



ChindaMT Translation API: Thai–English Accuracy and Rule Following in Production

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

Executive summary. We measured the ChindaMT Translation API as customers receive it, on the 1,012 sentences of the public FLORES-200 devtest benchmark in both directions. Without rules it scores chrF++ 54.6 (95% CI [53.9, 55.4]) from Thai to English and 45.7 [44.9, 46.5] from English to Thai, with no failed request. We then tested rules automatically. With glossary rules whose required terms come from the human reference, the required term appeared in 94.2% of 1,802 cases from English to Thai (82.5% without the rule) and in 83.7% of 1,692 cases from Thai to English (75.3% without), and the translations scored slightly higher, not lower. A rule to keep a proper name in English was followed in only 15.4% of 533 sentences: the model prefers the established Thai spelling of a name, and an explicit wording raised this to 38.0%. Through the public API one sentence took a median of 0.37 s, and a batch of 20 sentences 0.86 s.

1 Introduction

The ChindaMT Translation API translates text between Thai and English and follows instructions, called rules, for terms, tone, length and format. Its model, ChindaMT-4B, is described in a peer-reviewed paper [1], which compares it with open translation models of similar size using a judge model and native-speaker raters. This whitepaper does not repeat that study. It measures the deployed API, with reference-based metrics that anyone can recompute, and tests the rule feature with a check that needs no judge: whether a required term is present in the output.

2 Test Data

FLORES-200 devtest [2]: 1,012 English sentences from Wikinews, Wikijunior and Wikivoyage, translated by professional translators into Thai and the other languages of the benchmark (CC-BY-SA 4.0). A sentence has on average 130 characters in English and 125 in Thai. All sentences are used in both directions.

Rule tests built from FLORES-200 devtest by a deterministic script:

- **Glossary, both directions.** A source term and a target term form a pair if a public dictionary (the Thai entries of Wiktionary [7]) links them, the source term occurs in the source sentence and the target term occurs in the human reference. Up to two pairs per sentence become rules of the form “Translate *source term* as *target term*”, the form the API documentation uses. As each required term is one a professional translator used in that sentence, every rule can be met by a good translation; this is the reference-derived setting of the WMT terminology task [6].
- **Keep a name in English.** For English sentences with a capitalised name inside the sentence (a place, an organisation, a product, a historical period), the longest such name becomes the rule “Keep the name *Name* in English”. The reference, like most Thai writing, spells these names in Thai script, so the rule asks for something the model does not do by default.

Held out from training. The translation training set of ChindaMT is public (ChindaMT-Grounded, 1,973,358 records). We searched it for every devtest sentence, English and Thai, after removing case, spacing and punctuation, in windows of 32 characters. Sentences found in full: 0 English and 1 Thai, a

short generic sentence (“President George W. Bush welcomed the announcement”). Sentences sharing at least one window with a training record: 6 English and 3 Thai.

3 Method

Each test item 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: `source_lang`, `target_lang`, `text` and, for the rule tests, `rules`. Seeded samples were also sent through the public API (Section 6). The scores in Section 4 come from the direct measurements.

- **chrF++** [3]: 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 does not depend on word segmentation, which written Thai does not mark.
- **BLEU** [4]: 13a tokeniser for English output, the FLORES-200 SentencePiece tokeniser (spBLEU [2]) for Thai output. Both metrics are corpus-level and computed with sacreBLEU [5] (signatures in Section 8).
- **Term presence**. English output: the required term as a whole word, ignoring case; we also report a lenient match that accepts another form of the same word (*temporary* for *temporarily*). Thai output, which has no spaces between words: the term as a substring. A kept name must appear exactly, capitals included. “Without the rule” is the plain translation of the same sentence.
- **Confidence intervals**: 95% percentile bootstrap over sentences (2,000 resamples, seed 1337), recomputing the corpus score each time; paired for differences.

4 Results

4.1 Translation without rules

Table 1: FLORES-200 devtest, no rules. Failed: requests without a translation. Warned: answers with a warnings entry. Off-target: output not mainly in the target language.

Direction	Sentences	chrF++ [95% CI]	BLEU [95% CI]	Failed	Warned	Off-target
Thai → English	1,012	54.6 [53.9, 55.4]	27.1 [26.0, 28.1]	0	0	0
English → Thai	1,012	45.7 [44.9, 46.5]	41.2 [40.3, 42.2]	0	0	0

4.2 Rule following

Table 2: Required terms present in the output (%), without and with the rule. “Missed” counts the terms the plain translation did not contain; “Put in” is the share of those that the rule added.

Test	Items	Terms	Without	With rule [95% CI]	Missed	Put in
Glossary, Thai → English	951	1,692	75.3	83.7 [81.9, 85.5]	418	40.9
Glossary, English → Thai	978	1,802	82.5	94.2 [93.1, 95.3]	316	69.3
Keep a name, English → Thai	533	533	4.5	15.4 [12.6, 18.6]	509	12.0
multi-word names only	257	257	5.8	24.9		

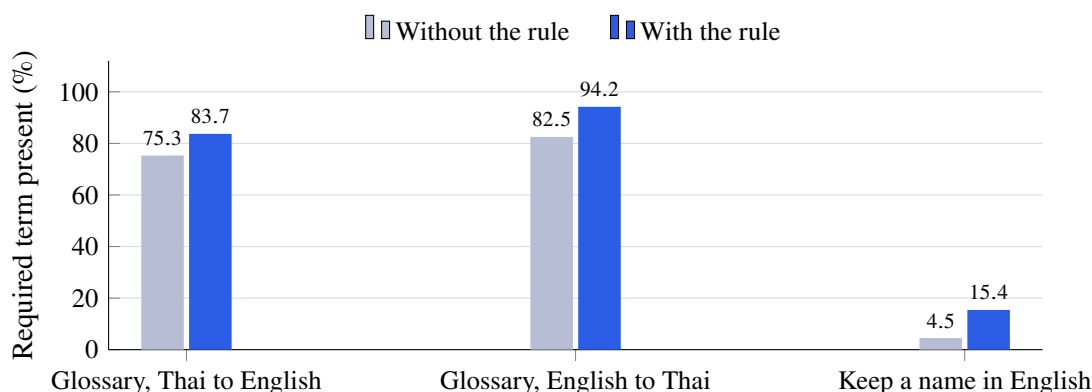


Figure 1: Share of required terms present in the output, without and with the rule.

Table 3: Translation quality on the same sentences without and with the rules. Δ : chrF++ with rules minus without, paired 95% CI.

Test	Items	chrF++ without	chrF++ with	Δ [95% CI]	Warned	Off-target
Glossary, Thai $\rightarrow$ English	951	54.8	56.1	+1.3 [+1.0, +1.7]	0	0
Glossary, English $\rightarrow$ Thai	978	45.8	46.8	+1.1 [+0.7, +1.4]	0	0
Keep a name, English $\rightarrow$ Thai	533	46.3	46.6	+0.3 [-0.1, +0.7]	0	0

Table 4: Keeping a name in English with two other wordings of the rule, on the same 150 seeded sentences (seed 1337). This check was added after the main run.

Wording of the rule	Items	Main (%)	This wording (%) [95% CI]
“Keep ‘Name’ in English. Do not translate or transliterate it into Thai.”	150	16.0	38.0 [30.0, 46.0]
“Keep names such as <i>Name</i> in English”	150	16.0	14.0 [8.7, 20.0]

5 Discussion

Glossary rules work, and more reliably into Thai. From English to Thai the required term was present in 94.2% of cases, and the rule put in 69.3% of the terms the plain translation lacked; every rule of a sentence was met in 89.6% of sentences. From Thai to English the strict figure is 83.7% (75.3% without). About half of the remaining misses are another form of the required word (*temporary* for *temporarily*); with the lenient match the figure is 91.8% (83.4% without). Many of the rest are loose dictionary pairings on very common words, which a customer’s own glossary would not contain.

Rules did not harm the translation. On the same sentences chrF++ rose by +1.1 and +1.3 points with glossary rules, and no answer was warned or off target. Part of the rise is expected, as the terms come from the reference; the point is that adding rules did not degrade the rest of the sentence.

Keeping a name in English is not reliable. With the rule, 15.4% of names were kept in Latin script (24.9% of multi-word names). The model keeps names that have no common Thai spelling (in this test, for example, *ZMapp* and *ADT Corporation*) and writes the established Thai form of the rest, as the reference does. An explicit wording that forbids translation and transliteration raised the rate to 38.0% (Table 4). Customers who need names in Latin script should check the output, or restore the names after translation.

Without rules, accuracy is stable and complete. All 1,012 sentences were translated in both directions, with no warning and no off-target answer.

6 Production Performance

On 28 October 2026 we sent seeded samples through the public endpoint <https://api.iapp.co.th/v3/store/text/mt/translate> from a client on the public internet: 200 plain FLORES sentences (seed 1337, both directions) and 100 rule-test items, then 80 further sentences (seed 2026) one at a time to time the round trip.

- **Latency.** One sentence: median 0.37 s (Thai to English 0.33 s, English to Thai 0.38 s), 90th percentile 0.54 s, 95th percentile 0.63 s, network time included.
- **Batch endpoint.** 6 requests of 20 sentences to `/translate/batch` took a median of 0.86 s, 0.043 s per sentence. For many short texts, one batch request is far faster than single requests.
- **Same quality, not always the same words.** 76.5% of the public-API answers were identical, character for character, to the direct answers for the same sentences (69.0% with rules; 78.9% between `/translate/batch` and `/translate`). The service is not strictly deterministic: sending the same text twice can return a slightly different wording. The quality is the same: chrF++ 52.3 through the API against 52.3 directly, difference 0.0 [−0.2, +0.2].
- **Failed requests:** 0 of 200 plain, 0 of 100 with rules, 0 of 80 timed and 0 of 6 batch requests.

7 Limitations

- **Automatic metrics** reward closeness to one reference translation; a correct translation worded differently scores lower. The paper [1] reports judge-model and human preference results.
- **One domain.** FLORES-200 is encyclopaedic and news text. Business e-mail, interface strings, legal and conversational text are not covered.
- **The rule test covers terminology and names only.** Tone, length and format rules cannot be checked by string matching and were not tested. The dictionary pairs some words loosely, and a Thai substring match can find a term inside a longer word, in the plain and the ruled output alike. As the terms come from the reference, the chrF++ change with rules is biased upward.
- **Training overlap** was checked against the published translation training set only; earlier, general-purpose training stages cannot be checked.
- **One run on one day.** The wording comparison in Table 4 was added after the main results were known and covers two wordings only.

8 Reproducibility

FLORES-200 and the Wiktionary extract are public. The folder of this paper holds the data fetcher, the rule-test builder, the benchmark client, the raw responses (one line per test item), the training-overlap check 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, Thai 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] T. Chay-intr, K. Harnchang, M. Thabua, K. Viriyayudhakorn and T. Theeramunkong. Reference-Grounded Data Curation for Instruction-Following Thai-English Machine Translation. ACL-IJCNLP 2026. <https://arxiv.org/abs/2609.34770>
- [2] NLLB Team. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672, 2022.
- [3] M. Popović. chrF++: words helping character n-grams. WMT 2017.
- [4] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. ACL 2002.
- [5] M. Post. A Call for Clarity in Reporting BLEU Scores. WMT 2018.
- [6] K. Semenov, V. Zouhar, T. Kocmi, D. Zhang, W. Zhou and Y. E. Jiang. Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. WMT 2023.
- [7] T. Ylonen. Wiktextextract: Wiktionary as Machine-Readable Structured Data. LREC 2022. Data: <https://kaikki.org>

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