Languages Still Left Behind: Toward a Better Multilingual Machine Translation Benchmark

Anonymous ACL submission

Abstract

Multilingual machine translation (MT) benchmarks play a central role in evaluating the capabilities of modern MT systems. Among them, the FLORES+ benchmark is widely used, offering English-to-many translation data for over 200 languages, curated with strict quality control protocols. However, we study data in four languages (Asante Twi, Japanese, Jinghpaw, and South Azerbaijani) and uncover critical shortcomings in the benchmark's suitability for truly multilingual evaluation. Human assessments reveal that many translations fall below the claimed 90% quality standard, and the annotators report that source sentences are often too domain-specific and culturally biased toward the English-speaking world. We further demonstrate that simple heuristics, such as copying named entities, can yield non-trivial BLEU scores, suggesting vulnerabilities in the evaluation protocol. Notably, we show that MT models trained on high-quality, naturalistic data perform poorly on FLORES+ while achieving significant gains on our domain-relevant evaluation set. Based on these findings, we advocate for multilingual MT benchmarks that use domain-general, named-entity-agnostic, and culturally neutral source texts to better reflect real-world translation challenges.

1 Introduction

Recent advances in natural language processing (NLP) have been increasingly multilingual in nature, and machine translation (MT) is no exception. To assess the multilingual capabilities of MT models, several evaluation benchmarks have been proposed, among which the FLORES+ benchmark (Guzmán et al., 2019; Goyal et al., 2022) has emerged as one of the most prominent. FLORES+ provides a large-scale English-to-many translation dataset covering over 200 languages. The creation process of the dataset adheres to strict quality control standards, including rigorous qualification requirements for both translators and reviewers, and the benchmark claims to maintain a translation quality score of over 90%.

In this paper, we carry out a re-evaluation of the FLORES+ data in four languages (Asante Twi, Japanese, Jinghpaw, and South Azerbaijani). Our study reveals significant concerns regarding the adequacy and fluency of the benchmark for truly multilingual MT evaluation. We observe that the translation quality in the FLORES+ dataset often falls short of the reported 90% threshold. Feedback from evaluators further highlights that many of the original English source sentences are overly domain-specific and frequently contain jargon or culturally specific expressions that are difficult or impossible to translate naturally into the target languages.

Moreover, we identify structural issues in the benchmark itself. Through a simple experiment, we show that models can achieve the average BLEU score of 0.29 merely by copying named entities (NEs) from the source sentence, calling into question the benchmark's robustness. More critically, our experiments on Jinghpaw MT demonstrate that models trained on naturalistic, high-quality translation data underperform on the FLORES+ benchmark but significantly outperform on a separate, out-of-domain evaluation set that we construct. These findings point to deep-rooted limitations in current benchmark design.

To address these shortcomings, we propose three essential directions for future multilingual MT benchmarks: (1) source sentences should be less domain-specific and technical; (2) benchmarks should minimize the influence of named entities in evaluation; and (3) the content should avoid being overly centered on English-speaking cultural and linguistic norms. These principles are crucial for the fair and meaningful MT evaluation across the

[*] Equal contribution.

world's languages.

2 Related work

A growing number of multilingual MT benchmarks have been proposed to assess the capabilities of modern MT systems across a wide range of languages. Among them, the FLoRes series of benchmarks have become particularly influential in recent years.

The original Flores benchmark (Guzmán et al., 2019) was introduced to evaluate low-resource MT, specifically targeting Nepali–English and Sinhala–English translation. It consisted of professionally translated Wikipedia sentences and aimed to highlight the performance gap between high- and low-resource languages. Despite its narrow language scope, this work played an important role in initiating systematic evaluation for low-resource MT.

Flores-101 (Goyal et al., 2022) extended this effort significantly by providing human-translated data for 3,001 English sentences in 101 languages, enabling over 10,000 translation directions. The source sentences are curated to be multi-domain (WikiNews,[1] WikiJunior,[2] and WikiVoyage[3]) and multi-topic (travel, politics, science, and crime, among others). The benchmark follows a rigorous quality control process involving multiple rounds of professional translation and review. It standardized multilingual MT evaluation by offering parallel test sets that are fully bitext-aligned, allowing direct comparison across language pairs.

FLORES-200 (NLLB Team et al., 2022) further expanded the benchmark to 200 languages, forming the backbone of the No Language Left Behind (NLLB) project (NLLB Team, 2024). With over 40,000 translation directions, it represents the most comprehensive multilingual MT test suite to date. The benchmark is currently managed and developed under the name of FLORES+,[4] available under the CC-BY-SA-4.0 license.

Recent work has aimed to further improve the quality and coverage of FLORES benchmarks. Abdulmumin et al. (2024) highlight systemic issues in the FLORES evaluation sets for four African languages, proposing corrections that enhance fluency and adequacy for more reliable MT evaluation. In parallel, several efforts have sought to extend the benchmark to additional underrepresented languages. For instance, Gordeev et al. (2024) introduce FLORES+ for Erzya, a severely low-resource Uralic language; Kuzhuget et al. (2024) augment FLORES with a high-quality Tuvan test set; Perez-Ortiz et al. (2024) expand coverage for Iberian languages; Ali et al. (2024) contribute new evaluation data for Emakhuwa, a Bantu language spoken in Mozambique; and Yankovskaya et al. (2023) focus on Finno–Ugric languages, addressing both translation quality and data availability.

[1] https://en.wikinews.org
[2] https://en.wikibooks.org/wiki/Wikijunior
[3] https://en.wikivoyage.org
[4] https://huggingface.co/datasets/openlanguagedata/flores_plus

3 Human re-evaluation

To investigate the translation quality of the FLORES+ benchmark, we manually re-evaluated the translation quality of the four genealogically, orthographically, and geographically diverse languages: Asante Twi, Japanese, Jinghpaw, and South Azerbaijani. See Table 1 for details of the target languages.

3.1 Setup

For each language, we recruited a native speaker who is also fluent in English to serve as an annotator. All of the annotators have had higher education in English and are currently affiliated with a university. Their academic backgrounds are linguistics (Japanese and South Azerbaijani), education (Jinghpaw), and biological sciences (Asante Twi). They have also had either professional or academic experience of translation involving their target language. Each annotator is given pairs of English and the target language from the development set of the FLORES+ dataset corresponding to their language. Their task is to assess the adequacy and fluency of each translation by comparing the English source sentence with its translation in the target language. As we proceeded with the assessment, the manual evaluation and correction turned out to be time-consuming for the annotators due to a large number of errors. For this reason, we restrict the examined sentences to be the first 50 sentences of the dev set for this study. The dataset version examined in this study is version 2.0 released on November 16, 2024.

Annotation guidelines. As per the original assessment guidelines in Flores-101 (Goyal et al., 2022), the annotators are instructed to categorize each sentence pair using the following labels: Correct, Wrong grammar, Wrong punctuation, Wrong spelling, Wrong capitalization, Inaccurate addi-

Language	Code	Language family	Writing system	Region
Asante Twi	twi_Latn_asan1239	Akan < Kwa	Latin	West Africa
Japanese	jpn_Jpan	Japonic	Hiragana, Katakana, Kanji	East Asia
Jinghpaw	kac_Latn	Tibeto–Burman < Sino–Tibetan	Latin	Southeast Asia
South Azerbaijani	azb_Arab	Oghuz < Turkic	Perso–Arabic	West Asia

Table 1: A list of the target languages and their characteristics. *Code* refers to the corresponding language code in the Flores dataset.

tion, Inaccurate omission, Mistranslation, Unnatural translation, Untranslated text, and Wrong register. Multiple labels can be selected per sentence to capture overlapping or compound issues. When the annotators identify at least one error in a sentence, they are asked to categorize the severity of the error(s) into Critical errors, Major errors, and Minor errors. The criteria for the error severity categorization also follow the guidelines in the original work (Goyal et al., 2022). In addition, annotators are optionally invited to provide their own translation as well as free-form comments explaining their judgments or highlighting specific concerns about the translation. This manual evaluation setup aims to identify not only obvious translation errors but also subtler issues such as fluency, naturalness, and contextual appropriateness, which automatic metrics often fail to capture.

Metric. The original work (Goyal et al., 2022) uses a metric called *Translation Quality Score* (TQS) to evaluate the translation quality; however, the detailed definition of this metric is not explained in their work. For this reason, we tentatively define TQS as follows:

$$\text{TQS} = \frac{3 \cdot C + 2 \cdot E_m + 1 \cdot E_M + 0 \cdot E_c}{3 \cdot (C + E_m + E_M + E_c)}$$

where C is the number of correct sentences, E_m the number of sentences identified as Minor errors, E_M the number of sentences identified as Major errors, and E_c the number of sentences identified as Critical errors. We also employ a modified version of Multidimensional Quality Metrics (MQM) (Lommel et al., 2013):

$$\text{TQS}_{\text{MQM}} = 1 - \frac{e_m + 5 \cdot e_M + 10 \cdot e_c}{W},$$

where e_m is the number of Minor errors, e_M the number of Major errors, e_c the number of Critical errors, and W the total word count. TQS_{MQM} generally yields higher scores than TQS, as it normalizes over word count, reducing the impact of errors.

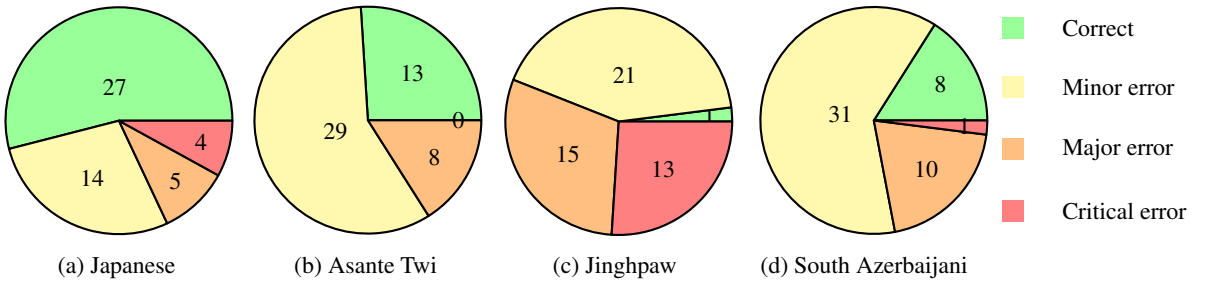
3.2 Results

Category	twi	jpn	kac	azb
Correct	27	34	1	12
Wrong grammar	4	0	13	3
Wrong punctuation	1	0	0	6
Wrong spelling	5	7	10	29
Wrong capitalization	0	0	1	0
Inaccurately added information	4	1	0	0
Inaccurately omitted information	3	2	20	5
Mistranslation	6	4	22	4
Unnatural translation	8	4	22	3
Untranslated text	1	25	9	1
Wrong register	1	0	1	0
Other	0	1	3	3

Table 2: Assessment statistics. The language codes correspond to the first three letters of the codes in Table 1. Multiple category selection is allowed for sentences that are evaluated not to be correct.

Table 2 summarizes the results of the human re-evaluation. Surprisingly, there was only one sentence in Jinghpaw that was evaluated to be correct. Figure 1 illustrates the error severities across the target languages. In terms of the error severities, while Asante Twi contains no critical errors, 4 critical errors were found in Japanese, 13 in Jinghpaw, and 1 in South Azerbaijani. Critical errors are "issues that render the content unfit for use" due to the original meaning seriously distorted "in such a way that it becomes incomprehensible or that the essence of the message is lost" (Goyal et al., 2022). Table 3 shows examples of the sentences identified as containing critical errors in Japanese and Jinghpaw. The Japanese sentence contains a number of errors such as *Inaccurate addition of information*, *Grammatical error*, *Untranslated text*, and *Mistranslation*, leading to the loss of the original meaning. The Jinghpaw translation contains a critical amount of *Inaccurate omission of information* and overall *Mistranslation* to the extent that it becomes incomprehensible.

Table 4 summarizes the TQS and TQS_{MQM} computed from the re-evaluation, as well as the character error rate (CER) and translation edit rate



(a) Japanese (b) Asante Twi (c) Jinghpaw (d) South Azerbaijani

Figure 1: Judgement distribution by error type for the four languages. Each pie chart shows the number of Correct, Minor error, Major error, and Critical error annotations.

English	Original	Backtranslation	Corrected
Hsieh also argued that the photogenic Ma was more style than substance.	(Japanese) 謝長廷は、馬英九について実物よりも写真写りが良いと主張していました。	Hsieh Chang-ting[Addition] was[Grammar] also[Untranslated] arguing[Grammar] that Ma Ying-jeou[Addition] was more photogenic than the actual appearance[Mistranslation].	シエ氏はまた、写真写りの良いマー氏は、見た目ばかりで中身が伴っていないとも主張しました。
At 11:20, the police asked the protesters to move back on to the pavement, stating that they needed to balance the right to protest with the traffic building up.	(Jinghpaw) Aten 11:20 hta dai masha ni hpe lam makau de sit na matu tsun wa ai, dai lam pat ai nta ni hpe atsawm lajang na matu ngu shana wa ai.	At 11:20, they told the people to move sideways, they informed them to organize the houses that blocked the way[Omission, Mistranslation].	Aten 11:20 hta, pyada ni hku na, dai ninghkap masha ni hpe lam makau de sit na matu tsun mat let, ninghkap na matu ahkaw ahkang nga ai hte maren, hkawm sa hkawm wa lam pat wa ai lam hpe mung, joi jang ra na matu shana wa ai.

Table 3: Examples of critical errors found in the Japanese and Jinghpaw sets. The English source sentences are in the *English* column, the original translations in the *Original* column. For readers' convenience, the backtranslation from the original translation into English is added in the *Backtranslation* column. The identified errors are highlighted with a red color with the error tags in square brackets. The *Corrected* column provides the reference translation corrected by the annotators.

Language	CER	TER	TQS	TQS_{MQM}
Jinghpaw	45.87	51.46	40.00	70.89
Asante Twi	15.66	18.84	70.00	94.07
Japanese	13.23	23.27	76.00	83.38
South Azerbaijani	17.32	35.03	64.00	85.83

Table 4: CER and TER between the reference sentences and the sentences corrected by the assessors, as well as TQS and TQS_{MQM}. All values are in percent.

(TER) between the gold translation sentences in FLORES+ and the sentences corrected by the annotators. TQS was substantially below the 90% threshold for all the languages we examined. For TQS_{MQM}, only Asante Twi exceeded 90%, while the other three languages remained below the threshold. In the worst case, the TER in the Jinghpaw set had more than 50% TER.

3.3 Feedback from the annotators

In addition to the quantitative evaluation, we collected qualitative feedback from human reevaluators of the translated sentences. Several recurring issues were observed in the evaluation across the target languages.

Named entities. The annotators commonly identified errors and unnaturalness regarding the translation and transliteration of NEs. For example, "Aldwych" (a toponym in London) is transliterated as アルドウィック /arudowikku/ in the benchmark translation, whereas it is commonly transliterated as オールドウィッチ /oorudowicchi/.

Domain-specific concepts. While FLORES+ is a multi-domain dataset covering wide variety of topics, the annotators often expressed concerns that some sentences are too domain-specific involving specialized vocabulary from various fields. These terms were sometimes unfamiliar even to fluent speakers. For example, several sport-related concepts caused a confusion among the annotators. The annotator for Jinghpaw noted that there is no widely accepted term in Jinghpaw for "net point"

used in tennis. The annotator for Twi noted that they did not know how to translate "season" in the context of sports. The annotator for Japanese reported not knowing the term related to auto racing コ・ドライバー *ko-doraibaa* (co-driver).

Grammar-specific issues. Japanese has an honorific register that is used when the speaker conveys politeness, respect, or humbleness. From its nature, the honorific register typically presupposes a hearer and is often used in spoken settings such as conversations and TV news. In contrast, the neutral register is primarily used in a written form, such as books and newspaper. The Japanese translations in FLORES+ were consistently rendered in the polite register, despite the fact that the sources of the sentences are from written media such as WikiNews. In addition, the annotators for Asante Twi and Jinghpaw noted that some long English sentences containing multiple clauses with syntactic coordinations and relativizations were difficult to translate directly. To maintain fluency, they often had to split these into multiple sentences.

Cultural issues. For Jinghpaw, fundamental lexical gaps from English were evident. The annotator noted the absence of native equivalents for certain concepts such as "spring" (as in four seasons) in the language because the region where Jinghpaw is spoken (Kachin State, Myanmar) falls into a (sub)tropical climate that does not have four seasons.

These limitations underscore the difficulty of translating domain-specific, technical English sentences into languages with different grammatical, lexical and conceptual systems and different cultural backgrounds.

4 Experiments

In this section, we conduct two experiments to address the issues of the benchmark. Based on the findings in the human re-evaluation that the sentences frequently contain NEs, the first experiment tests whether synthetically generated translations made of NEs copied from the English source sentences yield non-zero evaluation scores. In the second experiment, we evaluate MT models with FLORES+ and our in-house evaluation datasets and examine the implications of using the benchmark for evaluating low-resource MT models. The FLORES+ version used in the experiments is version 2.0 released on November 16, 2024. The in-house evaluation datasets will be made publicly available.

4.1 Named-entity copying

The objective of this experiment is to assess the extent to which NEs contribute to the evaluation scores in the FLORES+ benchmark. Our motivation stems from the observation that if the evaluation sentences contain a significant number of NEs, MT systems that merely copy those entities could achieve artificially inflated scores, thereby undermining the benchmark's ability to reflect true translation quality.

4.1.1 Setup

To simulate this scenario, we constructed a set of synthetic translations for each target language. For each source sentence, we extracted the NEs and used them as the "translated" output. To bypass the brevity penalty in BLEU (see below), we appended the string "dummy" 50 times after the extracted NEs. We used gpt-4o to identify the named entities.

We then computed the BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2015) scores of these synthetic translations against the official FLORES+ references in the languages with Latin-based orthographies. The BLEU score is calculated as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where BP is the brevity penalty, N the maximum n for counting n-grams, w_n the weight for each n-gram, and p_n the modified n-gram precision. The brevity penalty lowers the BLEU score when the hypothesis length c is shorter than the reference length c: $\text{BP} = \min(1, e^{1-\frac{r}{c}})$. The modified n-gram precision counts the number of matching n-grams between the two sentences; $p_n = 0$ when the predicted translation is empty. While BLEU only considers word-level matches, ChrF++ takes both word-level and character-level matches in account:

$$\text{ChrF++} = \frac{(1+\beta^2)P \cdot R}{\beta^2 P + R},$$

where P is the averaged character- and word-level precisions, and R is the averaged character- and word-level recalls. In the common implementation, $\beta = 2$, and the precision and recall considers 1 to 6-gram matches at the character level and 1 to 2-gram matches at the word level.

In our experiments, we consider two scenarios of the state of hypothesis strings: (1) the hypothesis

	Train	Dev	Test	Total
PARADISEC (P)	38113	755	743	39611
Dictionary (D)	3151	0	0	3151
Dialogue	0	0	222	222
NLLB (N)	121081	0	0	121081
FLORES+	0	997	1011	2008
Total	162345	1752	1976	166073

Table 5: The number of parallel sentences contained in the datasets used in this experiment. Note that we use the *Devtest* set of FLORES+ as the test set.

string is empty (*i.e.*, no NE is detected in the source sentence), and (2) the hypothesis string consists of at least one matching word (*i.e.*, NE). In case of (1), the BLEU and ChrF++ scores of the synthetic sentence is 0, while (2) yields BLEU and ChrF++ scores larger than 0. Under the assumption that the core objective of a translation task is not to just copy certain words but to test the ability of mapping the unique grammar and lexicon from one language to the other, the BLEU scores of those synthetic sentences in an multilingual MT evaluation benchmark should ideally be 0.

4.1.2 Results

As the results in Figure 2 demonstrate, even this trivial NE-copying baseline achieved non-zero scores across all the languages, indicating that the benchmark may reward surface-level lexical overlap rather than meaningful translation quality. These results call into question the adequacy of current evaluation practices in multilingual MT with FLORES+ because MT systems that are good at copying NEs yield a better score than those that are not.

4.2 Jinghpaw MT

As the second experiment, we train an MT model for Jinghpaw–English translation based on the pretrained NLLB models (NLLB Team et al., 2022) and see the implication of using the benchmark.

4.2.1 Setup

Datasets. The datasets used in the experiments include Jinghpaw narrative data from PARADISEC (Kurabe, 2013, 2017), example sentences listed in the *Jinghpaw Usage Dictionary* (Kurabe, 2020a), dialogues from the *Jinghpaw Reader* (Kurabe, 2020b), the Jinghpaw–English parallel sentences from the NLLB dataset,[5] and the FLORES+ benchmark; see Table 5 for details. Among these, due to the small number of samples, all the Dictionary sentences are used in the training data, and all the Dialogue sentences are used in the test data. Additionally, because the quality of the NLLB dataset is not guaranteed, we remove duplicate sentence pairs and sentence pairs with significantly mismatched lengths between the two languages based on the method used by Deguchi et al. (2023) and use the remaining data in the training data. The NLLB dataset originally contained 1,003,100 sentence pairs, which was reduced to 121,081 after filtering. Since FLORES+ is intended for evaluation, we use the dev set as validation data and the devtest set as test data and do not use it at all for training.

Model training. In this experiment, we set the baselines as the pretrained multilingual MT models NLLB-600M[6] and NLLB-1.3B[7] and fine-tune them for Jinghpaw↔English MT using the different sets of training data described in Table 5. We perform fine-tuning separately for both translation directions: from Jinghpaw to English and from English to Jinghpaw. The training objective is to minimize the cross-entropy loss, as in the original NLLB models.

For all model training runs, the batch size is set to 8, the learning rate to 0.0001, the number of warm-up steps to 500, the learning rate decay factor per epoch to 0.9, and the maximum number of epochs to 20. Other hyperparameters follow the default settings of the NLLB models. All model fine-tuning is conducted using a single A10 GPU.

For evaluation, we employ BLEU and ChrF++ as the evaluation metrics. After the training, the checkpoint with the highest BLEU score on the validation set is picked for evaluation. As the evaluation datasets, we use the test set of PARADISEC, the devtest set of FLORES+, and the Dialogue dataset.

4.2.2 Results

Results in Table 6b present the BLEU and ChrF++ scores for both translation directions: Jinghpaw→English and English→Jinghpaw. We compare the pretrained NLLB models (600M and 1.3B) against their fine-tuned variants using various subsets of our collected datasets: the Dictionary dataset (D), the PARADISEC dataset (P), and the filtered NLLB dataset (N).

[5] https://opus.nlpl.eu/NLLB/corpus/version/NLLB

[6] https://huggingface.co/facebook/nllb-200-distilled-600M

[7] https://huggingface.co/facebook/nllb-200-1.3B

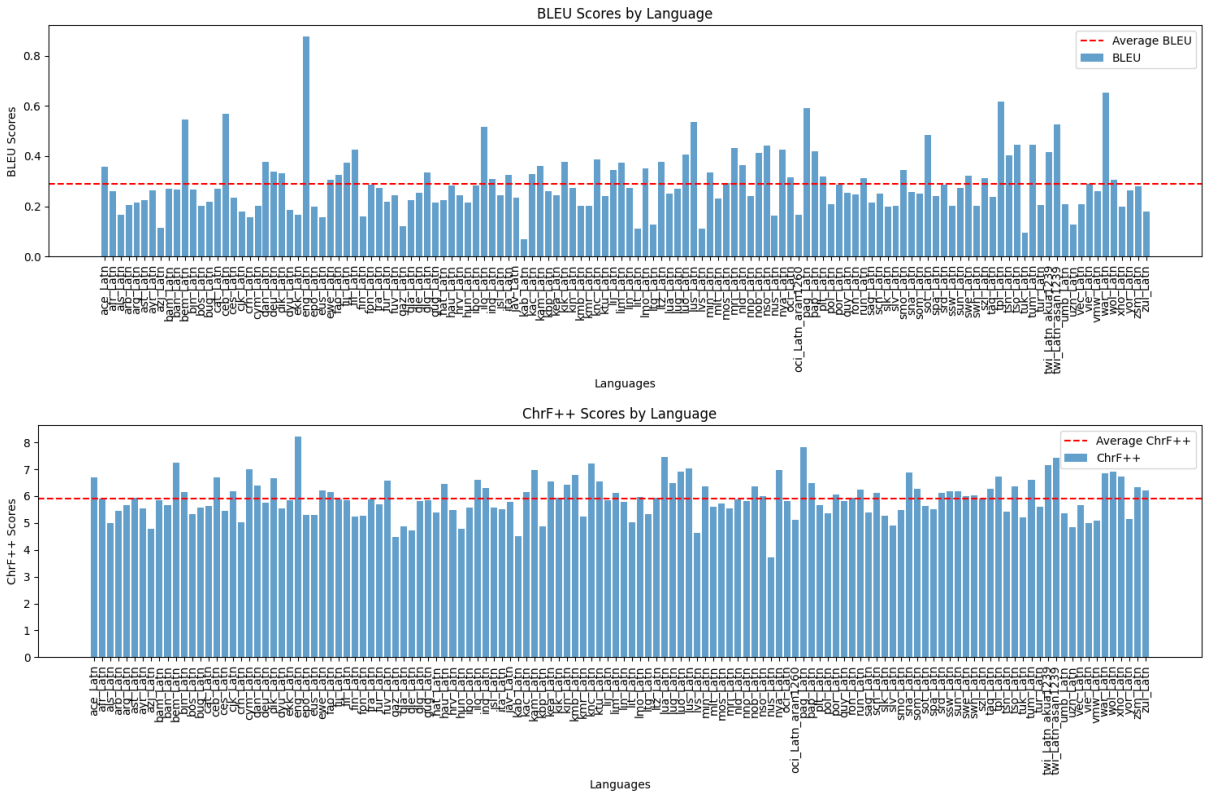


Figure 2: BLEU scores (top) and ChrF++ scores (bottom) of dummy translations with named entities copied from the source sentence. The tested languages are those with a Latin-based orthography.

Across both translation directions, we observe that the pretrained NLLB models achieve the highest scores on the FLORES+ set, particularly for the larger NLLB-1.3B model. However, their performance degrades substantially on more natural and simpler sentences from the PARADISEC and Dialogue test sets. For example, in the Jinghpaw→English setting, the NLLB-1.3B baseline achieves a BLEU score of only 2.29 on PARADISEC data, despite scoring 13.95 on FLORES+. This suggests that these models are tuned for data that share domains similar to FLORES+.

In contrast, fine-tuning on naturalistic datasets such as the PARADISEC corpus and the Dictionary examples substantially improves performance on the PARADISEC (test) data and the Dialogue data. For instance, the NLLB-600M model fine-tuned on P+D+N reaches 23.35 BLEU on the Dialogue test set, outperforming the baseline by a wide margin. A similar trend holds in the English→Jinghpaw direction, where models fine-tuned on the naturalistic datasets consistently outperform the baseline on non-FLORES+ test sets.

Notably, fine-tuning on PARADISEC and Dictionary typically results in a drop in FLORES+ performance, suggesting a mismatch between their domains. This performance trade-off raises a critical question: What does it mean to "perform well" on a benchmark in which the reference translations themselves are questionable? In our human evaluation, for instance, only one out of fifty Jinghpaw→English FLORES+ sentences was judged to be a correct translation by a native speaker.

The results of the two experiments in this section highlight two key insights: (1) FLORES+ may reward surface-level alignment with domain-specific data rather than generalizable translation ability; and (2) fine-tuning on small but naturalistic corpora can significantly enhance model performance on real-world test sets, albeit hurting the benchmark scores. These findings underscore the urgent need for more representative evaluation datasets that reflect the linguistic and cultural diversity of speaker communities.

5 Concluding remarks

In this study, we critically examined the FLORES+ multilingual MT benchmark through human re-evaluation and empirical analysis, with a partic-

Model	Training data	PARADISEC (test)	FLORES (devtest)	Dialogue	Average
NLLB-600M	Baseline	2.32 / 20.34	**12.77 / 35.47**	17.35 / 32.42	10.81 / 29.41
	D	4.58 / 22.52	9.73 / 31.44	18.08 / 34.43	10.80 / 29.46
	P	**12.58** / 31.35	5.91 / 27.34	22.41 / 39.78	13.63 / 32.82
	P+D	12.55 / **31.91**	6.04 / 27.12	22.30 / 40.38	13.63 / 33.14
	P+D+N	12.86 / 31.91	6.88 / 30.03	**23.35 / 41.67**	**14.36 / 34.54**
NLLB-1.3B	Baseline	2.29 / 19.72	**13.95 / 37.27**	16.66 / 33.44	10.97 / 30.14
	D	4.23 / 20.44	10.90 / 30.56	**24.57 / 41.71**	13.23 / 30.90
	P	12.95 / 32.11	5.48 / 26.59	18.09 / 34.84	12.17 / 31.18
	P+D	12.95 / 31.45	5.41 / 26.66	23.49 / 40.96	**13.95 / 33.02**
	P+D+N	**13.40 / 32.19**	4.87 / 27.25	22.44 / 39.81	13.57 / 33.08

(a) Results of Jinghpaw→English MT.

Model	Training data	PARADISEC (test)	FLORES (devtest)	Dialogue	Average
NLLB-600M	Baseline	3.67 / 25.32	**9.68 / 34.43**	13.05 / 39.61	8.80 / 33.12
	D	4.61 / 26.77	8.06 / 31.84	17.10 / 41.36	9.92 / 33.32
	P	10.99 / 32.92	1.65 / 21.76	18.90 / 44.28	10.51 / 32.99
	P+D	11.40 / 33.87	2.50 / 24.17	**21.75** / 44.82	11.88 / 34.29
	P+D+N	**11.43 / 34.35**	3.94 / 24.02	21.17 / **45.68**	**12.18 / 34.68**
NLLB-1.3B	Baseline	4.15 / 25.94	**10.97 / 36.00**	16.95 / 43.02	10.69 / 34.99
	D	4.48 / 26.05	6.12 / 29.68	20.17 / 44.34	10.26 / 33.36
	P	12.19 / 35.04	2.26 / 24.09	**24.41 / 48.21**	12.95 / 35.78
	P+D	12.30 / 35.75	3.74 / 27.45	22.75 / 48.01	12.93 / **37.07**
	P+D+N	**12.78 / 35.87**	3.60 / 25.16	23.01 / 47.82	**13.13** / 36.28

(b) Results of English→Jinghpaw MT.

Table 6: Results. Each row corresponds to a training setting, and each column shows performance on a specific evaluation set. D stands for the Dictionary dataset, P for the PARADISEC dataset, and N for the NLLB dataset. The Baselines are the pre-trained NLLB models without fine-tuning. In each cell, the score on the left-hand side represents the BLEU score, and the one on the right-hand side the ChrF++ score.

ular focus on low-resource languages. Our human re-evaluation revealed that the translation quality in these languages falls below the benchmark's reported standard, calling into question the reliability of its quality claims. Annotators consistently reported that many English source sentences were highly domain-specific or culturally loaded, making them difficult to translate naturally or at all in some languages. This raises fundamental concerns about the benchmark's linguistic inclusivity and representativeness.

Our experimental analyses further uncovered structural issues in the benchmark design. In the named-entity copying experiment, mock translations that merely reproduced named entities from the source achieved non-trivial BLEU and ChrF++ scores. This suggests that current evaluation protocols may reward superficial copying over true translational adequacy and fluency. Moreover, in our Jinghpaw MT experiments, we found that while the pretrained NLLB models performed best on the FLORES+ set, they performed poorly on more naturalistic evaluation datasets constructed from narratives and conversational texts. In contrast, models fine-tuned on authentic, community-sourced data demonstrated improved performance on these realistic test sets, though their FLORES+ scores declined.

Taken together, these findings highlight the misalignment between benchmark-centered evaluation and the realities of multilingual MT, particularly for low-resource and culturally diverse languages. To address these limitations, we propose the following directions for future benchmark development: (1) source sentences should be less domain-specific and more universally interpretable; (2) the influence of named entities should be minimized to better assess true translation capabilities; (3) content should avoid undue bias toward English-speaking cultural and linguistic norms; and (4) ongoing efforts to revise and expand existing datasets should be actively supported and valued.

Limitations

While our study reveals critical insights into the shortcomings of the FLORES benchmark and highlights challenges in multilingual MT evaluation, it is not without limitations.

First, our human re-evaluation was conducted on a relatively small subset of the benchmark (50 sentence pairs per language) and focused on a limited set of languages. Although the languages were chosen to represent typological and resource-level diversity, further large-scale evaluations across more language families are necessary to generalize our findings.

Second, while we critique the limitations of FLORES+, we do not propose a fully realized alternative benchmark. Future work should explore concrete methods for constructing domain-general, culturally inclusive evaluation sets that are robust to lexical and structural biases.

Last but not least, we fully acknowledge the substantial effort and coordination required to construct the original FLORES benchmark, particularly the translation of over 3,000 sentences into more than 200 languages. This undertaking has significantly advanced multilingual MT research and helped bring attention to low-resource languages in the NLP community. Our critiques are not meant to diminish this achievement, but rather to support its ongoing improvement. We hope our findings contribute to the broader conversation around building more inclusive, culturally aware, and contextually appropriate evaluation benchmarks.

Ethical considerations

This work aims to critically assess the limitations of multilingual MT benchmarks with a focus on underrepresented and low-resource languages. All human evaluations in this study were conducted by native speakers of the target languages who were fluent in English. Annotators were fully informed of the purpose of the study and participated voluntarily. No sensitive or offensive content was collected during the study, and any information that could identify annotators has been anonymized.

Our findings underscore the ethical implications of benchmark design in NLP and MT. Benchmarks that rely heavily on English-centric, domain-specific, or culturally narrow content can systematically disadvantage certain languages and communities. When such benchmarks are treated as gold standards for model evaluation, they risk misrepresenting the capabilities of MT systems and, more seriously, may perpetuate linguistic inequities, leaving many languages behind.

We emphasize the importance of language technologies that respect speaker communities, their linguistic norms, and their cultural contexts. Translation systems should not merely map words across languages and copy or transliterate proper nouns but should support meaningful, context-aware communication. This requires training and evaluating models on culturally informed, naturally occurring data rather than relying solely on synthetically generated or automatically translated corpora, which may introduce artifacts or distort local usage, leading to unnatural sentences or translationese.

The datasets used in our study, including the PARADISEC narrative corpus, the Jinghpaw Dictionary, and the Jinghpaw Reader, are all publicly available and were developed with care and proper documentation. We advocate for ongoing, collaborative partnerships with speaker communities in both data creation and model evaluation. Ethical NLP research must prioritize not just linguistic diversity in principle, but the lived realities, needs, and voices of language users.

While we used AI assistants such as ChatGPT to support the writing process in a non-native language, all content was carefully reviewed and edited to ensure accuracy, clarity, and alignment with our research objectives.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. *Preprint*, arXiv:2409.00626.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. Association for Computational Linguistics.

Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli, and Taro Watanabe. 2023. NAIST-NICT WMT'23 general MT task submission. In *Proceedings of the Eighth Conference on Machine Translation*, pages 110–118, Singapore. Association for Computational Linguistics.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. FLORES+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Keita Kurabe. 2013. Kachin folktales told in Jinghpaw.

Keita Kurabe. 2017. Kachin culture and history told in Jinghpaw.

Keita Kurabe. 2020a. *A Dictionary of Jinghpaw Usage*. The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Keita Kurabe. 2020b. *Jinghpaw Reader*. The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Ali Kuzhuget, Airana Mongush, and Nachyn-Enkhedorzhu Oorzhak. 2024. Enhancing tuvan language resources through the FLORES dataset. In *Proceedings of the Ninth Conference on Machine Translation*, pages 593–599, Miami, Florida, USA. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

NLLB Team, 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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the FLORES+ multilingual benchmark with translations for Aragonese, Aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.

A Implementation details of metrics

A.1 BLEU

For computing the BLEU scores, we used the implementation provided by the sacrebleu library (version 2.5.1) (Post, 2018). As given in the body of this paper, the general formula for calculating a BLEU score is:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where BP is the brevity penalty $\mathrm{BP} = \min(1, e^{1-\frac{r}{c}})$, N is the maximum value of n for counting n-grams, w_n the weight for each n, r the length (number of words, split by whitespaces) of the reference sentence, c the length of the hypothesis, and p_n the modified n-gram precision, which is the number of matches devided by the number of n-grams. The number of matches of a word considered in the modified n-gram precision is at

most the frequency of the word in the reference text. In our experiments, we used the default hyperparameters implemented by sacrebleu: $N = 4$ and $w_n = \frac{1}{4}$. To avoid the cases of $p_n = 0$, which is often the case in our experiments due to a number of empty hypothesis strings in the named-entity-copying dummy translations, we made sure that p_n is smoothed by exponential decay when $p_n = 0$. Exponential decay smoothing is the default setting in sacrebleu.

Algorithm 1 Exponential Decay for BLEU

Require: n-gram match counts: match$[1 \ldots N]$
Require: Total candidate n-grams: total$[1 \ldots N]$
Ensure: Smoothed n-gram precisions $p_1, \ldots, p_N$

```
s ← 1                                  ▷ Initialize decay factor
for n ← 1 to N do
    if correct[n] = 0 then
        s ← 2 · s                      ▷ Exponential penalty
        p_n ← 1 / (s · total[n]) · 100
    else
        p_n ← correct[n] / total[n] · 100
    end if
end for
```

A.2 ChrF++

We also used sacrebleu for computing ChrF++ score. The formula for computing ChrF++ score is:

$$\text{ChrF++} = \frac{(1+\beta^2)P \cdot R}{\beta^2 P + R},$$

In the typical implementation of ChrF++, which we followed in this study, uses $\beta = 2$, $N_w = 2$ and $N_c = 6$, where N_w is the maximum n-gram considered at the word level and N_c at the character level. The sentence-level precision P is the averaged value of character-level and word-level precisions, and the sentence-level recall R, similarly, is the averaged value of character-level and word-level recalls:

$$P = \frac{P_c + P_w}{N_c + N_w}, R = \frac{R_c + R_w}{N_c + N_w}.$$

The corner cases where the reference or hypothesis text is shorter than N_c or N_w are ignored; that is, if we have $N_w \leftarrow 2$ but the hypothesis length is 1, then the 2-gram is ignored and $N_w \leftarrow 1$. In our experiments, we removed whitespaces when computing P_c and R_c, which is the default configuration in sacrebleu. Also, we made sure that the ChrF++ calculation correctly considers word-level 2-grams by specifying `sacrebleu.CHRF(word_order=2)`.

B Prompt template for named-entity extraction

```
Please output the extract named entities
    concatenated by whitespace.

For example,
**Input:** On Sunday, May 4, in Peru's
    Pataz province in the northern
    Department of La Libertad region,
    near one of Peru's largest gold
    mines, police has found bodies of
    thirteen security guards who were
    kidnapped on Saturday, April 26,
    allegedly by individuals involved in
    illegal mining.
**Output:** Sunday May 4 Peru Pataz
    province Department of La Libertad
    region Peru Saturday April 26

{other_in_context_examples}

Now, it's your turn.
**Input:** {text}
**Output:**
```

C Sample sentences from the in-house datasets

Table 7 shows some sample sentences from the in-house datasets used in this study.

D Deetails of instructions given to the annotators

The data collection process was conducted in consultation with the Institutional Review Board at the first author's university.

	Jinghpaw	English
PARADISEC	Dai hka hpe rap na matu shan lahkawng gaw grai yak ai da.	It was really difficult for them to cross the river.
Dictionary	ndai nye a hkawhkam nang shayi hpe nang hpe ap sana	I'll leave this princess of mine to you
Dialogue	Manau poi hte seng ai laika buk ka na matu myit da ai.	I hope to write a book about the Manau festival.

Table 7: Sample sentences from the in-house datasets used in this study.

```
## Instruction
1. Open the Spreadsheet {link}
2. For each row, compare the sentences
    in English and {language}, and judge
    whether the {language} sentence is a
    correct translation of the English
    sentence. Specify  your judgment(s)
    in the "Fine-grained evaluation"
    column. Specify your judgment(s) in
    the "Evaluation" column. See the
    guidelines in Goyal et al. (2021).
3. If the sentence contains an error(s),
    specify the severity of the error(s)
    in the "Error severity" column. See
    the guidelines in Goyal et al.
    (2021).
4. If you think that the translation has
    a problem (the evaluation is not
    "Correct"), please provide your
    correct translation in the
    "Corrected {language} translation"
    column.
5. If you have any other comments, such
    as any additional clarification,
    questions, etc., feel free to write
    them in the "Comments" column.
6. Repeat steps 2 - 5 for the 50
    sentences (up to the 51st row in the
    spreadsheet)

## Acknowledgment
We aim to publish the results in an
    academic conference paper. To
    acknowledge your participation, we
    provide you with the following two
    options:
- To be a co-author of the paper.
- To receive a research participation
    reward in cash.
```