# Is Machine Translation Ready for the Next Crisis or Pandemic?

**Vipasha Bansal  William D. Lewis  Nicole Mihaylova  Elizabeth Brown  Benjamin Pong**
Computational Linguistics Program
University of Washington
{vipashab, wlewis2, nicolemi, egollhof, benpong}@uw.edu

## Abstract

Communication in times of crisis is essential. However, there is often a mismatch between the language of governments, aid providers, doctors, and those to whom they are providing aid. It is reasonable to turn to commercial MT tools in these scenarios. But how effective are these tools for translating to and from low resource languages, particularly in the crisis or medical domain? In this study, we evaluate 5 commercial systems used for MT (Google Translate, Microsoft Translator, GPT, Gemini and Claude) against the TICO-19 dataset, which contains pandemic-related sentences from a large set of high priority languages spoken by communities most likely to be adversely affected in the next pandemic. We assess the current degree of "readiness" for this next pandemic based on the usability of the output translations.

## 1 Introduction

COVID-19 was the worst pandemic in over a century. In order to slow the spread of the virus and to reduce hospitalizations and death, it was vital to communicate information to vulnerable populations. However, official communications from health agencies such the World Health Organization were only available in high resource languages.

The TICO-19 dataset (Anastasopoulos et al., 2020) was created during the pandemic to encourage MT research in under-resourced languages, specifically languages whose communities might be impacted by the pandemic but had not yet been touched by it. [1,2] Translators without Borders (TwB) was the driving force behind the choice of languages in TICO-19; TwB was, and still is, engaged in translating varied content for numerous languages spoken across Africa, Southeast Asia and other parts of the world.

The dataset contains n-way parallel test and dev sets across 36 languages. At the time of TICO-19's development, more than half of its languages had no commercial translation systems available (*e.g.,* Dari, Kurdish, Zulu). This situation has changed markedly in the ensuing years: 36 of the 37 languages in the TICO-19 test set are now supported by at least one commercial system. This change in MT availability raises the following questions: How good are commercial MT systems at translating into and out of these languages? In particular, how good are they for translating pandemic-related content? Would they be adequate for aid providers, medical professionals, or governments to communicate with speakers of these languages on the ground during a future public health crisis?

Our study measures the quality of five existing commercial systems in the context of pandemic response. We use the TICO-19 dataset to systematically evaluate two major MT-specific providers, Google Translate and Microsoft Translator. Given the global popularity of LLMs, including in MT research (Kocmi et al., 2024), and their potential competitiveness with existing translation services (Hendy et al., 2023; Robinson et al., 2023), we also evaluate three popular LLMs: GPT, Gemini and Claude. For the translation providers, we captured three snapshots over time, in 2023, 2025, and 2026, giving us a sense not only of their current performance, but also of how this performance has changed over the past three years.

We recognize the value of the output generated in our study to various fields, including Machine Translation, crisis response, medicine, and public health. We see this paper as the beginning and want to facilitate further research in this area. To this end, we are releasing all data collected in our study here: https://mt-crisis-ready.github.io.

[1] TICO-19 was created in mid-2020, early in the pandemic.

[2] A secondary goal of TICO-19 was to provide translation memories and glossaries for translators. These, along with the MT benchmark data used in this paper, are available for download from the TICO site: https://tico-19.github.io/.



*Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response*, pages 78–114
November 9, 2026 ©2026 Association for Computational Linguistics

## 2 Background and Related Work

### 2.1 The Use of MT and Other Language Technologies in Crisis Scenarios

There has been a large amount of research on the use of language technologies in crisis response over the past several years. Lewis et al., 2025, for instance, reviews 355 papers published over a five-year timeframe ending in 2025. These papers focus on various language tasks, albeit specific to social media use in crises. Most concentrate on classification, with a smattering of work on other technologies including NER, geo-location, summarization, and topic modeling. Oddly, given the prevalence of large-scale crises that occur in areas where English is not spoken (see Tin et al., 2024, and discussion in Lewis et al., 2025), almost none of the papers in the five-year study show work on translation or MT as it relates to crises and social media.

That is not to say that there has been no work on crises and translation. The INTERACT project[3], an EU-funded research consortium, produced a number of publications on translation in crises, focused on various aspects of *crisis communication* including translation workflows and training translators (Federici and Cadwell, 2018), translation needs assessments in times of crisis (O'Brien and Federici, 2019), and ethics and crisis translation (Hunt et al., 2019). Similarly, Bird, 2022 looks at crisis communication holistically, considering cross-cultural communication and the use of "trade" languages (*i.e.,* regional *linguae francae*). Further, Lankford and Way, 2024 examine the use of LLMs for low-resource MT in a pandemic, and Roussis, 2022 walks through the data collection and training of an end-to-end neural model dedicated to pandemic translations. And of course, Lewis, 2010 established the crucial need for commercial MT for low-resource languages in a crisis scenario.

Both commercial MT tools and LLMs have seen broad general-purpose usage and acceptance, and commercial MT systems are often used in medical settings, including for doctor-patient communication (Khoong et al., 2019; Khoong and Rodriguez, 2022). It is likely that MT is used similarly in crises, with some work showing MT being used by crisis managers in the United States (Wilson et al., 2026; Foubert et al., 2026).

[3]https://cordis.europa.eu/project/id/734211

### 2.2 Quality, Usability and Readiness

The end goal of much MT research is to design systems that produce high-quality translations across domains, *e.g.,* in bake-offs and competitions (Haddow et al., 2025; Salesky et al., 2026). Quality is often measured for competing systems using automated metrics, such as CHRF++ (Popović, 2017) and BLEU (Papineni et al., 2002), or human evaluation (Graham et al., 2013).

Our focus in this paper however, is to look at the readiness of competing systems in the context of pandemic response. Our goal is therefore to assess whether a translation system is good enough to be used in the field during an active crisis. The key to this is usability: it is essential that the translated information be *understood*, whether the translations are *to* a population from, say, aid workers or government bodies (O'Brien and Federici, 2019), or *from* a population, say, as posted to social media (Lewis et al., 2025; Imran et al., 2015).

So how does one measure usability, especially in the context of MT usage in a crisis scenario? One could ideally identify tasks that the MT would be used for in the field, *e.g.,* communicating directly with those affected, or sending emergency alerts to multilingual populations (Wilson et al., 2026; Foubert et al., 2026). Although usability is not the same as quality, there is a minimum threshold for quality that translations must meet, or any task dependent on those translations will fail. But how do we determine if output translations meet such a threshold? Further, how do we set this threshold?

We use four evaluation metrics against our translated output: BLEU (Papineni et al., 2002), CHRF++ (Popović, 2017), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020). Of these, BLEU and CHRF++ are n-gram based; despite their dependence on matching surface output with reference(s),[4] they have been shown to correlate well with human evaluation, the gold standard (Papineni et al., 2002; Popović, 2017).

BERTScore and COMET are "semantic" based, and instead focus on the holistic "meaning" of the output (Zhang et al., 2020). Unfortunately, these metrics do not adequately support many of the low resource languages in the TICO-19 dataset. This lack of coverage, along with issues with reliable performance even on supported low-resource lan-

[4]One of the main weaknesses of n-gram metrics is that an output could preserve meaning when compared with a reference but be assigned a low score because there is low n-gram overlap between the two (GoogleCloud, 2023).

guages and on specialized (or unseen) domains, has been noted in various studies (NLLB-Team et al., 2022; Tang et al., 2024; Falcão et al., 2024; Zouhar et al., 2024a,b). In particular, Zouhar et al. (2024b) notes the poor performance of COMET in the biomedical domain—a domain close to that of TICO-19.

Given the limited language coverage of BERTScore and COMET across the TICO-19 dataset, we turned to BLEU and CHRF++ to build our threshold. We follow a traditional rule-of-thumb to interpret BLEU scores, captured by Google's MT API documentation [5] (Google Cloud, 2025) and shown in Table 1. We see 30 as our usability threshold, as this indicates "understandable to good" quality translations (per the table). With CHRF++, we follow NLLB-Team et al. (2022), who identified a CHRF++ score of 45 as a minimum threshold. Although their task was focused on toxicity detection, we see this as a reasonable starting point. Section 4.3 provides further details on how we use these metrics for thresholding, and Section 5.2.3 outlines the results of a preliminary human evaluation to verify our proposal. [6]

| BLEU Score | Interpretation |
|---|---|
| < 10 | Almost useless |
| 10 - 19 | Hard to get the gist |
| 20 - 29 | Clear gist, major grammatical errors |
| 30 - 40 | Understandable to good translations |
| 40 - 50 | High quality translations |
| 50 - 60 | Very high quality, fluent translations |
| > 60 | Quality often better than human |

Table 1: BLEU Score Interpretation

## 3 The TICO-19 Dataset

The dataset consists of 3,071 sentences (971 dev, 2,100 test) drawn from Wikipedia, PubMed, and existing corpora such as the CMU English-Haitian Creole dataset (Anastasopoulos et al., 2020). The sentences were translated into 37 languages, making TICO uniquely n-ways parallel. There are 9 *pivot* languages, representing global or regional *linguae francae*; 20 *priority* languages identified by TwB based on translation requests during the pandemic or their importance to partner aid organizations such as the Red Cross; and 8 additional languages labeled *important*, included because they have large populations of speakers across Asia. Although these labels were not explicitly designed to map to language resource level, they align remarkably well: Pivot languages tend to be high resource, Priority languages low resource, and Important languages somewhere in between.

The dataset's structure allows for bidirectional evaluation that enables the measurement of MT quality in scenarios where English could be either the source, *e.g.,* of public health content, or the destination, *e.g.,* for translated reports or social media. Domain-source correspondence is shown in Table 2. We utilize these subdomains to examine potential training data contamination (see Appendix F).

Two languages were excluded from our study: the data for Congolese Swahili was unavailable for download, and the data file for Nuer was corrupted.

| Data Source | Domain | # Sentences |
|---|---|---|
| CMU | Conversational, medical | 104 |
| PubMed | Scientific, medical | 676 |
| Wikipedia | General | 1,153 |
| Wikivoyage | Travel | dev only |
| Wikinews | News | 66 |
| Wikisource | Announcements | 99 |

Table 2: TICO-19 Test Data Statistics by Domain

## 4 Methods

### 4.1 Dedicated Machine Translation Systems

We evaluated Google and Microsoft Translation services for three time frames: April-May 2023, April-July 2025, and May-June 2026. We used Microsoft Azure Cognitive Services Translator in 2023 and Azure Translator in 2025 and 2026[7] and Google Cloud's API throughout to access each model. The test sentences were identical across all three snapshots.[8]

### 4.2 LLM systems

We evaluated three GPT models (GPT-4o, GPT-5.4, and GPT-5.4-mini), Gemini-3.1-Flash-Lite, (for which translation is listed as a recommended use case),[9] and Claude-Haiku-4.5. The models were ac-

[5] https://cloud.google.com/translate/docs/advanced/automl-evaluate

[6] We realize that there is significant debate about using n-gram based metrics, particularly BLEU, for thresholding. Please refer to Appendix K for additional discussion on this topic and related issues.

[7] Although there were changes to the Azure API between 2023 and 2026, likely indicating reorganizations within Microsoft, the underlying translation service remained the same.

[8] The TICO-19 benchmark dataset has remained stable since its release in 2020, with no changes made to its test data.

[9] https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite

cessed through Azure's OpenAI Service, Google's Cloud API, and Anthropic's Claude Platform. [10]

Unlike the dedicated MT systems, which provide specific and documented language translation support, LLMs can generate text in multiple languages but are not explicitly optimized for translation. While an LLM can, in theory, be prompted to translate to and from *any* language, the lack of training data for low-resource languages means that, in reality, a limited number of languages are supported (Chua et al., 2025). We determined language support based on the inverse proportion of sentences where the model failed to produce a translation for a given language (see Appendix E). A failure consisted of blank output, a statement that the model could not do the task, output translations in the wrong language, or string repetitions.

For the three GPT models, some translations incorrectly triggered OpenAI's content filters (see Appendix G). While problematic for usability, unlike the failure types listed above, this was not a failure to produce a translation, but rather a failure to output it to the user. Therefore, filtered outputs were not used to determine language support.

When evaluating supported languages, failed outputs were not excluded from the final scoring; this ensured consistency and comparability of results across languages and systems. Instead, filtered sentences were treated as blank output, while the remaining failure types were retained and evaluated as is. Error handling was consistent across metrics.

### 4.3 Evaluation Metrics

BLEU and CHRF++ were calculated using the SacreBleu package (Post, 2018).[11] For Khmer, Myanmar, and Chinese, which do not use whitespace to mark word boundaries, we used the character-based CHRF metric (Popović, 2015) instead of CHRF++.[12,13] For BERTScore (Zhang et al., 2020), we used the `bert_score` Python package with default settings. Baseline rescaling—which normalizes scores to fall between 0–1—was applied to translations into English.[14] For COMET, we used the XLM-RoBERTa-based COMET model released as part of the WMT22 Quality Estimation shared task (Zerva et al., 2022). The model was loaded using the `comet` Python package.

To ensure consistency in evaluating longitudinal data, BLEU and CHRF++ scores for all 2023 and 2025 outputs were recalculated alongside 2026 outputs; therefore the same version of the SacreBleu toolkit (and the associated tokenizers) was used throughout. For the neural metrics (BERTScore and COMET), we used identical models and hyperparameters for all snapshots.

Kumar et al. (2026) recommend evaluating translations against both BLEU *and* CHRF++ together, particularly for low resource languages, as the metrics provide complementary insights. Following this, we use both in determining whether translations are usable. As outlined in Section 2.2, we use 30 and 45 as individual thresholds for BLEU and CHRF++, respectively. We consider translations that score above *both* thresholds to be usable, and translations that meet only one threshold *borderline usable*. Translations that do not pass either threshold are considered unusable. See Section 5.2.3 for a preliminary human evaluation of usability, and preliminary support for these thresholds.

## 5 Results

Figures 1 (a)-(f) show the CHRF++ scores for each system and language, separated by TICO classification (Pivot, Important, Priority) and translation direction. Example scores showing overall trends are in Table 3, and full results across all four metrics are in Tables 15–22, Appendix A.

### 5.1 System Comparison

#### 5.1.1 Coverage

All 36 languages that we evaluated are supported by GPT-5.4-mini in both translation directions. The full size version (GPT-5.4) provides partial support for Kurdish Kurmanji (in the EX direction, it produced all output in Kurdish Sorani instead, as did Gemini), and full support for the remainder. Similarly, Google Translate provides full support for all languages but Dari, which is partially supported.[15] Dinka, Nigerian Fulfulde, and Kanuri were problematic throughout; none are supported by Microsoft or Claude, Gemini provides only XE support for each, while GPT-4o does not support Dinka and provides only XE support for the others.

[10] See Appendix C for details on model configurations.

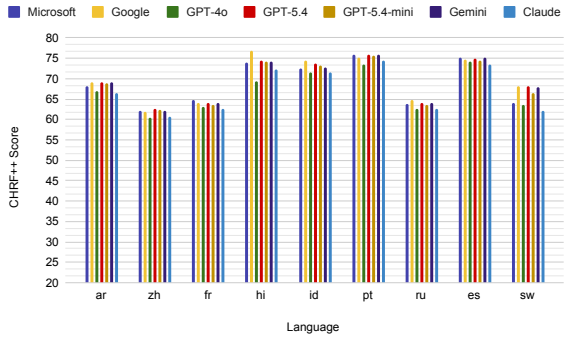
[11] https://github.com/mjpost/sacrebleu

[12] CHRF++ considers both word and character n-grams together (Popović, 2017)

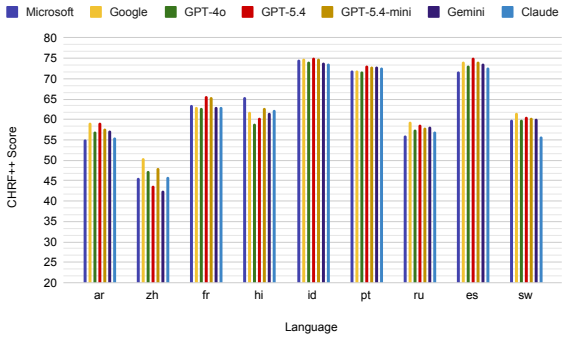
[13] See Appendix D for metric signatures.

[14] As rescaling is not supported for all languages, it was not applied in the EX direction, as it would prevent cross-lingual comparison of results.

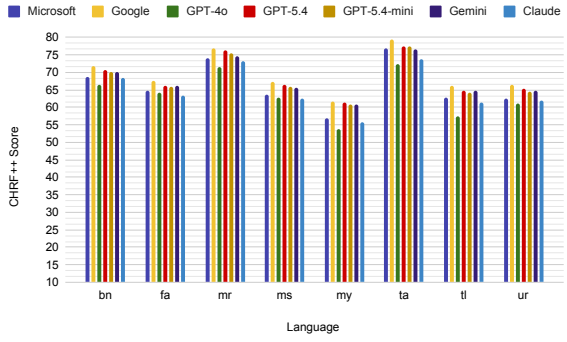
[15] While the Google Translate app supports Dari in both directions, the API does not show support for the language. However, we found that translations in the API *from* Dari were supported, but they were not supported *to* Dari.



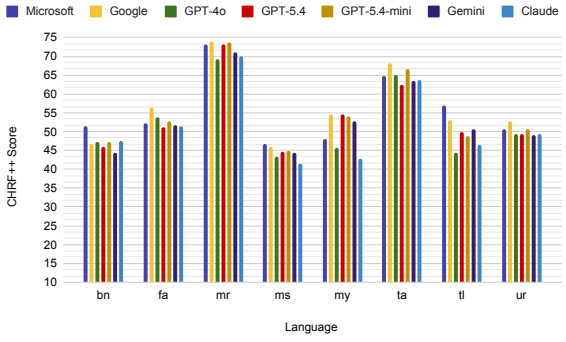

(a) X to E CHRF++ Pivot Languages



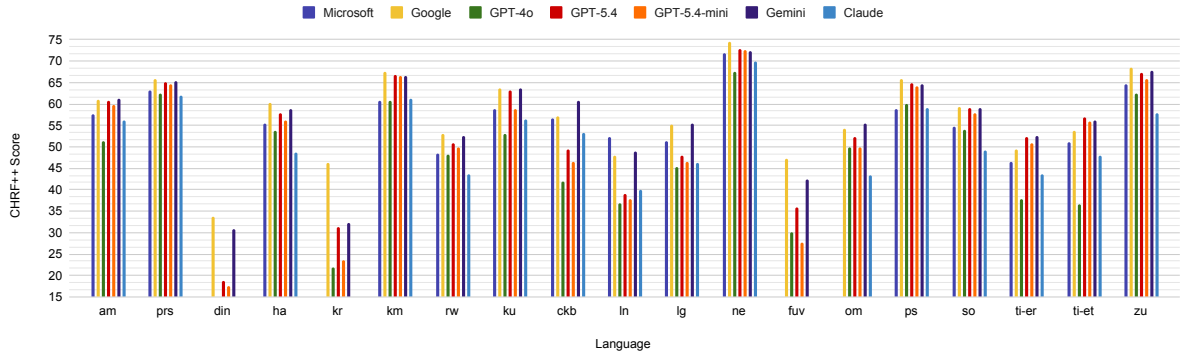

(b) E to X CHRF++ Pivot Languages



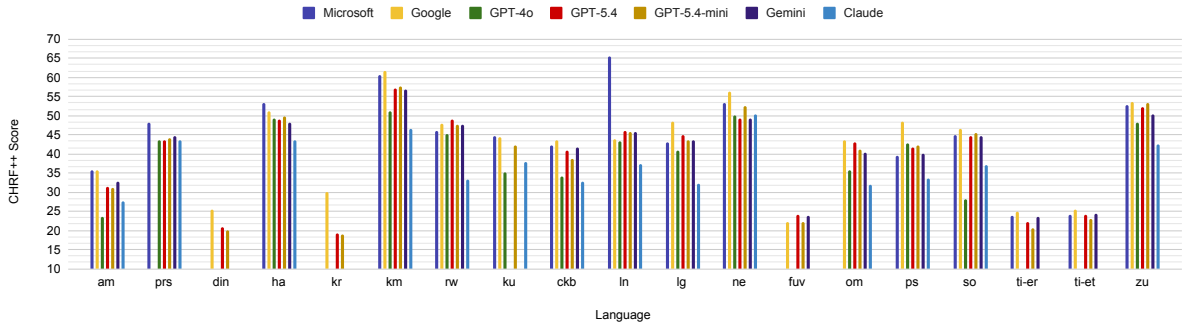

(c) X to E CHRF++ Important Languages




(d) E to X CHRF++ Important Languages




(e) X to E CHRF++ Priority Languages




(f) E to X CHRF++ Priority Languages

Figure 1: Graphs showing comparative 2026 CHRF++ scores across language pairs (E to X and X to E)

Further, both dialects of Tigrinya were supported in only the XE direction by Claude and Gemini, and Microsoft does not support Oromo.

#### 5.1.2 Quality

Google produced translations with the highest scores (in both BLEU and CHRF++) for most languages, in both translation directions. When translating *from* Pivot languages *to* English or *from* English *to* Important languages, there was an even split with Microsoft for best system. Otherwise, Microsoft generally had comparable or slightly lower scores than Google. GPT-5.4 was competitive when translating *from* English *to* Pivot languages, and produced the highest scoring translations for about half of these languages. Surprisingly, Microsoft's scores were closer to those of the lowest performing LLMs here (in contrast to its high performance on Pivot languages in the opposite (X to E) translation direction).

GPT-4o and Claude produced the lowest quality translations throughout (Claude did particularly poorly on Priority languages). Gemini achieved comparably low EX scores, but was more competitive with dedicated MT systems in the XE direction, often matching CHRF++ (but not BLEU) scores.

Exceptions to these trends (see Table 4) include Hindi and Hausa, where Microsoft outperformed Google (and GPT-5.4) in the EX direction, and Lingala, where Microsoft scored the highest in both directions (Microsoft's EX scores were 28 BLEU and 19 CHRF++ points above the next best system). Additionally, when translating into Myanmar and Kinyarwanda, GPT-5.4 produced the highest scoring translations, while in the XE direction, both GPT-5.4 models and Gemini outperformed both Microsoft and Google on Ethiopian Tigrinya. Gemini also achieved the highest scores for Arabic, Kurdish Sorani, Oromo, and Eritrean Tigrinya.

Table 5 shows the best MT system for each language, based on the highest scores on *both* metrics, alongside the usability of that system. For Pivot and Important languages, both BLEU and CHRF++ indicated good quality translations for all models. While LLMs can provide adequate translations for high resource languages, they were rarely the best system, and overall not competitive with dedicated MT systems for low-resource languages.

Some systems produced translations with extremely high scores (BLEU above 55) for a number of high resource languages. These scores are problematic, and suggest potential data contamination (see Appendix F). Likewise, some systems showed dismally low EX scores (BLEU below 5) for a subset of low-resource languages (*e.g.,* Dinka). For LLMs this indicates the presence of hallucinations (see Appendix J).

#### 5.1.3 Translation Direction

BLEU and CHRF++ scores were significantly higher for translations in the XE direction across all systems and languages, which is unsurprising given the vast quantity of English pretraining data available. This was especially true for low resource languages. Arabic had particularly large differences in scores between translation directions compared to other Pivot languages. Exceptions to this trend include Indonesian, where all systems (except Google) had lower XE scores, and Lingala, where Microsoft had significantly lower XE scores. [16]

### 5.2 Google & Microsoft Changes Over Time

#### 5.2.1 Coverage over Time

In 2023, many of the languages were not supported by one or both of Microsoft and Google. Coverage has improved since then: by 2025, all TICO-19 languages (except Congolese Swahili) were supported by at least one service. In 2026, the main change is Google's addition of partial support for Dari.

#### 5.2.2 Quality over Time

Table 6 shows notable changes in CHRF++ scores since 2023. Full results for 2023 and 2025 across the four metrics are shown in Tables 23–26, Appendix B. Microsoft's performance remained mostly stable when translating to English, with consistent improvements in Important and Priority languages. In particular, Myanmar, Tamil, and Khmer scores showed a strong increase in both metrics in 2025, as did Kurdish Kurmanji, Dari, and Ethiopian Tigrinya in 2026. In the EX direction, however, there was a significant decrease in quality for almost every language in 2025—especially Arabic, Indonesian, Malay, and Pashto—followed by near universal improvement in 2026. Hausa and Khmer were exceptions; both showed improvements throughout. It is also worth noting 2026 scores for Lingala, which improved by 16/7 BLEU/CHRF++ points, and Tagalog and Somali, which both dropped significantly.

[16] Chinese, Myanmar and Khmer appear to buck this trend in BLEU; they each do better in EX vs. XE. However, the EX direction for these languages was scored using character-based tokenizers provided by SacreBleu. Character-based measures tend to have inflated scores, and are not directly comparable.

| Language | EX/XE | Microsoft | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
|---|---|---|---|---|---|---|---|---|
| es-la | EX | 71.88 | 74.17 | 73.17 | **75.05** | 74.08 | 73.64 | 72.84 |
| | XE | **75.22** | 74.70 | 74.09 | 74.87 | 74.47 | **75.08** | 73.57 |
| sw | EX | 59.96 | **61.59** | 59.97 | 60.55 | 60.44 | 60.16 | 55.86 |
| | XE | 63.97 | **68.14** | 63.51 | 68.10 | 66.35 | 67.91 | 62.19 |
| fa | EX | 52.32 | **56.44** | 53.76 | 51.16 | 52.71 | 51.80 | 51.38 |
| | XE | 64.64 | **67.55** | 64.25 | 66.21 | 65.77 | 66.14 | 63.45 |
| bn | EX | **51.40** | 46.69 | 47.28 | 45.97 | 47.25 | 44.38 | 47.55 |
| | XE | 68.76 | **71.94** | 66.46 | 70.77 | 70.21 | 70.19 | 68.40 |
| am | EX | **35.84** | **35.78** | 23.51 | 31.47 | 31.23 | 32.62 | 27.60 |
| | XE | 57.54 | **60.98** | 51.41 | 60.71 | 59.82 | 61.11 | 56.14 |
| ne | EX | 53.42 | **56.19** | 50.06 | 49.15 | 52.39 | 49.35 | 50.25 |
| | XE | 71.96 | **74.38** | 67.49 | 72.92 | 72.50 | 72.42 | 69.83 |

Table 3: 2026 CHRF++ Scores showing example comparisons between providers for Pivot, Important, and Priority languages. See Tables 15–22 for complete scores.

| Language | EX/XE | Microsoft | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
|---|---|---|---|---|---|---|---|---|
| id | EX | 74.72 | 74.88 | 74.15 | **75.08** | 74.93 | 73.92 | 73.74 |
| | XE | 72.53 | **74.33** | 71.53 | 73.77 | 73.27 | 72.70 | 71.61 |
| hi | EX | **65.62** | 61.76 | 59.06 | 60.34 | 62.81 | 61.75 | 62.42 |
| ha | EX | **53.44** | 51.26 | 49.12 | 48.91 | 49.77 | 48.21 | 43.49 |
| rw | EX | 45.88 | 47.87 | 45.14 | **49.10** | 47.74 | 47.50 | 33.23 |
| ckb | XE | 56.53 | 57.09 | 41.84 | 49.45 | 46.52 | **60.66** | 53.37 |
| ln | EX | **65.59** | 43.85 | 43.42 | 46.07 | 45.62 | 45.68 | 37.34 |
| | XE | **52.21** | 47.85 | 36.95 | 39.07 | 37.69 | 48.86 | 40.07 |
| ti-et | XE | 51.17 | 53.70 | 36.56 | **56.83** | 55.97 | **56.17** | 47.83 |

Table 4: 2026 CHRF++ Scores showing notable exceptions to overall trends across languages and MT providers. See Tables 15–22 for complete scores.

| | |
|---|---|
| **E to X** | |
| Microsoft | **hi**, bn, mr, ta, am, ha, ku, **ln**, prs |
| Google | **ar**, **ru**, **sw**, **zh**, fa, **ms**, **tl**, ur, ckb, din, **km**, kr, lg, ne, om, ps, so, ti-er, ti-et, zu |
| GPT-5.4 | **es-la**, **fr**, **id**, **pt-br**, **my**, rw |
| GPT-5.4 -mini | **fr** |
| **X to E** | |
| Microsoft | **es-LA**, **fr**, **pt-br**, **zh**, **ln** |
| Google | **hi**, **id**, **ru**, **sw**, **bn**, **fa**, **mr**, **ms**, **my**, **ta**, **tl**, **ur**, din, fuv, **ha**, **km**, kr, **ku**, **lg**, **ne**, **prs**, **ps**, rw, **so**, **zu** |
| GPT-5.4 | **ti-et** |
| Gemini | **ar**, **pt-br**, **zh**, **ckb**, **om**, rw, ti-er, **ti-et** |

Table 5: Best System and Usability. Bold indicates a usable system, and underline a borderline one.

| System | | 2023 | 2025 | 2026 |
|---|---|---|---|---|
| Microsoft (XE) | km | 53.70 | 58.69 | 60.67 |
| | prs | 57.44 | 57.45 | 63.11 |
| Microsoft (EX) | ps | 46.64 | 37.84 | 39.47 |
| | ln | 59.00 | 58.19 | 65.59 |
| Google (XE) | ar | 69.61 | 70.87 | 69.17 |
| | so | 57.53 | 37.13 | 59.39 |
| Google (EX) | pt-br | 72.48 | 74.04 | 72.05 |
| | ta | 54.18 | 56.13 | 53.11 |

Table 6: Notable Changes in CHRF++ Scores between 2023, 2025, and 2026 (red indicates drops is quality).

Google showed similar instability in both directions for Pivot and Important languages, which saw small improvements in 2025, but universal drops in 2026. For Priority languages, Khmer and Kurdish Kurmanji showed good (XE) improvements in 2025 and 2026, respectively, while Somali (XE) was particularly unstable: in 2025 BLEU/CHRF++ dropped by 17/20 points, then recovered by 19/22 points in 2026. Scores for most remaining Priority languages were stable in both directions.

#### 5.2.3 Utility of the CHRF++ and BLEU Thresholds

We recognize that the CHRF++ and BLEU thresholds used in this paper, although based on existing literature or rules of thumb, are still somewhat arbitrary. In that vein, we did a limited human evaluation following the scale and methodology described by Picinini and Castilho (2025), with the Likert scale descriptors changed to meet the needs of the crisis scenario (see Appendix I). The evaluation was conducted over a small subset of XE language pairs to determine if the thresholds were reasonable, and fall within the ballpark range. Based on this human evaluation and subsequent statistical analysis, we determined that the com-

bined thresholds were reasonably accurate for the languages we evaluated, although we might lower the threshold for BLEU in the future. The details of the human evaluation, subsequent analysis, and discussion of measuring usability can be found in Appendix H. We save more detailed investigation for future work.

## 6 Discussion

### 6.1 MT Usability by Language

Table 5 shows the highest level of usability (and the best system) for each language. For all Pivot languages, in addition to Malay, Myanmar, and Tagalog, there was at least one usable system to translate both into and out of the language, making bidirectional translation possible. Additionally, for all Important languages and most Priority languages there was at least one system that could translate X to E. In the EX direction, all Important languages had at least a borderline usable system. Priority languages, however, are where we see gaps. With the exception of Lingala and Khmer, no Priority languages had usable EX systems, though many had at least one borderline system. Languages without usable systems in either direction included Dinka, Eritrean Tigrinya, Kinyarwanda, Nigerian Fulfulde, and Kanuri. Dinka was the only language without even a borderline usable XE system.

Claude and GPT-4o were particularly unreliable: for XE translations, they were usable for only 6 and 7 Priority languages respectively (as opposed to 12, 12, 10, 10 and 13 for Microsoft, Google, GPT-5.4, GPT-5.4-mini, and Gemini). Claude was often the only system without usable translations. Despite the quality variations, for high resource languages, all models produced usable, often high quality, translations. However for Priority languages, LLMs were not competitive with the traditional MT systems. Factors beyond quality—such as cost, model features, and content filters—that also impact usability are discussed in Appendix G.

### 6.2 MT Usability by Region

Tin et al., 2024 examined natural disasters globally between 1995 and 2022. In that period, 11,360 natural disasters occurred (on average 350 disasters per year). They further broke the disasters down by type (*e.g.,* hydrological, biological, meteorological, etc.) and by continent. Biological disasters are epidemic-related, and 10.6% of the disasters were biological. However, when we look at the prevalence of biological disasters by continent we see much higher numbers: in Africa, 32.2% of disasters were biological—far exceeding all other continents. Asia was 5.6%, followed by the Americas at 4.6%, then Europe at 2.9%.

If we treat Africa as a priority for pandemic response—the bulk of TwB's Priority languages in TICO-19 are African, and based on Tin et al., 2024, epidemics are far more common on that continent—MT is not ready. Only Swahili, Lingala, and Arabic are truly usable by our metric of usability, assuming the need for bidirectionality. If XE is all that is needed for a particular use case, then about half the African languages are usable. If borderline cases are included, that number is closer to 80%.

For other continents, readiness is much greater. Europe is well covered by the pivot languages (and arguably English in many contexts). Asia is a mixed bag: none of the Priority languages (except Khmer) are usable in the EX direction. However the Pivot and Important languages of Asia (*e.g.,* Hindi, Malay, etc.) are well-covered.

### 6.3 Overall Assessment of Readiness

In crisis response, when lives are often on the line, we must be opportunistic. It does not matter what provider produces a particular system, it only matters whether it will be usable in the context of our need. Our results indicate that Pivot languages are well supported by *all* translation systems, including the LLMs. The quality of these systems is high, and using them to translate content will likely result in understandable results.

For other languages, the task that the MT system would be used for on the ground——specifically whether that task requires bidirectionality—is relevant. For Important languages, the quality of XE translations seems high enough across systems to consider them ready, but in the EX direction, only Malay, Myanmar, and Tagalog have usable translations. All other languages are at best borderline.

Priority languages, for the most part, fall through the cracks. Other than Khmer and Lingala, translations to these languages were *at best* borderline. For the lowest resource languages, even XE translation was problematic (Dinka, Nigerian Fulfulde, Kanuri, Tigrinya)—in fact, only around 50% of Priority languages had fully usable XE translations. For Priority languages, model choice has a stronger impact—dedicated MT systems (particularly Google) have a higher likelihood of producing usable (or at least borderline) output, though

this not a guarantee, *especially* in the EX direction. Given that Priority languages are where translation need is likely to be the greatest—they were identified by TwB based on the situation on the ground in 2020—it would seem that MT is not yet ready to be used in this way.

For regions likely to be hardest hit by biological disasters, whether those be epidemics or pandemics, most notably Africa and Asia, a reliance on MT for higher resource trade or colonial languages may be necessary. However, this reliance may come at a cost in regions where these languages, even if official, are not spoken by a sufficiently large percentage of the population.

Another issue that can affect readiness is changes made by providers over time. These changes are not always positive (as we noted, there were drops from 2023 to 2025, and some drops from 2023 to 2026). It is good to be aware that this kind of instability is possible and can have both positive and (unfortunately) negative impacts on readiness.

## 7 Conclusion

In March 2020, at the start of the pandemic and beginning of the TICO-19 project, most priority languages were not covered by commercial translation engines. By 2025, the coverage was nearly complete. But language support is not enough—it is also important that the translations have sufficient quality to be usable in crises. Further, determining usability in the general domain is not enough; we need usable translations within the crisis (or pandemic) domain. That was the focus of our study.

Is MT ready for the next pandemic? For regions where high-resource languages are spoken, the answer is mostly yes. But there are significant gaps in usability when we look at lower resource languages, especially those spoken in Africa and Asia, regions that arguably are most likely to be hit hardest in the next pandemic (following Tin et al., 2024). There is still much work to be done for the lower-resource languages used in these parts of the world before we can consider MT truly ready.

## Limitations

One major limitation of our work is the absence of human evaluations across most of the languages and directions in our study. Alas, they are expensive to implement, and outside of the budgetary considerations for our project. It would also be difficult to source evaluators for some of the languages in TICO-19.

We also did not use mBERT or equivalent models in our evaluations. We felt that these models are still exploratory, and were concerned about whether the signal we might receive through these models could be relied on.

In doing this study, we translated to and from English. By assuming that we should only be measuring translation quality into and out of English, we have introduced an unfortunate bias: we are assuming that *all* aid providers, governing bodies, NGOs, etc., will be English-speaking. This is not at all true (Doctors without Borders is a prime example of a primarily non-English speaking aid provider), and marks a significant limitation of this work.

Nuer and Congolese Swahili were regrettably excluded from our study.

There were some mismatches between the locales represented in TICO-19 and in the translation services we used. Where locales were listed in the relevant API documentation that coincided with the locales in the TICO-19 dataset, *e.g.,* Brazilian Portuguese, Latin American Spanish, this locale information was passed to the API, *e.g.,* pt-BR, es-LA. For Tigrinya and Fulfulde, neither Microsoft nor Google indicated the locale that was supported by their translation engines, so no locale information was passed to the API. TICO-19 has data for two dialects of Tigrinya (Eritrean and Ethiopian) as well as the Nigerian dialect of Fulfulde. We also could not pass locales to the API when processing these languages.

## Ethical Considerations

All of the TICO-19 data, including the benchmark data we used, is publicly available. The APIs for each of the translation providers, Microsoft and Google, are publicly accessible for a fee. Likewise, for the three LLMs,Claude, GPT and Gemini. All code for processing the TICO-19 data and calling the various APIs, was done by members of our team. We are not aware of any ethical issues that arose while conducting this work.

## Acknowledgements

This work was partially supported by the National Science Foundation under the award CIRC 2346335. We want to thank the two anonymous reviewers for their thorough reviews and many help-

ful comments. Clearly both spent an enormous amount of time reviewing our paper, and we incorporated most of the changes they suggested. We would also like to thank the students of the Spring 2023, 2025 and 2026 LT4CPR graduate seminars at the University of Washington (Priyam Basu, Melody Bechler, Allison Burton, Alexandra Cassell, Jose Cols, Richard Forrester, Chelsea Kendrick, Lily Kurek, Vanesa Marar, Ijeoma Osakwe, Yongsin Park, Anantha Rao, Helen Salgi, and Natasha Schimka) for their helpful input and suggestions. Special thanks also goes to Jen Wilson, whose very closely related paper presented at the same LT4CPR workshop, along with her work with emergency managers across the United States, spurred many changes to this paper, making it more consumable by the very communities who would benefit most from its conclusions and suggestions. It is these communities—emergency managers, aid providers of all stripes, public information managers, communication staff officers, etc.—who were always in our thoughts throughout the work we put into this paper, ultimately spread over several years. It is for you, and the work you do every day saving lives, that we wrote this paper.

## References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. Proceedings of the 1st workshop on nlp for covid-19 (part 2) at emnlp 2020. In *NLP COVID-19 Workshop*, Online.

Ron Artstein. 2017. Inter-annotator agreement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer, Dordrecht.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Online. Association for Computational Linguistics.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual capabilities and knowledge barriers in multilingual large language models. *Preprint*, arXiv:2406.16135.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Federico Federici and Patrick Cadwell. 2018. Training citizen translators: Design and delivery of bespoke training on the fundamentals of translation for New Zealand Red Cross. *Translation Spaces*, 7(1):20–43.

Samsara Foubert, Minjung Park, Kanad Pardeshi, Thomas Manzini, Robert McDaniel, David Merrick, Robin Murphy, Aarti Singh, and Hoda Heidari. 2026. Investigating the role of ai in emergency management: Use cases, challenges, and opportunities. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, page 2055–2091, New York, NY, USA. Association for Computing Machinery.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. *Preprint*, arXiv:2308.08493.

Google Cloud. 2025. Evaluate models | cloud translation - google cloud. Accessed: 2025-05-25.

GoogleCloud. 2023. Evaluate models. Accessed on May 11, 2023.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors. 2025. *Proceedings of the Tenth Conference on Machine Translation*. Association for Computational Linguistics, Suzhou, China.

Amr Hendy, Vikas Raunak Mohamed Gabr Hitokazu Matsushita Young Jin Kim Mohamed Afify Mohamed Abdelrehim, Amr Sharaf, and Hany Hassan Awadalla. 2023. How Good are GPT Models at Machine Translation? a Comprehensive Evaluation. arXiv:2302.09210.

Matthew Hunt, Sharon O'Brien, Patrick Cadwell, and Dónal P O'Mathúna. 2019. Ethics at the intersection of crisis translation and humanitarian innovation. *Journal of Humanitarian Affairs*, 1(3):23–32.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4).

Elaine C. Khoong and Jorge A. Rodriguez. 2022. A Research Agenda for Using Machine Translation in Clinical Medicine. *Journal of General Internal Medicine*, 37(5):1275–1277.

Elaine C. Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine*, 179(4):580–582.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Marin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1—46, Miami, Florida. Association for Computational Linguistics.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2026. Evaluating extremely low-resource machine translation: A comparative study of ChrF++ and BLEU metrics. *Preprint*, arXiv:2602.17425.

Séamus Lankford and Andy Way. 2024. Leveraging LLMs for MT in crisis scenarios: a blueprint for low-resource languages. arXiv:2410.23890.

William Lewis. 2010. Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

William Lewis, Haotian Zhu, Keaton Strawn, and Fei Xia. 2025. Tapping into Social Media in Crisis: A Survey. In *Proceedings of the 4th Workshop on NLP for Positive Impact, EMNLP 2025*, vienna, Austria.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

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.

Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jonathan Washington, Nathaniel Oco, and Xiaobing Zhao, editors. 2026. *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*. Association for Computational Linguistics, Rabat, Morocco.

Sharon O'Brien and Federico Federici. 2019. Crisis translation: Considering language needs in multilingual disaster settings. *Disaster Prevention and Management*, 29(2):129–143.

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.

Silvio Picinini and Sheila Castilho. 2025. Context-aware monolingual evaluation of machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 138–149, Geneva, Switzerland. European Association for Machine Translation.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Machine Translation*.

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

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, Brussells, Belgium.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Dimitrios Roussis. 2022. Building End-to-End Neural Machine Translation Systems for Crisis Scenarios: The Case of COVID-19.

Elizabeth Salesky, Antonios Anastasopoulos, Matteo Negri, and Marcello Federico, editors. 2026. *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*. Association for Computational Linguistics, San Diego, USA (in-person and online).

Gongbo Tang, Oreen Yousuf, and Zeying Jin. 2024. Improving bertscore for machine translation evaluation through contrastive learning. *IEEE Access*.

D. Tin, L. Cheng, D. Le, R. Hata, and G.Ciottone. 2024. Natural disasters: a comprehensive study using EM-DAT database 1995–2022. *Public Health*, 226:255–260.

Jen Wilson, Vipasha Bansal, and William D. Lewis. 2026. Communicating with the Entire Population: Machine Translation of Emergency Alerts and Warnings. In *Proceedings of the Language Technologies for Crisis Preparedness and Response (LT4CPR Workshop)*, Hengqin, Zhuhai, China. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Andy K. Zhang, Kevin Klyman, Yifan Mai, Yoav Levine, Yian Zhang, Rishi Bommasani, and Percy Liang. 2025. Language model developers should report train-test overlap. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

# Appendices

## A 2026 Full BLEU, CHRF++, BERTScore, and COMET Results

Tables 15-22 show the full BLEU, CHRF++, BERTScore, and COMET[17] results for all seven systems evaluated in 2026. Figures 2 and 3 plot all 2026 BLEU and CHRF++ scores against one another, with each threshold marked. These tables and figures can be found at the end of the appendices.

## B BLEU, CHRF++, BERTScore, and COMET Results for 2023 and 2025 for Microsoft and Google

Tables 23, 24, 25, and 26 show the full BLEU, CHRF++, BERTScore, and COMET results for Microsoft and Google from 2023 and 2025, in both translation directions. These illustrate the change in the performance of these two systems during this period. These tables can be found at the end of the appendices.

## C Model Configurations

Due to time and funding constraints, reasoning_effort was set to 'none' for both GPT-5.4 checkpoints. GPT-4o, Gemini, and Claude were run with temperature=0. All systems were given the following, identical prompt: "Translate the following text from {source_language} to {target_language} in one line without any transliterations or explanations".

## D SacreBleu Signatures

Table 7 shows the metric signatures for the BLEU and CHRF++ (and CHRF) metrics, run using SacreBleu. (1) shows the BLEU signature for Khmer and Myanmar in the EX direction, (2) shows the BLEU signature for Chinese in the EX direction, (3) shows the CHRF signature for Khmer, Myanmar, and Chinese in the EX direction, while (4) and (5) show the BLEU and CHRF++ signatures respectively for all remaining languages in the EX direction and all languages in the XE direction. These are the same for all longitudinal data as well.

[17] Only supported languages are included in COMET results. The list of supported languages can be found here: github.com/Unbabel/COMET

| | |
|---|---|
| (1) | nrefs:1/case:mixed/eff:no/tok:char/smooth:exp/version:2.3.1 |
| (2) | nrefs:1/case:mixed/eff:no/tok:zh/smooth:exp/version:2.3.1 |
| (3) | nrefs:1/case:mixed/eff:yes/nc:6/nw:0/space:no/version:2.3.1 |
| (4) | nrefs:1/case:mixed/eff:no/tok:13a/smooth:exp/version:2.3.1 |
| (5) | nrefs:1/case:mixed/eff:yes/nc:6/nw:2/space:no/version:2.3.1 |

Table 7: Metric Signatures

## E LLM Translation Failures

There were a number of translation failures across the LLMs; these took on a number of different forms, depending on the model. The most common forms were blank output and 'refusals', and occurred almost exclusively when translating to and from Dinka, Nigerian Fulfulde, and Kanuri. The proportions for both of these error types are shown in Table 8.

| | System | Error Type | Number |
|---|---|---|---|
| | | **E to X** | |
| din | Claude | "Refusal" | 943 (45%) |
| | Gemini | Blank Output | 1154 (55%) |
| | GPT-4o | "Refusal" | 498 (24%) |
| fuv | Claude | "Refusal" | 83 (4%) |
| | Gemini | Blank Output | 80 (4%) |
| kr | Claude | "Refusal" | 773 (37%) |
| | Gemini | Blank Output | 1649 (79%) |
| | GPT-4o | "Refusal" | 187 (9%) |
| | | **X to E** | |
| din | Claude | "Refusal" | 1300 (62%) |
| | GPT-4o | "Refusal" | 73 (3%) |
| | | Blank Output | 237 (11%) |
| fuv | Claude | "Refusal" | 385 (18%) |
| | GPT-4o | Blank Output | 81 (4%) |
| kr | Claude | "Refusal" | 731 (35%) |
| | GPT-4o | "Refusal" | 154 (7%) |

Table 8: LLM Translation Failures

A 'refusal' was output that overtly stated that the model was unable to translate to or from the requested language. Claude and GPT-4o did this in both translation directions for these three languages. An example of a refusal from GPT-4o is shown in (a) below. Claude's 'refusals' tended to be more descriptive, with examples shown in (b) and (c).

1. (a) "My apologies, but I can't provide a translation for that text."

   (b) "I don't have reliable knowledge of Dinka language to provide an accurate translation of this phrase. Dinka is a Nilotic language spoken in South Sudan, and without verified linguistic resources,

I cannot confidently translate this line without risking significant errors."

(c) "I don't have reliable knowledge of the Dinka language to provide an accurate translation of this text. Dinka is a complex language with specific grammatical structures and vocabulary that requires expertise to translate correctly. I'd recommend consulting a Dinka language expert or speaker for an accurate translation."

Gemini did not 'refuse' to translate any sentences; instead it failed to produce any output at all for a large number of sentences in the EX direction specifically. GPT-4o was more variable, producing both 'refusals' and blank output in both directions.

In some cases, we also saw random string repetitions in the output, particularly in the EX direction. Table 9 shows examples of this from GPT-4o and Claude when translating into Nigerian Fulfulde.

## F Data Contamination

A significant issue in assessing commercial MT systems and LLMs is the risk of training data contamination. This occurs when evaluation benchmarks or test sets, or a dataset from which they are drawn, are used as training data for models, leading to inflated evaluation scores and misleading conclusions about the model's real generalization capabilities (Golchin and Surdeanu, 2024). This is especially concerning when benchmarking in zero-shot settings (specifically for LLMs) where there is no fine-tuning for task-specific data. If near perfect or unrealistically high scores are achieved, there is the potential for training data contamination, resulting in falsely reported numbers due to memorization instead of generalization.[18]

Although it is not possible for us to tell whether any particular provider has trained on the TICO-19 test data, some of the highest BLEU scores we see in our results *may* indicate some overlap. MT providers typically do not train on test or dev data (to preserve the value of test data over time), but LLM providers do not necessarily follow this stricture. That said, since some of the TICO-19 data is drawn from public sources, it is possible that MT providers trained on these same sources, thus inadvertently introducing overlap. We hypothesize that this overlap would be possible, and potentially evident, with certain portions of the TICO-19 benchmark dataset, specifically for those sources or domains which are publicly available. In opposition to this, we suspect non-public sources would not show this effect.

Table 10 provides potential evidence for this hypothesis. For Spanish, the Scientific Medical (SM), General (G), and News (N) domains show exceptionally high BLEU scores across the systems (excepting GPT-4o). The BLEU scores in the high 50s to 60s suggest potential overlap. It is completely possible that the sources of these data (Wikipedia, Wikinews, even PubMed) had been consumed and trained on by these engines (potentially even without the providers having had to translate content, *e.g.,* PubMed content is translated to Spanish, as are the other data sources). Notably, the Conversational Medical (MD) data, which comes from CMU, and is not public, does not show such large BLEU scores. Also note that the language Farsi, which is lesser resourced and less likely to be translated into, also has lower scores, and the relative differences in scores between domains is also lower.

There are confounds here, of course, that might only be teased apart through further analysis. First, the number of sentences in each TICO-19 subdomain is quite small (domain labels and counts are shown in Table 2). This could lead to significant variability in the BLEU scores calculated. Likewise, Spanish is much higher resource than Farsi, so just the availability of resources for the former would lead to much larger BLEU scores (the higher scores being indicative of higher quality not contamination). But would these scores be in the 50-60 point range?

Overlap such as that described previously certainly plagues any evaluation data drawn from public sources, *e.g.,* news, social media, etc., since nothing would constrain MT providers from drawing from the same sources, potentially translating them, and adding them into their training data. Overtly adding test data into training is frowned on, but not completely avoided (Zhang et al., 2025). A thorough analysis of data contamination (and the TICO-19 domains) will be left to future research.

[18] As noted by one of the reviewers, high BLEU scores do *not*, in and of themselves, necessarily indicate data contamination; the systems being measured could just be really good. However, based on our experience with developing MT systems, when inflated BLEU scores were observed for particular systems or domains but not others, we suspected data contamination.

| **GPT-4o** | **— English to Nigerian Fulfulde** |
|---|---|
| **Source** | and can you tell me what other symptoms are you having along with this? |
| **Reference** | e footi a limta mandolji nyau feere je a don mari fahin na? |
| **Output** | Mido waawi wadugo dum, ko dum wadi dum didi dum wadi dum wadi dum wadi dum wadi dum wadi dum wadi dum. |
| **Claude** | **—English to Nigerian Fulfulde** |
| **Source** | These enzymes are expressed extensively in multiple organs, especially in the heart and liver, and are released during tissue damage. |
| **Reference** | Neebamji ndu don matinama ha nder kujeji baandu jur, burna ha nder bernde be hengre, e wurtan to bana laaral man seeki. |
| **Output** | Enzimoji didi wadi ko'e di-di ko'e ko'e ko'e ko'e ko'e ko'e ko'e [...] ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko'e ko |

Table 9: Example string repetition errors in GPT-4o and Claude output when translating from English to Nigerian Fulfulde.

| | **Domain** | **Microsoft** | **Google** | **GPT-4o** | **GPT-5.4** | **GPT-5.4-mini** | **Gemini** | **Claude** |
|---|---|---|---|---|---|---|---|---|
| | | | | **English to X** | | | | |
| es-LA | CM | 32.62 | 38.80 | 33.03 | **45.16** | 41.44 | 36.18 | 34.5 |
| | SM | 53.90 | **59.04** | 55.45 | **59.10** | 56.87 | 57.42 | 54.61 |
| | G | 51.16 | 55.42 | 53.36 | **56.25** | 55.09 | 53.19 | 52.09 |
| | N | 56.19 | 54.85 | 59.51 | **60.63** | 59.55 | 58.17 | 58.67 |
| | A | 40.83 | 42.58 | 41.52 | 43.89 | **44.10** | **44.40** | 41.51 |
| fa | CM | 18.81 | 25.24 | **31.40** | 28.29 | 23.41 | **31.81** | 20.7 |
| | SM | 26.53 | **33.72** | 29.01 | 26.62 | 24.34 | 26.67 | 26.32 |
| | G | 25.19 | **28.27** | 25.52 | 22.56 | 22.92 | 23.83 | 22.26 |
| | N | 21.70 | **28.18** | 23.94 | 20.02 | 19.77 | 22.45 | 18.94 |
| | A | **28.16** | 26.99 | 21.58 | 19.28 | 19.80 | 19.46 | 19.00 |
| | | | | **X to English** | | | | |
| es-LA | CM | 35.17 | 29.52 | 29.84 | **36.01** | 35.06 | 30.94 | 29.12 |
| | SM | **55.10** | 54.67 | 54.22 | 53.71 | 54.52 | **55.34** | 52.80 |
| | G | **55.19** | 53.74 | 52.77 | 53.55 | 52.47 | 54.04 | 50.26 |
| | N | **56.31** | 50.88 | 53.14 | 52.57 | 52.88 | **56.83** | 50.41 |
| | A | **48.21** | 44.12 | 46.37 | 46.61 | 44.58 | **48.36** | 42.29 |
| fa | CM | **27.35** | 25.25 | 24.92 | 26.63 | 26.24 | 23.68 | 22.70 |
| | SM | 40.49 | **45.87** | 40.84 | 41.64 | 42.18 | 42.65 | 30.00 |
| | G | 39.01 | **43.62** | 38.63 | 40.33 | 39.49 | 41.93 | 35.24 |
| | N | 34.88 | 37.60 | 35.51 | 35.70 | 35.00 | **39.03** | 32.68 |
| | A | 39.17 | **43.05** | 37.49 | 37.94 | 38.60 | 40.89 | 34.69 |

Table 10: Domain BLEU Scores for Spanish (LA) and Farsi. Bold text indicates best system within each specific domain. Legend for domains: Conversational Medical (CM), Scientific Medical (SM), General (G), News (N), Announcements (A).

## G Factors Beyond Quality That Can Affect Usability

In addition to output quality, there are a number of practical factors, such as cost, runtime, and content filters, that can impact usability, particularly when utilizing LLMs. GPT-5.4-mini is a smaller version of the full size GPT-5.4 model and is therefore faster and cheaper. One might assume that the smaller of the two would show lower quality, yet our results show that the two systems have comparable scores. There are many languages, such as EX Hindi, where the mini actually outperforms the full size model. One notable example is Kurdish Kurmanji: the GPT5.4 full size model (and Gemini) produced output in Kurdish Sorani instead, which does not even share the same script. GPT-5.4-mini, however, produced translations in the correct language and orthography. This of course is one more element in the usability of LLMs—they may produce unexpected or unwanted output, particularly in low resource languages.

From a usability standpoint then, the GPT-5.4-mini is much more cost effective. However, despite its being a reasoning model, we very quickly found that the model was not usable for translating into low resource languages unless we set reasoning to 'None'. When active, reasoning caused the model to be excessively slow and use a large amount of tokens, resulting in blank output combined with much higher costs. Therefore, one of the main features of the model, *i.e.,* reasoning, cannot easily be applied in the crisis scenario. Still, reasoning is not the only issue at play here: when compared to an older, but more widespread, non-reasoning model,

GPT-4o, GPT-4o consistently produced lower quality translations.

Another factor we encountered was OpenAI's content filtering. Many of the sentences in the dataset incorrectly triggered various content filters, such as 'hate speech'. This mostly occurred when translating from a low resource language *into*. It is unclear if this is because the model is producing mistranslations or hallucinations that are, in fact, hate speech (see Appendix J for examples), or if it is the content filter itself that is not performing well on certain languages. These issues severely restrict the usability of all three GPT models in the medical and crisis domain, even if the output translations are good quality; good quality is not useful if the model refuses to provide output for certain crisis-related input in the first place. We did not observe this issue with either Claude or Gemini. The proportion of sentences that were filtered in the GPT models for each language are shown in Table 11.

As outlined in Section 4.2, our usability analysis treated filtered outputs as blank or empty when scoring across metrics. From a practical standpoint, the two are the same: in both cases, the user does not receive a translation. However, in terms of evaluating translation quality specifically, it is also useful to look at the scores with the filtered sentences excluded. These are shown in Table 12. In many cases—such as all three GPT models' performance on Luganda—the difference in scores is substantial.

| | GPT-4o | GPT-5.4 | GPT-5.4-mini |
|---|---|---|---|
| ckb | 427 (20%) | 362 (17%) | 431 (21%) |
| din | — | 598 (28%) | 707 (34%) |
| fuv | — | 91 (4%) | 558 (27%) |
| ha | 112 (5%) | 107 (5%) | 111 (5%) |
| kr | 406 (19%) | — | 408 (19%) |
| lg | 381 (18%) | 383 (18%) | 386 (18%) |
| ln | 578 (28%) | 574 (27%) | 563 (27%) |
| om | — | 153 (7%) | 161 (8%) |

Table 11: GPT Content Filtering Rates (XE only)

Finally, a known issue with LLMs is hallucinations—we observed this across the models we evaluated, and, unsurprisingly, hallucinations were much more frequent and more severe in low resource languages. Unfortunately, many of these hallucinations produced output that was racist, violent, or potentially contained misinformation, and would have consequences far beyond poor quality translation, thus further affecting the model's usability. Fortunately, frequent hallucinations often result in low BLEU/CHRF++ scores (which means that thresholding against these scores can potentially identify output containing a high number of hallucinations, at least for LLMs).

| | GPT-4o | GPT-5.4 | GPT-5.4-mini |
|---|---|---|---|
| ckb | 28.17/51.77 | 34.04/58.65 | 32.95/57.63 |
| din | — | 9.35/26.94 | 9.54/27.20 |
| fuv | — | 17.84/37.31 | 17.43/36.29 |
| ha | 34.13/56.36 | 38.49/60.43 | 37.00/58.86 |
| kr | 11.95/29.64 | — | 11.64/29.28 |
| lg | 31.97/53.59 | 34.72/56.82 | 32.58/54.97 |
| ln | 26.18/48.16 | 28.28/50.73 | 26.46/48.67 |
| om | — | 32.14/55.99 | 29.94/53.75 |

Table 12: BLEU/CHRF++ Scores for GPT Models with filtered sentences excluded (XE only).

## H Human Evaluation and CHRF++ and BLEU Thresholds

We performed a limited, blind human evaluation on seven samples of output produced by two MT systems (Microsoft and Gemini), when translating into English from four languages ((Dinka, Eritrean Tigrinya, Lingala, Hindi). The evaluation was done only in the XE direction, with the primary focus on establishing the utility of the CHRF++ and BLEU thresholds discussed in Sections 2.2 and 4.3.

The four languages were selected based on output quality—we wanted to ensure that a range of BLEU/CHRF++ scores were represented in our analysis. Output translations from Dinka (Gemini only) were unusable, while those from Eritrean Tigrinya were borderline for both systems. Lingala output was borderline for Gemini and just about passed the usability thresholds for Microsoft, and both systems had high performance for Hindi.

The human evaluation was conducted by four evaluators; each evaluator received 42 reference-translation pairs from each of the seven sets of output (one for each system and language.)[19]

In order to establish the level of inter-annotator agreement, each evaluator received the same 42 sentences from the Dinka-English and Hindi-English outputs. For the four remaining outputs (Eritrean Tigrinya-English and Lingala-English from both Gemini and Microsoft), each evaluator was assigned a distinct set of 42 sentences. They then scored reference-translation pairs (of the same 42 sentences) across each of the four sets of output. Therefore, a total of 168 distinct sentence pairs

[19] Dinka was unsupported by Microsoft.

were evaluated for each of these four systems-language pairs. Given the high inter-annotator agreement established by Dinka-English and Hindi-English evaluations (to be presented shortly), this decreased the margin of error from 14.97% (if all evaluators scored the same 42 sentence pairs) to 7.33% (168 distinct sentence pairs scored).

After all four language pairs were sampled (with a total of 294 sentence pairs from across the seven system-language sets assigned to each evaluator), the sentences were placed in a random order, and the evaluators reviewed the English reference and English MT output (we assumed that none of the evaluators could speak the source languages, so the input sentences were not presented). The evaluators were asked to evaluate the translated output using the modified Picinini and Castilho (2025) Likert scale shown in Appendix I, and assign an integer score ranging from 1-5. Based on this scale, we considered a score of 3 as the threshold for usable translation in the crisis scenario.

As different numbers of distinct sentences were evaluated for various outputs, a statistical analysis assuming maximum variability (p=0.5) indicates varying margins of error at a 95% confidence level. We established inter-annotator agreement using evaluator scores for Hindi-English (Gemini and Microsoft) and Dinka-English (Gemini), (since, as noted, all evaluators scored the same sentences across each of these outputs). Krippendorff's Alpha over all evaluations of these three extracts is 0.7925, suggesting good inter-annotator agreement (Artstein, 2017). Tables 13 and 14 show the results of the human evaluation, including margins of error.

At a 95% confidence level, Hindi-English (Microsoft and Gemini), Lingala-English (Microsoft and Gemini), and Eritrean Tigrinya-English (Gemini only) met the human evaluation threshold of 3, while Dinka-English (Gemini) did not. At a 95% confidence level, it is inconclusive whether Eritrean Tigrinya-English (Microsoft) met the threshold (given the corresponding margin of error).

We ran an Ordinary Least Squares regression to establish a linear relationship between BLEU and our human evaluations (Pearson r = 0.9411 with p = 0.0016) and between CHRF++ and our human evaluations (Pearson r = 0.9703 with p = 0.0003), subsequently identifying the BLEU and CHRF++ values corresponding to a human evaluation of 3.0 from the linear regression. The results suggest thresholds of 23.49 for BLEU and 45.11 for CHRF++. Thus, the results of the linear regression suggest lowering the BLEU score threshold, while keeping a similar CHRF++ threshold.

Changing the BLEU threshold to 23.49 in the XE direction would suggest that a number of systems and languages currently classified as borderline would actually be considered usable. No languages classified as fully unusable would change in their usability categorization. However, with only seven data points, this is a preliminary analysis; it suggests the thresholds we used in our analysis are within a reasonable range of accuracy. A full threshold sensitivity analysis in both translation directions along with additional human evaluation (particularly in the EX direction) is warranted and would be a valuable area for future research.

Overall, the strong linear correlation of BLEU and CHRF++ with our human evaluation suggests that these automatic metrics may be good predictors of minimum usability as we defined it. Exceeding these thresholds suggests higher degrees of usability.

## I Human Evaluation Scale

The Likert scale described by Picinini and Castilho (2025) used in our human evaluation, with our modifications for the crisis scenario italicized, can be found below.

1. Critical/Incomprehensible:
   When the translation:
   - Is incomprehensible (hallucinations, meanings that make no sense in the text),
   - Is completely not in the target language,
   - Contains critical errors,
   - Can be a misleading mistranslation,
   - May carry health, safety, legal, reputation, religious or financial implications,
   - Contains profanities/insulting words not appropriate in the context.
2. Significant errors/effort:
   Translation contains significant errors and may be incomplete (such as significant mistranslations, significant portions untranslated, significant omissions) and requires significant effort to understand it. It would require a significant amount of change to be usable.
3. Some errors/some effort:
   There are errors in translation, such as untranslated words and mistranslations (wrong

| Language | X to English | | | |
|---|---|---|---|---|
| | Microsoft | Gemini | Margin of Error | % Margin of Error |
| Hindi | 4.5298 | 4.7679 | 0.5988 | 14.97% |
| Dinka | — | 1.8453 | 0.5988 | 14.97% |
| Lingala | 3.5536 | 3.5774 | 0.2932 | 7.33% |
| Tigrinya (Eritrean) | 3.0774 | 3.7262 | 0.2932 | 7.33% |

Table 13: 2026 XE average human evaluation scores and margin of error for all seven evaluated sets. See Appendix I for scale used in evaluation.

| **Language** | **% Critical Errors (score of 1)** | | **% Usable Sentences (score of 3-5)** | |
|---|---|---|---|---|
| | **Microsoft** | **Gemini** | **Microsoft** | **Gemini** |
| **Hindi** | 0.0/0.0/0.0/2.4 | 0.0/0.0/0.0/2.4 | 97.6/95.2/ 100.0/95.2 | 100.0/100.0/ 100.0/97.6 |
| **Dinka** | — | 59.5/54.8/ 52.4/38.1 | — | 19.1/19.1/ 19.1/33.3 |
| **Lingala** | 7.1 | 7.1 | 75.6 | 81.0 |
| **Tigrinya (Eritrean)** | 11.3 | 6.6 | 65.5 | 85.1 |

Table 14: Proportion of sentences that were rated as having a critical error and the proportion that were considered usable by human evaluators for each language and system. For Dinka and Hindi, where each evaluator had the same set of sentences, the proportions for each evaluator's responses are shown separately.

but not misleading translations). The translation requires some effort to understand it and change it. *The translation would be partially or borderline usable in the crisis context (*e.g., *the necessary information would be conveyed and understood), but might require subsequent verification.*

4. Do not affect understanding:
   There are minor grammar errors: wrong agreement, word or der, missing prepositions, minor omissions or other errors that do not affect understanding. It would require only minor changes. *The translation would still be usable in the crisis context (*e.g., *the necessary information would be conveyed and understood).*

5. Complete/fluent:
   Translation is (or seems to be) correct, complete, fluent, easy to read, and contains no errors. It looks like (almost) human translation and requires no changes. *The translation would be usable in the crisis context and would be fully understood.*

It is important to note here, that the use of the word 'borderline' for entry 3 above differs from our use of the word in our overall analysis. As described previously, we defined systems that were 'borderline usable' as producing output translations that scored higher than the threshold for one metric but not both, indicating that they were almost, but not quite, of high enough quality to be usable. In this human evaluation scale however, we use 'borderline' to mean just barely usable—the translation is low quality, and if it were reviewed by a human post-editor, they would likely make changes. This distinction was unintentional and was a source of confusion for at least one of our reviewers. We regret this error and would fix this in future human evaluation work.

## J Qualitative Error Analysis and Hallucinations

We also conducted a preliminary error analysis on LLM output translations for a subset of languages in the XE translation direction. We looked first at some of the lowest scoring systems, (which had BLEU scores of around 10). We then looked at Gemini's output when translating from Nigerian Fulfulde to English—while scores here were still too low to be usable, they were much closer to reasonable (BLEU/CHRF++ were 21.21/42.42). Finally, we examined some of the errors produced by systems that were considered borderline usable.

In the very low performing languages—Dinka and Kanuri consistently fell into this category across models—hallucinations were prevalent. The conversational medical domain was particularly

prone to this; in many cases, the output for the entire domain is fully hallucinated. We identified several types of hallucinations: output that was completely irrelevant or random, output that was unrelated to the source sentence but on topic, and output where the source sentence was translated, but with extra, hallucinated content included. Many of the hallucinations included religious themes, and in a few instances, centered around one key word in the source. Unfortunately, there were also many hallucinations (or severe mistranslations) that were problematic, containing content that was racist, sexist, actively spreading misinformation, or were otherwise aggressive or violent. Examples of various hallucinations are shown in 1-7 below:

1. **Irrelevant Output, GPT-5.4-full, Kanuri-English**

   **Reference:** extensions of monitoring and surveillance would have to have sunset clauses;

   **Output:** The people have ten fingers and ten toes, and all of them are counted together.

2. **Extra Hallucinated Content, Gemini, Dinka-English**

   **Reference:** As of 19 March drive in tests were offered in several large cities.

   **Output:** On the 19th of the month of Kol, many people were tested for the disease, and *they were all remembered in prayer.*

3. **Religious Hallucinations, GPT-4o, Nigerian Fulfulde-English**

   **Reference:** since you have a fever

   **Output:** Prepare your mat, we are about to pray.

4. **Key Word Based Hallucination, Gemini, Kanuri-English**

   Reference: *Panama* has made it obligatory to wear a face mask whenever going outside, while also recommending the manufacture of a homemade face mask for those who cannot purchase face masks.

   **Output:** *The Panama Papers* revealed how the wealthy hide their assets within the community, and also how various types of shell companies are used to consult with those who wish to evade taxes.

5. **Problematic Hallucination: Aggressive, GPT-5.4-mini, Dinka-English**

   **Reference:** Patients with fever and/or respiratory symptoms and acute fever, even without pulmonary imaging abnormalities, should be screened for the virus for early diagnosis.

   **Output:** People are lazy, foolish, and full of anger and hatred, so if they are not careful, they will end up making mistakes and being deceived.

6. **Problematic Hallucination/Mistranslation: Violent, Gemini, Dinka-English**

   **Reference:** "after patient care.The addition of antiseptic chemicals to soap (‘medicated’ or ‘antimicrobial’ soaps) confers killing action to a hand-washing agent."

   **Output:** If there is no one sick in Akoor to be helped with soap, then they should be killed immediately.

7. **Problematic Hallucination/Mistranslaion: Spreads Misinformation, Gemini, Dinka-English**

   **Reference:** Other types of respirators can provide greater protection and improve worker comfort.The WHO does not recommend coveralls, as COVID-19 is a respiratory disease rather than being transmitted through bodily fluids.

   **Output:** Some of the best ways to help save lives have not been implemented, and the World Health Organization has not yet approved the use of masks to prevent the spread of COVID-19.

Other common errors among the lowest performing outputs include: difficulty with medical terminology, archaic word choice, meaning of translation is the opposite of the intended/original meaning, and unusual phrasing. Some examples are shown in 1-4 below.

1. **Medical Terminology, Gemini, Dinka-English**

   **Reference:** and with your history of *diabetes*

   **Output:** What is the name you have been given in the *sugar*?

2. **Medical Terminology, Gemini, Dinka-English**

   **Reference:** Instead, they may cause severe side effects, especially avascular *osteonecrosis*, dramatically affecting the prognosis.

   **Output:** There is something that is done in the body, just as death is the *bone of death*, and it is in that which is found to be bad

3. **Archaic/Unusual Phrasing, Gemini, Dinka-English**

   **Reference:** Many different hand dryer manufacturers exist, and hand driers have been compared against drying with paper towels.

   **Output:** The people of the village of the hand did not do it, but they were found in the village of the hand wearing clothes that covered them well.

4. **Opposite Meaning, Gemini, Kanuri-English**

   **Reference:** In addition, Turkey seized hundreds of ventilators destined for Spain.

   **Output:** "Because of that, Turkey is sending the medical equipment called a 'ventilator to' Spain, which has been hit hard by the world pandemic."

We then turned to Gemini's output when translating from Nigerian Fulfulde to English. As noted earlier, the BLEU/CHRF++ scores were still too low to be considered usable, but more reasonable than those for Dinka or Kanuri. In general, we observed the same error types as with the very low performing languages, but with fewer numbers of problematic outputs and errors that did not seem as severe. The conversational medical domain was not fully hallucinated here; instead, the medical terms specifically were hallucinated or mistranslated. The mistranslations tended to be consistent (*e.g.,* "frog" was present in several translations where the reference sentence referred to a "fever"). Some of the mistranslations from this output could be inferred from context (*e.g.,* "shortness of breath" in the reference sentence aligned with "breath runs out" in the translation). However, there are still some extremely problematic mistranslations and hallucinations present in the output. See 1-5 in the list below for examples.

1. **Poor Quality, Mistranslation, Gemini, Nigerian Fulfulde-English**

   **Reference:** are you having any shortness of breath?

   **Output:** Is it okay if your breath runs out?

2. **Poor Quality, Gemini, Nigerian Fulfulde-English**

   **Reference:** are there other people sick as you at home with your same symptoms?

   **Output:** Are there other people in your town who are sick with the same kind of illness as you?

3. **Medical Terminology, Gemini, Nigerian Fulfulde-English**

   **Reference:** I've very concerned of this chest pain

   **Output:** My heart is aching and I am very troubled.

4. **Opposite/Untrue Meaning, Gemini, Nigerian Fulfulde-English**

   **Reference:** The White House Coronavirus Task Force was established on 29 January.

   **Output:** The White House will release the coronavirus report on January 29th.

5. **Problematic Mistranslation: Spreads Misinformation, Gemini, Nigerian Fulfulde-English**

   **Reference:** Several measures are commonly used to quantify mortality.

   **Output:** Many tricks are being used to inflate the death toll.

Finally, we turn to systems that produced borderline usable translations, focusing on various Lingala-English and Oromo-English outputs across systems. Translations were significantly better than the lower-performing, unusable, systems, especially in the conversational medical domain. In many cases, medical or technical terms were simplified, but not necessarily incorrect. We also see mismatches in tone between the reference and output (or some added dramatization), and translations that were almost or partially correct (*i.e.,* some parts were conveyed correctly, but some information was mistranslated or omitted). We do still note

some archaic phrasing or word choices, and some problematic mistranslations. Some examples are shown in 1-10 in the list below.

1. **Simplifying Technical Terms, Gemini, Lingala-English**

   **Reference:** digital surveillance would have to avoid exacerbating discrimination and marginalisation;

   **Output:** Computer-based monitoring should not be done with malice toward others.

2. **Simplifying Technical Terms, GPT-5.4-mini, Lingala-English**

   **Reference:** Patients often develop an atypical pneumonia, acute lung injury, and acute respiratory distress syndrome (ARDS).

   **Output:** Often, sick people had lung disease, severe lung inflammation, and severe breathing difficulty (ARDS).

3. **Tone Difference or Dramatization, Gemini, Lingala-English**

   **Reference:** But it's almost impossible to make any sensible projection right now about when that will be.

   **Output:** But there is no way to know how things will be these days.

4. **Tone Difference or Dramatization, GPT-5.4-full, Oromo-English**

   **Reference:** *The measure was taken* to enforce quarantine and protect those who may come into contact with infected citizens.

   **Output:** *The punishment is imposed* to influence the duration of isolation and also to prevent people from touching citizens infected with the disease.

5. **Tone Difference or Dramatization, Gemini, Lingala-English**

   **Reference:** Many consider an economic recession to be likely.

   **Output:** Many people think that the economy will truly collapse.

6. **Archaic Phrasing, GPT-5.4-mini, Lingala-English**

   **Reference:** *As of March 19*, these three had already recovered.

   **Output:** *On the 19th day of the third month*, three people survived.

7. **Archaic Word Choice, GPT-5.4-mini, Lingala-English**

   **Reference:** *Doctors* have said they are using the drug when "there's no other option."

   **Output:** The *healers* say they use that medicine only when there is no other medicine.

8. **Partially Correct, GPT-5.4-full, Oromo-English**

   **Reference:** The study failed to find evidence of viral replication in *pigs, ducks,* and chickens.

   **Output:** The study could not find evidence that the virus reproduces in *sheep, donkeys,* and chickens.

9. **Problematic Mistranslation, Gemini, Lingala-English**

   **Reference:** Women who are pregnant or breastfeeding or who do not have an effective contraceptive method *are not eligible*.

   **Output:** Women who are pregnant, breastfeeding, or not using effective contraception *are not permitted to enter*.

10. **Mistranslation, Gemini, Lingala**

    **Reference:** Pakistan has said that it will *not be evacuating any citizens* from China.

    **Output:** Pakistan said that it would *launch its own person into space* with China.

## K The Choice of Evaluation Metric(s)

The choice of evaluation metric(s) is always an issue in a study such as ours. While human evaluation is the gold standard, it is both time consuming and expensive to implement (Bugliarello et al., 2020). In fact, expense aside, locating reviewers for many of the lower resource languages in TICO-19 would be difficult at best. As an example, translators and reviewers were a significant problem for these languages during the development of TICO-19 a few years ago.

Still, the choice of evaluation metric is important and should not be taken lightly. Yet, the options available for a study such as ours are limited, given the languages that we focused on.

On the surface, given the underlying technology used for translation today, neural metrics such as BERTScore or COMET would make the most sense. We knew this going in and made sure that we calculated scores for these two metrics for almost every output across our three-year study. But we could not use these metrics for all data points. Alas, these metrics do not work well for lower-resource languages and specialized domains (as noted in Section 2.2, per (NLLB-Team et al., 2022; Tang et al., 2024; Falcão et al., 2024; Zouhar et al., 2024a,b)). Further, there is no clear consensus on how to interpret these metrics directly.

BLEU has had a long and, dare we say, sordid past in the field of Machine Translation. It is a metric that everyone loves to hate, and most would freely admit that it has outlived its shelf-life. That said, it still manages to cling to life, and is still used widely in the field, even in industrial settings (*e.g.,* for hill climbing during development). Why won't BLEU just die already? The reasons vary greatly, but we feel it is mostly due to the fact that it is cheap and easy to implement, that it is universal (*i.e.,* it can be applied to *any* language pair), that we understand the scores and what they mean, and that it provides some sort of continuity across the lengthy history of our field.

In an earlier version of this paper, which was submitted to and rejected by a major conference, we used BLEU for most of our analysis (we still had BERTScore and COMET, but could not apply these broadly, as noted). Concerns about BLEU dominated in the reviews and in the final acceptance assessment. Some of the reviewers suggested we should use CHRF++ instead, which we subsequently adopted wholeheartedly.

CHRF++ (Popović, 2017) is favored in low-resource MT workshops, such as LoResMT (Ojha et al., 2026), now in its ninth year. CHRF++ is still an n-gram based metric, so it will not catch semantic similarities across texts (something that the neural methods are more adept at). But, because it uses a composite word and character-based n-gram score, it is a more effective metric for evaluating the output of typologically diverse languages, and much more effective than BLEU for morphologically rich languages in particular (Popović, 2017, 2015).

Because of BLEU's lengthy history in our field, and because of CHRF++'s frequent use in low-resource contexts, we felt a combination of the two metrics was warranted. This follows other work in the field for low-resource languages that show that a composite score across the two metrics provides complementary insights into the text being measured (*e.g.,* Kumar et al. (2026)), and is in line with the recommendation made by Marie et al. (2021) for the use of additional metrics alongside BLEU.

There is still the question of comparability between languages. BLEU has been shown to not be fully comparable between languages (*e.g.,* Marie et al. (2021)), although we could not find a reference saying the same for CHRF++. This issue has been noted by a couple of reviewers. Yet, we feel it is important to recognize that these measures are not just producing white noise; there *is* signal that is returned, and that signal can be interpreted, as long as we ensure that we are using standard methods and practices, *e.g.,* SacreBLEU (Post, 2018). Incorporating both measures into one threshold, and testing that this threshold is at least in the ballpark (*e.g.,* through human eval), gives us enough signal to be able to say that certain language pairs may be usable in the context of crisis preparedness and response and others not.

It is crucial to note that the ultimate goal of this paper was to offer insights to crisis managers, aid providers and the like about what tools *might* be usable in the field, and for which languages. If a particular language pair is well below our threshold, it will likely be necessary to pursue other options. Outside of other available information, these communities and practitioners are left completely in the dark, in that they have no information about what tools they should be using, what the quality of these tools are, and whether the tools are usable for certain languages versus others. [20] We hope to give these communities at least some glimmer of light in this dark about the tools that so many of them are seeking to use or have already become dependent on. As a field, burying our heads in the sand and saying that we have no way to measure quality for these language pairs is not an adequate response. The development of high-performing and broadly applicable evaluation metrics that fully include low-resource languages is a vital area for further research (though this is a distinct goal from those of this study), but we cannot wait for this

[20] In fact, some emergency managers, outside of additional information, have determined overall and cross-lingual usability of particular tools, such as Google Translate, based on the evaluation of *one* language, for which they had a native speaker in house (Wilson et al., 2026).

work to be completed before evaluating translation tools on low resource languages, particularly when such tools are likely already being used. We owe these communities more than that.

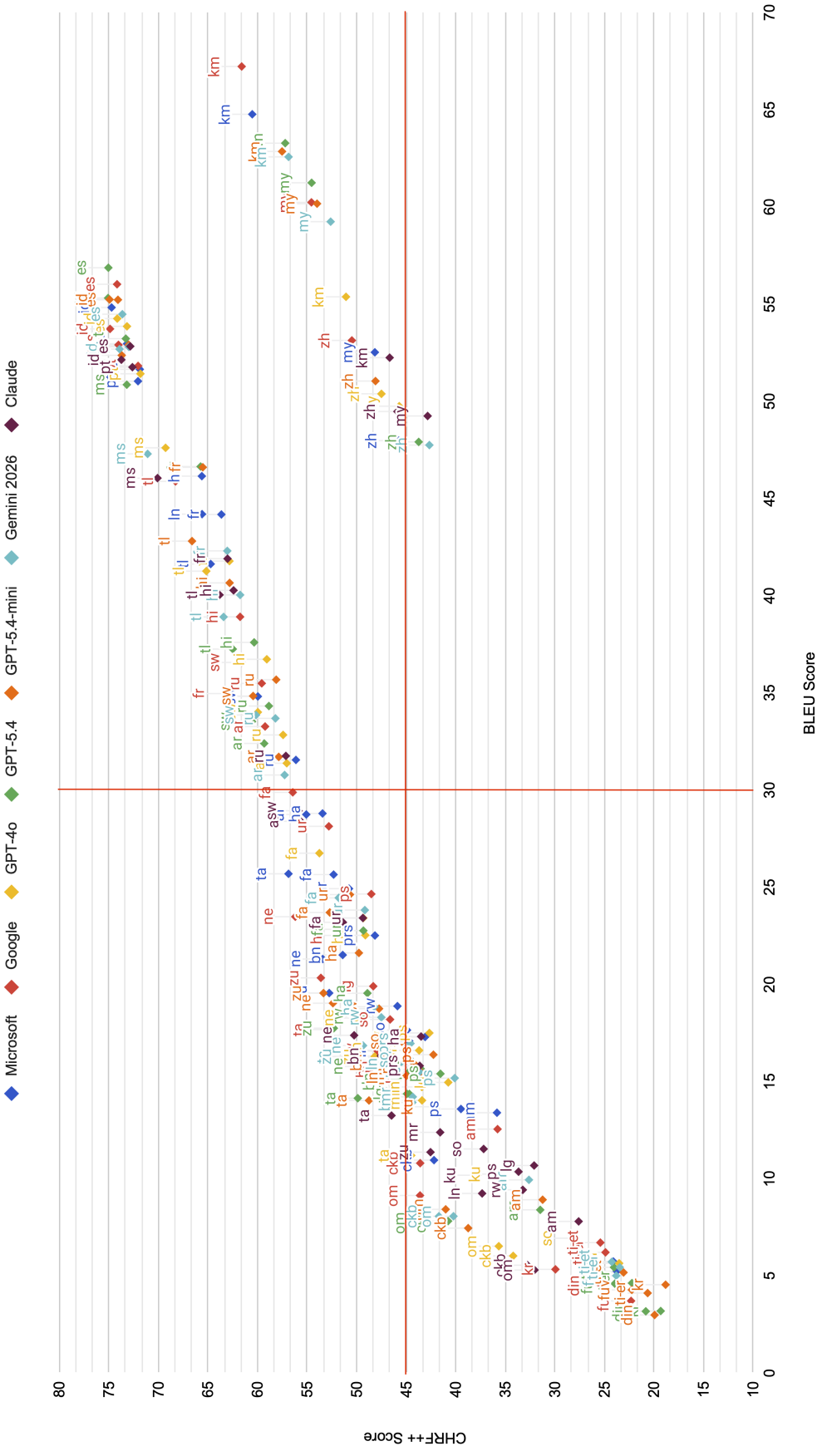



Figure 2: E to X BLEU x CHRF++ Scatter Plot for all TICO languages

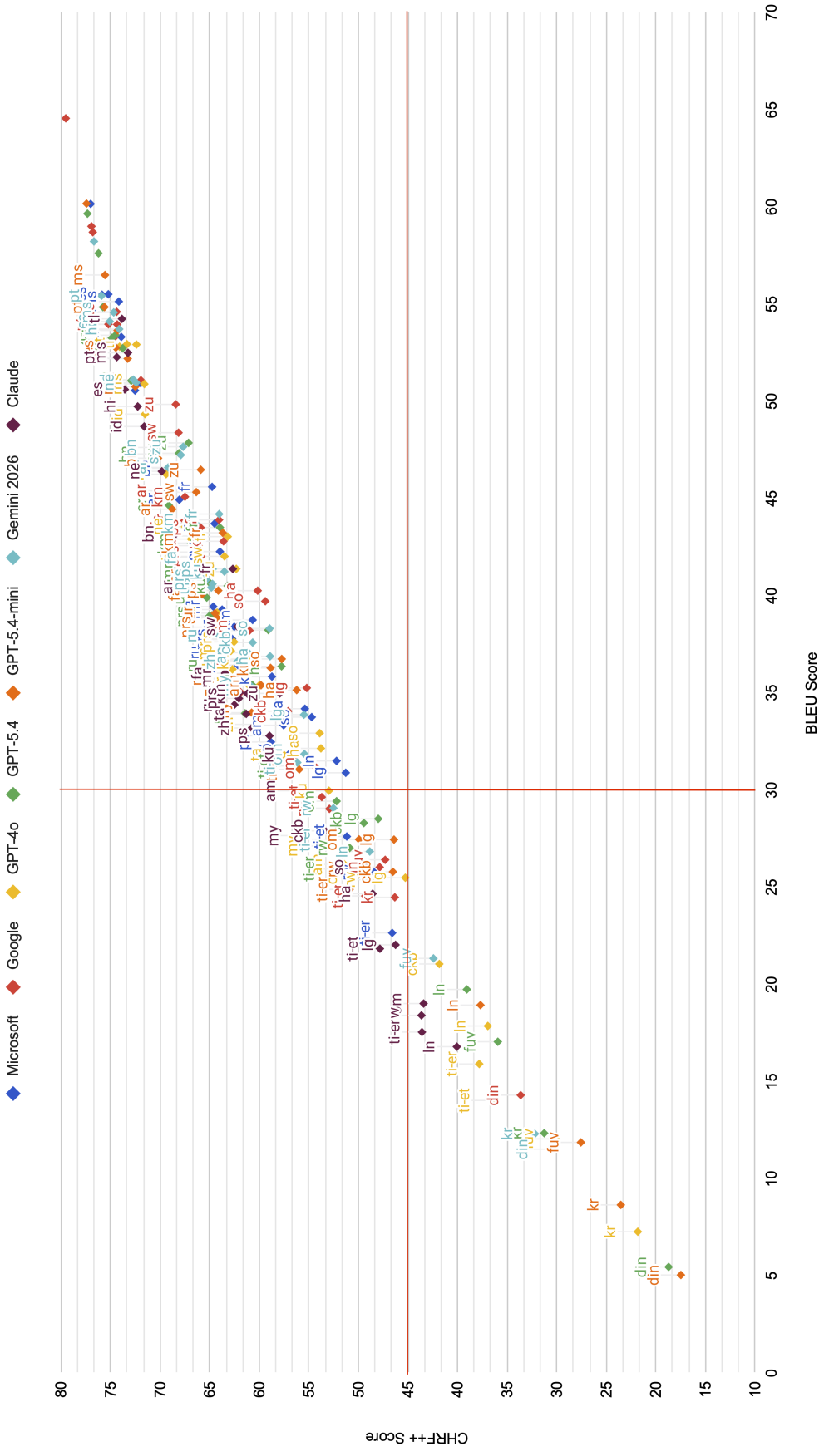



Figure 3: X to E BLEU x CHRF++ Scatter Plot for all TICO languages

| Language | English to X | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 28.71 | **33.24** | 31.35 | 32.37 | 31.68 | 30.75 | 28.36 |
| Chinese | 47.95 | **53.10** | 50.36 | 47.88 | 51.01 | 47.72 | 49.45 |
| French | 44.15 | 34.94 | 41.75 | **46.61** | 46.58 | 42.28 | 41.87 |
| Hindi | **46.12** | 38.88 | 36.70 | 37.57 | 40.63 | 40.01 | 40.24 |
| Indonesian | 54.80 | 53.70 | 54.23 | **55.29** | 55.21 | 52.66 | 52.11 |
| Portuguese (BR) | 51.01 | 51.79 | 51.39 | **53.20** | 52.88 | 52.77 | 51.73 |
| Russian | 31.52 | 35.46 | 32.81 | 34.29 | **35.65** | 33.66 | 31.73 |
| Spanish (LA) | 51.62 | 55.99 | 53.83 | **56.84** | 55.19 | 54.45 | 52.80 |
| Swahili | 34.79 | **36.54** | 33.98 | 33.50 | 34.81 | 33.83 | 28.91 |
| **——Important——** | | | | | | | |
| Bengali | **21.49** | 15.61 | 16.58 | 14.96 | 15.97 | 13.90 | 16.30 |
| Farsi | 25.62 | **29.85** | 26.72 | 22.81 | 23.67 | 24.43 | 23.17 |
| Malay | **52.93** | 52.88 | 47.58 | 50.83 | 52.34 | 47.27 | 46.02 |
| Marathi | **16.20** | 15.18 | 14.00 | 14.34 | 15.27 | 14.19 | 12.36 |
| Myanmar | 52.49 | 60.21 | 49.72 | **61.21** | 60.14 | 59.21 | 49.22 |
| Tagalog | 41.60 | **45.85** | 41.24 | 37.22 | 42.78 | 38.88 | 40.01 |
| Tamil | **25.66** | 17.65 | 11.17 | 14.12 | 14.00 | 16.26 | 13.23 |
| Urdu | 24.90 | **28.11** | 23.39 | 22.74 | 24.59 | 23.80 | 23.39 |
| **——Priority——** | | | | | | | |
| Amharic | **13.37** | 12.53 | 5.62 | 8.37 | 8.89 | 9.92 | 7.78 |
| Dari | **22.49** | — | 16.58 | 15.60 | 15.91 | 16.93 | 15.78 |
| Dinka | — | **4.48** | — | 3.15 | 2.96 | — | — |
| Hausa | **28.77** | 22.49 | 22.50 | 19.52 | 21.59 | 18.82 | 17.29 |
| Kanuri | — | **5.31** | — | 3.17 | 4.52 | — | — |
| Khmer | 64.74 | **67.20** | 55.35 | 63.25 | 62.83 | 62.55 | 52.22 |
| Kinyarwanda | **18.86** | 16.54 | 17.14 | 18.33 | 18.72 | 18.28 | 9.40 |
| Kurdish (Kurmanji) | **17.06** | 15.40 | 10.21 | — | 13.68 | — | 10.16 |
| Kurdish (Sorani) | **10.93** | 10.78 | 6.00 | 7.80 | 7.43 | 8.04 | 5.53 |
| Lingala | **44.17** | 15.18 | 14.70 | 15.12 | 15.17 | 15.89 | 9.22 |
| Luganda | 17.27 | **19.88** | 14.93 | 14.34 | 15.48 | 15.48 | 10.66 |
| Nepali | 21.28 | **23.45** | 18.24 | 15.75 | 19.01 | 16.83 | 17.35 |
| Nigerian Fulfulde | — | 3.69 | — | 4.58 | 4.26 | **4.98** | — |
| Oromo | — | **9.11** | 6.50 | 7.71 | 8.40 | 8.04 | 5.28 |
| Pashto | 13.56 | **24.62** | 17.47 | 15.37 | 16.36 | 15.15 | 10.34 |
| Somali | 17.62 | **18.17** | 6.92 | 15.49 | 17.03 | 16.09 | 11.51 |
| Tigrinya (Eritrean) | 5.22 | **6.19** | — | 4.61 | 4.09 | 5.42 | — |
| Tigrinya (Ethiopian) | 5.71 | **6.69** | — | 5.42 | 5.15 | 5.70 | — |
| Zulu | 19.53 | **20.31** | 16.26 | 17.72 | 19.53 | 16.27 | 11.34 |

Table 15: 2026 EX BLEU scores. The best system for each language is in bold.

| Language | English to X | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 55.08 | 59.23 | 57.05 | **59.32** | 57.86 | 57.27 | 55.64 |
| Chinese | 45.74 | **50.48** | 47.50 | 43.74 | 48.10 | 42.66 | 45.89 |
| French | 63.64 | 63.13 | 62.83 | **65.74** | 65.52 | 63.05 | 63.02 |
| Hindi | **65.62** | 61.76 | 59.06 | 60.34 | 62.81 | 61.75 | 62.42 |
| Indonesian | 74.72 | 74.88 | 74.15 | **75.08** | 74.93 | 73.92 | 73.74 |
| Portuguese (BR) | 72.06 | 72.05 | 71.81 | **73.29** | 72.99 | 72.99 | 72.62 |
| Russian | 56.13 | **59.57** | 57.42 | 58.85 | 58.11 | 58.19 | 57.14 |
| Spanish (LA) | 71.88 | 74.17 | 73.17 | **75.05** | 74.08 | 73.64 | 72.84 |
| Swahili | 59.96 | **61.59** | 59.97 | 60.55 | 60.44 | 60.16 | 55.86 |
| **——Important——** | | | | | | | |
| Bengali | **51.40** | 46.69 | 47.28 | 45.97 | 47.25 | 44.38 | 47.55 |
| Farsi | 52.32 | **56.44** | 53.76 | 51.16 | 52.71 | 51.80 | 51.38 |
| Malay | 73.16 | **74.03** | 69.28 | 73.19 | 73.70 | 71.09 | 70.09 |
| Marathi | **46.80** | 46.05 | 43.40 | 44.67 | 44.97 | 44.33 | 41.57 |
| Myanmar | 48.16 | **54.56** | 45.70 | 54.54 | 53.99 | 52.62 | 42.84 |
| Tagalog | 64.72 | **68.27** | 65.17 | 62.50 | 66.61 | 63.42 | 63.83 |
| Tamil | **56.87** | 53.11 | 44.49 | 49.88 | 48.75 | 50.68 | 46.48 |
| Urdu | 50.77 | **52.80** | 49.32 | 49.32 | 50.67 | 49.17 | 49.37 |
| **——Priority——** | | | | | | | |
| Amharic | **35.84** | 35.78 | 23.51 | 31.47 | 31.23 | 32.62 | 27.60 |
| Dari | **48.15** | — | 43.69 | 43.47 | 44.23 | 44.56 | 43.63 |
| Dinka | — | **25.31** | — | 20.82 | 19.93 | — | — |
| Hausa | **53.44** | 51.26 | 49.12 | 48.91 | 49.77 | 48.21 | 43.49 |
| Kanuri | — | **29.94** | — | 19.32 | 18.83 | — | — |
| Khmer | 60.53 | **61.60** | 51.07 | 57.20 | 57.52 | 56.86 | 46.66 |
| Kinyarwanda | 45.88 | 47.87 | 45.14 | **49.10** | 47.74 | 47.50 | 33.23 |
| Kurdish (Kurmanji) | **44.70** | 44.38 | 35.29 | — | 42.09 | — | 37.80 |
| Kurdish (Sorani) | 42.19 | **43.59** | 34.20 | 40.77 | 38.74 | 41.70 | 32.75 |
| Lingala | **65.59** | 43.85 | 43.42 | 46.07 | 45.62 | 45.68 | 37.34 |
| Luganda | 43.09 | **48.32** | 40.75 | 44.92 | 43.53 | 43.64 | 32.09 |
| Nepali | 53.42 | **56.19** | 50.06 | 49.15 | 52.39 | 49.35 | 50.25 |
| Nigerian Fulfulde | — | 22.31 | — | **23.96** | 22.23 | 23.80 | — |
| Oromo | — | **43.60** | 35.66 | 42.89 | 41.01 | 40.23 | 32.05 |
| Pashto | 39.47 | **48.52** | 42.66 | 41.54 | 42.25 | 40.10 | 33.66 |
| Somali | 44.92 | **46.62** | 28.05 | 44.68 | 45.44 | 44.53 | 37.18 |
| Tigrinya (Eritrean) | 23.75 | **24.89** | — | 22.22 | 20.64 | 23.45 | — |
| Tigrinya (Ethiopian) | 24.09 | **25.41** | — | 24.03 | 23.07 | 24.25 | — |
| Zulu | 52.76 | **53.61** | 48.18 | 52.30 | 53.34 | 50.31 | 42.54 |

Table 16: 2026 EX CHRF++ scores. The best system for each language is in bold.

| Language | X to English | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 44.91 | 45.36 | 42.88 | 44.62 | 44.42 | **46.57** | 40.40 |
| Chinese | 36.40 | 35.00 | 33.35 | 35.38 | 35.21 | **36.59** | 33.16 |
| French | **45.57** | 43.87 | 43.01 | 43.48 | 43.21 | 44.17 | 41.35 |
| Hindi | 53.28 | **58.67** | 46.22 | 53.37 | 53.62 | 53.69 | 49.70 |
| Indonesian | 50.53 | **53.93** | 49.31 | 52.70 | 52.17 | 51.08 | 48.67 |
| Portuguese (BR) | **55.46** | 53.94 | 52.91 | 54.82 | 54.82 | 55.40 | 52.25 |
| Russian | 37.06 | **38.87** | 36.18 | 36.39 | 35.69 | 37.93 | 34.36 |
| Spanish (LA) | **55.48** | 53.22 | 52.78 | 53.23 | 52.73 | 54.08 | 50.58 |
| Swahili | 42.24 | **48.35** | 42.00 | 47.30 | 45.30 | 47.21 | 38.38 |
| **——Important——** | | | | | | | |
| Bengali | 46.53 | **51.06** | 42.22 | 47.26 | 47.05 | 47.60 | 43.17 |
| Farsi | 39.41 | **43.84** | 39.10 | 40.41 | 40.04 | 41.94 | 35.99 |
| Malay | 55.11 | **58.98** | 50.86 | 57.59 | 56.47 | 54.54 | 52.47 |
| Marathi | 39.25 | **43.69** | 37.13 | 41.14 | 40.85 | 40.97 | 35.72 |
| Myanmar | 31.82 | **36.66** | 27.10 | 33.92 | 33.95 | 35.34 | 27.63 |
| Tagalog | 60.13 | **64.53** | 52.89 | 59.63 | 60.14 | 58.20 | 54.21 |
| Tamil | 37.75 | **42.55** | 31.81 | 38.96 | 38.82 | 40.58 | 33.89 |
| Urdu | 38.37 | **43.04** | 35.42 | 39.87 | 39.07 | 40.39 | 34.98 |
| **——Priority——** | | | | | | | |
| Amharic | 33.31 | **38.18** | 26.15 | 35.38 | 35.36 | 37.04 | 29.94 |
| Dari | 37.72 | **41.99** | 37.60 | 38.90 | 38.35 | 40.79 | 34.68 |
| Dinka | — | **14.27** | — | 5.44 | 5.02 | 11.50 | — |
| Hausa | 34.16 | **40.23** | 32.11 | 36.34 | 35.12 | 36.85 | 24.64 |
| Kanuri | — | **24.45** | 7.25 | 12.32 | 8.63 | 12.29 | — |
| Khmer | 38.72 | **45.06** | 36.71 | 42.66 | 42.68 | 43.71 | 34.93 |
| Kinyarwanda | 25.74 | 29.01 | 25.36 | 26.99 | 25.89 | **29.04** | 18.38 |
| Kurdish (Kurmanji) | 35.81 | **42.77** | 29.94 | 40.48 | 36.26 | 41.21 | 31.80 |
| Kurdish (Sorani) | 31.78 | 34.03 | 21.02 | 28.28 | 25.76 | **37.56** | 27.88 |
| Lingala | **31.47** | 26.00 | 17.83 | 19.71 | 18.91 | 26.81 | 16.77 |
| Luganda | 30.86 | **35.22** | 25.46 | 28.49 | 27.42 | 33.85 | 22.01 |
| Nepali | 50.90 | **54.58** | 43.52 | 51.03 | 50.73 | 50.93 | 46.37 |
| Nigerian Fulfulde | — | **26.38** | 12.03 | 17.02 | 11.85 | 21.31 | — |
| Oromo | — | 31.12 | 25.69 | 29.40 | 27.43 | **31.83** | 18.99 |
| Pashto | 32.44 | **43.48** | 35.37 | 40.61 | 40.24 | 41.33 | 32.76 |
| Somali | 33.73 | **39.69** | 32.90 | 38.20 | 36.72 | 38.28 | 26.00 |
| Tigrinya (Eritrean) | 22.62 | 24.52 | 15.88 | 25.91 | 24.86 | **27.44** | 17.52 |
| Tigrinya (Ethiopian) | 27.58 | 29.61 | 14.00 | 31.21 | 31.04 | **31.39** | 21.81 |
| Zulu | 43.68 | **49.81** | 41.35 | 47.83 | 46.45 | 47.63 | 34.94 |

Table 17: 2026 XE BLEU scores. The best system for each language is in bold.

| Language | X to English | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 68.06 | 69.17 | 66.88 | 69.12 | 68.78 | **69.19** | 66.51 |
| Chinese | 62.16 | 61.87 | 60.40 | **62.53** | 62.35 | 62.20 | 60.76 |
| French | **64.76** | 64.06 | 63.18 | 63.97 | 63.68 | 64.04 | 62.67 |
| Hindi | 73.92 | **76.79** | 69.39 | 74.52 | 74.29 | 74.13 | 72.26 |
| Indonesian | 72.53 | **74.33** | 71.53 | 73.77 | 73.27 | 72.70 | 71.61 |
| Portuguese (BR) | 75.86 | 75.21 | 73.35 | 75.77 | 75.61 | **75.91** | 74.37 |
| Russian | 63.73 | **64.70** | 62.71 | 64.03 | 63.54 | 64.07 | 62.51 |
| Spanish (LA) | **75.22** | 74.70 | 74.09 | 74.87 | 74.47 | 75.08 | 73.57 |
| Swahili | 63.97 | **68.14** | 63.51 | 68.10 | 66.35 | 67.91 | 62.19 |
| **——Important——** | | | | | | | |
| Bengali | 68.76 | **71.94** | 66.46 | 70.77 | 70.21 | 70.19 | 68.40 |
| Farsi | 64.64 | **67.55** | 64.25 | 66.21 | 65.77 | 66.14 | 63.45 |
| Malay | 74.17 | **76.92** | 71.59 | 76.21 | 75.56 | 74.68 | 73.24 |
| Marathi | 63.75 | **67.19** | 62.86 | 66.43 | 65.76 | 65.49 | 62.58 |
| Myanmar | 57.02 | **61.79** | 53.90 | 61.43 | 60.78 | 60.87 | 55.82 |
| Tagalog | 77.01 | **79.52** | 72.39 | 77.33 | 77.42 | 76.67 | 73.85 |
| Tamil | 62.74 | **66.14** | 57.57 | 64.78 | 64.25 | 64.72 | 61.31 |
| Urdu | 62.53 | **66.55** | 61.18 | 65.30 | 64.48 | 64.84 | 62.01 |
| **——Priority——** | | | | | | | |
| Amharic | 57.54 | 60.98 | 51.41 | 60.71 | 59.82 | **61.11** | 56.14 |
| Dari | 63.11 | **65.88** | 62.54 | 65.09 | 64.67 | 65.36 | 62.05 |
| Dinka | — | **33.63** | — | 18.71 | 17.48 | 30.69 | — |
| Hausa | 55.40 | **60.16** | 53.78 | 57.75 | 56.23 | 58.91 | 48.57 |
| Kanuri | — | **46.33** | 21.82 | 31.26 | 23.53 | 32.21 | — |
| Khmer | 60.67 | **67.50** | 60.85 | 66.87 | 66.43 | 66.46 | 61.21 |
| Kinyarwanda | 48.40 | **52.93** | 48.08 | 50.88 | 49.87 | 52.52 | 43.64 |
| Kurdish (Kurmanji) | 58.73 | **63.61** | 52.97 | 63.19 | 58.86 | 63.52 | 56.32 |
| Kurdish (Sorani) | 56.53 | 57.09 | 41.84 | 49.45 | 46.52 | **60.66** | 53.37 |
| Lingala | **52.21** | 47.85 | 36.95 | 39.07 | 37.69 | 48.86 | 40.07 |
| Luganda | 51.28 | 55.21 | 45.27 | 47.99 | 46.41 | **55.49** | 46.26 |
| Nepali | 71.96 | **74.38** | 67.49 | 72.92 | 72.50 | 72.42 | 69.83 |
| Nigerian Fulfulde | — | **47.30** | 30.12 | 35.94 | 27.56 | 42.43 | — |
| Oromo | — | 54.24 | 49.91 | 52.21 | 49.94 | **55.45** | 43.42 |
| Pashto | 58.82 | **65.92** | 59.97 | 64.95 | 64.14 | 64.71 | 58.96 |
| Somali | 54.71 | **59.39** | 53.90 | 59.09 | 57.74 | 58.96 | 49.26 |
| Tigrinya (Eritrean) | 46.59 | 49.32 | 37.82 | 52.21 | 50.89 | **52.64** | 43.59 |
| Tigrinya (Ethiopian) | 51.17 | 53.70 | 36.56 | **56.83** | 55.97 | 56.17 | 47.83 |
| Zulu | 64.51 | **68.42** | 62.32 | 67.13 | 65.90 | 67.68 | 57.93 |

Table 18: 2026 XE CHRF++ scores. The best system for each language is in bold.

| Language | English to X | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 0.8744 | 0.8800 | 0.8759 | **0.8830** | 0.8788 | 0.8797 | 0.8718 |
| Chinese | 0.8801 | **0.8882** | 0.8834 | 0.8710 | 0.8843 | 0.8716 | 0.8773 |
| French | 0.8772 | 0.8440 | 0.8692 | **0.8801** | 0.8791 | 0.8703 | 0.8701 |
| Hindi | **0.9038** | 0.8876 | 0.8809 | 0.8863 | 0.8982 | 0.8915 | 0.8956 |
| Indonesian | **0.9301** | 0.9249 | 0.9283 | 0.9292 | 0.9297 | 0.9273 | 0.9258 |
| Portuguese (BR) | 0.9310 | 0.9251 | 0.9213 | 0.9321 | 0.9314 | **0.9332** | 0.9320 |
| Russian | 0.8723 | **0.8781** | 0.8737 | 0.8765 | 0.8764 | 0.8764 | 0.8735 |
| Spanish (LA) | 0.9287 | 0.9284 | 0.9311 | **0.9348** | 0.9318 | 0.9329 | 0.9302 |
| Swahili | 0.8722 | **0.8737** | 0.8728 | 0.8690 | 0.8713 | 0.8694 | 0.8556 |
| **——Important——** | | | | | | | |
| Bengali | **0.8668** | 0.8564 | 0.8602 | 0.8542 | 0.8627 | 0.8481 | 0.8576 |
| Farsi | 0.8747 | **0.8851** | 0.8773 | 0.8670 | 0.8756 | 0.8700 | 0.8389 |
| Malay | **0.9269** | 0.9236 | 0.8855 | 0.9234 | 0.9264 | 0.9176 | 0.9154 |
| Marathi | **0.8400** | 0.8362 | 0.8296 | 0.8330 | 0.8347 | 0.8327 | 0.8229 |
| Myanmar | 0.8187 | **0.8375** | 0.7726 | 0.8317 | 0.8355 | 0.8275 | 0.7865 |
| Tagalog | **0.8910** | 0.8902 | 0.8882 | 0.8705 | 0.8894 | 0.8797 | 0.8809 |
| Tamil | **0.8673** | 0.8480 | 0.8159 | 0.8403 | 0.8369 | 0.8433 | 0.8293 |
| Urdu | 0.8582 | **0.8613** | 0.8464 | 0.8484 | 0.8558 | 0.8473 | 0.8513 |
| **——Priority——** | | | | | | | |
| Amharic | 0.9508 | 0.9544 | 0.9545 | **0.9588** | 0.9576 | 0.9569 | 0.9330 |
| Dari | **0.8537** | — | 0.8161 | 0.8388 | 0.8441 | 0.8433 | 0.8381 |
| Dinka | — | **0.7552** | — | 0.7328 | 0.7265 | — | — |
| Hausa | **0.8379** | 0.8206 | 0.8251 | 0.8173 | 0.8257 | 0.8149 | 0.8022 |
| Kanuri | — | **0.7318** | — | 0.6999 | 0.6888 | — | — |
| Khmer | 0.9488 | 0.9413 | 0.9373 | 0.9270 | **0.9507** | 0.9264 | 0.9118 |
| Kinyarwanda | 0.7959 | 0.7995 | 0.7799 | **0.8140** | 0.8109 | 0.8128 | 0.7551 |
| Kurdish (Kurmanji) | **0.8161** | 0.8118 | 0.7862 | — | 0.8038 | — | 0.7909 |
| Kurdish (Sorani) | **0.8648** | 0.8615 | 0.8463 | 0.8540 | 0.8509 | 0.8570 | 0.8311 |
| Lingala | **0.8667** | 0.7822 | 0.7922 | 0.7940 | 0.7932 | 0.7926 | 0.7668 |
| Luganda | 0.7829 | **0.8095** | 0.7797 | 0.7921 | 0.7904 | 0.7886 | 0.7386 |
| Nepali | 0.8676 | **0.8754** | 0.8409 | 0.8527 | 0.8668 | 0.8566 | 0.8614 |
| Nigerian Fulfulde | — | 0.7300 | — | **0.7396** | 0.7295 | 0.7131 | — |
| Oromo | — | **0.8754** | 0.7681 | 0.7940 | 0.7874 | 0.7875 | 0.7518 |
| Pashto | 0.8200 | **0.8421** | 0.8232 | 0.8248 | 0.8285 | 0.8175 | 0.7847 |
| Somali | 0.8211 | **0.8216** | 0.7516 | 0.8081 | 0.8146 | 0.8092 | 0.7823 |
| Tigrinya (Eritrean) | 0.9483 | — | — | 0.9514 | **0.9523** | 0.9466 | — |
| Tigrinya (Ethiopian) | 0.9542 | 0.9462 | — | 0.9533 | **0.9563** | 0.9510 | — |
| Zulu | **0.8478** | 0.8385 | 0.7969 | 0.8330 | 0.8394 | 0.8269 | 0.7982 |

Table 19: 2026 EX BERTScores. The best system for each language is in bold.

| Language | X to English | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 0.7486 | 0.7416 | 0.7406 | 0.7514 | 0.7506 | **0.7641** | 0.7285 |
| Chinese | 0.7148 | 0.6959 | -0.5152 | 0.7098 | 0.7118 | **0.7160** | 0.6920 |
| French | **0.6865** | 0.6627 | 0.6702 | 0.6762 | 0.6744 | 0.6810 | 0.6700 |
| Hindi | 0.8113 | **0.8225** | -0.7376 | 0.8118 | 0.8162 | 0.8206 | 0.7948 |
| Indonesian | 0.8087 | 0.8108 | 0.8005 | 0.8155 | 0.8099 | **0.8167** | 0.8030 |
| Portuguese (BR) | 0.8307 | 0.8058 | -0.0443 | 0.8332 | 0.8317 | **0.8343** | 0.8181 |
| Russian | 0.7173 | 0.7133 | 0.7018 | 0.7096 | 0.7058 | **0.7233** | 0.7041 |
| Spanish (LA) | 0.8199 | 0.7965 | 0.8134 | 0.8191 | 0.8166 | **0.8204** | 0.8057 |
| Swahili | 0.7236 | 0.7574 | -0.4285 | 0.7447 | 0.7310 | **0.7756** | 0.7111 |
| **——Important——** | | | | | | | |
| Bengali | 0.7711 | 0.7744 | 0.7398 | 0.7665 | 0.7661 | **0.7919** | 0.7605 |
| Farsi | 0.7426 | 0.7535 | 0.7395 | 0.7467 | 0.7470 | **0.7611** | 0.7237 |
| Malay | 0.8220 | 0.8267 | -0.4093 | 0.8331 | 0.8285 | **0.8344** | 0.8141 |
| Marathi | 0.7319 | 0.7456 | 0.7208 | 0.7436 | 0.7400 | **0.7499** | 0.7059 |
| Myanmar | 0.6585 | 0.6929 | -0.7524 | 0.6982 | 0.6908 | **0.7001** | 0.6259 |
| Tagalog | 0.8380 | 0.8394 | -0.3387 | 0.8471 | **0.8472** | 0.8468 | 0.8079 |
| Tamil | 0.7033 | 0.7126 | -0.9708 | 0.6792 | 0.6738 | **0.7322** | 0.6792 |
| Urdu | 0.7104 | 0.7375 | -0.5403 | 0.7314 | 0.7273 | **0.7421** | 0.7019 |
| **——Priority——** | | | | | | | |
| Amharic | 0.6584 | 0.6944 | 0.6055 | 0.6933 | 0.6915 | **0.7103** | 0.6346 |
| Dari | 0.7166 | 0.7214 | -0.5382 | 0.7306 | 0.7261 | **0.7390** | 0.7016 |
| Dinka | — | **0.3791** | — | -1.2368 | -1.4992 | 0.3156 | — |
| Hausa | 0.6388 | 0.6850 | 0.3746 | 0.4153 | 0.3907 | **0.6986** | 0.5458 |
| Kanuri | — | **0.4763** | -0.7413 | 0.2510 | -0.7590 | 0.2883 | — |
| Khmer | 0.6914 | 0.7431 | -0.8413 | 0.7518 | 0.7496 | **0.7563** | 0.6946 |
| Kinyarwanda | 0.5777 | 0.6236 | -0.6382 | 0.4456 | 0.4462 | **0.6426** | 0.5005 |
| Kurdish (Kurmanji) | 0.6673 | 0.7046 | 0.4258 | 0.7096 | 0.4806 | **0.7238** | 0.6259 |
| Kurdish (Sorani) | 0.6267 | 0.6448 | -0.5270 | -0.3058 | -0.4956 | **0.6883** | 0.5896 |
| Lingala | **0.6376** | 0.5847 | -0.9227 | -0.8905 | -0.8785 | 0.6056 | 0.4390 |
| Luganda | 0.6038 | 0.6555 | -0.3596 | -0.3446 | -0.3695 | **0.6655** | 0.5135 |
| Nepali | 0.7943 | 0.8004 | -0.7168 | 0.7972 | 0.7978 | **0.8091** | 0.7706 |
| Nigerian Fulfulde | — | **0.5271** | -0.5379 | 0.1528 | -1.0359 | 0.4766 | — |
| Oromo | — | 0.6211 | -0.5438 | 0.2383 | 0.1954 | **0.6474** | 0.4658 |
| Pashto | 0.6427 | 0.7312 | -0.5797 | 0.7239 | 0.7172 | **0.7379** | 0.6623 |
| Somali | 0.6139 | 0.6601 | 0.6208 | 0.6291 | 0.6148 | **0.6667** | 0.5355 |
| Tigrinya (Eritrean) | 0.5596 | 0.5740 | 0.4233 | 0.6063 | 0.5983 | **0.6208** | 0.4677 |
| Tigrinya (Ethiopian) | 0.6141 | 0.6262 | -0.7240 | 0.6680 | 0.6628 | **0.6706** | 0.5311 |
| Zulu | 0.7397 | 0.7619 | -0.4587 | 0.6760 | 0.6737 | **0.7823** | 0.6637 |

Table 20: 2026 XE BERTScores (rescaled). The best system for each language is in bold.

| Language | English to X | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 0.8485 | 0.8750 | 0.8683 | **0.8764** | 0.8688 | 0.8761 | 0.8649 |
| Chinese | 0.8749 | **0.8953** | 0.8840 | 0.8753 | 0.8862 | 0.8779 | 0.8804 |
| French | 0.8154 | 0.8242 | 0.8199 | **0.8245** | 0.8229 | 0.8239 | 0.8199 |
| Hindi | 0.8076 | 0.8203 | 0.8064 | 0.8061 | **0.8217** | 0.8177 | 0.8148 |
| Indonesian | 0.9189 | **0.9254** | 0.9234 | 0.9251 | 0.9237 | 0.9253 | 0.9212 |
| Portuguese (BR) | 0.8994 | 0.9037 | 0.8965 | 0.9041 | 0.9020 | **0.9053** | 0.9022 |
| Russian | 0.8775 | **0.9027** | 0.8889 | 0.8939 | 0.8898 | 0.8996 | 0.8898 |
| Spanish (LA) | 0.8839 | **0.8932** | 0.8886 | 0.8923 | 0.8879 | 0.8903 | 0.8880 |
| Swahili | 0.8236 | 0.8392 | 0.8406 | 0.8431 | 0.8400 | **0.8448** | 0.8139 |
| **——Important——** | | | | | | | |
| Bengali | 0.8513 | **0.8749** | 0.8624 | 0.8629 | 0.8686 | 0.8634 | 0.8565 |
| Farsi | 0.8782 | **0.8901** | 0.8822 | 0.8716 | 0.8825 | 0.8806 | 0.8476 |
| Malay | 0.9071 | 0.9065 | 0.8838 | 0.9074 | **0.9086** | 0.9068 | 0.9016 |
| Marathi | 0.7216 | **0.7460** | 0.7240 | 0.7366 | 0.7332 | 0.7409 | 0.7133 |
| Myanmar | 0.8719 | 0.8849 | 0.8407 | 0.8946 | 0.8906 | **0.8961** | 0.8016 |
| Tagalog | 0.8672 | 0.8682 | **0.8685** | 0.8449 | 0.8655 | 0.8638 | 0.8543 |
| Tamil | 0.8751 | **0.8911** | 0.8503 | 0.8787 | 0.8706 | 0.8851 | 0.8409 |
| Urdu | 0.8230 | 0.8172 | 0.8181 | 0.8210 | **0.8282** | 0.8247 | 0.8194 |
| **——Priority——** | | | | | | | |
| Amharic | 0.8533 | 0.8656 | 0.7891 | 0.8612 | 0.8526 | **0.8696** | 0.8057 |
| Hausa | 0.7993 | 0.8087 | 0.8053 | 0.8060 | 0.8052 | **0.8108** | 0.7592 |
| Khmer | 0.8560 | 0.8734 | 0.8364 | 0.8683 | 0.8662 | **0.8744** | 0.7787 |
| Kinyarwanda | 0.6271 | 0.6473 | 0.6327 | **0.6549** | 0.6496 | 0.6536 | 0.6060 |
| Kurdish (Kurmanji) | 0.8033 | **0.8116** | 0.7359 | — | 0.7917 | — | 0.7630 |
| Nepali | 0.8159 | 0.8358 | 0.8112 | 0.8203 | **0.8368** | 0.8328 | 0.8227 |
| Oromo | — | **0.8358** | 0.7683 | 0.8040 | 0.7917 | 0.7995 | 0.7507 |
| Pashto | 0.7877 | **0.8128** | 0.7999 | 0.8115 | 0.8120 | 0.8106 | 0.7561 |
| Somali | 0.8008 | **0.8070** | 0.6765 | 0.7950 | 0.7996 | 0.8029 | 0.7389 |

Table 21: 2026 EX COMET. The best system for each language is in bold. Only supported languages are included.

| Language | X to English | | | | | | |
|---|---|---|---|---|---|---|---|
| | MS | Google | GPT-4o | GPT-5.4 | GPT-5.4-mini | Gemini | Claude |
| **——Pivot——** | | | | | | | |
| Arabic | 0.8725 | 0.8779 | 0.8765 | 0.8788 | 0.8784 | **0.8820** | 0.8716 |
| Chinese | 0.8751 | 0.8765 | 0.6173 | 0.8778 | 0.8781 | **0.8790** | 0.8728 |
| French | 0.8360 | 0.8353 | 0.8352 | **0.8370** | 0.8362 | 0.8367 | 0.8334 |
| Hindi | 0.9070 | **0.9150** | 0.4515 | 0.9121 | 0.9125 | 0.9135 | 0.9075 |
| Indonesian | 0.9032 | **0.9086** | 0.9047 | 0.9079 | 0.9067 | 0.9072 | 0.9036 |
| Portuguese (BR) | 0.9018 | 0.9009 | 0.7162 | **0.9042** | 0.9039 | 0.9041 | 0.8990 |
| Russian | 0.8637 | **0.8684** | 0.8640 | 0.8664 | 0.8652 | **0.8684** | 0.8628 |
| Spanish (LA) | 0.8981 | 0.8966 | 0.8975 | 0.8989 | 0.8984 | **0.8995** | 0.8948 |
| Swahili | 0.8459 | 0.8679 | 0.5409 | 0.8685 | 0.8633 | **0.8722** | 0.8472 |
| **——Important——** | | | | | | | |
| Bengali | 0.8995 | 0.9048 | 0.9005 | 0.9062 | 0.9054 | **0.9087** | 0.9016 |
| Farsi | 0.8846 | 0.8924 | 0.8889 | 0.8906 | 0.8901 | **0.8930** | 0.8823 |
| Malay | 0.8984 | 0.9061 | 0.6778 | **0.9067** | 0.9051 | 0.9060 | 0.8992 |
| Marathi | 0.8806 | **0.8903** | 0.8838 | 0.8894 | 0.8879 | 0.8899 | 0.8778 |
| Myanmar | 0.8696 | 0.8876 | 0.3951 | 0.8894 | 0.8868 | **0.8897** | 0.8689 |
| Tagalog | 0.8899 | 0.8970 | 0.6188 | **0.8997** | 0.8982 | 0.8988 | 0.8850 |
| Tamil | 0.8710 | 0.8816 | 0.4526 | 0.8798 | 0.8784 | **0.8869** | 0.8731 |
| Urdu | 0.8694 | 0.8854 | 0.4548 | 0.8838 | 0.8816 | **0.8856** | 0.8743 |
| **——Priority——** | | | | | | | |
| Amharic | 0.8569 | 0.8782 | 0.8486 | 0.8784 | 0.8748 | **0.8813** | 0.8576 |
| Hausa | 0.7886 | 0.8243 | 0.7936 | 0.8064 | 0.7956 | **0.8302** | 0.7569 |
| Khmer | 0.8647 | 0.8910 | 0.4306 | 0.8938 | 0.8923 | **0.8944** | 0.8750 |
| Kinyarwanda | 0.7233 | 0.7621 | 0.2931 | 0.7519 | 0.7460 | **0.7693** | 0.6950 |
| Kurdish (Kurmanji) | 0.8139 | 0.8403 | 0.7938 | 0.8439 | 0.8131 | **0.8468** | 0.8063 |
| Nepali | 0.9113 | 0.9173 | 0.4189 | 0.9175 | 0.9170 | **0.9188** | 0.9103 |
| Oromo | — | 0.7876 | 0.3212 | 0.7731 | 0.7595 | **0.8080** | 0.7250 |
| Pashto | 0.8499 | 0.8752 | 0.3743 | 0.8749 | 0.8716 | **0.8765** | 0.8532 |
| Somali | 0.7857 | **0.8250** | 0.8101 | 0.8182 | 0.8105 | 0.6990 | 0.7607 |

Table 22: 2026 XE COMET. The best system for each language is in bold. Only supported languages are included.

| | English to X | | | | X to English | | | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Microsoft** | | **Google** | | **Microsoft** | | **Google** | |
| | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 |
| **——Pivot——** | | | | | | | | |
| Arabic | 30.26 | 25.16 | 29.15 | 32.88 | 44.88 | 44.87 | 46.90 | 49.28 |
| Portuguese (BR) | 52.42 | 47.83 | 52.27 | 54.26 | 54.81 | 55.33 | 55.19 | 57.63 |
| Chinese | 52.89 | 50.24 | 54.29 | 54.27 | 36.46 | 36.48 | 37.71 | 39.56 |
| French | 45.61 | 40.35 | 34.14 | 45.42 | 45.25 | 45.32 | 44.25 | 46.03 |
| Hindi | 43.24 | 40.27 | 46.00 | 43.48 | 53.17 | 53.28 | 56.24 | 58.80 |
| Indonesian | 54.39 | 46.85 | 55.07 | 54.82 | 51.73 | 50.29 | 51.15 | 54.61 |
| Spanish (LA) | 56.06 | 51.33 | 56.86 | 57.12 | 54.75 | 54.59 | 54.15 | 55.88 |
| Russian | 36.88 | 32.95 | 36.71 | 37.35 | 36.06 | 36.84 | 37.14 | 40.14 |
| Swahili | 33.10 | 30.74 | 33.88 | 33.87 | 41.04 | 41.92 | 45.56 | 49.21 |
| **——Important——** | | | | | | | | |
| Bengali | 19.29 | 17.18 | 22.95 | 21.14 | 44.38 | 46.08 | 49.46 | 51.33 |
| Farsi | 27.38 | 25.25 | 29.20 | 29.88 | 38.97 | 39.35 | 39.65 | 44.45 |
| Malay | 51.06 | 45.78 | 53.33 | 53.31 | 50.88 | 54.08 | 57.63 | 59.70 |
| Marathi | 15.52 | 13.53 | 16.75 | 16.09 | 37.21 | 37.34 | 42.30 | 44.43 |
| Myanmar | 53.55 | 56.15 | 56.95 | 56.62 | 23.82 | 31.44 | 32.39 | 37.33 |
| Tagalog | — | 49.40 | 48.46 | 48.82 | — | 59.92 | — | 65.55 |
| Tamil | 24.17 | 24.84 | 21.53 | 22.76 | 34.21 | 40.73 | 39.38 | 43.11 |
| Urdu | 23.50 | 20.41 | 28.14 | 28.12 | 35.93 | 35.88 | 40.61 | 43.82 |
| **——Priority——** | | | | | | | | |
| Amharic | 8.67 | 7.04 | 11.88 | 11.79 | 26.49 | 33.02 | 36.51 | 38.60 |
| Dari | 17.32 | 15.21 | — | — | 31.21 | 31.22 | — | — |
| Dinka | — | — | — | 4.54 | — | — | — | 14.47 |
| Eritrean Tigrinya | 3.39 | 3.47 | 6.13 | 6.13 | 18.07 | 18.00 | 24.83 | — |
| Ethiopian Tigrinya | 3.83 | 3.98 | 6.64 | 6.69 | 21.60 | 21.52 | 30.33 | 29.94 |
| Hausa | 15.98 | 25.23 | 25.87 | 28.17 | — | 33.92 | 36.17 | 40.65 |
| Kanuri | — | — | — | 5.51 | — | — | — | 26.08 |
| Khmer | 61.20 | 65.93 | 66.50 | 66.08 | 27.41 | 37.04 | 36.31 | 45.76 |
| Kinyarwanda | — | 18.75 | — | 17.88 | — | 25.39 | — | 29.49 |
| Kurdish Kurmanji | — | 11.82 | — | 10.05 | — | 28.17 | — | 28.80 |
| Kurdish Sorani | — | 10.38 | 10.66 | 10.85 | — | 31.87 | 34.18 | 34.13 |
| Lingala | 32.07 | 28.35 | 15.08 | 15.17 | — | 29.14 | 26.04 | 26.22 |
| Luganda | — | 17.26 | 19.71 | 20.64 | — | 30.52 | — | 35.45 |
| Nepali | 16.71 | 14.20 | 25.44 | 23.56 | 48.38 | 48.38 | 53.76 | 55.16 |
| Nigerian Fulfulde | — | — | — | 3.70 | — | — | — | 26.97 |
| Oromo | — | — | 9.06 | 10.23 | — | — | 31.36 | 31.52 |
| Pashto | 21.12 | 12.79 | 26.09 | 24.63 | 27.57 | 32.16 | 40.75 | 44.07 |
| Somali | 26.01 | 26.02 | 17.90 | 17.56 | 33.03 | 33.12 | 37.61 | 19.99 |
| Zulu | 19.76 | 20.28 | 19.60 | 19.71 | 40.26 | 43.40 | 47.58 | 50.63 |

Table 23: BLEU Score Differences between 2023 and 2025 for Microsoft and Google

| | English to X | | | | X to English | | | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Microsoft** | | **Google** | | **Microsoft** | | **Google** | |
| | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 |
| **——Pivot——** | | | | | | | | |
| Arabic | 57.73 | 56.41 | 56.70 | 59.18 | 67.74 | 67.73 | 69.61 | 70.87 |
| Portuguese (BR) | 72.79 | 71.12 | 72.48 | 74.04 | 75.43 | 75.88 | 75.45 | 76.79 |
| Chinese | 39.76 | 45.07 | 48.72 | 45.33 | 62.08 | 62.69 | 39.76 | 64.60 |
| French | 64.72 | 64.32 | 62.09 | 65.16 | 64.50 | 64.74 | 64.02 | 65.25 |
| Hindi | 62.72 | 64.56 | 65.04 | 64.83 | 73.50 | 73.67 | 75.63 | 77.15 |
| Indonesian | 74.32 | 72.03 | 74.28 | 75.10 | 72.89 | 72.53 | 72.73 | 74.71 |
| Spanish (LA) | 74.39 | 73.16 | 74.47 | 75.30 | 75.13 | 75.33 | 74.97 | 75.48 |
| Russian | 59.89 | 58.91 | 59.56 | 60.61 | 63.50 | 63.75 | 63.73 | 65.27 |
| Swahili | 58.41 | 57.99 | 59.18 | 59.34 | 62.44 | 63.95 | 66.22 | 68.58 |
| **——Important——** | | | | | | | | |
| Bengali | 49.04 | 48.5 | 52.4 | 52.0 | 68.01 | 69.17 | 70.84 | 72.53 |
| Farsi | 53.52 | 53.58 | 54.69 | 56.49 | 63.61 | 64.57 | 64.84 | 67.88 |
| Malay | 72.78 | 71.73 | 73.76 | 74.0 | 71.99 | 74.15 | 76.12 | 77.53 |
| Marathi | 46.34 | 44.68 | 46.08 | 46.40 | 62.3 | 62.34 | 65.83 | 67.53 |
| Myanmar | 37.89 | 39.86 | 43.36 | 43.21 | 50.36 | 57.57 | 59.35 | 62.10 |
| Tagalog | — | 70.41 | 69.36 | 69.93 | — | 77.06 | — | 80.40 |
| Tamil | 56.68 | 57.22 | 54.18 | 56.13 | 61.83 | 64.78 | 63.9 | 66.76 |
| Urdu | 50.03 | 48.96 | 52.99 | 53.15 | 60.91 | 60.9 | 64.76 | 66.89 |
| **——Priority——** | | | | | | | | |
| Amharic | 30.12 | 29.61 | 33.92 | 33.83 | 51.10 | 57.00 | 59.63 | 61.38 |
| Dari | 42.20 | 41.55 | — | — | 57.44 | 57.45 | — | — |
| Dinka | — | — | — | 25.37 | — | — | — | 33.82 |
| Eritrean Tigrinya | 19.30 | 19.31 | 25.00 | 25.12 | 42.98 | 42.93 | 49.69 | — |
| Ethiopian Tigrinya | 19.49 | 19.52 | 25.81 | 25.99 | 46.29 | 46.36 | 54.67 | 54.19 |
| Hausa | 40.87 | 52.66 | 52.52 | 52.71 | — | 55.44 | 57.37 | 60.58 |
| Kanuri | — | — | — | | — | — | — | 30.16 |
| Khmer | 45.82 | 52.38 | 53.10 | 52.97 | 53.7 | 58.69 | 60.26 | 67.88 |
| Kinyarwanda | — | 45.85 | — | 45.18 | — | 48.35 | — | 53.13 |
| Kurdish Kurmanji | — | 39.87 | — | 33.49 | — | 51.81 | — | 47.14 |
| Kurdish Sorani | — | 41.83 | 43.61 | 43.77 | — | 56.70 | 57.29 | 57.26 |
| Lingala | 59.00 | 58.18 | 43.83 | 43.86 | — | 50.49 | 47.99 | 48.02 |
| Luganda | — | 43.07 | 48.24 | 48.53 | — | 51.23 | — | 55.67 |
| Nepali | 49.30 | 48.72 | 57.01 | 56.55 | 70.46 | 70.4 | 73.75 | 74.95 |
| Nigerian Fulfulde | — | — | — | 22.38 | — | — | — | 48.12 |
| Oromo | — | — | 43.75 | 44.00 | — | — | 54.42 | 54.48 |
| Pashto | 46.63 | 37.84 | 49.11 | 48.76 | 54.45 | 58.83 | 63.51 | 66.21 |
| Somali | 53.01 | 53.01 | 45.16 | 45.97 | 54.30 | 54.37 | 57.53 | 37.13 |
| Zulu | 53.43 | 53.46 | 53.21 | 53.13 | 60.78 | 64.51 | 67.02 | 69.21 |

Table 24: CHRF++ Score Differences between 2023 and 2025 for Microsoft and Google

| | English to X | | | | X to English | | | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Microsoft** | | **Google** | | **Microsoft** | | **Google** | |
| | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 |
| **——Pivot——** | | | | | | | | |
| Arabic | 0.8841 | 0.8375 | 0.8791 | 0.8833 | 0.7661 | 0.7465 | 0.7701 | 0.7987 |
| Portuguese (BR) | 0.9363 | 0.8750 | 0.9339 | 0.9521 | 0.8434 | 0.8302 | 0.8248 | 0.8437 |
| Chinese | 0.8930 | 0.8619 | 0.8971 | 0.9450 | 0.7373 | 0.7234 | 0.7271 | 0.7400 |
| French | 0.8837 | 0.8300 | 0.8465 | 0.8838 | 0.7078 | 0.6863 | 0.6930 | 0.7164 |
| Hindi | 0.8945 | 0.8570 | 0.9023 | 0.9649 | 0.8156 | 0.8103 | 0.8166 | 0.8423 |
| Indonesian | 0.9325 | 0.8705 | 0.9308 | 0.9352 | 0.8300 | 0.8123 | 0.8114 | 0.8451 |
| Spanish (LA) | 0.9379 | 0.8773 | 0.9349 | 0.9394 | 0.8362 | 0.8278 | 0.8191 | 0.8427 |
| Russian | 0.8869 | 0.8339 | 0.8873 | 0.8907 | 0.7373 | 0.7214 | 0.7291 | 0.7584 |
| Swahili | 0.8706 | 0.8263 | 0.8720 | 0.9281 | 0.7282 | 0.7235 | 0.7564 | 0.7814 |
| **——Important——** | | | | | | | | |
| Bengali | 0.8475 | 0.8192 | 0.8625 | 0.8755 | 0.7748 | 0.7702 | 0.7878 | 0.8169 |
| Farsi | 0.8778 | 0.8377 | 0.8820 | 0.8904 | 0.7528 | 0.7400 | 0.7428 | 0.7900 |
| Malay | 0.9283 | 0.8691 | 0.9286 | 0.9494 | 0.8151 | 0.8214 | 0.8342 | 0.8465 |
| Marathi | 0.8417 | 0.8044 | 0.8413 | 0.8461 | 0.7302 | 0.7123 | 0.7487 | 0.7822 |
| Myanmar | 0.7982 | 0.7835 | 0.8288 | 0.9441 | 0.5944 | 0.6627 | 0.6788 | 0.7104 |
| Tagalog | — | 0.9400 | 0.9017 | 0.9406 | — | 0.8423 | — | 0.7427 |
| Tamil | 0.8674 | 0.8352 | 0.8588 | 0.9647 | 0.6981 | 0.7203 | 0.7083 | 0.7465 |
| Urdu | 0.8491 | 0.8095 | 0.8623 | 0.9514 | 0.7104 | 0.6927 | 0.7349 | 0.7536 |
| **——Priority——** | | | | | | | | |
| Amharic | 0.9503 | 0.8473 | 0.9412 | 0.9422 | 0.5900 | 0.6559 | 0.6915 | 0.7347 |
| Dari | 0.8365 | 0.7956 | — | — | 0.6642 | 0.6578 | — | — |
| Dinka | — | — | — | 0.8674 | — | — | — | 0.3917 |
| Eritrean Tigrinya | 0.9366 | 0.8418 | 0.9431 | 0.9375 | 0.5233 | 0.5212 | 0.6006 | 0.6394 |
| Ethiopian Tigrinya | 0.9408 | 0.8449 | 0.9464 | 0.9374 | 0.5687 | 0.5227 | 0.6668 | 0.6394 |
| Hausa | 0.7840 | 0.7853 | 0.8287 | 0.8359 | — | 0.6398 | 0.6719 | 0.7234 |
| Kanuri | — | — | — | 0.8330 | — | — | — | 0.7800 |
| Khmer | 0.8755 | 0.8017 | 0.9480 | 0.9600 | 0.6098 | 0.7234 | 0.6976 | 0.8200 |
| Kinyarwanda | — | 0.8800 | — | 0.8000 | — | 0.5752 | — | 0.6358 |
| Kurdish Kurmanji | — | 0.9300 | — | 0.8128 | — | 0.6337 | — | 0.7466 |
| Kurdish Sorani | — | 0.8700 | 0.8643 | 0.8674 | — | 0.5620 | 0.6651 | 0.6773 |
| Lingala | 0.8481 | 0.7961 | 0.7924 | 0.7936 | — | 0.6126 | 0.6075 | 0.6247 |
| Luganda | — | 0.8700 | 0.8112 | 0.8147 | — | 0.6091 | — | 0.6972 |
| Nepali | 0.8534 | 0.8180 | 0.8804 | 0.9595 | 0.7879 | 0.7823 | 0.8087 | 0.8293 |
| Nigerian Fulfulde | — | — | — | 0.8400 | — | — | — | 0.5623 |
| Oromo | — | — | 0.7963 | 0.8885 | — | — | 0.6477 | 0.6318 |
| Pashto | 0.8472 | 0.7908 | 0.8517 | 0.9456 | 0.5942 | 0.6430 | 0.7229 | 0.7420 |
| Somali | 0.7808 | 0.7487 | 0.7543 | 0.8635 | 0.3793 | 0.3405 | 0.3945 | 0.3474 |
| Zulu | 0.8418 | 0.7876 | 0.8403 | 0.9136 | 0.7179 | 0.7396 | 0.7738 | 0.7536 |

Table 25: BERTScore Differences between 2023 and 2025 for Microsoft and Google

| | English to X | | | | X to English | | | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Microsoft** | | **Google** | | **Microsoft** | | **Google** | |
| | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 | 2023 | 2025 |
| **——Pivot——** | | | | | | | | |
| Arabic | 0.8450 | 0.8253 | 0.8486 | 0.8813 | 0.8710 | 0.8400 | 0.8777 | 0.8891 |
| Portuguese (BR) | 0.8978 | 0.8698 | 0.8976 | 0.8962 | 0.9007 | 0.8790 | 0.9002 | 0.8983 |
| Chinese | 0.8824 | 0.8652 | 0.8924 | 0.8910 | 0.8738 | 0.8090 | 0.8755 | 0.8752 |
| French | 0.8139 | 0.7967 | 0.8099 | 0.8910 | 0.8351 | 0.8400 | 0.8333 | 0.9028 |
| Hindi | 0.7908 | 0.7835 | 0.8035 | 0.8272 | 0.9038 | 0.8700 | 0.9094 | 0.9136 |
| Indonesian | 0.9156 | 0.8894 | 0.9168 | 0.9239 | 0.9042 | 0.9000 | 0.9032 | 0.9097 |
| Spanish (LA) | 0.8858 | 0.8583 | 0.8835 | 0.8918 | 0.8979 | 0.8600 | 0.8962 | 0.9025 |
| Russian | 0.8843 | 0.8633 | 0.8819 | 0.9000 | 0.8621 | 0.8100 | 0.8633 | 0.8724 |
| Swahili | 0.8119 | 0.8007 | 0.8200 | 0.8412 | 0.8380 | 0.8200 | 0.8588 | 0.8649 |
| **——Important——** | | | | | | | | |
| Bengali | 0.8465 | 0.8328 | 0.8556 | 0.8751 | 0.8952 | 0.8700 | 0.9046 | 0.9124 |
| Farsi | 0.8578 | 0.8490 | 0.8704 | 0.8933 | 0.8801 | 0.8300 | 0.8836 | 0.8931 |
| Malay | 0.8964 | 0.8736 | 0.9010 | 0.8977 | 0.8901 | 0.8560 | 0.9032 | 0.9016 |
| Marathi | 0.7255 | 0.7113 | 0.7164 | 0.7402 | 0.8741 | 0.8800 | 0.8858 | 0.8930 |
| Myanmar | — | 0.8629 | — | 0.8673 | — | 0.8840 | — | 0.8801 |
| Tagalog | — | 0.8140 | — | 0.8503 | — | 0.8200 | — | 0.9005 |
| Tamil | — | 0.9348 | — | 0.92420 | — | 0.8600 | — | 0.8540 |
| Urdu | 0.8062 | 0.7892 | 0.8168 | 0.8125 | 0.8618 | 0.8400 | 0.8770 | 0.8821 |
| **——Priority——** | | | | | | | | |
| Amharic | 0.8390 | 0.8306 | 0.8400 | 0.8641 | 0.8161 | 0.8500 | 0.8691 | 0.8722 |
| Dari | — | — | — | — | — | — | — | — |
| Dinka | — | — | — | — | — | — | — | — |
| Eritrean Tigrinya | — | — | — | — | — | — | — | — |
| Ethiopian Tigrinya | — | — | — | — | — | — | — | — |
| Hausa | — | 0.8548 | — | 0.7957 | — | 0.8100 | — | 0.8217 |
| Kanuri | — | — | — | — | — | — | — | — |
| Khmer | — | 0.7848 | — | 0.7928 | — | 0.8900 | — | 0.8854 |
| Kinyarwanda | — | — | — | — | — | — | — | — |
| Kurdish Kurmanji | — | 0.4583 | — | 0.8049 | — | 0.8090 | — | 0.8370 |
| Kurdish Sorani | — | — | — | — | — | — | — | — |
| Lingala | — | — | — | — | — | — | — | — |
| Luganda | — | — | — | — | — | — | — | — |
| Nepali | 0.8013 | 0.7854 | 0.8227 | 0.8413 | 0.9050 | 0.8600 | 0.9162 | 0.9192 |
| Nigerian Fulfulde | — | — | — | — | — | — | — | — |
| Oromo | — | — | — | 0.7912 | — | — | — | 0.9030 |
| Pashto | 0.7828 | 0.7617 | 0.7921 | 0.8112 | 0.8192 | 0.8430 | 0.8631 | 0.8720 |
| Somali | 0.7189 | 0.7077 | 0.6995 | 0.7018 | 0.6617 | 0.8200 | 0.6831 | 0.4749 |
| Zulu | — | — | — | — | — | — | — | — |

Table 26: COMET Score Differences between 2023 and 2025 for Microsoft and Google