



Thai Question Answering: Accuracy on Two Public Thai Reading-Comprehension Benchmarks

Endpoint /v3/store/nlp/question/answer/v3 · Measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. The iApp Thai Question Answering API reads a Thai passage and a question about it and returns the answer. On the held-out test split of the iApp Wikipedia QA dataset — 739 questions on 192 Thai Wikipedia passages — its answers reach a token F1 of **82.4%** (95% CI 80.5–84.3) against the human answers. On 500 questions from XQuAD, a professionally translated Thai benchmark, the F1 is 81.2% (CI 78.2–84.0), the same within the margin of error. Exact match is lower, 53.2% and 64.2%, mostly because the answer and the reference differ in extent rather than in content: one is a shorter or longer form of the other. When the passage does not contain the answer, the service says so in 90.5% of cases (CI 86.5–94.5). No request failed, and the public API returned the same answer as the production service for all 120 questions sent through it, with a median response time of 1.70 s.

1 Introduction

The Thai Question Answering API (POST /v3/store/nlp/question/answer/v3) takes a Thai passage and a question in JSON and returns an answer drawn from that passage. Typical uses are answering customer questions from a product manual, a policy or a contract, and pulling single facts out of long documents. This paper measures how often the answer is right, on public data that anyone can use to repeat the measurement, and where it goes wrong.

2 Test Data

iApp Wikipedia QA, test split

739 questions written by human annotators about 192 passages from Thai Wikipedia, each with one reference answer marked in the passage [1]. MIT licence. The test split is the partition its publishers hold out for evaluation.

XQuAD, Thai

Professional Thai translations of passages, questions and answers from the English SQuAD v1.1 development set [2]. CC BY-SA 4.0. A seeded random sample of 500 of the 1,190 questions (seed 1337), on 218 passages.

iApp did not train or tune the service on either dataset. Both are drawn from Wikipedia and have been public for several years, so we cannot rule out that some of their text was seen when the underlying model was built (Section 7).

3 Method

Each question was sent with its passage, in the request format of the API documentation (`{"question", "document"}`), to the production service — the same instances that `answer.api.iapp.co.th` — on 28

October 2026, with at most two requests in flight. A seeded sample of 120 test questions was also sent through `api.iapp.co.th` (Section 6).

- **Token F1:** the overlap of words between the answer and the reference, as the harmonic mean of word precision and recall, as in SQuAD [3]. Words are segmented with the *newmm* dictionary segmenter of PyThaiNLP [4] after Unicode NFKC normalisation, lower-casing and removal of punctuation and symbols.
- **Exact match:** the normalised answer equals the normalised reference, ignoring spaces.
- **Contains reference:** the normalised reference appears, whole, inside the normalised answer. It credits correct answers that add words.
- **95% confidence intervals:** percentile bootstrap, 2,000 resamples, seed 1337, resampling passages rather than questions, because questions about the same passage are not independent.
- **Unanswerable questions:** 200 test questions, each sent with the passage of a different, randomly chosen article (seed 1338). The right behaviour is to say that the passage does not answer the question.

4 Results

Table 1: Accuracy (%) with 95% confidence intervals.

Test set	Questions	Token F1	Exact match	Contains reference
iApp Wikipedia QA, test	739	82.4 [80.5, 84.3]	53.2 [49.4, 56.8]	68.2 [64.3, 72.0]
XQuAD, Thai	500	81.2 [78.2, 84.0]	64.2 [59.3, 68.5]	85.0 [81.5, 88.2]

No request failed (0 errors on the test split, 0 on XQuAD). Empty answers: 1 on the test split, 0 on XQuAD. Questions the service wrongly said the passage could not answer: 1 on the test split, 2 on XQuAD.

4.1 By question type

Table 2: Exact match and token F1 (%) by question type, classified from the question’s Thai interrogative words.

Question type	iApp Wikipedia QA, test			XQuAD, Thai		
	Questions	EM	F1	Questions	EM	F1
Yes / no	1	0.0	50.0	2	50.0	92.9
Who	72	79.2	94.0	50	84.0	89.0
When	115	59.1	87.2	50	72.0	87.8
How many / how much	85	50.6	77.8	76	47.4	76.7
Where	63	47.6	76.7	25	76.0	89.1
What / which	367	47.4	80.6	286	62.9	78.9
Other	36	58.3	83.6	11	63.6	86.3

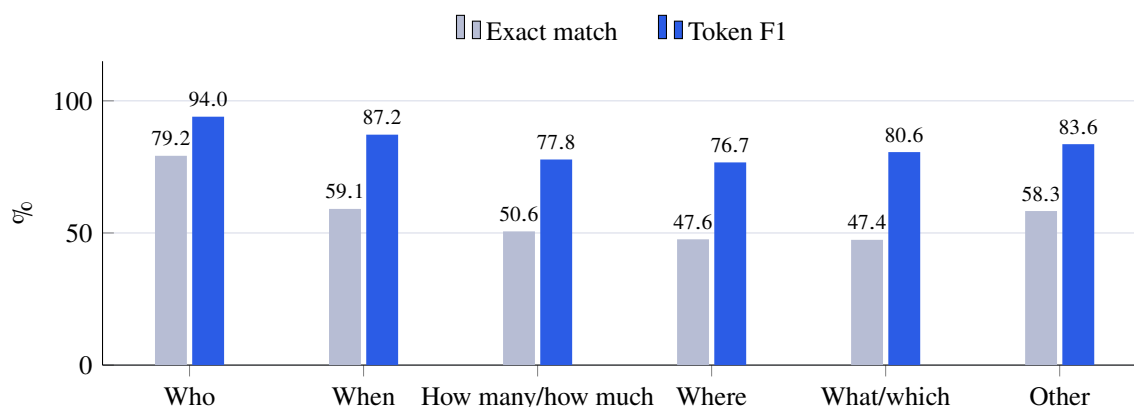


Figure 1: Exact match and token F1 by question type, iApp Wikipedia QA test split (types with at least 20 questions).

4.2 Why exact match is lower than F1

On the iApp test split, 346 answers are not exact matches. Of these, 183 are a shorter part of the reference — the references in this dataset often carry a label or extra words, as in “scientific name: . . .” where the service returns only the name — and 111 contain the reference and add to it, such as a unit or an abbreviation in brackets. The other 52 overlap with the reference partly or not at all; they include answers that point to a different but defensible part of the passage, and references that are themselves cut off or misspelt. On XQuAD the pattern reverses: of 179 non-exact answers, 104 contain the reference and add to it and 23 are shorter. Answers are short: the median answer is 3 words against 4 for the reference, and only 6.1% are more than twice as long as the reference.

4.3 By passage length

Table 3: Accuracy (%) by passage length in characters, iApp Wikipedia QA test split.

Passage length	Questions	EM	F1
under 500	248	56.0	82.9
500–999	267	52.1	82.1
1,000–1,999	166	51.2	81.9
2,000 or more	58	51.7	82.8

4.4 When the passage does not contain the answer

Sent with the passage of an unrelated article, the service stated that the passage does not give the answer for 181 of 200 questions (90.5%, 95% CI 86.5–94.5). For 5 it returned the question itself as the answer, and for 14 it returned an answer that the passage does not support: some were correct from general knowledge, others were names or figures taken from the unrelated passage. Declines were detected from Thai and English phrases such as “no information”, “not stated” and “unrelated”.

5 Discussion

The service finds the right fact. An F1 of 82.4% and 81.2% means that, for most questions, the words a person marked as the answer are in the returned text. Exact match understates this, because the two datasets draw the boundaries of an answer differently and the service does not copy either convention.

Accuracy does not depend on passage length within the range tested (passages of up to 4,636 characters): F1 is flat across the length bands. It is lowest for counting and quantity questions (*how many*, *how much*) and, on the iApp set, for *where* questions.

It rarely invents an answer. Given a passage that does not answer the question, the service declines in 90.5% of cases. An application that must never show an unsupported answer should still check the returned text against the passage; echoing the question is a failure mode worth filtering.

Native and translated text score alike. XQuAD’s passages are translations from English, with transliterated names; the F1 intervals of the two sets overlap.

6 Production Performance

The 120 questions sent through `api.iapp.co.th` returned 120 answers (status codes: 200: 120); 120 (100.0%) were identical to the answers the production service gave the same questions during the full run, so the token F1 on the sample is the same on both paths (82.7%). End to end from a client in Bangkok, the median time was 1.70 s and the 90th percentile 1.89 s; inside the data centre the service took a median of 1.75 s. The public-API check cost 313 IC, as expected at 1 IC per 400 characters of passage and question (313 IC for these 120 questions).

7 Limitations

- **Exposure.** Both datasets are public and drawn from Wikipedia. We cannot verify that none of their text was seen while the underlying model was built; scores on private documents may differ.
- **One reference per question.** Both datasets give a single reference answer, so a correct answer phrased differently scores below 100%. Some references in the iApp test split are truncated or misspelt.
- **Domain.** Encyclopaedia text only; contracts, manuals, chat logs and scanned documents were not measured.
- **Question types** are assigned by keyword and are approximate.
- **Unanswerable questions** were simulated with unrelated passages; questions about a related passage that lacks the specific fact are harder and were not measured.

8 Reproducibility

Both datasets are public. The benchmark and scoring scripts and the raw answer to every question are available from iApp Technology on request, sale@iapp.co.th.

References

- [1] K. Viriyayudhakorn and C. Polpanumas. *iapp_wiki_qa_squad*. Zenodo, 2021. https://huggingface.co/datasets/iapp/iapp_wiki_qa_squad
- [2] M. Artetxe, S. Ruder and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *Proc. ACL*, 2020.

- [3] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, 2016.
- [4] W. Phatthiyaphaibun et al. PyThaiNLP: Thai natural language processing in Python. In *Proc. NLP-OSS*, 2023.

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