



OpenThai 2.0 on Thai National Examinations, Instructions and Documents

Production build v2.0.3 · Measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. OpenThai 2.0 is iApp’s open-weight Thai vision-language model, served through an OpenAI-compatible chat API. We measured the production service on public Thai test sets. On the 443 Thai national examination questions of OpenThaiEval that a text-only model can score (A-Level, TGAT, TPAT-1 and O-NET at grades 9 and 12), it chose the correct option on 83.1% (95% CI 79.5–86.7). On a seeded 30 of the 95 questions of the Investment Consultant licence examination in ThaiExam it scored 73.3% (56.7–86.7). It followed every verifiable instruction in 90.0% of a seeded 20 of the 215 IFEval-TH prompts (75.0–100.0), read the 104-image OCR-Eval-104 set at a character error rate of 0.080 and Thai handwriting at 0.247. Its weakest papers are TPAT-1 (60.0%), the medical-school aptitude test, and O-NET Thai language (73.3%), where 9 answers were still reasoning when they reached the 16,384-token output limit. Through the public API, answers matched the model service’s, but a request that is not streamed fails with HTTP 504 once the model has reasoned for 180 s, which 15.9% of the examination answers needed; streaming avoids this. We also traced the model page’s published figures to their raw outputs: the OpenThaiEval figure there is computed on a 500-question set in which 53% of the questions are translated international benchmarks and none is from O-NET.

1 Introduction

The OpenThai 2.0 API (`/v3/11m/openthai2p0/chat/completions`) serves the production build v2.0.3 of OpenThai 2.0, a 27-billion-parameter model that answers in Thai, reads images and calls tools. Its model page reports benchmark figures measured on the training servers before release. This paper measures the service that customers actually call, on public test sets, and states for each figure the raw outputs it comes from. It covers three of the model’s claims: Thai knowledge as tested by national examinations, instruction following, and document reading. It does not re-measure tool use (Section 7).

2 Test Data

Table 1: Test sets. All were used for evaluation only; examination content is not reproduced in this paper or its result files.

Set	Items	Source and licence
OpenThaiEval, Thai national exams	443	iapp/openthaieval: A-Level, TGAT, TPAT-1, O-NET M3 and M6 (Thai, social studies, mathematics, science). Of 515 rows, 72 are excluded as the dataset card specifies: no single answer key, a missing figure, or an image the text cannot carry. Compilation Apache-2.0; the items are NIETS and CUPT examination papers.
ThaiExam, Investment Consultant	30	scb10x/thai_exam (Apache-2.0), IC test split: the investment-consultant licence examination, levels 1 to 3; a seeded sample of its 95 questions. The other 470 ThaiExam questions are A-Level, TGAT, TPAT-1 and O-NET, the exam types above, and were not run (Section 7).
IFEval-TH	20	scb10x/ifeval-th (Apache-2.0): Thai prompts with instructions a program can verify (word counts, formats, keywords); a seeded sample of its 215 prompts, carrying 27 instructions.
OCR-Eval-104	104	openthaigpt/thai-ocr-evaluation (CC BY-SA 4.0): five categories of Thai and English images.
Handwriting, text-disjoint	916	sivakorn-su/thai-handwriting-textdisjoint-v1, test split (Apache-2.0); no transcription in it occurs in the training split.

Contamination check. We searched every fine-tuning file of the production build’s lineage for the test questions. Each question was cut into 20-character windows of unbroken text (about four to six Thai words), 4,668 windows in all; windows shared by three or more test items were dropped as generic exam phrasing. Of the 364 OpenThaiEval, 95 Investment Consultant and 213 IFEval-TH items long enough to check, 0, 0 and 0 have all their windows in the training data. 3 OpenThaiEval and 6 Investment Consultant questions share half or more of their windows with training rows; on inspection the shared text is domain terms (types of financial risk, names of agencies and royal projects) and Wikipedia-derived questions about the same facts, not the examination items. Excluding them changes the combined accuracy from 82.5% to 82.4%. Pretraining data of the base model cannot be inspected; these examinations have been public since 2023 or earlier.

3 Method

Service. The examination, instruction and reading sets were run in full against the production model service, the same instance that answers API calls, without passing through the billing gateway, at most four requests at a time. A seeded sample of 20 examination questions (seed 1337) was then sent through <https://api.iapp.co.th> with an iApp API key (Section 6).

Settings. Questions were sent as the model’s documentation recommends: no system prompt, reasoning on (the default), an output limit of 16,384 tokens, repetition penalty 1.05, and temperature 0.0 with a fixed seed so that runs repeat. Multiple-choice items use the prompt of the model’s release battery, which asks for the answer in the form “the correct answer is (n)”. Document reading follows the model card: a fixed Markdown transcription prompt, reasoning off, temperature 0.0, images at most 1,800 pixels on the longer side.

Scoring. The chosen option is the number after “the correct answer is”, or the only parenthesised option number in the answer. A reply with no extractable option, an empty reply and an HTTP error each count as wrong and are reported separately. IFEval-TH is scored with the IFEval rule checker as its dataset

card requires for Thai: word and sentence counts use a Thai tokenizer when the text is Thai. Character error rate (CER) is the edit distance to the reference over the reference length, after Unicode normalisation and whitespace collapsing, capped at 1 per item, as in the release battery. Intervals are seeded percentile bootstrap 95% intervals over items.

4 Results

4.1 Thai examinations

Table 2: Accuracy (%) of the production service by examination and subject, with 95% bootstrap intervals.

Set	Examination	Items	Accuracy	95% CI
OpenThaiEval	A-Level	101	96.0	[92.1, 99.0]
OpenThaiEval	TGAT	50	76.0	[64.0, 88.0]
OpenThaiEval	TPAT-1	40	60.0	[45.0, 75.0]
OpenThaiEval	O-NET M6 Thai	65	72.3	[61.5, 83.1]
OpenThaiEval	O-NET M6 social studies	55	81.8	[70.9, 90.9]
OpenThaiEval	O-NET M6 science	28	89.3	[78.6, 100.0]
OpenThaiEval	O-NET M6 mathematics	17	100.0	[100.0, 100.0]
OpenThaiEval	O-NET M3 Thai	25	76.0	[60.0, 92.0]
OpenThaiEval	O-NET M3 social studies	20	100.0	[100.0, 100.0]
OpenThaiEval	O-NET M3 science	26	84.6	[69.2, 96.2]
OpenThaiEval	O-NET M3 mathematics	16	87.5	[68.8, 100.0]
ThaiExam	Investment Consultant	30	73.3	[56.7, 86.7]
OpenThaiEval	All Thai national exams	443	83.1	[79.5, 86.7]
Both sets		473	82.5	[78.9, 85.8]

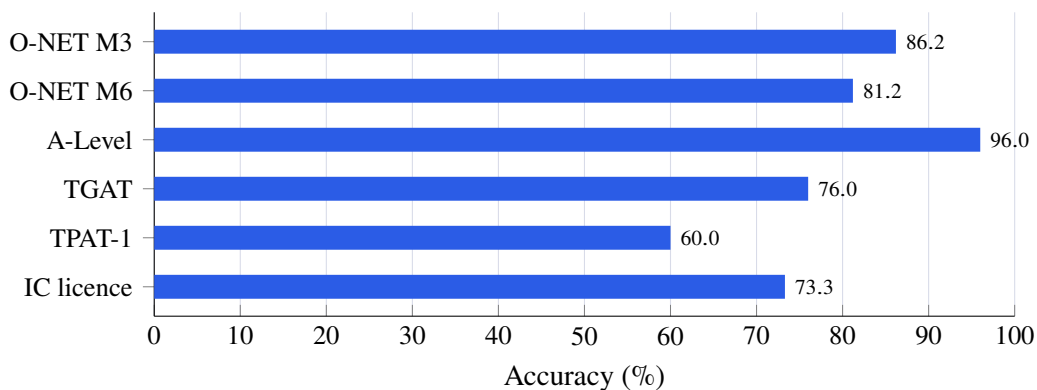


Figure 1: Accuracy by examination. O-NET, A-Level, TGAT and TPAT-1 from OpenThaiEval; the Investment Consultant (IC) licence examination from ThaiExam.

Failures among the 473 examination items, all counted as wrong: HTTP errors, 0; answers cut off at the output limit before choosing, 19; replies with no extractable option, 0. Answers scored from the parenthesised option alone, without the requested phrase: 1. The median answer used 672 output tokens, mostly reasoning (90th percentile 6,658, longest 16,384); the longest tenth of the answers account for 56% of all output tokens. Of the answers that reached the 16,384-token cap without answering, 1 repeated a phrase until the cap and 18 were still working through the problem: 9 O-NET Thai questions and 4 TGAT logic puzzles among them.

4.2 Instruction following

Table 3: IFEval-TH, production service. Prompt level: every instruction in the prompt followed. Loose: also accepted after removing Markdown emphasis or the first or last line.

Measure	Strict	Loose
Prompt level (20 prompts)	90.0 [75.0, 100.0]	90.0
Instruction level (27 instructions)	88.9	88.9

On the same 20 prompts, the v2.0.3 release outputs score 85.0%; at this sample size the interval is wide and the two agree within it. The model page quotes IFEval-TH figures from the release battery over all 215 prompts. Its scorer imported the English word and sentence counter, which counts a Thai sentence written without spaces as one or two words; scored that way, the production outputs above would read 85.0%. Re-scored with the Thai counter, the release outputs of v2.0.3 score 80.0% over all prompts (84.2% with the English counter), a solo re-run 81.9% (85.1%), and the v2.0.0 build the model page describes 77.2% (80.5%).

4.3 Document reading

Table 4: Character error rate, lower is better: OCR-Eval-104 by category, and the handwriting test split. “Release run” is the v2.0.3 release battery on the same images.

Set	Items	Production	Release run
Printed documents	22	0.042	0.042
Thai and English documents	12	0.059	0.059
Document photographs	24	0.051	0.060
Scene text	24	0.035	0.035
Handwriting (words)	22	0.213	0.218
OCR-Eval-104, all	104	0.080 [0.043, 0.126]	0.084
Handwriting, text-disjoint	916	0.247 [0.228, 0.264]	0.250

4.4 The release battery and the published figures

The OpenThaiEval figure on the model page (84.2%, shown there as a fraction, build v2.0.0) and the one in the v2.0.3 release record (86.0%) come from the release battery, whose raw outputs we located and re-scored with the extraction rule above; our rule agrees with the battery’s own on 1,500 of 1,500 items. The battery asked the first 500 rows of the dataset in its loading order: 210 A-Level, TGAT and TPAT-1 questions, 25 Investment Consultant questions and 265 questions from Belebele and XCOPA, which are Thai translations of international benchmarks. It contains no O-NET question. The same battery was run on Typhoon 2.5 (30B-A3B), the comparison named on the model page.

Table 5: Release battery, located raw outputs, accuracy (%). Not the production service; the same 500 questions and settings for each model.

Model	National exams (210)	IC (25)	Belebele + XCOPA (265)	All (500)
OpenThai 2.0 v2.0.3	80.5	88.0	90.2	86.0
OpenThai 2.0 v2.0.0	78.1	84.0	89.1	84.2
Typhoon 2.5	58.6	60.0	87.9	74.2

v2.0.3 leads Typhoon 2.5 by 11.8 points over the 500 questions (95% CI 7.8 to 15.6) and by 21.9 points on the national-exam questions (14.8 to 29.0). On the 191 national-exam questions that both the release run and our production run asked, the production service scored 83.2% against 83.2% (difference 95% CI -3.7 to 3.7 points), giving the same verdict on 179 of them.

5 Discussion

Thai knowledge. The model is strongest on A-Level (96.0%) and on O-NET mathematics (93.9%) and social studies (86.7%). It is weakest on TPAT-1 (60.0%), the medical-school aptitude paper of reasoning and medical-ethics questions, and on Thai language (73.3% on O-NET Thai). Many of the Thai-language misses are not wrong choices but answers that never arrived: on a question about tone marks, for example, the model weighed the tone of each word in turn until it reached the output limit. On the 81 O-NET Thai questions it did answer, it was right on 81.5%; over all of OpenThaiEval, 86.4% of the 426 answered questions were correct. The scored items are exactly those a text-only model can answer; questions that need a figure were excluded rather than counted as wrong.

Two ways to read the published figure. The release battery’s 500 questions are a fair comparison between models, since every model received the same items, but more than half of them measure translated reading comprehension and commonsense reasoning rather than Thai national examinations, and they omit O-NET. The full Thai national examination subset measured here is the better description of what “Thai national exams” means to a reader.

Instruction following depends on how Thai words are counted. 45 of IFEval-TH’s 306 instructions, in 39 of its prompts, set a length in words or sentences. Counting Thai text with an English tokenizer changes the prompt-level score from 90.0% to 85.0% on the production outputs; we report the Thai counting the dataset’s authors specify, and the English-counted figure only for comparison with earlier releases.

Document reading on the production service matches the release run on the same images. Printed pages read at 0.042 CER and document photographs at 0.051; the hard case is handwriting, 0.213 on OCR-Eval-104’s handwritten words and 0.247 on the handwriting test split, about one character in four. The model page names scene text as the weakest reading task; that figure comes from a scene-text question-answering set not re-run here, while OCR-Eval-104’s 24 scene-text images read at 0.035.

6 Production Performance

A seeded 20 of the examination questions were sent through the public API, two at a time, on 28 October 2026, without streaming. Of these, 19 returned an answer; on those, accuracy was 78.9% through the API against 73.7% from the model service, and 18 of 19 chose the same option. Because the model reasons before it answers, latency follows the length of the reasoning: the median call took 30.4 s through the API (90th percentile 92.9 s), against 23.6 s (204.9 s) on the model service for the same questions. Across the full runs, the median examination answer took 28.8 s (90th percentile 276.9 s); a handwriting image took 4.93 s and an OCR-Eval-104 page 1.6 s.

Long answers time out without streaming. Of the 20 questions, 1 failed with HTTP 504 after 180 s on each of three attempts: the API gateway closes a request that has sent nothing for that long, and the model service had needed 376 s to answer it. The billing log, which records every attempt, shows 1 more timed-out call: a question that timed out once and answered on the retry with shorter reasoning. On the model service, 15.9% of the examination answers (75) took longer than 180 s, so a non-streaming client asking questions of this kind should expect a similar share of timeouts. With streaming the same question returned HTTP 200 after 570 s, because the reasoning tokens keep the connection active (on that run the model reached the 16,384-token limit and did not answer). Clients that leave reasoning on should stream,

with `stream_options.include_usage` set so that the call is billed by its tokens. Timed-out calls were not charged.

Cost. The billing records show 24 calls for this paper, 4 of them timeouts at 0.00 IC, and 1.18 IC in all, the amount the token counts give at 0.01 IC per 1,000 input and 0.02 IC per 1,000 output tokens (7,838 input and 37,624 output tokens: 1.18 IC).

7 Limitations

- **One run, temperature 0.** The model’s documentation recommends temperature 1.0 for open questions; at that setting single answers vary. Batched serving is not bit-exact even at temperature 0: the release run and this run gave the same verdict on 179 of 191 shared questions.
- **Examination content may be in pretraining data.** Our check covers the fine-tuning data only. ThaiExam and OpenThaiEval have been public since 2023 and 2024.
- **Not re-measured here.** Tool use (the model page’s BFCL figure) needs the official function-calling harness and its several thousand cases, including multi-turn episodes, run against the production service with its tool-call parser; at the four-request limit this study kept to, that run did not fit in the measurement window. Reading of books and Royal Gazette pages uses a set iApp built from Wikisource scans that is not yet published. Both remain release-battery figures until they are measured on the production service.
- **Samples.** OpenThaiEval, OCR-Eval-104 and the handwriting test split were run in full. The Investment Consultant questions, IFEval-TH and the public-API comparison are seeded samples (seed 1337) of the sizes stated, and ThaiExam’s A-Level, TGAT, TPAT-1 and O-NET splits, which repeat the exam types of OpenThaiEval, were not run: at four requests at a time, with answers that reason for up to the 16,384-token cap, the full sets did not fit the measurement window. The samples’ intervals are correspondingly wider.
- **Small subsets.** Several examination subjects have fewer than 30 items; their intervals are wide.
- **The release-battery comparison** with Typhoon 2.5 is a located result on the training servers, not a production measurement.

8 Reproducibility

All test sets are public on Hugging Face at the revisions listed in `revisions.txt`. `bench.py` sends them to the service and stores one JSON line per item; `gen_results.py` computes every figure in this paper from those lines; `contamination.py` is the training-data search. The scripts, the per-item results and the vendored rule checker are available from iApp Technology on request: sale@iapp.co.th.

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