

Regional LLM versus Dedicated Neural MT for Waray–English: A Bidirectional FLORES-200 Evaluation of a Low-Resource Philippine Language

Richelle R. Esporas*, Mary Anne C. Sedanza

Leyte Normal University, College of Education,
Philippines

richelle.esporas, maryanne.sedanza@lnu.edu.ph

Abstract. Waray (Winaray), an Austronesian language of the eastern Philippines with about three million speakers, is computationally low-resourced and rarely evaluated in machine translation (MT). This study compares two MT approaches on Waray–English using FLORES-200: NLLB-200, a dedicated multilingual neural model that supports Waray, and Gemma-SEA-LION-v4-27B-IT, a Southeast Asian regional large language model (LLM) for which Waray is unsupported. Both were evaluated bidirectionally with spBLEU and chrF++, and a blind two-rater expert evaluation assessed the into-Waray direction. A directional asymmetry emerged: the LLM exceeded the dedicated model by 16.5 chrF++ into English but trailed by 4.5 into Waray. Yet two native experts found no significant difference in the systems' Waray output, suggesting the metric gap may largely reflect single-reference divergence rather than quality. A consistent domain split favored the LLM on news and the dedicated model on instructional text. The findings caution against surface-metric-only evaluation for low-resource generation.

Keywords. Waray, low-resource machine translation, FLORES-200, large language models, evaluation.

1 Introduction

Machine translation has advanced dramatically for high-resource languages, yet the great majority of the world's languages remain underserved. Low-resource languages—those lacking large parallel corpora, curated lexicons, and dedicated tools—pose persistent challenges for both the development and the evaluation of translation systems. Among them are many regional languages of the Philippines, a

linguistically rich archipelago whose national language technologies have largely centred on Tagalog-based Filipino, leaving major regional languages comparatively neglected.

Waray, also called Winaray, is one such language. An Austronesian language of the Central Philippine subgroup, it is spoken by approximately three million people in the Eastern Visayas. Despite this substantial speaker base, Waray has minimal presence in language technology: parallel data are scarce, few systems target it explicitly, and it rarely appears in MT evaluation. The recent emergence of two distinct classes of multilingual systems makes a careful evaluation timely. On one hand, dedicated massively multilingual neural MT models such as NLLB-200 now include Waray among their supported languages. On the other hand, instruction-tuned large language models (LLMs), including regionally specialized variants such as the SEA-LION family, have shown surprising multilingual translation ability—even for languages outside their declared support sets—raising the question of how a general-purpose regional LLM compares with a purpose-built MT system on a language like Waray.

Evaluating such systems for low-resource languages is itself difficult. The dominant automatic metrics—BLEU and its tokenization-robust variant spBLEU, and the character-level chrF++—compare system output against one or a few human references. When references are limited to a single translation, valid but lexically divergent outputs are penalized indistinguishably from errors, a problem that

is especially acute for morphologically rich, synonym-flexible low-resource languages.

Prior work on LLM translation reports that such models can be competitive for high-resource pairs while lagging on low-resource ones, but findings depend heavily on language, direction, and the evaluation instrument, and human validation for specific low-resource languages remains rare.

This study addresses that gap for Waray through a bidirectional, multi-method evaluation. Using FLORES-200, it compares a dedicated NMT model (NLLB-200) with a regional LLM (Gemma-SEA-LION-v4-27B-IT) on Waray–English translation in both directions, combining automatic metrics with a blind expert human evaluation of the lower-resource direction.

Three questions guide the work: (a) How does a dedicated multilingual NMT system perform on naturalistic, multi-domain Waray–English translation? (b) How does a regional LLM compare, and does the comparison depend on translation direction? (c) When automatic metrics and human judgments diverge, what does that reveal about evaluating low-resource generation? The study contributes one of the first systematic Waray MT evaluations on a standard benchmark, evidence of a direction-dependent complementarity between regional LLMs and dedicated MT systems, and a demonstration of the limits of single-reference surface metrics for low-resource targets.

2 Related Work

2.1 Low-Resource Neural Machine Translation

Neural machine translation has become the dominant paradigm, and recent efforts have extended it to hundreds of languages [15, 23]. The No Language Left Behind initiative scaled a single model to 200 languages and released the FLORES-200 evaluation benchmark, providing professionally translated, multi-domain test sets for many previously unevaluated languages, Waray among them [8, 15]. Surveys of low-resource MT document the persistent gap between high- and low-resource pairs and the central role of

data scarcity, transfer from related languages, and evaluation difficulty [9].

Yet research and model coverage remain heavily skewed toward a small set of high-resource languages [11], and individual low-resource languages such as Waray are rarely the focus of dedicated evaluation.

2.2 Large Language Models for Translation

General-purpose LLMs show strong but uneven translation ability across languages. Comprehensive evaluations find that GPT-class models rival commercial systems for high-resource pairs [10], while controlled studies report that such models are competitive for high- but not low-resource languages [22].

Empirical analyses of multilingual LLM translation similarly document sharp degradation as resource level falls [26]. Instruction-tuned multilingual models such as Aya [24] and regionally focused models such as SeaLLMs [14] and SEA-LION [1] aim to narrow this gap for under-served language groups, including those of Southeast Asia. How these regional models behave on a Philippine regional language outside their declared support set, however, has not been systematically established, and direction-dependent behaviour is rarely isolated.

2.3 Evaluation Metrics and Their Limits

Automatic MT evaluation relies heavily on n-gram overlap metrics—BLEU [16] and the character-level chrF and chrF++ [17, 18]—with sacreBLEU providing standardized, reproducible computation [19]. spBLEU extends BLEU with a language-agnostic SentencePiece tokenizer for fair scoring across scripts [8].

These metrics are efficient but reference-bound: with a single reference, valid paraphrases are scored as errors, a limitation widely noted for morphologically rich and low-resource languages. Neural metrics such as COMET [20] and reference-free quality estimation such as CometKiwi [21], along with LLM-based evaluators [12], partly address this but require models that may not cover low-resource languages

well. Studies of human-evaluation reliability [5] and metric meta-evaluation at WMT [6] underscore that automatic scores should be validated against human judgment, particularly where references are sparse.

2.4 Philippine and Waray Language Technology

Philippine NLP has concentrated on Tagalog-based Filipino, with comparatively little work on the country's other major languages [3].

Resources and evaluations for Waray specifically are scarce, and to our knowledge no prior study has systematically benchmarked Waray-English MT on a standard test set or compared a dedicated NMT system with a regional LLM for the language. The present study addresses this gap directly.

Table 1. NLLB-200 (Distilled 600M) on the full FLORES-200 Waray-English devtest

Direction	Domain	n	spBLEU	chrF++
Waray → English	All	1,012	28.17	48.85
	wikibooks	351	25.03	46.80
	wikinews	341	32.49	52.36
	wikivoyage	320	26.96	47.47
English → Waray	All	1,012	28.26	52.84
	wikibooks	351	27.05	52.57
	wikinews	341	29.20	53.22
	wikivoyage	320	28.63	52.75

Note. spBLEU = sacreBLEU with the FLORES-200 SentencePiece tokenizer; chrF++ = chrF with word order 2; single human references.

3 Method

3.1 Research Design

This study employed a comparative MT evaluation design combining automatic metrics with blind expert human evaluation. Two systems—a dedicated multilingual neural MT model and a Southeast Asian regional LLM—were evaluated bidirectionally on the Waray-English pair using FLORES-200.

The design proceeded in three phases: (a) full-benchmark automatic evaluation of the neural MT baseline, (b) a head-to-head automatic comparison of both systems on a stratified

random sample under identical conditions, and (c) a blind, two-rater human evaluation of the lower-resource direction to assess the validity of the automatic metrics.

3.2 Language and Evaluation Benchmark

The target language, Waray (Winaray; ISO 639-3: war), is an Austronesian language of the Central Philippine subgroup spoken by approximately three million people in the Eastern Visayas [4]. Evaluation used FLORES-200 [8, 15], which provides professionally produced human translations aligned across 200 languages.

The Waray (war_Latn) and English (eng_Latn) portions were drawn from the development-test (devtest) split: 1,012 sentence pairs spanning three domains—journalistic prose (wikinews; $n = 341$), instructional/encyclopedic text (wikibooks; $n = 351$), and travel-guide material (wikivoyage; $n = 320$). FLORES-200 supplies a single human reference per sentence and is distributed under CC BY-SA 4.0. The 997-sentence development split was reserved for prompt selection and not used for reported scoring, preserving devtest as a held-out test set.

3.3 Evaluated Systems

The dedicated baseline was NLLB-200 [15], in its distilled 600-million-parameter variant (facebook/nllb-200-distilled-600M). NLLB-200 explicitly supports Waray and is a standard open baseline for low-resource MT. The distilled 600-million-parameter checkpoint was chosen over the larger 1.3B and 3.3B variants for two reasons: it is the most widely reported open NLLB baseline in low-resource MT studies, aiding comparability, and it is the largest variant that runs comfortably on the single NVIDIA T4 GPU available to this study, so the reported baseline remains reproducible under the modest computational budgets typical of low-resource language communities. Because larger NLLB checkpoints generally score higher on FLORES-200, the dedicated-model side of the comparison should be read as a lower bound (see Section 5.5).

Inference used the Hugging Face Transformers library with greedy decoding (single beam,

no sampling), batch size 16, maximum input length 256 tokens, and a maximum of 128 newly generated tokens; the target language was enforced via the model's forced beginning-of-sequence language token. Inference ran on a single NVIDIA T4 GPU under PyTorch.

The LLM system was Gemma-SEA-LION-v4-27B-IT (aisingapore/Gemma-SEA-LION-v4-27B-IT), an instruction-tuned LLM built on the Gemma-3 27B architecture [7] and specialized for Southeast Asian languages [1], accessed through the SEA-LION API (an OpenAI-compatible endpoint) in June 2026. Crucially, Waray is not among SEA-LION's officially supported languages (which include Filipino in the sense of Tagalog); its evaluation here is exploratory. Translations were elicited zero-shot with the fixed instruction: "Translate the following:

source language

sentence into:

target language

Output ONLY the translation, with no notes, quotes, or explanation," followed by the source sentence. No in-context examples were provided. A single generation was drawn per sentence using the API's default decoding parameters; because the endpoint did not expose deterministic decoding controls, temperature and sampling were not overridden, and outputs therefore reflect realistic out-of-the-box deployment conditions rather than enforced deterministic decoding.

Because decoding was consequently non-deterministic, replication runs of the SEA-LION condition may produce slightly different outputs and correspondingly small deviations in metric scores; the figures reported here should be read as one realization of the system's out-of-the-box behaviour rather than a fixed deterministic value (revisited in Section 5.5).

3.4 Evaluation Procedure and Metrics

NLLB-200 was first evaluated on the complete devtest set in both directions, with scores

Table 2. System comparison on the stratified sample ($n = 300$ per direction)

Direction	System	spBLEU	chrF++
Waray → English	SEA-LION-v4-27B	47.54	64.88
	NLLB-200-600M	27.54	48.40
English → Waray	SEA-LION-v4-27B	21.99	48.04
	NLLB-200-600M	27.65	52.55

Note. Identical source sentences and references. SEA-LION via the SEA-LION API; NLLB run locally (greedy decoding, batch 16).

disaggregated by domain. The two systems were then compared on a stratified random sample of 300 sentences (100 per domain), drawn with a fixed random seed (42) for reproducibility and translated by both systems in both directions; sampling reflected the access constraints of the LLM API, and the close correspondence between NLLB's sample and full-corpus scores confirmed representativeness.

Both systems received identical sources and were scored against identical references. Quality was quantified with sacreBLEU [19]. spBLEU used the FLORES-200 SentencePiece tokenizer (signature nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp), and chrF++ used word order 2 (signature nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no). chrF++ served as the primary comparison metric because its character basis is more robust to morphological variation and single-reference penalties in low-resource settings [18].

Both were computed at the corpus level against the single FLORES-200 reference. Software versions and exact signatures are recorded in the supplementary materials.

3.5 Human Evaluation

A blind comparative human evaluation was conducted on a stratified subsample of 45 sentences (15 per domain) drawn from the 300-sentence sample. For each sentence, the two systems' Waray outputs were displayed side by side as "System A" and "System B," with system-to-label assignment randomized per item

Table 3. Blind two-rater expert ratings of English→Waray output ($n = 45$)

Measure	SEA-LION M (SD)	NLLB M (SD)	Diff.	p
Adequacy	3.83 (0.73)	3.88 (0.84)	-0.04	.666
Fluency	3.38 (0.89)	3.64 (0.85)	-0.27	.109

Note. Pooled across two native Waray-speaking expert raters (mean per item), blind to system identity; A/B order randomized and counterbalanced. Quadratic-weighted kappa = .44 across 180 paired judgments. Difference = SEA-LION minus NLLB. Higher = better (max = 5).

and counterbalanced across the instrument (23 vs. 22) and concealed from raters.

Two native Waray-speaking experts rated the outputs independently, without consultation.

Adequacy measured how much of the English source meaning was conveyed by the Waray (5 = all; 1 = none), judged against the source; fluency measured the naturalness and grammaticality of the Waray, judged on its own terms (5 = flawless, native; 1 = incomprehensible). Raters could also record free-text observations.

3.6 Statistical Analysis

Automatic scores were summarized at corpus and domain levels. The two raters' scores were averaged per item to yield pooled values. Inter-rater reliability used quadratic-weighted Cohen's kappa [2].

Since the rating data were ordinal and paired, system differences were tested with the Wilcoxon signed-rank test [25] on pooled item means; domain patterns were examined descriptively.

Differences in corpus-level automatic metrics between systems were tested with paired bootstrap resampling (1,000 resamples) as implemented in sacreBLEU [13, 19]. Alpha was set at .05.

4 Results

Results are reported in three stages: full-benchmark automatic evaluation of NLLB, the head-to-head comparison on the stratified sample, and the blind two-rater human evaluation of English→Waray output.

4.1 NLLB on the Full Benchmark

On the full devtest set ($N = 1,012$), NLLB achieved spBLEU 28.17 and chrF++ 48.85 (Waray→English) and spBLEU 28.26 and chrF++ 52.84 (English→Waray) (Table 1). Into English, journalistic text was easiest (chrF++ = 52.36) and instructional text hardest (chrF++ = 46.80), a 5.56-point spread; into Waray, domain differences were flat (52.57–53.22).

4.2 Head-to-Head: Regional LLM versus Dedicated NMT

On the 300-sentence sample, NLLB's scores tracked its full-corpus values within 0.7 spBLEU points, supporting representativeness. A pronounced directional asymmetry emerged (Table 2). Into English, SEA-LION exceeded NLLB by 20.00 spBLEU (47.54 vs. 27.54) and 16.48 chrF++ (64.88 vs. 48.40), in every domain. Into Waray, the order reversed: NLLB led by 5.66 spBLEU (27.65 vs. 21.99) and 4.51 chrF++ (52.55 vs. 48.04), also consistently. The regional LLM thus dominated when generating the high-resource language but trailed when generating the low-resource language.

To confirm that these metric differences exceed sampling variation, paired bootstrap resampling (1,000 resamples; sacreBLEU implementation) [13, 19] was applied to the sentence-aligned outputs of the two systems in each direction. All four contrasts were statistically significant: SEA-LION's into-English advantage and NLLB's into-Waray advantage each held for both spBLEU and chrF++ at $p < .001$, with non-overlapping 95% bootstrap confidence intervals in every case (into English, spBLEU 47.5 ± 2.4 vs. 27.6 ± 2.4 and chrF++ 64.9 ± 1.8 vs. 48.4 ± 1.9 ; into Waray, spBLEU 22.0 ± 1.5 vs. 27.7 ± 1.6 and chrF++ 48.0 ± 1.2 vs. 52.6 ± 1.3).

The directional asymmetry is therefore robust under resampling; the question is what the metrics measure, which the human evaluation addresses next.

As a reference-free complement to the surface metrics, the Waray→English outputs were additionally scored with the quality-estimation metric CometKiwi (wmt22-cometkiwi-da) [21].

The reference-free scores corroborated the surface metrics in this direction: SEA-LION attained a system-level CometKiwi of 0.677 against NLLB's 0.604, indicating that the LLM's into-English advantage is not an artifact of reference wording. CometKiwi's underlying encoder does not cover Waray, so quality-estimation scores for the English→Waray direction would not be interpretable and are not reported; this coverage gap is itself symptomatic of the evaluation problem the study documents (Section 5.2).

4.3 Expert Human Evaluation (English → Waray)

Two native Waray experts rated 45 sentences blind (180 paired judgments each). Inter-rater reliability (quadratic-weighted kappa) was .44—moderate—with raters within one scale point on 82.8% of judgments; the second rater was systematically more lenient (e.g., SEA-LION adequacy 4.22 vs. 3.44), an offset weighted kappa captures. Pooled across raters, the systems were statistically indistinguishable (Table 3): adequacy 3.83 vs. 3.88, $W = 306.0$, $p = .666$; fluency 3.38 vs. 3.64, $W = 308.0$, $p = .109$.

These human gaps (0.04 and 0.27 points, favoring NLLB) are small against the 4.51-point chrF++ and 5.66-point spBLEU advantages the metrics assigned to NLLB in the same direction, suggesting that much of the metric gap may reflect reference-wording divergence rather than human-perceived quality.

Beneath this equivalence lay a domain split reproduced by both raters: on instructional wikibooks text NLLB was clearly preferred (fluency 3.90 vs. 2.90; adequacy 3.97 vs. 3.50), while on journalistic wikinews text SEA-LION led (adequacy 4.07 vs. 3.67; fluency 3.80 vs. 3.53); wikivoyage

text was rated equivalently. The aggregate "tie" thus masks two systems excelling in different domains, with the LLM weakest exactly where automatic scores were lowest.

4.4 Qualitative Observations

SEA-LION generated genuine Waray, not Tagalog: characteristic Waray function-word rates in its output (37.6%) matched the references (39.8%), with Tagalog markers no more frequent than in the references.

Much divergence was valid lexical variation—e.g., *katapos nga bahin* for the reference's *urhi nga parte*, or *langgam* for *katamsihan* ("birds")—penalized by single-reference scoring as if erroneous. Occasional cross-Visayan leakage appeared (Cebuano *karon*, "now," for Waray *yana*). Both systems faltered on low-frequency items (for *yatot*, "mice," NLLB produced *bug-os*, SEA-LION *irat*), and NLLB produced confident lexical errors into English (*ika-siyam*, "ninth," as "seventh"). Notably, SEA-LION's English stayed fluent even when wrong—a fluently-incorrect profile with deployment implications.

The moderate inter-rater reliability ($\kappa = .44$) is also informative at item level. Disagreements clustered where fluency and adequacy pulled apart, and where the raters weighted cross-Visayan lexicon differently. Three examples illustrate the pattern. On a wikibooks item ("Many paleontologists today believe that one group of dinosaurs survived... We call them birds"), NLLB omitted the second sentence entirely while SEA-LION rendered it with Cebuano-influenced forms (*Tinatawag kita hira nga mga langgam*); one rater penalized the borrowing heavily (adequacy 2, fluency 2) while the other, crediting the preserved meaning, scored it 5 and 4, noting the output "has some Cebuano forms but keeps the meaning."

On a wikivoyage item about box jellyfish, the raters split in the opposite direction over NLLB's output, which dropped the final source sentence but otherwise read smoothly: one rater scored its fluency 2, the other 5, disagreeing over how far an omission should bleed into the fluency judgment.

And on a wikinews item ("a plane to be hijacked"), where both systems chose imperfect verbs (NLLB *gin-kidnap*; SEA-LION *ginkawat*, "stolen"), one rater scored both systems 2 on both dimensions while the other gave 3–4s—a global severity offset rather than a preference reversal, consistent with the leniency difference that weighted kappa captures.

These cases suggest the raters at times applied different implicit weightings of naturalness against completeness of meaning transfer, a known source of variance in adequacy–fluency protocols [5], and they reinforce the need for error-type annotation in future panels (Section 5.6).

5 Discussion

Three findings stand out: a direction-dependent asymmetry between the systems, a divergence between automatic metrics and human judgment when generating Waray, and a consistent domain-level complementarity.

5.1 A Direction-Dependent Asymmetry

System superiority depended on direction. Translating Waray into English, the regional LLM was decisively better; generating Waray, the dedicated model led.

This is coherent with each system's design: the LLM's vast English generation capacity plus sufficient regional pretraining to comprehend Waray makes it excellent at producing fluent English, whereas generating well-formed text in an unsupported low-resource language is harder for it than for a model trained with explicit Waray supervision and a constrained decoder.

For low-resource communities, the best tool may therefore differ by direction—LLMs for understanding and into-English tasks, dedicated MT for generating the local language.

5.2 The Validity of Automatic Metrics for Low-Resource Generation

The most consequential finding is the disagreement between metrics and human judgment into Waray. Surface metrics awarded NLLB a clear edge, yet two native-speaker experts found no significant difference. Qualitative analysis explains why: much of the LLM's deficit stemmed from valid lexical and morphological variation that single-reference scoring treats as error.

The metrics are not useless—they tracked human judgment well into English—but single-reference spBLEU and chrF++ can understate fluent, variation-rich systems when the target is a morphologically flexible low-resource language. Such scores should be treated as provisional and validated against human judgment or multiple references, or supplemented with neural and reference-free metrics [20, 21], before strong superiority claims are made.

5.3 Domain-Dependent Complementarity

Beneath the aggregate human "tie" lay a robust domain split: the dedicated model was preferred on instructional text, the LLM on journalistic text. Procedural prose, with fixed terminology and conventional phrasing, appears to reward reference-faithful supervised output, while the freer register of news suits the LLM's fluency.

Domain, not just direction, should inform system choice, and domain-routed or hybrid deployment could outperform either system alone.

5.4 Fluent Incorrectness and Deployment Risk

The LLM tended to remain fluent even when wrong, whereas the dedicated model's errors were often visibly disfluent and thus self-signalling. In low-resource contexts, where users may lack a second system or proficient reviewer, fluent incorrectness is hazardous.

This argues for caution in unsupervised LLM deployment for low-resource languages and for evaluation that probes adequacy explicitly rather than treating fluency as a proxy for correctness.

5.5 Limitations

Evaluation relied on FLORES-200's single reference; the metric–human divergence would be mitigated by multi-reference evaluation. The LLM was queried with a single generation under default, non-deterministic decoding, reflecting realistic deployment but introducing minor run-to-run variability that multiple sampled generations would quantify.

The human evaluation covered 45 sentences with moderate inter-rater reliability (weighted kappa = .44), reflecting the subjectivity of adequacy and fluency judgments; larger, calibrated panels would strengthen it, though the central null held across both raters with 82.8% within-one-point concordance. Only one variant of each system was tested (distilled 600M NLLB; one SEA-LION version), so results may differ for larger checkpoints or other LLMs. SEA-LION's exploratory status for Waray means its behaviour reflects incidental rather than targeted competence.

Finally, neural and quality-estimation metrics could be applied only where their encoders cover the language involved; the absence of Waray from such models' coverage restricts reference-free validation of the into-Waray direction and is itself a resource gap future metric development should close.

5.6 Future Directions

Future work should add multi-reference and neural or reference-free metrics (e.g., COMET, CometKiwi [20, 21]) better suited to morphologically rich low-resource languages; broaden the system set to additional sizes and LLMs; enlarge the human panel with error-type annotation; and probe few-shot and fine-tuning strategies for improving LLM generation of Waray.

More broadly, the demonstrated complementarity invites combining the comprehension strengths of regional LLMs with the generation fidelity of dedicated MT for Philippine and other low-resource languages.

6 Conclusion

This study provides one of the first systematic, bidirectional evaluations of machine translation for Waray, comparing a dedicated multilingual NMT model with a Southeast Asian regional LLM on FLORES-200.

The results reveal a clear direction-dependent asymmetry—the LLM superior into English, the dedicated model superior into Waray on automatic metrics—together with a critical caveat: blind expert evaluation found no real difference in the systems' Waray output, exposing how single-reference surface metrics can understate a fluent system's quality for a low-resource target.

A consistent domain split further showed the two approaches excelling on different text types. For Waray and languages like it, these findings recommend choosing tools by direction and domain, validating automatic scores against human judgment, and treating fluent low-resource output with caution. Beyond Waray, the study offers a replicable template—full-benchmark metrics, controlled comparison, and blind expert evaluation—for honest assessment of MT in low-resource settings.

Ethical and Deployment Considerations

This study used only publicly available, professionally translated benchmark data (FLORES-200, CC BY-SA 4.0) and involved no personal or sensitive data; the two expert raters participated voluntarily.

The findings carry a deployment caveat. Because both systems produced usable but imperfect Waray and the LLM in particular can be fluently incorrect, machine translation should not be used for legal, medical, educational, governmental, or other high-stakes communication in Waray or comparable low-resource languages without review by a competent native speaker.

Responsible deployment should pair automatic translation with human oversight, communicate uncertainty to users, and involve language communities in evaluating and governing the tools built for their languages.

Data Availability

The evaluation corpus (FLORES-200 Waray–English devtest with domain metadata), the complete bidirectional system outputs, the stratified sample indices, the blind human-evaluation instrument and ratings, and all scoring scripts are available as supplementary materials. FLORES-200 is distributed under CC BY-SA 4.0.

System outputs generated through the SEA-LION API are shared in accordance with the provider's terms of use; should any redistribution restriction apply, those outputs are available from the corresponding author on reasonable request.

References

1. **AI Singapore (2024).** SEA-LION: Southeast Asian Languages in One Network [large language model]. <https://sea-lion.ai>.
2. **Cohen, J. (1968).** Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, Vol. 70, No. 4, pp. 213–220. DOI: 10.1037/h0026256.
3. **textbfCruz, J. C. B., Cheng, C. (2020).** Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*.
4. **Eberhard, D. M., Simons, G. F., Fennig, C. D. (2024).** *Ethnologue: Languages of the World*, 27th ed. SIL International, Dallas. <https://www.ethnologue.com>.
5. **Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W. (2021).** Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1460–1474.
6. **Freitag, M., Rei, R., Mathur, N., Lo, C.-K., Stewart, C., Avramidis, E., et al. (2022).** Results of WMT22 metrics shared task. *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, pp. 46–68.
7. **Gemma Team (2024).** Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.
8. **Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., et al. (2022).** The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 522–538.
9. **Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., Birch, A. (2022).** Survey of low-resource machine translation. *Computational Linguistics*, Vol. 48, No. 3, pp. 673–732.
10. **Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., et al. (2023).** How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.
11. **Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M. (2020).** The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293.
12. **Kocmi, T., Federmann, C. (2023).** Large language models are state-of-the-art evaluators of translation quality. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203.
13. **Koehn, P. (2004).** Statistical significance tests for machine translation evaluation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 388–395.

14. **Nguyen, X.-P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., et al. (2023).** SeaLLMs: Large language models for Southeast Asian languages. arXiv preprint arXiv:2312.00738.
15. **NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., et al. (2024).** Scaling neural machine translation to 200 languages. *Nature*, Vol. 630, No. 8018, pp. 841–846. DOI: 10.1038/s41586-024-07335-x.
16. **Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002).** BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.
17. **Popović, M. (2015).** chrF: Character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395.
18. **Popović, M. (2017).** chrF++: Words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, pp. 612–618.
19. **Post, M. (2018).** A call for clarity in reporting BLEU scores. *Proceedings of the Third Conference on Machine Translation*, pp. 186–191.
20. **Rei, R., Stewart, C., Farinha, A. C., Lavie, A. (2020).** COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702.
21. **Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., et al. (2022).** CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, pp. 634–645.
22. **Robinson, N., Ogayo, P., Mortensen, D. R., Neubig, G. (2023).** ChatGPT MT: Competitive for high- (but not low-) resource languages. *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, pp. 392–418.
23. **Stahlberg, F. (2020).** Neural machine translation: A review. *Journal of Artificial Intelligence Research*, Vol. 69, pp. 343–418. DOI: 10.1613/jair.1.12007.
24. **Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., et al. (2024).** Aya model: An instruction finetuned open-access multilingual language model. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15894–15939.
25. **Wilcoxon, F. (1945).** Individual comparisons by ranking methods. *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83. DOI: 10.2307/3001968.
26. **Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., et al. (2024).** Multilingual machine translation with large language models: Empirical results and analysis. *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781.

Article received on 18/06/2026; accepted on 17/09/2026.

**Corresponding author is Richelle R. Esporas.*