46.8	41.1	19.6	44.0	33.7	34.6	37.7	49.9	39.2
Overall	35.6	60.4	36.4	42.9	38.1	55.6	49.7	38.6	43.1	39.7	42.1	51.9	37.9	37.9	46.0	44.6	25.5	30.3	32.1	38.8	44.3	43.0	45.0	58.0	30.0	39.5	46.1	46.4	25.4	46.3	35.3	37.5	42.8	41.5	41.4

Table 26: NLU evaluation results in weighted F1-score per language.

	ACE	BAN	BBC	BJN	BUG	CEB	FIL	HMV	ILO	IND	JAV	KAC	KHM	LAO	LJL	LUS	MAD	MIN	MYA	NIJ	PAG	SHN	SUN	THA	VIE	WAR	ZSM	Overall
GPT-4	32.9	40.7	28.8	42.0	24.1	66.5	65.6	50.0	52.9	59.3	54.2	16.7	29.8	41.9	10.0	33.3	29.2	46.1	21.5	27.7	37.0	14.5	50.0	28.8	47.5	66.4	59.6	39.9
Command-R	19.6	26.1	16.4	30.0	16.0	44.3	52.5	16.8	29.4	57.9	32.6	8.8	8.7	14.2	6.0	19.5	17.2	31.6	9.5	18.4	20.4	8.9	27.5	24.3	46.8	34.4	50.1	25.5
Mistral	12.4	15.0	10.0	13.9	11.1	28.5	37.2	10.2	15.9	28.6	15.4	7.3	8.7	10.8	4.2	11.7	9.5	18.0	5.7	12.4	17.5	9.5	14.8	15.1	25.1	22.4	31.1	15.6
Llama3	11.0	12.3	8.1	13.8	7.6	25.1	33.2	7.6	18.4	21.9	17.0	4.8	6.5	5.8	3.2	9.6	8.5	16.4	4.5	9.5	11.8	6.3	15.1	9.6	21.7	20.5	25.2	13.2
Falcon	7.3	9.5	8.2	8.3	7.9	18.6	23.6	6.6	9.7	15.3	7.7	6.0	3.1	3.1	4.2	9.3	6.6	11.8	1.8	8.7	12.9	4.5	7.7	2.4	13.5	13.5	17.0	9.2
mT0	4.8	5.6	3.7	5.7	3.1	4.6	6.8	4.5	3.8	29.3	5.8	2.1	4.3	6.1	1.7	3.4	3.6	6.5	5.0	3.5	3.6	3.5	6.8	9.4	19.6	6.1	9.1	6.4
BLOOMZ	3.8	4.6	2.8	5.3	2.9	4.1	5.1	3.4	4.2	32.3	4.9	3.0	1.5	2.4	1.5	4.0	2.7	5.7	1.2	3.2	4.9	2.6	4.6	3.3	24.1	5.4	10.1	5.7
BactrianX-Llama	10.9	11.6	8.9	12.3	8.8	22.0	32.1	8.5	12.1	25.1	11.4	6.9	6.4	8.2	4.1	10.9	8.7	14.1	4.3	8.4	15.2	8.0	11.4	10.8	19.4	16.6	23.4	12.6
AYA-23	9.3	10.5	8.0	11.6	6.9	14.2	17.5	5.6	8.3	18.3	11.3	5.7	4.0	5.9	2.7	8.1	7.6	12.2	3.3	9.0	8.8	6.5	10.4	6.8	24.3	10.6	17.7	9.8
AYA-101	26.4	26.8	14.6	21.6	12.6	49.3	46.6	33.3	25.8	49.5	38.8	12.2	25.9	37.2	4.4	17.8	13.4	29.7	17.6	13.2	23.3	20.4	35.6	22.2	36.5	36.9	41.9	27.2
SEA-LION	7.2	8.1	6.5	9.3	5.8	12.5	17.1	4.9	7.0	13.9	7.9	5.3	7.0	9.6	2.0	7.6	6.0	9.5	4.8	6.6	8.4	4.9	8.0	5.9	21.2	10.3	14.1	8.6
SeaLLM v2.5	15.2	20.2	11.7	19.5	11.5	37.1	49.1	14.5	26.8	43.0	26.6	7.5	17.8	22.2	4.7	15.1	12.2	26.8	9.2	14.6	19.2	9.4	22.0	21.6	36.7	28.8	45.7	21.8
Sailor	19.2	24.5	15.3	23.1	14.6	29.0	39.7	8.6	13.5	46.8	30.6	7.1	12.5	24.4	6.2	10.5	16.0	28.8	5.8	19.1	16.5	9.0	26.7	22.0	41.1	21.5	49.9	21.6
Cendol-mT5	8.3	11.4	14.2	11.6	6.9	7.2	8.4	4.7	5.5	35.8	17.5	4.0	6.3	8.5	2.0	5.2	6.1	10.5	2.9	8.8	6.6	4.1	17.1	5.5	4.4	6.4	20.5	9.3
Cendol-Llama2	8.6	10.0	14.4	19.3	6.6	6.9	8.2	6.4	6.4	36.1	19.1	5.5	3.0	4.3	4.1	4.5	14.1	22.0	1.9	17.5	5.4	4.8	17.3	3.4	8.1	7.6	22.0	10.6
Merak	7.4	10.3	6.7	11.3	7.1	8.2	12.8	6.3	6.7	29.5	9.6	3.7	3.8	5.9	3.2	8.0	6.5	12.5	2.4	8.0	8.2	5.6	10.6	5.9	7.2	7.4	20.4	8.7
WangchanX-Llama3	19.8	24.4	14.3	28.9	13.4	42.2	48.6	12.7	29.4	50.1	29.4	7.7	18.1	19.7	6.0	17.6	15.6	30.0	10.4	18.1	22.4	13.9	28.0	25.1	39.2	35.5	45.4	24.7
Malaysian Llama3	15.2	17.3	12.3	22.2	11.1	19.7	24.0	8.7	12.6	38.6	19.4	7.2	6.7	9.0	5.9	10.6	12.4	23.5	4.2	14.3	13.9	8.3	19.0	14.2	17.3	15.6	44.4	15.8
Overall	13.3	16.1	11.4	17.2	9.9	24.4	29.3	11.8	16.0	35.1	20.0	6.7	9.7	13.3	4.2	11.5	10.9	19.8	6.4	12.3	14.2	8.0	18.5	13.1	25.2	20.3	30.4	15.9

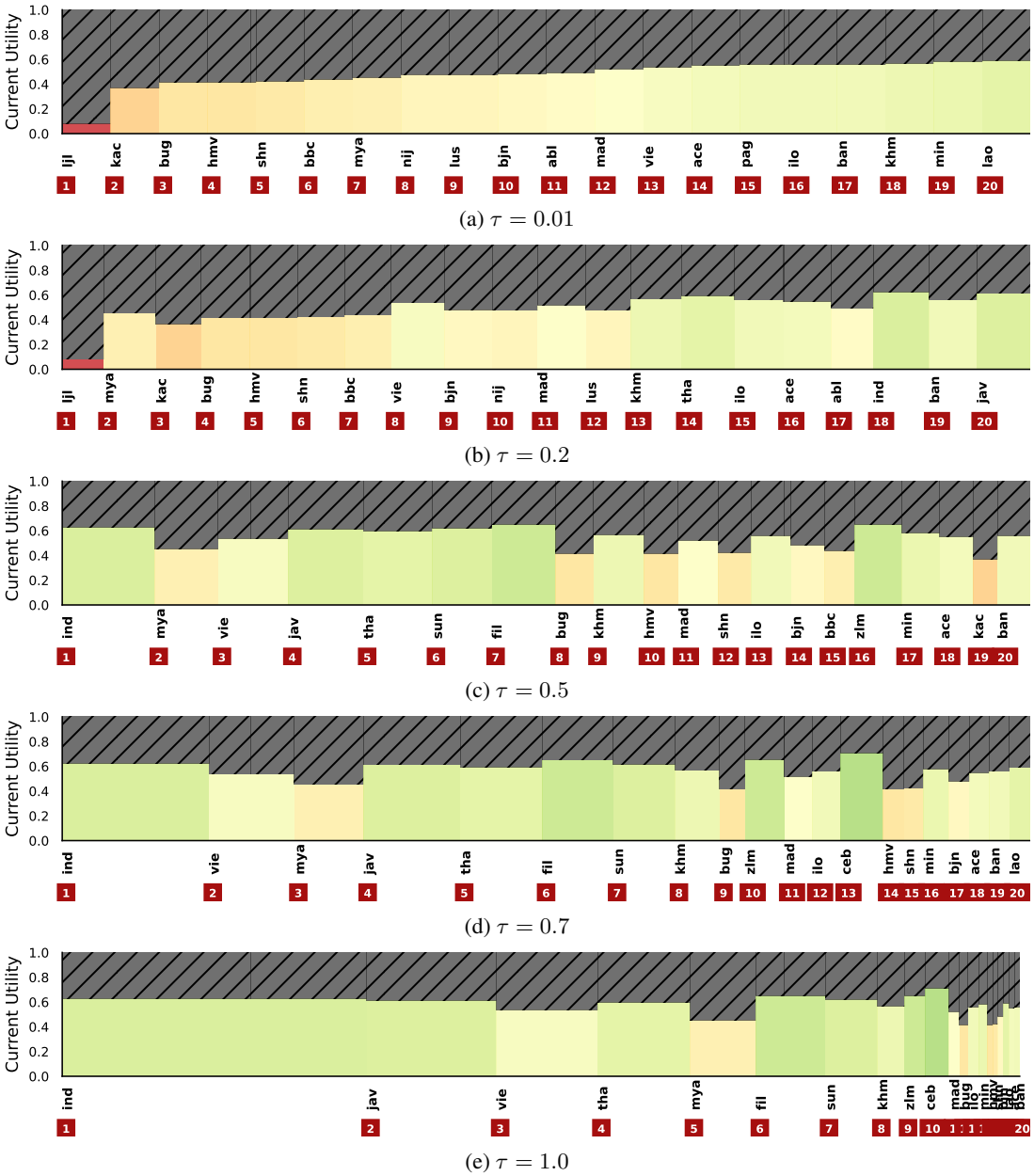
Table 27: NLG evaluation results in ROUGE-L per language.

Lang.	Subset	Original Task	Domain	# Samples
		Translationese		
ENG	emotes_3k_eng_seacrowd_t2t	Commonsense Reasoning	Ethics	2000
ENG	aya_evaluation_suite_eng_seacrowd_t2t	Instruction Tuning	General	400
IND	belebele_ind_latn_seacrowd_qa	QA	General	1969
IND	parallel_asian_treebank_ind_eng_seacrowd_t2t	Machine Translation	News	31
IND	aya_evaluation_suite_ind_seacrowd_t2t	Instruction Tuning	General	4
IND	bactrian_x_id_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1972
IND	seaeval_cross_logiqa_ind_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	16
IND	seaeval_cross_mmlu_ind_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	8
KHM	belebele_khm_khmr_seacrowd_qa	QA	General	399
KHM	khmer_alt_pos_seacrowd_seq_label	POS Tagging	News	1595
KHM	parallel_asian_treebank_khm_eng_seacrowd_t2t	Machine Translation	News	6
KHM	aya_evaluation_suite_khm_seacrowd_t2t	Instruction Tuning	General	8
KHM	bactrian_x_km_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1992
LAO	belebele_lao_laoo_seacrowd_qa	QA	General	1969
LAO	parallel_asian_treebank_lao_eng_seacrowd_t2t	Machine Translation	News	31
LAO	aya_evaluation_suite_lao_seacrowd_t2t	Instruction Tuning	General	400
MYA	belebele_mya_mymr_seacrowd_qa	QA	General	1969
MYA	parallel_asian_treebank_mya_eng_seacrowd_t2t	Machine Translation	News	31
MYA	aya_evaluation_suite_mya_seacrowd_t2t	Instruction Tuning	General	8
MYA	bactrian_x_my_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1992
FIL	belebele_tgl_latn_seacrowd_qa	QA	General	2000
FIL	bactrian_x_tl_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	2000
THA	belebele_tha_thai_seacrowd_qa	QA	General	1969
THA	parallel_asian_treebank_tha_eng_seacrowd_t2t	Machine Translation	News	31
THA	aya_evaluation_suite_tha_seacrowd_t2t	Instruction Tuning	General	8
THA	bactrian_x_th_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1992
VIE	belebele_vie_latn_seacrowd_qa	QA	General	1969
VIE	parallel_asian_treebank_vie_eng_seacrowd_t2t	Machine Translation	News	31
VIE	aya_evaluation_suite_vie_seacrowd_t2t	Instruction Tuning	General	4
VIE	bactrian_x_vi_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1972
VIE	seaeval_cross_logiqa_vie_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	16
VIE	seaeval_cross_mmlu_vie_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	8
ZLM	belebele_zsm_latn_seacrowd_qa	QA	General	1969
ZLM	parallel_asian_treebank_zlm_eng_seacrowd_t2t	Machine Translation	News	31
ZLM	aya_evaluation_suite_zsm_seacrowd_t2t	Instruction Tuning	General	400
ZLM	seaeval_cross_logiqa_zlm_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	1056
ZLM	seaeval_cross_mmlu_zlm_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	300
		Natural		
ENG	cosem_seacrowd_ssp	Language Modeling	Social media	2000
IND	sea_bench_ind_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	200
KHM	gklmip_newsclass_seacrowd_text	Sentiment Analysis	E-commerce	1436
KHM	sea_bench_khm_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
LAO	sea_bench_lao_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
MYA	gklmip_sentiment_seacrowd_text	Sentiment Analysis	E-commerce	716
MYA	sea_bench_mya_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
FIL	sea_bench_tgl_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
THA	sea_bench_tha_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	40
THA	vistec_tp_th_21_seacrowd_seq_label	NER	Social media	1960
VIE	sea_bench_vie_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	200
ZLM	sea_bench_zlm_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160

Table 28: Train data used in the translationese classifier experiment.

Lang.	Subset	Original Task	Domain	# Samples
		Translationese		
ENG	emotes_3k_eng_seacrowd_t2t	Commonsense Reasoning	Ethics	2000
ENG	aya_evaluation_suite_eng_seacrowd_t2t	Instruction Tuning	General	400
IND	belebele_ind_latn_seacrowd_qa	QA	General	1969
IND	parallel_asian_treebank_ind_eng_seacrowd_t2t	MT	News	31
IND	aya_evaluation_suite_ind_seacrowd_t2t	Instruction Tuning	General	4
IND	bactrian_x_id_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1972
IND	seaeval_cross_logiqa_ind_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	16
IND	seaeval_cross_mmlu_ind_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	8
KHM	belebele_khm_khmr_seacrowd_qa	QA	General	399
KHM	khmer_alt_pos_seacrowd_seq_label	POS Tagging	News	1595
KHM	parallel_asian_treebank_khm_eng_seacrowd_t2t	MT	News	6
KHM	aya_evaluation_suite_khm_seacrowd_t2t	Instruction Tuning	General	8
KHM	bactrian_x_km_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1992
LAO	belebele_lao_laoo_seacrowd_qa	QA	General	1969
LAO	parallel_asian_treebank_lao_eng_seacrowd_t2t	MT	News	31
LAO	aya_evaluation_suite_lao_seacrowd_t2t	Instruction Tuning	General	400
MYA	belebele_mya_mymr_seacrowd_qa	QA	General	1969
MYA	parallel_asian_treebank_mya_eng_seacrowd_t2t	MT	News	31
MYA	aya_evaluation_suite_mya_seacrowd_t2t	Instruction Tuning	General	8
MYA	bactrian_x_my_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1992
FIL	belebele_tgl_latn_seacrowd_qa	QA	General	2000
FIL	bactrian_x_tl_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	2000
THA	belebele_tha_thai_seacrowd_qa	QA	General	1969
THA	parallel_asian_treebank_tha_eng_seacrowd_t2t	MT	News	31
THA	aya_evaluation_suite_tha_seacrowd_t2t	Instruction Tuning	General	8
THA	bactrian_x_th_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1992
VIE	belebele_vie_latn_seacrowd_qa	QA	General	1969
VIE	parallel_asian_treebank_vie_eng_seacrowd_t2t	MT	News	31
VIE	aya_evaluation_suite_vie_seacrowd_t2t	Instruction Tuning	General	4
VIE	bactrian_x_vi_seacrowd_t2t	Instruction Tuning	Mixed, Multi-domain, Wikipedia	1972
VIE	seaeval_cross_logiqa_vie_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	16
VIE	seaeval_cross_mmlu_vie_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	8
ZLM	belebele_zsm_latn_seacrowd_qa	QA	General	1969
ZLM	parallel_asian_treebank_zlm_eng_seacrowd_t2t	MT	News	31
ZLM	aya_evaluation_suite_zsm_seacrowd_t2t	Instruction Tuning	General	400
ZLM	seaeval_cross_logiqa_zlm_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	1056
ZLM	seaeval_cross_mmlu_zlm_seacrowd_qa	Commonsense Reasoning, QA	Commentary, General, Multi-domain, Culture & heritage	300
		Natural		
ENG	cosem_seacrowd_ssp	Language Modeling	Social media	2000
IND	sea_bench_ind_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	200
KHM	gklmip_newsclass_seacrowd_text	Sentiment Analysis	E-commerce	1436
KHM	sea_bench_khm_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
LAO	sea_bench_lao_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
MYA	gklmip_sentiment_seacrowd_text	Sentiment Analysis	E-commerce	716
MYA	sea_bench_mya_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
FIL	sea_bench_tgl_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160
THA	sea_bench_tha_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	40
THA	vistec_tp_th_21_seacrowd_seq_label	NER	Social media	1960
VIE	sea_bench_vie_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	200
ZLM	sea_bench_zlm_seacrowd_t2t	Instruction Tuning	Commentary, General, Multi-domain, Culture & heritage	160

Table 29: Test data used in the translationese classifier experiment.



(a) $\tau = 0.01$

(b) $\tau = 0.2$

(c) $\tau = 0.5$

(d) $\tau = 0.7$

(e) $\tau = 1.0$

Figure 10: Top-20 SEA indigenous languages to be prioritized based on their potential demand and current utility.

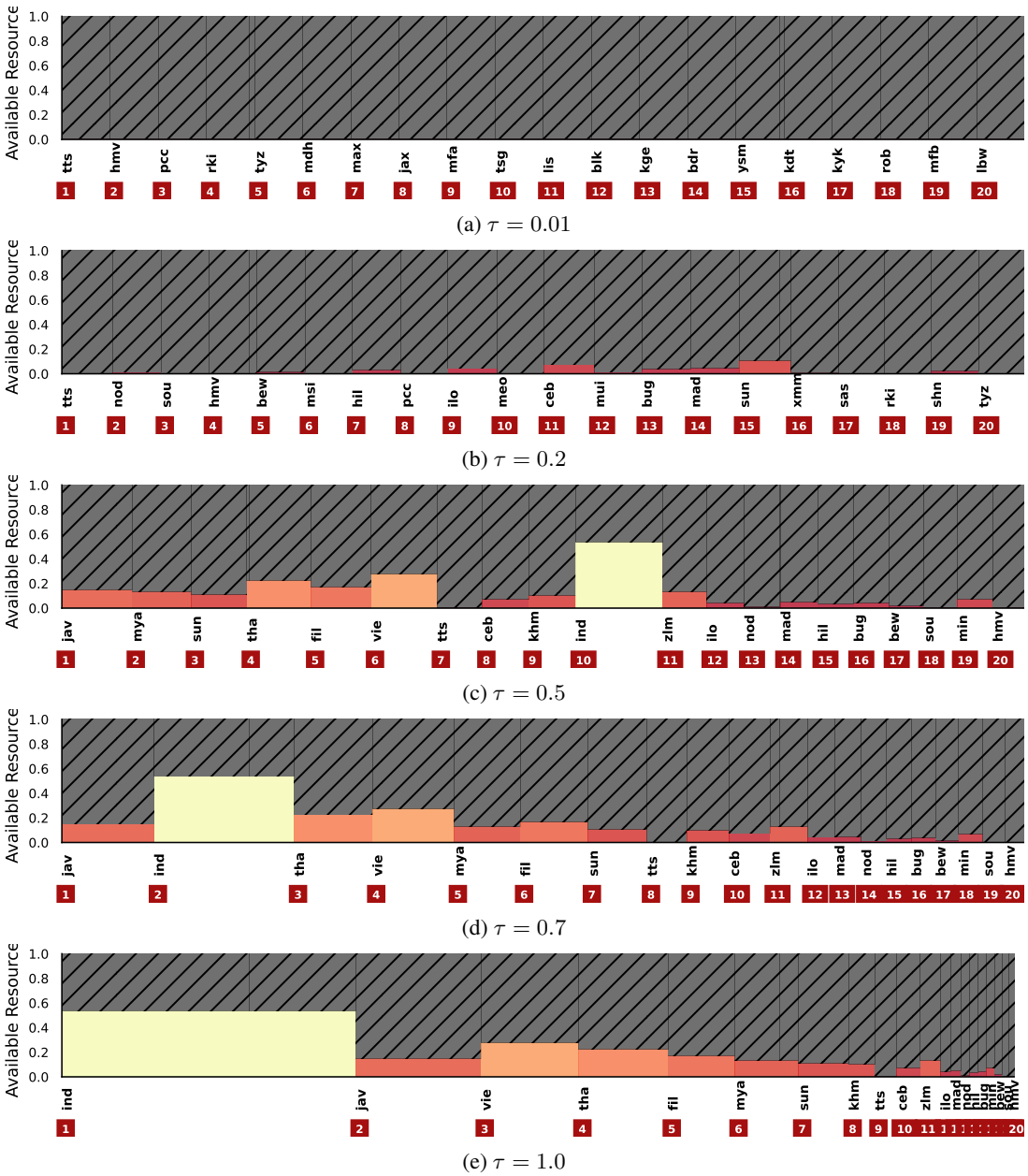


Figure 11: Top-20 SEA indigenous languages to be prioritized based on their potential demand and data availability.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	IND	Indonesian	Indonesia	<1B
2	JAV	Javanese	Indonesia	<100M
3	VIE	Vietnamese	Vietnam	<100M
4	THA	Thai	Thailand, Cambodia	<100M
5	FIL	Filipino	Philippines	<100M
6	MYA	Burmese	Myanmar	<100M
7	SUN	Sunda	Indonesia	<100M
8	TGL	Tagalog	Philippines	<100M
9	KHM	Khmer	Cambodia, Vietnam	<100M
10	CEB	Cebuano	Philippines	<100M
11	TTS	Northeastern Thai	Thailand	<100M
12	ZLM	Malay	Malaysia	<100M
13	ZSM	Standard Malay	Malaysia, Brunei, Singapore	<100M

Table 30: SEA indigenous languages with ≥10M speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	ILO	Ilocano	Philippines	<10M
2	MAD	Madura	Indonesia	<10M
3	NOD	Northern Thai	Laos, Thailand	<10M
4	HIL	Hiligaynon	Philippines	<10M
5	MIN	Minangkabau	Indonesia	<10M
6	BUG	Bugis	Indonesia	<10M
7	BEW	Betawi	Indonesia	<10M
8	SOU	Southern Thai	Thailand	<10M
9	LAO	Lao	Cambodia, Laos	<10M
10	HMV	Hmong Dô	Vietnam	<10M
11	ACE	Aceh	Indonesia	<10M
12	BJN	Banjar	Indonesia	<10M
13	BAN	Bali	Indonesia	<10M
14	SHN	Shan	Myanmar, Thailand	<10M
15	MUI	Musi	Indonesia	<10M
16	MSI	Sabah Malay	Malaysia	<10M
17	MEO	Kedah Malay	Malaysia, Thailand	<10M
18	PCC	Giáy	Vietnam	<10M
19	WAR	Waray-Waray	Philippines	<10M
20	MAK	Makasar	Indonesia	<10M
21	BCL	Central Bikol	Philippines	<10M
22	XMM	Manado Malay	Indonesia	<10M
23	SAS	Sasak	Indonesia	<10M
24	BBC	Batak Toba	Indonesia	<10M
25	PAM	Kapampangan	Philippines	<10M
26	RKI	Rakhine	Myanmar	<10M
27	TYZ	Tày	Vietnam	<10M
28	ABS	Ambonese Malay	Indonesia	<10M
29	PSE	Central Malay	Indonesia	<10M
30	IBA	Iban	Brunei, Indonesia, Malaysia	<10M
31	KXM	Northern Khmer	Thailand	<10M
32	KHG	Khams Tibetan	Myanmar	<10M
33	KSW	S'gaw Karen	Myanmar, Thailand	<10M
34	BTD	Batak Dairi	Indonesia	<10M
35	BTS	Batak Simalungun	Indonesia	<10M
36	CBK	Chavacano	Philippines	<10M
37	PAG	Pangasinan	Philippines	<10M
38	MTQ	Muong	Vietnam	<10M
39	BTM	Batak Mandailing	Indonesia	<10M
40	MDH	Maguindanaon	Philippines	<10M
41	PMY	Papuan Malay	Indonesia	<10M
42	GOR	Gorontalo	Indonesia	<10M
43	JAX	Jambi Malay	Indonesia	<10M
44	KJP	Pwo Eastern Karen	Myanmar, Thailand	<10M
45	MAX	North Moluccan Malay	Indonesia	<10M
46	MFA	Pattani Malay	Thailand	<10M
		Not in SEACrowd		
47	MFP	Makassar Indonesian	Indonesia	<10M

Table 31: SEA indigenous languages with <10M speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	NUT	Nung	Vietnam	<1M
2	KAC	Jingpho	Myanmar	<1M
3	TSG	Tausug	Philippines	<1M
4	NIJ	Ngaju	Indonesia	<1M
5	LJP	Lampung Api	Indonesia	<1M
6	MQY	Manggarai	Indonesia	<1M
7	MRW	Maranao	Philippines	<1M
8	NIA	Nias	Indonesia	<1M
9	AKB	Batak Angkola	Indonesia	<1M
10	SDA	Toraja-Sa'dan	Indonesia	<1M
11	MNW	Mon	Myanmar, Thailand	<1M
12	HNI	Hani	Laos, Vietnam	<1M
13	KJG	Khmu	Laos, Thailand, Vietnam	<1M
14	AOZ	Uab Meto	Indonesia	<1M
15	BLT	Tai Dam	Laos, Vietnam	<1M
16	LUS	Mizo Chin	Myanmar	<1M
17	CPS	Capiznon	Philippines	<1M
18	BTX	Batak Karo	Indonesia	<1M
19	LIS	Lisu	Myanmar	<1M
20	MSB	Masbatenyo	Philippines	<1M
21	BLK	Pa'o	Myanmar, Thailand	<1M
22	TDD	Tai Nüa	Myanmar	<1M
23	DAY	Land Dayak	Indonesia	<1M
24	XDY	Malayic Dayak	Indonesia	<1M
25	BHP	Bima	Indonesia	<1M
26	IBG	Ibanag	Philippines	<1M
27	ZMI	Negeri Sembilan Malay	Malaysia	<1M
28	MDR	Mandar	Indonesia	<1M
29	KGE	Komering	Indonesia	<1M
30	BDR	West Coast Bajau	Malaysia	<1M
31	KDT	Kuay	Cambodia, Laos, Thailand	<1M
32	PRK	Parauk Wa	Myanmar	<1M
33	SGD	Surigaonon	Philippines	<1M
34	TET	Tetun	East Timor, Indonesia	<1M
35	BTO	Rinconada Bikol	Philippines	<1M
36	TDT	Tetun Dili	East Timor	<1M
37	IUM	Iu Mien	Laos, Vietnam	<1M
38	KRJ	Kinaray-a	Philippines	<1M
39	KYK	Kamayo	Philippines	<1M
40	LEW	Ledo Kaili	Indonesia	<1M
41	MKN	Kupang Malay	Indonesia	<1M
42	REJ	Rejang	Indonesia	<1M
43	MFB	Bangka	Indonesia	<1M
44	ROB	Tae'	Indonesia	<1M
45	LBW	Tolaki	Indonesia	<1M
46	KNX	Kendayan	Indonesia, Malaysia	<1M
47	GAY	Gayo	Indonesia	<1M
48	MNB	Muna	Indonesia	<1M
49	RBL	Miraya Bikol	Philippines	<1M
50	SMW	Sumbawa	Indonesia	<1M
51	KXD	Brunei	Brunei	<1M
52	KHB	Lü	Laos, Myanmar	<1M
53	LHU	Lahu	Laos, Myanmar	<1M
54	TWH	Tai Dón	Laos, Vietnam	<1M
55	YSM	Myanmar Sign Language	Myanmar	<1M
56	DTP	Kadazan Dusun	Malaysia	<1M
57	FBL	West Albay Bikol	Philippines	<1M
58	KVR	Kerinci	Indonesia	<1M
59	PCE	Ruching Palaung	Myanmar	<1M
60	MRY	Mandaya	Philippines	<1M
61	NBE	Konyak Naga	Myanmar	<1M
62	TCZ	Thado Chin	Myanmar	<1M
63	JRA	Jarai	Cambodia, Vietnam	<1M
64	XBR	Kambera	Indonesia	<1M
65	MOG	Mongondow	Indonesia	<1M
66	PWO	Pwo Western Karen	Myanmar	<1M
67	CJA	Western Cham	Cambodia, Vietnam	<1M
68	AHK	Akha	Laos, Myanmar, Thailand	<1M
69	SSB	Southern Sama	Philippines	<1M
70	SXN	Sangir	Indonesia	<1M

Table 32: (1/2) SEA indigenous languages with <1M speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
71	BTZ	Batak Alas-Kluet	Indonesia	<1M
72	CTD	Tedim Chin	Myanmar	<1M
73	SRV	Southern Sorsoganon	Philippines	<1M
74	ABL	Lampung Nyo	Indonesia	<1M
75	DNW	Western Dani	Indonesia	<1M
76	KTP	Kaduo	Laos	<1M
77	SLP	Lamaholot	Indonesia	<1M
78	RAD	Rade	Vietnam	<1M
79	SKI	Sika	Indonesia	<1M
80	KPM	Koho	Vietnam	<1M
81	BDQ	Bahnar	Vietnam	<1M
82	BDL	Indonesian Bajau	Indonesia	<1M
83	BPR	Koronadal Blaan	Philippines	<1M
84	CCP	Chakma	Myanmar	<1M
85	KNE	Kankanaey	Philippines	<1M
86	KYU	Western Kayah	Myanmar	<1M
87	MHY	Ma'anyan	Indonesia	<1M
88	TNT	Tontemboan	Indonesia	<1M
89	PLL	Shwe Palaung	Myanmar	<1M
90	DAW	Davawenyo	Philippines	<1M
91	CNH	Hakha Chin	Myanmar	<1M
92	SYB	Central Subanen	Philippines	<1M
93	RBB	Rumai Palaung	Myanmar	<1M
94	PMF	Pamona	Indonesia	<1M
95	BLN	Southern Catanduanes Bikol	Philippines	<1M
96	ITV	Itawit	Philippines	<1M
97	PDU	Kayan	Myanmar	<1M
98	MGM	Mambae	East Timor	<1M
99	BHQ	Tukang Besi South	Indonesia	<1M
100	SLY	Selayar	Indonesia	<1M
101	MVP	Duri	Indonesia	<1M
102	BGZ	Banggai	Indonesia	<1M
103	KJC	Coastal Konjo	Indonesia	<1M
104	SUC	Western Subanon	Philippines	<1M
105	CYO	Cuyonon	Philippines	<1M
106	KHC	Tukang Besi North	Indonesia	<1M
107	LHI	Lahu Shi	Myanmar	<1M
108	MEL	Central Melanau	Malaysia	<1M
109	IBL	Ibaloi	Philippines	<1M
110	END	Ende	Indonesia	<1M
111	HVN	Hawu	Indonesia	<1M
112	KKV	Kangean	Indonesia	<1M
113	YKA	Yakan	Philippines	<1M
114	LJL	Li'o	Indonesia	<1M
115	MKZ	Makasae	East Timor	<1M
116	BKD	Binukid	Philippines	<1M
117	BKR	Bakumpai	Indonesia	<1M
118	EKG	Ekari	Indonesia	<1M
119	HNJ	Hmong Njua	Laos, Thailand, Vietnam	<1M
120	KAK	Kalanguya	Philippines	<1M
121	KKH	Khün	Myanmar	<1M
122	LBX	Lawangan	Indonesia	<1M
123	MHX	Lhao Vo	Myanmar	<1M
124	MQJ	Mamasa	Indonesia	<1M
125	PSP	Filipino Sign Language	Philippines	<1M
126	TGN	Tandaganon	Philippines	<1M
		Not in SEACrowd		
127	RHG	Rohingya	Myanmar	<1M
128	PHT	Phu Thai	Laos, Thailand, Vietnam	<1M
129	TVN	Tavoyan	Myanmar	<1M
130	OSI	Osing	Indonesia	<1M
131	ILP	Iranun	Philippines	<1M
132	KZS	Sugut Dusun	Malaysia	<1M
133	VKT	Tenggarong Kutai Malay	Indonesia	<1M
134	PHU	Phuan	Laos, Thailand	<1M
135	CSH	Asho Chin	Myanmar	<1M
136	MLC	Cao Lan	Vietnam	<1M
137	KJK	Highland Konjo	Indonesia	<1M
138	LIW	Col	Indonesia	<1M
139	SSS	So	Laos, Thailand	<1M
140	DNV	Danu	Myanmar	<1M
141	SDQ	Semandang	Indonesia	<1M
142	TJL	Tai Laing	Myanmar	<1M

Table 33: (2/2) SEA indigenous languages with <1M speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	ADR	Adonara	Indonesia	<100K
2	SED	Sedang	Vietnam	<100K
3	BLF	Buol	Indonesia	<100K
4	TBL	Tboli	Philippines	<100K
5	HRE	Hre	Vietnam	<100K
6	ROL	Romblomanon	Philippines	<100K
7	AKL	Aklanon	Philippines	<100K
8	TDN	Tondano	Indonesia	<100K
9	BPS	Sarangani Blaan	Philippines	<100K
10	KQR	Kimaragang	Malaysia	<100K
11	SML	Central Sama	Philippines	<100K
12	TXS	Tonsea	Indonesia	<100K
13	STB	Northern Subanen	Philippines	<100K
14	BKS	Northern Sorsoganon	Philippines	<100K
15	KEI	Kei	Indonesia	<100K
16	KLG	Tagakaulo	Philippines	<100K
17	TLD	Talaud	Indonesia	<100K
18	ATB	Zaiwa	Myanmar	<100K
19	SSE	Balangingih Sama	Philippines	<100K
20	TES	Tengger	Indonesia	<100K
21	TYR	Tai Daeng	Laos, Vietnam	<100K
22	CIA	Cia-Cia	Indonesia	<100K
23	GBI	Galela	Indonesia	<100K
24	OTD	Ot Danum	Indonesia	<100K
25	CTS	Northern Catanduanes Bikol	Philippines	<100K
26	LOE	Saluan	Indonesia	<100K
27	BNO	Bantoanon	Philippines	<100K
28	CMR	Mro-Khimi	Myanmar	<100K
29	UBL	Buhi'non Bikol	Philippines	<100K
30	CJM	Eastern Cham	Vietnam	<100K
31	BKX	Baikeno	East Timor	<100K
32	AAZ	Amarasi	Indonesia	<100K
33	BHW	Biak	Indonesia	<100K
34	KQE	Kalagan	Philippines	<100K
35	XNN	Northern Kankanay	Philippines	<100K
36	XSB	Sambal	Philippines	<100K
37	CFM	Falam Chin	Myanmar	<100K
38	LBL	Libon Bikol	Philippines	<100K
39	WLO	Wolio	Indonesia	<100K
40	BTH	Biatah Bidayuh	Indonesia, Malaysia	<100K
41	KEM	Kemak	East Timor, Indonesia	<100K
42	RAW	Rawang	Myanmar	<100K
43	TFT	Ternate	Indonesia	<100K
44	ZOM	Zo	Myanmar	<100K
45	CNK	Khumi Chin	Myanmar	<100K
46	MQX	Mamuju	Indonesia	<100K
47	MSM	Agusan Manobo	Philippines	<100K
48	NST	Tangshang Naga	Myanmar	<100K
49	NXG	Ngad'a	Indonesia	<100K
50	OBO	Obo Manobo	Philippines	<100K
51	PWW	Pwo Northern Karen	Thailand	<100K
52	SYA	Siang	Indonesia	<100K
53	TOM	Tombulu	Indonesia	<100K
54	XML	Malaysian Sign Language	Malaysia	<100K
55	MBS	Sarangani Manobo	Philippines	<100K
56	MWV	Mentawai	Indonesia	<100K
57	MSK	Mansaka	Philippines	<100K
58	SMK	Bolinao	Philippines	<100K
59	BFN	Bunak	East Timor, Indonesia	<100K
60	BGI	Bagobo-Klata	Philippines	<100K
61	DRG	Rungus	Malaysia	<100K
62	KZF	Da'a Kaili	Indonesia	<100K
63	WEW	Wejewa	Indonesia	<100K
64	ROG	Northern Roglai	Vietnam	<100K
65	ILK	Bogkalot	Philippines	<100K
66	KTV	Eastern Katu	Vietnam	<100K
67	DNT	Mid Grand Valley Dani	Indonesia	<100K
68	FRD	Fordata	Indonesia	<100K
69	MBT	Matigsalug Manobo	Philippines	<100K
70	NXE	Nage	Indonesia	<100K
71	PTT	Enrekang	Indonesia	<100K

Table 34: (1/5) SEA indigenous languages with <100K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
72	TIY	Teduray	Philippines	<100K
73	TJG	Tunjung	Indonesia	<100K
74	WMM	Maiwa	Indonesia	<100K
75	SDO	Bukar-Sadong Bidayuh	Indonesia, Malaysia	<100K
76	KYP	Kang	Laos	<100K
77	TVO	Tidore	Indonesia	<100K
78	HOS	Ho Chi Minh City Sign Language	Vietnam	<100K
79	MHS	Buru	Indonesia	<100K
80	STI	Bulo Stieng	Cambodia, Vietnam	<100K
81	LAW	Lauje	Indonesia	<100K
82	BGS	Tagabawa	Philippines	<100K
83	SJM	Mapun	Philippines	<100K
84	BLR	Blang	Myanmar, Thailand	<100K
85	RGS	Southern Roglai	Vietnam	<100K
86	SMR	Simeulue	Indonesia	<100K
87	CZT	Zotung Chin	Myanmar	<100K
88	KVQ	Geba Karen	Myanmar	<100K
89	MTD	Mualang	Indonesia	<100K
90	XXK	Ke'o	Indonesia	<100K
91	TKD	Tukudede	East Timor	<100K
92	KIX	Khiamniungan Naga	Myanmar	<100K
93	BSB	Brunei Bisaya	Brunei, Malaysia	<100K
94	DAO	Daai Chin	Myanmar	<100K
95	DDG	Fataluku	East Timor	<100K
96	MQN	Moronene	Indonesia	<100K
97	GES	Geser-Gorom	Indonesia	<100K
98	PHO	Phunoi	Laos	<100K
99	SLM	Pangutaran Sama	Philippines	<100K
100	HRO	Haroi	Vietnam	<100K
101	IVV	Ivatan	Philippines	<100K
102	MRH	Mara Chin	Myanmar	<100K
103	BTW	Butuanon	Philippines	<100K
104	CMA	Maa	Vietnam	<100K
105	SBL	Botolan Sambal	Philippines	<100K
106	CMO	Central Mnong	Cambodia, Vietnam	<100K
107	BLZ	Balantak	Indonesia	<100K
108	TPU	Tampuan	Cambodia	<100K
109	BLJ	Bulungan	Indonesia	<100K
110	CGC	Kagayanen	Philippines	<100K
111	CLU	Caluyanun	Philippines	<100K
112	CML	Koneq-koneq	Indonesia	<100K
113	GAD	Gaddang	Philippines	<100K
114	HLT	Matu Chin	Myanmar	<100K
115	IFK	Tuwali Ifugao	Philippines	<100K
116	IFU	Mayoyao Ifugao	Philippines	<100K
117	KNB	Lubuagan Kalinga	Philippines	<100K
118	KSX	Kedang	Indonesia	<100K
119	LCF	Lubu	Indonesia	<100K
120	LSI	Lacid	Myanmar	<100K
121	MBA	Higaonon	Philippines	<100K
122	MNG	Eastern Mnong	Vietnam	<100K
123	MRO	Mru	Myanmar	<100K
124	MTA	Cotabato Manobo	Philippines	<100K
125	SET	Sentani	Indonesia	<100K
126	TMN	Taman	Indonesia	<100K
127	TWU	Termanu	Indonesia	<100K
128	TXM	Tomini	Indonesia	<100K
129	ULM	Ulumanda'	Indonesia	<100K
130	WOW	Wawonii	Indonesia	<100K
131	SNE	Bau Bidayuh	Indonesia, Malaysia	<100K
132	TDF	Talieng	Laos	<100K
133	LBO	Laven	Laos	<100K
134	ACN	Ngochang	Myanmar	<100K
135	TLB	Tobelo	Indonesia	<100K
136	IFA	Amganad Ifugao	Philippines	<100K
137	ITD	Southern Tidung	Indonesia, Malaysia	<100K
138	PHA	Pa-Hng	Vietnam	<100K
139	ATD	Ata Manobo	Philippines	<100K
140	BRU	Eastern Bru	Laos, Vietnam	<100K
141	KZP	Kaidipang	Indonesia	<100K
142	ABX	Inabaknon	Philippines	<100K

Table 35: (2/5) SEA indigenous languages with <100K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
143	AOL	Alor	Indonesia	<100K
144	JMD	Yamdena	Indonesia	<100K
145	LAA	Southern Subanen	Philippines	<100K
146	LMY	Lamboya	Indonesia	<100K
147	TXE	Totoli	Indonesia	<100K
148	OYB	Oy	Laos	<100K
149	MLF	Mal	Laos, Thailand	<100K
150	LND	Lundayeh	Brunei, Indonesia, Malaysia	<100K
151	PRH	Porohanon	Philippines	<100K
152	BRB	Brao	Cambodia, Laos, Vietnam	<100K
153	LBN	Rmeet	Laos	<100K
154	ILM	Iranun	Malaysia	<100K
155	PTU	Bambam	Indonesia	<100K
156	VKL	Kulisusu	Indonesia	<100K
157	BLW	Balangao	Philippines	<100K
158	BSY	Sabah Bisaya	Malaysia	<100K
159	KRR	Krung	Cambodia	<100K
160	DTB	Labuk-Kinabatangan Kadazan	Malaysia	<100K
161	AYZ	Mai Brat	Indonesia	<100K
162	BAC	Badui	Indonesia	<100K
163	BRV	Western Bru	Laos, Thailand	<100K
164	BWP	Mandobo Bawah	Indonesia	<100K
165	DNA	Upper Grand Valley Dani	Indonesia	<100K
166	DNI	Lower Grand Valley Dani	Indonesia	<100K
167	DTR	Lotud	Malaysia	<100K
168	DUN	Dusun Deyah	Indonesia	<100K
169	KJE	Kisar	Indonesia	<100K
170	KLI	Kalumpang	Indonesia	<100K
171	KOD	Kodi	Indonesia	<100K
172	LLG	Lole	Indonesia	<100K
173	LRT	Larantuka Malay	Indonesia	<100K
174	MNZ	Moni	Indonesia	<100K
175	PEA	Peranakan Indonesian	Indonesia	<100K
176	PPK	Uma	Indonesia	<100K
177	PRT	Prai	Laos, Thailand	<100K
178	TMM	Tai Thanh	Vietnam	<100K
179	TNW	Tonsawang	Indonesia	<100K
180	TWY	Tawoyan	Indonesia	<100K
181	TXQ	Tii	Indonesia	<100K
182	WLW	Walak	Indonesia	<100K
183	SKH	Sikule	Indonesia	<100K
184	LBK	Central Bontok	Philippines	<100K
185	CJE	Chru	Vietnam	<100K
186	HNN	Hanunoo	Philippines	<100K
187	TLU	Tulehu	Indonesia	<100K
188	WMH	Waima'a	East Timor	<100K
189	HRK	Haruku	Indonesia	<100K
190	LEX	Luang	Indonesia	<100K
191	PUO	Puoc	Vietnam	<100K
192	REN	Rengao	Vietnam	<100K
193	ALP	Alune	Indonesia	<100K
194	BWE	Bwe Karen	Myanmar	<100K
195	TLT	Sou Nama	Indonesia	<100K
196	ZYP	Zyphe Chin	Myanmar	<100K
197	ABZ	Abui	Indonesia	<100K
198	AKG	Anakalangu	Indonesia	<100K
199	HAD	Hatam	Indonesia	<100K
200	HTU	Hitu	Indonesia	<100K
201	NLC	Nalca	Indonesia	<100K
202	PAC	Pacoh	Laos, Vietnam	<100K
203	YOG	Yogad	Philippines	<100K
204	MXD	Modang	Indonesia	<100K
205	JEH	Jeh	Laos, Vietnam	<100K
206	KYN	Northern Binukidnon	Philippines	<100K
207	PHG	Phuong	Vietnam	<100K
208	AGN	Agutaynen	Philippines	<100K
209	CNW	Ngawn Chin	Myanmar	<100K
210	ILA	Ile Ape	Indonesia	<100K
211	KRD	Kairui-Midiki	East Timor	<100K
212	LOA	Loloda	Indonesia	<100K
213	MBB	Western Bukidnon Manobo	Philippines	<100K
214	MWQ	Müün Chin	Myanmar	<100K
215	NXA	Nauete	East Timor	<100K
216	PRF	Paranan	Philippines	<100K

Table 36: (3/5) SEA indigenous languages with <100K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
217	SNL	Sangil	Philippines	<100K
218	TBY	Tabaru	Indonesia	<100K
219	TEA	Temiar	Malaysia	<100K
220	YLI	Angguruk Yali	Indonesia	<100K
221	MEJ	Meyah	Indonesia	<100K
222	MBI	Ilianen Manobo	Philippines	<100K
223	PLW	Brooke's Point Palawano	Philippines	<100K
224	DUU	Drung	Myanmar	<100K
225	HEG	Helong	Indonesia	<100K
226	MZQ	Mori Atas	Indonesia	<100K
227	UHN	Damal	Indonesia	<100K
228	XMZ	Mori Bawah	Indonesia	<100K
229	KJM	Kháng	Vietnam	<100K
230	HAL	Salang	Laos, Vietnam	<100K
231	IDT	Idaté	East Timor	<100K
232	DOK	Dondo	Indonesia	<100K
233	GAL	Galolen	East Timor, Indonesia	<100K
234	KSC	Southern Kalinga	Philippines	<100K
235	TXA	Tombonuo	Malaysia	<100K
236	NGT	Kriang	Laos	<100K
237	KMK	Limos Kalinga	Philippines	<100K
238	ALO	Larike-Wakasihu	Indonesia	<100K
239	YNO	Yong	Thailand	<100K
240	RIL	Riang Lang	Myanmar	<100K
241	ATQ	Aralle-Tabulahan	Indonesia	<100K
242	CEK	Eastern Khumi Chin	Myanmar	<100K
243	CUA	Cua	Vietnam	<100K
244	MNX	Sougb	Indonesia	<100K
245	MQS	West Makian	Indonesia	<100K
246	NUF	Nusu	Myanmar	<100K
247	PLC	Central Palawano	Philippines	<100K
248	PLV	Southwest Palawano	Philippines	<100K
249	RGU	Rikou	Indonesia	<100K
250	SZW	Sawai	Indonesia	<100K
251	TDJ	Tajio	Indonesia	<100K
252	XKL	Mainstream Kenyah	Indonesia, Malaysia	<100K
253	YIN	Riang Lai	Myanmar	<100K
254	LCL	Lisela	Indonesia	<100K
255	LRA	Rara Bakati'	Indonesia, Malaysia	<100K
256	BVE	Berau Malay	Indonesia	<100K
257	KML	Tanudan Kalinga	Philippines	<100K
258	BEU	Blagar	Indonesia	<100K
259	XEM	Mateq	Indonesia	<100K
260	LEV	Western Pantar	Indonesia	<100K
261	PTN	Patani	Indonesia	<100K
262	OOG	Ong	Laos	<100K
263	SPR	Saparua	Indonesia	<100K
264	AMK	Ambai	Indonesia	<100K
265	IFB	Batad Ifugao	Philippines	<100K
266	AAX	Mandobo Atas	Indonesia	<100K
267	BEP	Behoa	Indonesia	<100K
268	BVY	Baybayanon	Philippines	<100K
269	CSY	Siyin Chin	Myanmar	<100K
270	DBJ	Ida'an	Malaysia	<100K
271	EMB	Embaloh	Indonesia	<100K
272	IRY	Iraya	Philippines	<100K
273	JAK	Jakun	Malaysia	<100K
274	JAQ	Yaqay	Indonesia	<100K
275	KPS	Tehit	Indonesia	<100K
276	KVB	Kubu	Indonesia	<100K
277	KXF	Kawyaw	Myanmar	<100K
278	KYT	Kayagar	Indonesia	<100K
279	LJE	Rampi	Indonesia	<100K
280	LUR	Loura	Indonesia	<100K
281	MBD	Dibabawon Manobo	Philippines	<100K
282	MBF	Baba Malay	Singapore	<100K
283	MKY	East Makian	Indonesia	<100K
284	MVD	Mamboru	Indonesia	<100K
285	NDX	Nduga	Indonesia	<100K
286	PEZ	Eastern Penan	Brunei, Malaysia	<100K
287	PLE	Palu'e	Indonesia	<100K
288	SEA	Semai	Malaysia	<100K
289	SSQ	So'a	Indonesia	<100K

Table 37: (4/5) SEA indigenous languages with <100K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
290	SZB	Ngalum	Indonesia	<100K
291	TBK	Calamian Tagbanwa	Philippines	<100K
292	TBW	Tagbanwa	Philippines	<100K
293	TXX	Tatana	Malaysia	<100K
294	WNK	Wanukaka	Indonesia	<100K
295	YVA	Yawa	Indonesia	<100K
		Not in SEACrowd		
296	INT	Intha	Myanmar	<100K
297	LOC	Inonhan	Philippines	<100K
298	MQG	Kota Bangun Kutai Malay	Indonesia	<100K
299	BFX	Bantayanon	Philippines	<100K
300	TOU	Tho	Vietnam	<100K
301	NCQ	Northern Katang	Laos	<100K
302	BVU	Bukit Malay	Indonesia	<100K
303	BYD	Benyadu'	Indonesia	<100K
304	TSQ	Thai Sign Language	Thailand	<100K
305	NYW	Nyaw	Thailand	<100K
306	RIR	Ribun	Indonesia	<100K
307	SCG	Sanggau	Indonesia	<100K
308	SCT	Southern Katang	Laos	<100K
309	STT	Budeh Stieng	Vietnam	<100K
310	TCO	Taungyo	Myanmar	<100K
311	VKK	Kaur	Indonesia	<100K
312	HAB	Hanoi Sign Language	Vietnam	<100K
313	DJO	Jangkang	Indonesia	<100K
314	SBX	Seberuang	Indonesia	<100K
315	LSO	Laos Sign Language	Laos	<100K
316	SEZ	Senthang Chin	Myanmar	<100K
317	SOA	Thai Song	Thailand	<100K
318	KNL	Keninjal	Indonesia	<100K
319	TTH	Upper Ta'oih	Laos, Vietnam	<100K
320	APG	Ampanang	Indonesia	<100K
321	MNN	Southern Mnong	Vietnam	<100K
322	PEL	Pekal	Indonesia	<100K
323	ZKD	Kadu	Myanmar	<100K
324	BKZ	Bungku	Indonesia	<100K
325	MKX	Kinamiging Manobo	Philippines	<100K
326	BNU	Bentong	Indonesia	<100K
327	KXY	Kayong	Vietnam	<100K
328	MHP	Balinese Malay	Indonesia	<100K
329	UNZ	Unde Kaili	Indonesia	<100K
330	BLD	Bolango	Indonesia	<100K
331	KUF	Western Katu	Laos	<100K
332	DNK	Dengka	Indonesia	<100K
333	MVV	Tagal Murut	Indonesia, Malaysia	<100K
334	SKN	Kolibugan Subanon	Philippines	<100K
335	SZN	Sula	Indonesia	<100K
336	CNB	Uppu Chin	Myanmar	<100K
337	BHV	Bahau	Indonesia	<100K
338	ITT	Maeng Itneg	Philippines	<100K
339	HJI	Haji	Indonesia	<100K
340	GHK	Geko Karen	Myanmar	<100K
341	KVL	Kayaw	Myanmar	<100K
342	TTO	Lower Ta'oih	Laos	<100K
343	BDB	Basap	Indonesia	<100K
344	CLJ	Laitu Chin	Myanmar	<100K
345	CLT	Lautu Chin	Myanmar	<100K
346	DUP	Duano	Indonesia, Malaysia	<100K
347	KYB	Butbut Kalinga	Philippines	<100K
348	STG	Trieng	Vietnam	<100K
349	CBW	Kinabalian	Philippines	<100K
350	CSV	Sumtu Chin	Myanmar	<100K
351	RIU	Riung	Indonesia	<100K
352	SRG	Sulod	Philippines	<100K
353	ITY	Moyadan Itneg	Philippines	<100K
354	KKG	Mabaka Valley Kalinga	Philippines	<100K
355	BNE	Bintauna	Indonesia	<100K
356	NLK	Ninia Yali	Indonesia	<100K
357	HIK	Seit-Kaitetu	Indonesia	<100K
358	KSN	Kasiguranin	Philippines	<100K
359	TSL	Ts'ün-Lao	Vietnam	<100K
360	XAO	Khao	Vietnam	<100K

Table 38: (5/5) SEA indigenous languages with <100K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	XTE	Ketengban	Indonesia	<10K
2	BNA	Bonerate	Indonesia	<10K
3	BKU	Buhid	Philippines	<10K
4	AWS	South Awyu	Indonesia	<10K
5	WOO	Manombai	Indonesia	<10K
6	ASC	Casuarina Coast Asmat	Indonesia	<10K
7	TIH	Timugon Murut	Malaysia	<10K
8	ASL	Asilulu	Indonesia	<10K
9	SGB	Mag-antsi Ayta	Philippines	<10K
10	EKY	Eastern Kayah	Myanmar, Thailand	<10K
11	IFY	Keley-i Kallahan	Philippines	<10K
12	INL	Indonesian Sign Language	Indonesia	<10K
13	KGQ	Kamoro	Indonesia	<10K
14	KHT	Khamti	Myanmar	<10K
15	KPQ	Korupun-Sela	Indonesia	<10K
16	KTI	North Muyu	Indonesia	<10K
17	LCP	Western Lawa	Thailand	<10K
18	MTJ	Moskona	Indonesia	<10K
19	SLU	Selaru	Indonesia	<10K
20	TMW	Temuan	Malaysia	<10K
21	TXT	Citak	Indonesia	<10K
22	WHK	Wahau Kenyah	Indonesia	<10K
23	TXN	West Tarangan	Indonesia	<10K
24	DRO	Daro-Matu Melanau	Malaysia	<10K
25	AWU	Central Awyu	Indonesia	<10K
26	ITB	Binongan Itneg	Philippines	<10K
27	LTI	Leti	Indonesia	<10K
28	SAJ	Sahu	Indonesia	<10K
29	KVV	Kola	Indonesia	<10K
30	KVU	Yinbaw	Myanmar	<10K
31	AKC	Mpur	Indonesia	<10K
32	CNS	Central Asmat	Indonesia	<10K
33	CRW	Chrau	Vietnam	<10K
34	LWL	Eastern Lawa	Thailand	<10K
35	LZN	Lainong Naga	Myanmar	<10K
36	MRZ	Marind	Indonesia	<10K
37	ROW	Dela-Oenale	Indonesia	<10K
38	SFE	Eastern Subanen	Philippines	<10K
39	TTD	Tutong	Brunei	<10K
40	IWO	Morop	Indonesia	<10K
41	TWB	Tawbuid	Philippines	<10K
42	BHZ	Bada	Indonesia	<10K
43	PWM	Molbog	Malaysia, Philippines	<10K
44	PSA	Asue Awyu	Indonesia	<10K
45	EBK	Eastern Bontok	Philippines	<10K
46	TRE	East Tarangan	Indonesia	<10K
47	NPY	Napu	Indonesia	<10K
48	GDG	Ga'dang	Philippines	<10K
49	GIR	Red Gelao	Vietnam	<10K
50	KLL	Kagan Kalagan	Philippines	<10K
51	LWT	Lewotobi	Indonesia	<10K
52	MOO	Monom	Vietnam	<10K
53	PNP	Pancana	Indonesia	<10K
54	TDR	Todrah	Vietnam	<10K
55	WEO	Wemale	Indonesia	<10K
56	WOI	Kamang	Indonesia	<10K
57	WRP	Waropen	Indonesia	<10K
58	LHA	Laha	Vietnam	<10K
59	KVO	Dobel	Indonesia	<10K
60	MTG	Una	Indonesia	<10K
61	INN	Isinay	Philippines	<10K
62	IHP	Iha	Indonesia	<10K
63	JKA	Kaera	Indonesia	<10K
64	MYL	Moma	Indonesia	<10K
65	MMN	Minamanwa	Philippines	<10K
66	NXR	Ninggerum	Indonesia	<10K
67	BLX	Mag-Indi Ayta	Philippines	<10K
68	DUW	Dusun Witu	Indonesia	<10K
69	KGW	Karon Dori	Indonesia	<10K
70	KYO	Klon	Indonesia	<10K
71	LBT	Lachi	Vietnam	<10K
72	MLI	Malimpung	Indonesia	<10K
73	NFA	Dhao	Indonesia	<10K
74	PDO	Padoe	Indonesia	<10K
75	RAZ	Rahambuu	Indonesia	<10K
76	TPG	Kula	Indonesia	<10K
77	URK	Urak Lawoi'	Thailand	<10K
78	WAD	Wamesa	Indonesia	<10K
79	WOD	Wolani	Indonesia	<10K
80	WUL	Silimo	Indonesia	<10K

Table 39: (1/6) SEA indigenous languages with <10K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
81	YAC	Pass Valley Yali	Indonesia	<10K
82	YOY	Yoy	Laos, Thailand	<10K
83	AND	Ansus	Indonesia	<10K
84	MXN	Moi Kelim	Indonesia	<10K
85	TLV	Taliabu	Indonesia	<10K
86	BTY	Bobot	Indonesia	<10K
87	DUQ	Dusun Malang	Indonesia	<10K
88	UMS	Pendau	Indonesia	<10K
89	VBB	Southeast Babar	Indonesia	<10K
90	BAJ	Barakai	Indonesia	<10K
91	BGR	Bawm Chin	Myanmar	<10K
92	IRR	Ir	Laos	<10K
93	NBQ	Nggem	Indonesia	<10K
94	BQR	Burusu	Indonesia	<10K
95	KVD	Kui	Indonesia	<10K
96	BNY	Bintulu	Malaysia	<10K
97	RKA	Kraol	Cambodia	<10K
98	JAH	Jah Hut	Malaysia	<10K
99	KYS	Baram Kayan	Malaysia	<10K
100	SMU	Somray	Cambodia	<10K
101	SZA	Semelai	Malaysia	<10K
102	ALK	Alak	Laos	<10K
103	ANL	Anu-Khongso Chin	Myanmar	<10K
104	BEI	Bakati'	Indonesia	<10K
105	IRH	Irarutu	Indonesia	<10K
106	KTA	Katua	Vietnam	<10K
107	KTS	South Muyu	Indonesia	<10K
108	KZI	Kelabit	Indonesia, Malaysia	<10K
109	LMR	Lamalera	Indonesia	<10K
110	MWT	Moken	Myanmar, Thailand	<10K
111	NTX	Tangkhul Naga	Myanmar	<10K
112	ROR	Rongga	Indonesia	<10K
113	SDU	Sarudu	Indonesia	<10K
114	SLZ	Ma'ya	Indonesia	<10K
115	SRE	Sara Bakati'	Indonesia	<10K
116	TGB	Tobilung	Malaysia	<10K
117	TWE	Teiwa	Indonesia	<10K
118	TYN	Kombai	Indonesia	<10K
119	WAH	Watubela	Indonesia	<10K
120	NEV	Nyaheun	Laos	<10K
121	KLZ	Kabola	Indonesia	<10K
122	AWY	Edera Awyu	Indonesia	<10K
123	ABD	Manide	Philippines	<10K
124	TNM	Tabla	Indonesia	<10K
125	SKB	Saek	Laos, Thailand	<10K
126	KVW	Wersing	Indonesia	<10K
127	XOD	Kokoda	Indonesia	<10K
128	BPQ	Banda Malay	Indonesia	<10K
129	BAY	Batuley	Indonesia	<10K
130	KGX	Kamaru	Indonesia	<10K
131	KHE	Korowai	Indonesia	<10K
132	LKJ	Remun	Malaysia	<10K
133	PKU	Paku	Indonesia	<10K
134	SAW	Sawi	Indonesia	<10K
135	TCG	Tamagario	Indonesia	<10K
136	PNE	Western Penan	Malaysia	<10K
137	XKS	Kumbewaha	Indonesia	<10K
138	PGU	Pagu	Indonesia	<10K
139	TPO	Tai Pao	Laos, Vietnam	<10K
140	ZRS	Mairasi	Indonesia	<10K
141	KZZ	Kalabra	Indonesia	<10K
142	BLS	Balaesang	Indonesia	<10K
143	KUV	Kur	Indonesia	<10K
144	REE	Rejang Kayan	Malaysia	<10K
145	ABP	Abellen Ayta	Philippines	<10K
146	ADN	Adang	Indonesia	<10K
147	AHH	Aghu	Indonesia	<10K
148	BND	Banda	Indonesia	<10K
149	BNQ	Bantik	Indonesia	<10K
150	CKH	Chak	Myanmar	<10K
151	DUE	Umiray Dumaget Agta	Philippines	<10K
152	EIP	Lik	Indonesia	<10K
153	KGR	Abun	Indonesia	<10K
154	KIG	Kimaghima	Indonesia	<10K
155	NSY	Nasal	Indonesia	<10K
156	SWT	Sawila	Indonesia	<10K
157	TMG	Ternateño	Indonesia	<10K
158	WMS	Wambon	Indonesia	<10K
159	MHE	Mah Meri	Malaysia	<10K
160	BGL	Bo	Laos	<10K

Table 40: (2/6) SEA indigenous languages with <10k speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
161	BPV	Bian Marind	Indonesia	<10K
162	GZN	Gane	Indonesia	<10K
163	DMR	East Damar	Indonesia	<10K
164	OBK	Southern Bontok	Philippines	<10K
165	BZL	Boano	Indonesia	<10K
166	HBU	Habun	East Timor	<10K
167	ZNG	Mang	Vietnam	<10K
168	GEI	Gebe	Indonesia	<10K
169	SPB	Sepa	Indonesia	<10K
170	AGV	Remontado Dumagat	Philippines	<10K
171	BZQ	Buli	Indonesia	<10K
172	BRP	Barapasi	Indonesia	<10K
173	CBL	Bualkhaw Chin	Myanmar	<10K
174	GRS	Gresi	Indonesia	<10K
175	JMN	Makuri Naga	Myanmar	<10K
176	KMT	Kemtuik	Indonesia	<10K
177	KWE	Kwerba	Indonesia	<10K
178	SKO	Seko Tengah	Indonesia	<10K
179	WRS	Waris	Indonesia	<10K
180	KYI	Kiput	Malaysia	<10K
181	NRM	Narom	Malaysia	<10K
182	KLW	Tado	Indonesia	<10K
183	SPU	Sapuan	Laos	<10K
184	JEI	Yei	Indonesia	<10K
185	SQQ	Sou	Laos	<10K
186	AWV	Jair Awyu	Indonesia	<10K
187	BUP	Busoa	Indonesia	<10K
188	KKL	Kosarek Yale	Indonesia	<10K
189	ZKA	Kaimbulawa	Indonesia	<10K
190	KJR	Kurudu	Indonesia	<10K
191	ALJ	Alangan	Philippines	<10K
192	ASY	Yaosakor Asmat	Indonesia	<10K
193	DMS	Dampelas	Indonesia	<10K
194	ENR	Emem	Indonesia	<10K
195	HNU	Hung	Laos, Vietnam	<10K
196	KWT	Kwesten	Indonesia	<10K
197	KYJ	Karao	Philippines	<10K
198	LAU	Laba	Indonesia	<10K
199	LEY	Limola	Indonesia	<10K
200	MQF	Momuna	Indonesia	<10K
201	MQO	Modole	Indonesia	<10K
202	NIR	Nimboran	Indonesia	<10K
203	PMO	Pom	Indonesia	<10K
204	SGE	Segai	Indonesia	<10K
205	SZC	Semaq Beri	Malaysia	<10K
206	TGT	Central Tagbanwa	Philippines	<10K
207	TTY	Sikaritai	Indonesia	<10K
208	BGK	Bit	Laos	<10K
209	GRM	Kota Marudu Talantang	Malaysia	<10K
210	SRL	Isirawa	Indonesia	<10K
211	WBW	Woi	Indonesia	<10K
212	SIB	Sebop	Malaysia	<10K
213	BNB	Bookan Murut	Malaysia	<10K
214	LLM	Lasalimu	Indonesia	<10K
215	RMM	Roma	Indonesia	<10K
216	PCB	Pear	Cambodia	<10K
217	ABC	Ambala Ayta	Philippines	<10K
218	NXX	Nafri	Indonesia	<10K
219	LWH	White Lachi	Vietnam	<10K
220	URY	Orya	Indonesia	<10K
221	IRX	Kamberau	Indonesia	<10K
222	ATK	Ati	Philippines	<10K
223	BGB	Bobongko	Indonesia	<10K
224	BVZ	Bauzi	Indonesia	<10K
225	BZP	Kemberano	Indonesia	<10K
226	CBN	Nyahkur	Thailand	<10K
227	DBF	Edopi	Indonesia	<10K
228	ENO	Enggano	Indonesia	<10K
229	MKM	Moklen	Thailand	<10K
230	NXL	South Nuaulu	Indonesia	<10K
231	VKO	Kodeoha	Indonesia	<10K
232	WBB	Wabo	Indonesia	<10K
233	YIR	North Awyu	Indonesia	<10K
234	ZBC	Central Berawan	Malaysia	<10K
235	BYA	Batak	Philippines	<10K

Table 41: (3/6) SEA indigenous languages with <10K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
236	BDG	Bonggi	Malaysia	<10K
237	FAU	Fayu	Indonesia	<10K
238	ILU	Ili'uun	Indonesia	<10K
239	YET	Yetfa	Indonesia	<10K
240	DMY	Sowari	Indonesia	<10K
241	DDW	Dawera-Daweloor	Indonesia	<10K
242	JHI	Jehai	Malaysia	<10K
243	XMT	Matbat	Indonesia	<10K
244	BEG	Belait	Brunei	<10K
245	IVB	Ibatan	Philippines	<10K
246	OIA	Oirata	Indonesia	<10K
247	BKL	Berik	Indonesia	<10K
248	DUO	Dupaninan Agta	Philippines	<10K
249	KDW	Koneraw	Indonesia	<10K
250	MSF	Mekwei	Indonesia	<10K
251	NQM	Ndom	Indonesia	<10K
252	SBG	Moi Lemas	Indonesia	<10K
253	SEU	Serui-Laut	Indonesia	<10K
254	TVE	Te'un	Indonesia	<10K
255	TZN	Tugun	Indonesia	<10K
256	WNG	Wanggom	Indonesia	<10K
257	BNJ	Bangon	Philippines	<10K
258	SNV	Sa'ban	Indonesia, Malaysia	<10K
259	BDW	Baham	Indonesia	<10K
260	RAN	Riantana	Indonesia	<10K
261	RNN	Roon	Indonesia	<10K
262	SZP	Suabo	Indonesia	<10K
263	ZBE	East Berawan	Malaysia	<10K
264	SCB	Chut	Laos, Vietnam	<10K
265	TVM	Tela-Masbuar	Indonesia	<10K
266	UDJ	Ujir	Indonesia	<10K
267	AGY	Southern Alta	Philippines	<10K
268	AIR	Airoran	Indonesia	<10K
269	AQM	Atohwaim	Indonesia	<10K
270	ASI	Buruwai	Indonesia	<10K
271	ATT	Pamplona Atta	Philippines	<10K
272	BCD	North Babar	Indonesia	<10K
273	BNF	Masiwang	Indonesia	<10K
274	BTQ	Batek	Malaysia	<10K
275	CTH	Thaiphum Chin	Myanmar	<10K
276	DEM	Dem	Indonesia	<10K
277	DMG	Upper Kinabatangan	Malaysia	<10K
278	DNU	Danau	Myanmar	<10K
279	ETZ	Semimi	Indonesia	<10K
280	JBJ	Arandai	Indonesia	<10K
281	KBV	Dla	Indonesia	<10K
282	KPU	Kafoa	Indonesia	<10K
283	KVY	Yintale	Myanmar	<10K
284	MSG	Moraid	Indonesia	<10K
285	NKS	North Asmat	Indonesia	<10K
286	PNX	Phong-Kniang	Laos	<10K
287	SOB	Sobei	Indonesia	<10K
288	WGO	Ambel	Indonesia	<10K
289	WNO	Wano	Indonesia	<10K
290	XSE	Sempan	Indonesia	<10K
291	ZBW	West Berawan	Malaysia	<10K
		Not in SEACrowd		
292	RBK	Northern Bontok	Philippines	<10K
293	KVT	Lahta	Myanmar	<10K
294	LBG	Laopang	Laos	<10K
295	STU	Samtao	Myanmar	<10K
296	KXK	Zayein	Myanmar	<10K
297	ITI	Inlaud Itneg	Philippines	<10K
298	NQQ	Chen-Kayu Naga	Myanmar	<10K
299	PNC	Pannei	Indonesia	<10K
300	ZKN	Kanan	Myanmar	<10K
301	MLZ	Malaynon	Philippines	<10K
302	KHF	Khuen	Laos	<10K
303	KKX	Kohin	Indonesia	<10K
304	LMJ	West Lembata	Indonesia	<10K
305	DKR	Kuijau	Malaysia	<10K
306	EBC	Beginci	Indonesia	<10K
307	MTW	Southern Binukidnon	Philippines	<10K
308	MQK	Rajah Kabunsuwan Manobo	Philippines	<10K
309	CSX	Cambodian Sign Language	Cambodia	<10K
310	TIS	Masadiit Itneg	Philippines	<10K
311	CSJ	Songlai Chin	Myanmar	<10K
312	MQC	Mangole	Indonesia	<10K
313	BPZ	Bilba	Indonesia	<10K
314	LMF	South Lembata	Indonesia	<10K
315	WHA	Sou Upaa	Indonesia	<10K
316	LKC	Kucong	Vietnam	<10K
317	MQA	Maba	Indonesia	<10K
318	LCQ	Luhu	Indonesia	<10K
319	MJB	Makalero	East Timor	<10K

Table 42: (4/6) SEA indigenous languages with <10K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		Not in SEACrowd		
320	KRV	Kavet	Cambodia	<10K
321	CEY	Ekai Chin	Myanmar	<10K
322	KJT	Phrae Pwo Karen	Thailand	<10K
323	KUK	Kepo'	Indonesia	<10K
324	PUT	Putoh	Indonesia	<10K
325	RJG	Rajong	Indonesia	<10K
326	SJB	Sajau Basap	Indonesia	<10K
327	TKZ	Takua	Vietnam	<10K
328	AMV	Ambelau	Indonesia	<10K
329	WLH	Welaun	East Timor, Indonesia	<10K
330	PLZ	Paluan Murut	Malaysia	<10K
331	JKP	Paku Karen	Myanmar	<10K
332	ADB	Atauran	East Timor	<10K
333	NEA	Eastern Ngad'a	Indonesia	<10K
334	NTD	Northern Tidung	Malaysia	<10K
335	PHH	Phula	Vietnam	<10K
336	REB	Rembong	Indonesia	<10K
337	SKX	Seko Padang	Indonesia	<10K
338	SWU	Suwawa	Indonesia	<10K
339	TGR	Tareng	Laos	<10K
340	WEU	Rawngtu Chin	Myanmar	<10K
341	SAU	Saleman	Indonesia	<10K
342	THI	Tai Long	Laos	<10K
343	LOW	Tampias Lobu	Malaysia	<10K
344	NPG	Ponyo-Gongwang Naga	Myanmar	<10K
345	UKK	Muak Sa-aak	Myanmar	<10K
346	TLQ	Tai Loi	Laos, Myanmar	<10K
347	HKN	Mel-Khaonh	Cambodia	<10K
348	JKM	Mobwa Karen	Myanmar	<10K
349	LMQ	Lamatuka	Indonesia	<10K
350	LVU	Levuka	Indonesia	<10K
351	LWE	Lewoeleng	Indonesia	<10K
352	RTC	Rungtu Chin	Myanmar	<10K
353	RUU	Lanas Lobu	Malaysia	<10K
354	TIU	Adasen	Philippines	<10K
355	UMN	Paungnyuan Naga	Myanmar	<10K
356	LHH	Laha	Indonesia	<10K
357	BJX	Vanaw Kalinga	Philippines	<10K
358	BVT	Bati	Indonesia	<10K
359	KQV	Okolod	Indonesia, Malaysia	<10K
360	XKK	Kachok	Cambodia	<10K
361	IWK	I-wak	Philippines	<10K
362	LKA	Lakalei	East Timor	<10K
363	BZN	Boano	Indonesia	<10K
364	SBR	Sembakung Murut	Indonesia, Malaysia	<10K
365	BFG	Busang Kayan	Indonesia	<10K
366	HAP	Hupla	Indonesia	<10K
367	KXI	Keningau Murut	Malaysia	<10K
368	LLQ	Lolak	Indonesia	<10K
369	ROC	Cacgia Roglai	Vietnam	<10K
370	SLS	Singapore Sign Language	Singapore	<10K
371	STE	Liana-Seti	Indonesia	<10K
372	ULU	Uma' Lung	Indonesia	<10K
373	WLI	Waioli	Indonesia	<10K
374	WRX	Wae Rana	Indonesia	<10K
375	XHV	Khua	Laos, Vietnam	<10K
376	TDY	Tadyawan	Philippines	<10K
377	ZBT	Batui	Indonesia	<10K
378	SWS	Seluwasan	Indonesia	<10K
379	PNI	Aoheng	Indonesia	<10K
380	TUJ	Tugutil	Indonesia	<10K
381	NPS	Nipsan	Indonesia	<10K
382	UAN	Kuan	Laos	<10K
383	VBK	Southwestern Bontok	Philippines	<10K
384	DMV	Dumpas	Malaysia	<10K
385	XKO	Kiorr	Laos	<10K
386	KVE	Kalabakan Murut	Malaysia	<10K
387	MCM	Malaccan Portuguese Creole	Malaysia	<10K
388	LTU	Latu	Indonesia	<10K
389	GEF	Gerai	Indonesia	<10K
390	CNC	Côông	Vietnam	<10K
391	BPO	Anasi	Indonesia	<10K
392	HLD	Halang Doan	Laos, Vietnam	<10K
393	NXK	Kokak Naga	Myanmar	<10K
394	PUJ	Punan Tubu	Indonesia	<10K
395	XKN	Kayan River Kayan	Indonesia	<10K
396	YCP	Chepya	Laos	<10K
397	LCS	Lisabata-Nuniali	Indonesia	<10K
398	HAF	Haiphong Sign Language	Vietnam	<10K
399	SLT	Sila	Laos, Vietnam	<10K

Table 43: (5/6) SEA indigenous languages with <10K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		Not in SEACrowd		
400	KVH	Komodo	Indonesia	<10K
401	APF	Pahanan Agta	Philippines	<10K
402	BZB	Andio	Indonesia	<10K
403	JAL	Yalahatan	Indonesia	<10K
404	MVR	Marau	Indonesia	<10K
405	AGZ	Mt. Iriga Agta	Philippines	<10K
406	DKK	Dakka	Indonesia	<10K
407	GAK	Gamkonora	Indonesia	<10K
408	KMD	Majukayang Kalinga	Philippines	<10K
409	MQP	Manipa	Indonesia	<10K
410	PZN	Jejara Naga	Myanmar	<10K
411	XKD	Mendalam Kayan	Indonesia	<10K
412	XAY	Kayan Mahakam	Indonesia	<10K
413	XKY	Uma' Lasan	Indonesia, Malaysia	<10K
414	MQQ	Minokok	Malaysia	<10K
415	NEO	Ná-Meo	Vietnam	<10K
416	TLN	Talondo'	Indonesia	<10K
417	BQY	Kata Kolok	Indonesia	<10K
418	MXR	Murik	Malaysia	<10K
419	NTY	Mantsi	Vietnam	<10K
420	TEV	Teor	Indonesia	<10K
421	TTP	Tombelala	Indonesia	<10K
422	AYT	Magbukun Ayta	Philippines	<10K
423	CKN	Kaang Chin	Myanmar	<10K
424	CNO	Con	Laos	<10K
425	GOQ	Gorap	Indonesia	<10K
426	HOV	Hovongan	Indonesia	<10K
427	LPN	Long Phuri Naga	Myanmar	<10K
428	NLQ	Lao Naga	Myanmar	<10K
429	NQY	Akyaung Ari Naga	Myanmar	<10K
430	NUO	Ngoaun	Laos, Vietnam	<10K
431	PSG	Penang Sign Language	Malaysia	<10K
432	UES	Kioko	Indonesia	<10K

Table 44: (6/6) SEA indigenous languages with <10K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	SOW	Sowanda	Indonesia	<1K
2	DUV	Duvle	Indonesia	<1K
3	HMU	Hamap	Indonesia	<1K
4	KTT	Ketum	Indonesia	<1K
5	MPZ	Mpi	Thailand	<1K
6	TVW	Sedoa	Indonesia	<1K
7	SYO	Su'ung	Cambodia	<1K
8	MGK	Mawes	Indonesia	<1K
9	MSS	West Masela	Indonesia	<1K
10	DIJ	Dai	Indonesia	<1K
11	DRN	West Damar	Indonesia	<1K
12	LJI	Laiyolo	Indonesia	<1K
13	MTH	Munggui	Indonesia	<1K
14	PSN	Panasuan	Indonesia	<1K
15	RET	Reta	Indonesia	<1K
16	TWG	Tereweng	Indonesia	<1K
17	BPG	Bonggo	Indonesia	<1K
18	AGT	Central Cagayan Agta	Philippines	<1K
19	KVZ	Tsaukambo	Indonesia	<1K
20	SKP	Sekapan	Malaysia	<1K
21	BSM	Busami	Indonesia	<1K
22	BZI	Bisu	Thailand	<1K
23	KZM	Kais	Indonesia	<1K
24	MHZ	Mor	Indonesia	<1K
25	NKJ	Nakai	Indonesia	<1K
26	PRU	Puragi	Indonesia	<1K
27	SKV	Skou	Indonesia	<1K
28	LAQ	Qabiao	Vietnam	<1K
29	SSM	Semnam	Malaysia	<1K
30	SLG	Selungai Murut	Indonesia, Malaysia	<1K
31	TPF	Tarpia	Indonesia	<1K
32	VTO	Vitou	Indonesia	<1K
33	WSA	Warembori	Indonesia	<1K
34	DGC	Casiguran Dumagat Agta	Philippines	<1K
35	BFE	Betaf	Indonesia	<1K
36	KGB	Kawe	Indonesia	<1K
37	KWH	Kowiai	Indonesia	<1K
38	PPM	Papuma	Indonesia	<1K
39	TDI	Tomadino	Indonesia	<1K
40	TMU	Iau	Indonesia	<1K
41	UKA	Kaburi	Indonesia	<1K
42	BKN	Bukitan	Indonesia, Malaysia	<1K
43	IMR	Imroing	Indonesia	<1K
44	TGQ	Tring	Malaysia	<1K
45	TLK	Taloki	Indonesia	<1K
46	ERT	Eritai	Indonesia	<1K
47	LPE	Lepki	Indonesia	<1K
48	VME	East Masela	Indonesia	<1K
49	MXZ	Central Masela	Indonesia	<1K
50	AOS	Taikat	Indonesia	<1K
51	COG	Chong	Thailand	<1K
52	DPP	Papar	Malaysia	<1K
53	JET	Manem	Indonesia	<1K
54	KAG	Kajaman	Malaysia	<1K
55	KGI	Selangor Sign Language	Malaysia	<1K
56	KLY	Kalao	Indonesia	<1K
57	KND	Konda	Indonesia	<1K
58	KUC	Kwinsu	Indonesia	<1K
59	LVI	Lavi	Laos	<1K
60	NBN	Kuri	Indonesia	<1K
61	NER	Yahadian	Indonesia	<1K
62	ONI	Onin	Indonesia	<1K
63	ORZ	Ormu	Indonesia	<1K
64	PKT	Maleng	Laos, Vietnam	<1K
65	RTH	Ratahan	Indonesia	<1K
66	SBT	Kimki	Indonesia	<1K
67	TCM	Tanahmerah	Indonesia	<1K
68	TRT	Tunggare	Indonesia	<1K
69	WTW	Wotu	Indonesia	<1K
70	XKQ	Koroni	Indonesia	<1K
71	CWG	Cheq Wong	Malaysia	<1K
72	BPP	Kaure	Indonesia	<1K
73	ISD	Isnag	Philippines	<1K
74	PNA	Punan Bah-Biau	Malaysia	<1K
75	SKZ	Sekar	Indonesia	<1K
76	THM	Aheu	Thailand	<1K
77	TOY	Topoiyo	Indonesia	<1K
78	DBE	Dabe	Indonesia	<1K
79	BVK	Bukat	Indonesia	<1K
80	DEI	Demisa	Indonesia	<1K

Table 45: (1/3) SEA indigenous languages with <1K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
81	JEL	Yelmek	Indonesia	<1K
82	NUN	Anong	Myanmar	<1K
83	OPK	Kopkaka	Indonesia	<1K
84	PAS	Papasena	Indonesia	<1K
85	TMJ	Samarokena	Indonesia	<1K
86	URN	Uruangnirin	Indonesia	<1K
87	XAU	Kauwera	Indonesia	<1K
88	KDY	Keijar	Indonesia	<1K
89	AUU	Auye	Indonesia	<1K
90	AUW	Awyi	Indonesia	<1K
91	FLH	Foau	Indonesia	<1K
92	GOP	Yeretuar	Indonesia	<1K
93	JAU	Yaur	Indonesia	<1K
94	LHN	Lahanan	Malaysia	<1K
95	PEE	Taje	Indonesia	<1K
96	PHQ	Phana'	Laos	<1K
97	TNZ	Ten'edn	Malaysia, Thailand	<1K
98	WRU	Waru	Indonesia	<1K
99	SVE	Serili	Indonesia	<1K
100	BGV	Warkay-Bipim	Indonesia	<1K
101	BHC	Biga	Indonesia	<1K
102	BQB	Bagusa	Indonesia	<1K
103	BSA	Abinomn	Indonesia	<1K
104	CCM	Malaccan Malay Creole	Malaysia	<1K
105	GIQ	Green Gelao	Vietnam	<1K
106	KJA	Mlap	Indonesia	<1K
107	KZV	Komyandaret	Indonesia	<1K
108	MRF	Elseng	Indonesia	<1K
109	SWR	Saweru	Indonesia	<1K
110	TAD	Tause	Indonesia	<1K
111	TBP	Diebroud	Indonesia	<1K
112	TMO	Temoq	Malaysia	<1K
113	TYH	O'du	Laos, Vietnam	<1K
114	WUY	Wauyai	Indonesia	<1K
115	XWR	Kwerba Mamberamo	Indonesia	<1K
116	RMH	Murkim	Indonesia	<1K
117	TML	Tamnim Citak	Indonesia	<1K
118	WET	Perai	Indonesia	<1K
119	BQQ	Biritai	Indonesia	<1K
120	BRS	Baras	Indonesia	<1K
121	BZU	Burmeso	Indonesia	<1K
122	EMW	Emplawas	Indonesia	<1K
123	KIQ	Kosare	Indonesia	<1K
124	KIY	Kirikiri	Indonesia	<1K
125	KNS	Kensiu	Malaysia, Thailand	<1K
126	LCC	Legenyem	Indonesia	<1K
127	MSO	Mombum	Indonesia	<1K
128	MVX	Meoswar	Indonesia	<1K
129	SAO	Sause	Indonesia	<1K
130	SNU	Viid	Indonesia	<1K
131	TLG	Tofanma	Indonesia	<1K
132	KGV	Karas	Indonesia	<1K
133	LNH	Lanoh	Malaysia	<1K
134	ASZ	As	Indonesia	<1K
135	KBI	Kaptiau	Indonesia	<1K
136	MSL	Molof	Indonesia	<1K
137	WFG	Zorop	Indonesia	<1K
138	DMU	Tebi	Indonesia	<1K
139	LLK	Lelak	Malaysia	<1K
140	TCQ	Kaiy	Indonesia	<1K
141	AQN	Northern Alta	Philippines	<1K
142	BNV	Beneraf	Indonesia	<1K
143	ENC	En	Vietnam	<1K
144	ERW	Erokwanas	Indonesia	<1K
145	JBR	Jofotek-Bromnya	Indonesia	<1K
146	KHH	Kehu	Indonesia	<1K
147	KHP	Kapauri	Indonesia	<1K
148	KXN	Kanowit-Tanjong Melanau	Malaysia	<1K
149	MMB	Momina	Indonesia	<1K
150	NEC	Nedebang	Indonesia	<1K
151	NYL	Nyeu	Thailand	<1K
152	RAC	Rasawa	Indonesia	<1K
153	TNU	Tai Khang	Laos	<1K
154	WAI	Wares	Indonesia	<1K
155	YKI	Yoke	Indonesia	<1K
156	BED	Bedoanas	Indonesia	<1K
157	MZT	Mintil	Malaysia	<1K
158	AGF	Arguni	Indonesia	<1K
159	APX	Aputai	Indonesia	<1K
160	KCD	Ngkâlmpw Kanum	Indonesia	<1K

Table 46: (2/3) SEA indigenous languages with <1K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
161	UGO	Ugong	Thailand	<1K
162	WBE	Waritai	Indonesia	<1K
163	MRA	Mlabri	Laos, Thailand	<1K
164	AFZ	Obokuitai	Indonesia	<1K
165	MGF	Maklew	Indonesia	<1K
166	TTN	Towei	Indonesia	<1K
167	KNQ	Kintaq	Malaysia	<1K
168	ULF	Usku	Indonesia	<1K
169	AWH	Awbono	Indonesia	<1K
170	BTI	Burate	Indonesia	<1K
171	BYL	Bayono	Indonesia	<1K
172	DIY	Diuwe	Indonesia	<1K
173	KPI	Kofei	Indonesia	<1K
174	KRZ	Sota Kanum	Indonesia	<1K
175	KWR	Kwer	Indonesia	<1K
176	TFO	Tefaro	Indonesia	<1K
177	TKX	Tangko	Indonesia	<1K
178	TTI	Tobati	Indonesia	<1K
		Not in SEACrowd		
179	LCD	Lola	Indonesia	<1K
180	ORS	Orang Seletar	Malaysia	<1K
181	KPD	Koba	Indonesia	<1K
182	TRX	Tringgus-Sembaan Bidayuh	Malaysia	<1K
183	KQT	Klias River Kadazan	Malaysia	<1K
184	ATP	Pudtol Atta	Philippines	<1K
185	TCP	Tawr Chin	Myanmar	<1K
186	KYD	Karey	Indonesia	<1K
187	PYY	Pyen	Myanmar	<1K
188	TTW	Long Wat	Malaysia	<1K
189	XMX	Salawati	Indonesia	<1K
190	YMN	Sunum	Indonesia	<1K
191	WKD	Mo	Indonesia	<1K
192	ABF	Abai Sungai	Malaysia	<1K
193	ESY	Eskayan	Philippines	<1K
194	KZB	Kaibobo	Indonesia	<1K
195	NJS	Nisa	Indonesia	<1K
196	NNI	North Nuaulu	Indonesia	<1K
197	WHU	Wahau Kayan	Indonesia	<1K
198	XKE	Kereho	Indonesia	<1K
199	LCE	Sekak	Indonesia	<1K
200	SDX	Sibu Melanau	Malaysia	<1K
201	BFK	Ban Khor Sign Language	Thailand	<1K
202	KAX	Kao	Indonesia	<1K
203	SRK	Serudung Murut	Malaysia	<1K
204	PUD	Punan Aput	Indonesia	<1K
205	BGY	Benggoi	Indonesia	<1K
206	KZD	Kadai	Indonesia	<1K
207	KVP	Kompane	Indonesia	<1K
208	AUQ	Anus	Indonesia	<1K
209	AZT	Faire Atta	Philippines	<1K
210	HUD	Huaulu	Indonesia	<1K
211	LGH	Laghuu	Vietnam	<1K
212	TIP	Trimuris	Indonesia	<1K
213	TYJ	Tai Yo	Laos, Vietnam	<1K
214	TYS	Tày Sa Pa	Vietnam	<1K
215	MQI	Mariri	Indonesia	<1K
216	PDN	Fedan	Indonesia	<1K
217	MNQ	Minriq	Malaysia	<1K
218	DAZ	Dao	Indonesia	<1K
219	GNQ	Gana	Malaysia	<1K
220	LRN	Lorang	Indonesia	<1K
221	BSU	Bahonsuai	Indonesia	<1K
222	PUC	Punan Merap	Indonesia	<1K
223	RMX	Romam	Vietnam	<1K
224	TYL	Thu Lao	Vietnam	<1K
225	YRS	Yarsun	Indonesia	<1K
226	ATL	Mt. Iraya Agta	Philippines	<1K
227	PUF	Punan Merah	Indonesia	<1K
228	UMI	Ukit	Malaysia	<1K
229	JVD	Javindo	Indonesia	<1K
230	SRT	Sauri	Indonesia	<1K

Table 47: (3/3) SEA indigenous languages with <1K speakers.

No.	ISO 639-3	Language	Region(s)	Population
		In SEACrowd		
1	MNU	Mer	Indonesia	<100
2	ITX	Itik	Indonesia	<100
3	KXQ	Smärky Kanum	Indonesia	<100
4	LIX	Liabuku	Indonesia	<100
5	AWR	Awera	Indonesia	<100
6	BDX	Budong-Budong	Indonesia	<100
7	IRE	Yeresiam	Indonesia	<100
8	TDS	Doutai	Indonesia	<100
9	MRX	Dineor	Indonesia	<100
10	AMQ	Amahai	Indonesia	<100
11	KZU	Kayupulau	Indonesia	<100
12	MOK	Morori	Indonesia	<100
13	PLH	Paulohi	Indonesia	<100
14	SGU	Salas	Indonesia	<100
15	AIP	Burumakok	Indonesia	<100
16	DBN	Duriankere	Indonesia	<100
17	DUL	Inagta Alabat	Philippines	<100
18	MOQ	Mor	Indonesia	<100
19	NAA	Namla	Indonesia	<100
20	MVS	Massep	Indonesia	<100
21	AEM	Arem	Laos, Vietnam	<100
22	MQR	Mander	Indonesia	<100
23	XKW	Kembra	Indonesia	<100
24	KKB	Kwerisa	Indonesia	<100
25	ATZ	Arta	Philippines	<100
26	IBH	Bih	Vietnam	<100
27	KHD	Bädi Kanum	Indonesia	<100
28	NUL	Nusa Laut	Indonesia	<100
29	SCQ	Chung	Cambodia	<100
30	MQT	Mok	Myanmar, Thailand	<10
31	BTJ	Bacanese Malay	Indonesia	<10
32	WOR	Woria	Indonesia	<10
33	SPI	Saponi	Indonesia	<10
34	DSN	Dusner	Indonesia	<10
35	LGI	Lengilu	Indonesia	<10
36	BTN	Ratagnon	Philippines	<10
37	TNI	Tandia	Indonesia	<10
38	HUW	Hukumina	Indonesia	<10
39	KZL	Kayeli	Indonesia	<10
40	SXM	Samre	Cambodia, Thailand	<10
41	HPO	Hpon	Myanmar	<10
42	MPY	Mapia	Indonesia	<10
43	NIL	Nila	Indonesia	<10
44	SBO	Sabüm	Malaysia	<10
45	SRW	Serua	Indonesia	<10
46	TAS	Tay Boi	Vietnam	<10
47	XBN	Kenaboi	Malaysia	<10
48	XXT	Tambora	Indonesia	<10
		Not in SEACrowd		
49	ORN	Orang Kanaq	Malaysia	<100
50	LVA	Makuva	East Timor	<100
51	SPG	Sihan	Malaysia	<100
52	IBU	Ibu	Indonesia	<100
53	PNM	Punan Batu	Malaysia	<100
54	CSD	Chiangmai Sign Language	Thailand	<100
55	AYS	Sorsogon Ayta	Philippines	<100
56	LIO	Liki	Indonesia	<100
57	PEY	Petjo	Indonesia	<100
58	HTI	Hoti	Indonesia	<100
59	HUK	Hulung	Indonesia	<100
60	ISM	Masimasi	Indonesia	<100
61	KZX	Kamarian	Indonesia	<100
62	PNS	Ponosakan	Indonesia	<100
63	AGK	Katubung Agta	Philippines	<10
64	NAE	Naka'ela	Indonesia	<10
65	ATM	Ata	Philippines	<10
66	IHB	Iha Based Pidgin	Indonesia	<10
67	TVY	Timor Pidgin	East Timor	<10
68	DUY	Dicamay Agta	Philippines	<10
69	DYG	Villa Viciosa Agta	Philippines	<10
70	LOX	Loun	Indonesia	<10
71	ONX	Onin Based Pidgin	Indonesia	<10
72	TCL	Taman	Myanmar	<10
73	VMS	Moksela	Indonesia	<10
74	WEA	Wewaw	Myanmar	<10

Table 48: SEA indigenous languages with <100 speakers.