



Multilingual Translation API: Accuracy on FLORES-200 for Thai and Asian Language Pairs

Production service measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. We measured the iApp Multilingual Translation API on the public FLORES-200 devtest benchmark: all 1,012 sentences in each of 13 directions, 13,156 translations, none of which failed. From Thai to English it scores chrF++ 56.1 (95% CI [55.4, 56.8]), on a par with Chinese to English (55.5) and Japanese to English (54.0) on the same reference sentences. Into Thai, English gives the best result (46.6); Chinese, Korean and Japanese give 40.1, 39.8 and 38.7. Burmese is the weak point: chrF++ 29.2 from English and 27.4 from Thai, and 55.4% and 53.0% of the answers contain fragments of unrelated languages. Across all directions, 89 answers were not in the target language, 48 of them left in the source language. Through the public API the median round trip was 0.75 s per sentence.

1 Introduction

The iApp Multilingual Translation API translates text between 28 languages, one text or a batch per request. This whitepaper measures the production service on a public, professionally translated benchmark, in the directions most used with Thai and in the further pairs that the product page reports. It reports every direction measured, including the weaker ones, with confidence intervals, and states the limits of the evidence in Section 7.

2 Test Data

FLORES-200 devtest [1]: 1,012 English sentences from Wikinews, Wikijunior and Wikivoyage, translated by professional translators into some 200 languages, with the same sentences in the same order in every language (CC-BY-SA 4.0). Because every language holds the same sentences, scores of different directions are measured on the same content. We used all 1,012 sentences in each of 13 directions, 13,156 translations in all:

- **Thai pairs:** Thai to and from English, Chinese (Simplified), Japanese and Korean, and Thai to Burmese.
- **Further pairs reported on the product page:** English to Burmese, English to Japanese, Japanese to English and Chinese to English.

The training data of the translation model is not public, so we cannot test whether FLORES-200, public since 2022, was seen in training (Section 7).

3 Method

Each sentence was sent once to the production service, unbilled and from inside iApp's network, at most two requests at a time, with the fields a customer sends (`text`, `source_lang`, `target_lang`) and the result cache switched off, so that every answer is a fresh translation. A seeded sample was then sent through the public API (Section 6).

- **chrF++** [2]: the F-score of character n-grams (up to 6) and word n-grams (up to 2) between output and reference, 0–100. It is the headline metric because it works for languages that do not separate words with spaces (Thai, Chinese, Japanese, Burmese).
- **BLEU** [3]: the 13a tokeniser for English output and the FLORES-200 SentencePiece tokeniser (spBLEU [1]) for every other target language, so that scores into Thai, Chinese, Japanese, Korean and Burmese are tokenised the same way. Both metrics are corpus-level and computed with sacreBLEU [4] (signatures in Section 8).
- **Off-target**: an answer not mainly written in the script of the target language, judged on its letters outside the Latin alphabet (for example a Thai sentence returned unchanged when Chinese was requested).
- **Mixed script**: the share of answers containing any letter of a script that belongs to neither the target nor the source language nor the Latin alphabet — a fragment of an unrelated language inside the answer.
- **Latin kept**: among sentences whose reference contains no word in Latin script, the share of answers that do — mostly names and terms left in their original spelling rather than written in the target script.
- **Confidence intervals**: 95% percentile bootstrap over sentences (2,000 resamples, seed 1337), recomputing the corpus score each time.

4 Results

Table 1: FLORES-200 devtest, 1,012 sentences per direction; 0 of 13,156 requests failed. Off-target (answers), mixed script and Latin kept as defined in Section 3. Time: median processing time per sentence on the server.

Direction	n	chrF++ [95% CI]	BLEU [95% CI]	Off-target	Mixed (%)	Latin (%)	Time (s)
<i>Thai pairs</i>							
Thai → English	1,012	56.1 [55.4, 56.8]	27.5 [26.5, 28.4]	0	0.0	–	0.54
English → Thai	1,012	46.6 [45.9, 47.4]	41.1 [40.1, 42.0]	0	0.4	6.7	0.78
Thai → Chinese	1,012	23.7 [22.9, 24.5]	26.9 [25.9, 27.9]	20	0.5	6.0	0.57
Chinese → Thai	1,012	40.1 [39.4, 40.7]	31.6 [30.7, 32.5]	6	1.0	11.4	0.83
Thai → Japanese	1,012	26.3 [25.2, 27.6]	21.9 [21.0, 22.8]	0	0.1	3.9	0.66
Japanese → Thai	1,012	38.7 [38.0, 39.4]	31.2 [30.4, 32.1]	13	1.1	4.7	0.82
Thai → Korean	1,012	28.4 [27.7, 29.1]	20.4 [19.5, 21.3]	0	0.5	5.6	0.73
Korean → Thai	1,012	39.8 [39.1, 40.5]	32.2 [31.3, 33.1]	8	0.5	9.1	0.81
Thai → Burmese	1,012	27.4 [26.9, 27.9]	8.3 [7.8, 8.8]	26	53.0	30.4	1.29
<i>Further pairs reported on the product page</i>							
English → Burmese	1,012	29.2 [28.7, 29.8]	10.5 [9.9, 11.0]	16	55.4	32.6	1.17
English → Japanese	1,012	32.3 [30.9, 34.0]	29.7 [28.7, 30.8]	0	0.4	5.0	0.65
Japanese → English	1,012	54.0 [53.2, 54.7]	25.0 [24.1, 26.0]	0	0.0	–	0.54
Chinese → English	1,012	55.5 [54.9, 56.3]	27.1 [26.2, 28.2]	0	0.0	–	0.54

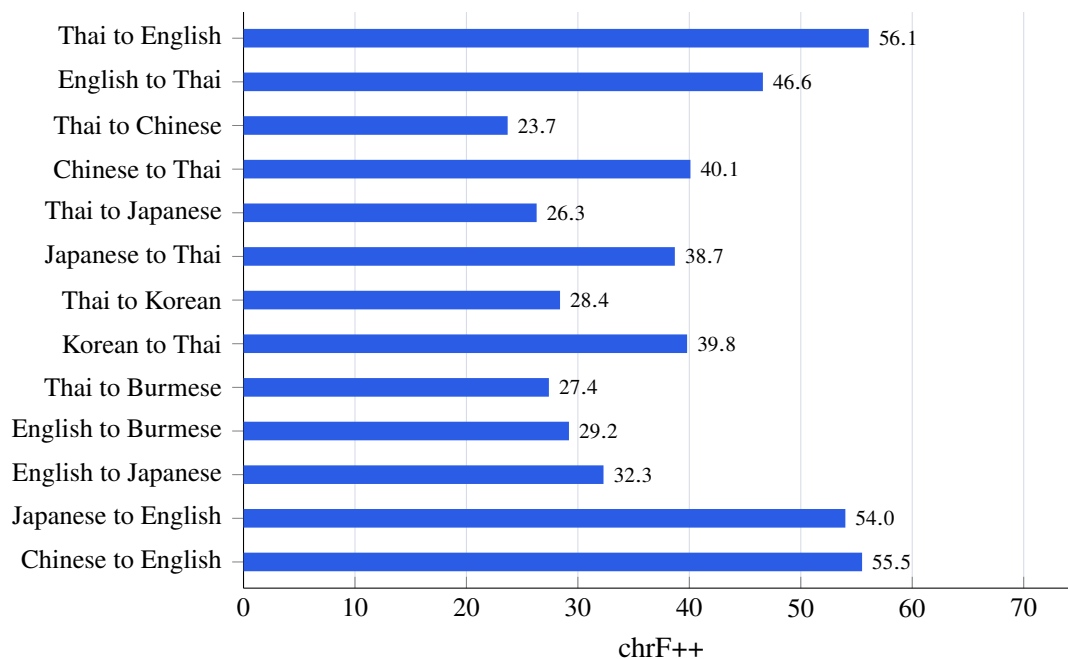


Figure 1: chrF++ by direction on FLORES-200 devtest, Thai pairs first.

5 Discussion

Into English, the three source languages perform alike. FLORES-200 holds the same sentences in every language, so directions with the same target share one reference and can be compared directly. Into English, Thai (56.1), Chinese (55.5) and Japanese (54.0) are within a few points of one another. Scores into different target languages cannot be compared this way: chrF++ is lower by construction for Chinese and Japanese, which do not separate words.

Into Thai, English is the strongest source. English to Thai scores 46.6, against 40.1 from Chinese, 39.8 from Korean and 38.7 from Japanese on the same Thai reference. Likewise into Japanese, English (32.3) scores above Thai (26.3).

Burmese output is not reliable. 55.4% of the answers from English and 53.0% from Thai contain fragments of unrelated languages — for example Tamil, Telugu, Lao or Khmer words inside a Burmese sentence — and 42 of the 1,012 answers in the two directions together are mainly in another language. In every other direction the share of mixed answers is at most 1.1%. Burmese output should be reviewed before use.

Some answers are left in the source language. Of the 89 off-target answers, 48 repeat the source text, or most of it, instead of translating it; most are Thai sentences returned in Thai when Chinese was requested (20 off-target answers in that direction) and Japanese returned in Japanese when Thai was requested (13). The other 41 are in another language or a mixture, nearly all of them Burmese.

Names are often kept in their original spelling. When the service is not sure of an accepted local form, it writes a name or term in Latin script: 6.7% of English to Thai answers do so where the Thai reference has none, and 32.6% of English to Burmese answers. This is often what customers want for brand and product names, but it lowers reference-based scores, as the references transliterate every name.

6 Production Performance

On 28 October 2026 we sent a seeded sample of 300 sentences (seed 1337, all 13 directions) through the public endpoint <https://api.iapp.co.th/v3/store/nlp/multilingual-translation>, two requests at a time, from a client on the public internet, with the result cache switched off.

- **Latency.** Median round trip 0.75 s per sentence, 90th percentile 1.45 s, 95th percentile 1.70 s. The service's own processing time had a median of 0.68 s, so the gateway and the network added a median of 0.06 s.
- **Failures.** 0 of 300 requests.
- **Same quality, not always the same words.** 61.0% of the answers were identical, character for character, to the answers measured directly. Repeating a request can return slightly different wording, with the same quality: chrF++ 38.8 through the API against 38.9 directly (mean over directions; pooled difference -0.1 [-0.4 , $+0.2$]).
- **Batch endpoint.** The service translates up to eight texts of a batch at the same time, so a batch takes far less server time per sentence than single requests (0.12 s against 0.70 s, all 2,024 sentences of both Thai–English directions in requests of 20). End to end, 40 Thai sentences sent one at a time took 0.70 s each, and the same sentences in 2 batches 0.10 s each, 86% less. Translated side by side, the wording differs more often from that of single requests (50.1% identical), with the same quality: chrF++ 56.2 from Thai to English and 46.4 from English to Thai, against 56.1 and 46.6 one at a time. The batch figures were measured after the service update of 28 October 2026 that introduced this parallel translation; single requests are unchanged by it.

7 Limitations

- **Possible exposure to the test set.** FLORES-200 has been public since 2022 and the training data of the translation model is not public, so we cannot rule out that some test sentences were seen in training, which would raise the scores.
- **Automatic metrics** reward closeness to one reference translation; a correct translation worded differently scores lower, and keeping a name in its original spelling is counted as an error against a reference that transliterates it. Scores of different target languages are not directly comparable, as chrF++ depends on the script and the morphology of the language.
- **One domain.** FLORES-200 is encyclopaedic and news text, one sentence at a time. Long documents, conversational text and specialised terminology are not covered.
- **Coverage.** 13 of the 756 possible directions between the 28 languages were measured; the product page's semantic-similarity figures were not re-measured.
- **One run on one day.**

8 Reproducibility

FLORES-200 is public. The folder of this paper holds the data fetcher, the benchmark client, the raw responses (one line per sentence) and the scoring script that writes every number in this document. Metric signatures:

- chrF++: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.6.0
- BLEU, English output: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0

- BLEU, other output: nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.6.0

The files are available from iApp Technology on request: sale@iapp.co.th.

References

- [1] NLLB Team. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672, 2022.
- [2] M. Popović. chrF++: words helping character n-grams. WMT 2017.
- [3] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. ACL 2002.
- [4] M. Post. A Call for Clarity in Reporting BLEU Scores. WMT 2018.

About iApp Technology. iApp Technology Co., Ltd. is a Thai artificial-intelligence company providing OCR, document understanding, speech, identity verification and large-language-model services for the Thai language. Web: iapp.co.th · E-mail: sale@iapp.co.th · Tel: +66 2 124 4041