



OpenThai 2.0 Legal: Statute Citation Accuracy on NitiBench Through the Production API

Production service as of 28 October 2026 · Measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. OpenThai 2.0 Legal answers Thai legal questions and names the statute sections the answer rests on, after retrieving them from its index of Thai law. We sent the public NitiBench benchmark through the production service with retrieval on, as customers receive it, and scored the cited sections with NitiBench's own citation-F1 metric. Given only the question, the service scored 0.772 (95% CI [0.721, 0.820]) on 200 NitiBench-CCL questions, 0.444 [0.348, 0.547] on the 50 Revenue Department rulings of NitiBench-Tax, and 0.218 [0.119, 0.332] on 38 newer rulings held out from training. Retrieval decides the outcome: on CCL it found the governing section for 95.0% of questions, and those answers scored 0.812; on Tax rulings, which rest on several sections at once, it found all of them for 46.0%. With the right sections supplied, the model cites them with F1 1.000 on CCL. The deployed model matched the released weights on all seven release-card protocols, and the public API returned the same citations as the service for 94 of 100 sampled questions.

1 Introduction

OpenThai 2.0 Legal is iApp's open-weight Thai legal language model, offered as an OpenAI-compatible API at <https://api.iapp.co.th/v3/llm/openthai2p0-legal>. For each question the API retrieves the most relevant statute sections from its index, places them in the prompt, generates the answer, and checks every section the answer cites against the retrieved text before returning it, together with the sections it retrieved. This paper asks one question: when a customer sends a Thai legal question, does the answer cite the right law and section? It measures the whole service as customers receive it, separates retrieval from generation, and repeats the published model-card protocols on the production service to confirm that the deployed model is the released one.

2 Test Data

Table 1: Test sets. Gold citations are NitiBench's `relevant_laws` for every set.

Set	Questions	Source
NitiBench-CCL	200	200 of the 3,729 questions of the CCL split, on 21 Thai laws (civil, commercial, financial and corporate), one gold section each. They are the questions pinned by the model's release evaluation (seed 42, SHA-256 manifest), so every figure here pairs with the release run.
NitiBench-Tax	50	All 50 real Revenue Department rulings from 2021 or later, each resting on 1 to 9 sections of the Revenue Code (2.6 on average). 24 of them quote at least one governing section in the question.
Held-out Tax	38	Newer Revenue Department rulings in NitiBench format, built by the OpenThai team and kept out of every training stage.

NitiBench [1] is published under the MIT licence. **Overlap with training.** The model was fine-tuned on the training split of the dataset NitiBench-CCL derives from, after a decontamination pass against the

NitiBench test questions; CCL is therefore within the model’s training distribution, though its questions were not trained on. Revenue Department rulings from 2021 onwards entered late training rounds, so NitiBench-Tax is not fully clean either. The held-out set is the generalisation figure.

3 Method

Paths. Every arm was run against the production service directly (unbilled), at most two requests at a time, at temperature 0.0, seed 1337, reasoning off. A seeded sample of 100 end-to-end questions (seed 1337) was also sent through the public API, paced to its limit of 30 requests a minute, to confirm that it returns the same answers. Every response is stored.

Arms.

- *End to end, JSON* (headline): retrieval on, the question alone, with the system prompt the model was trained on, which asks for the answer and its citations as JSON. This is the setting the API reference recommends for machine-readable citations.
- *End to end, chat*: retrieval on, the question alone with no system prompt, as a chat client sends it; citations are read from the Thai prose.
- *Model alone*, the release-card protocols, with retrieval and the answer checks off: the gold sections supplied in the prompt; the gold sections mixed with NitiBench’s near-miss distractors (6.9 sections on average on Tax); and no sections at all, citing from memory.

Metric. Citation F1 per question between the cited and the gold (law, section) pairs, averaged over questions (macro F1), computed with NitiBench’s scoring code unchanged. Extra citations lower precision, so citing many sections is penalised. The model-alone arms are scored exactly as in the release evaluation. For the end-to-end arms, law names are compared in canonical form without their year on both sides, because the retrieval index and NitiBench spell some names differently; requiring the year to match as well gives the same figures (CCL 0.772, Tax 0.444). Prose answers are parsed by attributing each “section N” to the nearest law named in the answer. Confidence intervals are seeded percentile bootstraps over questions (2,000 resamples). A refused or failed request scores zero and is counted separately.

4 Results

4.1 End to end, retrieval on

Table 2: End-to-end citation F1 through the production service, JSON citations. Retrieval recall: share of gold sections among the eight retrieved. All gold: questions for which every gold section was retrieved. Refused: answers the API replaced with its standard refusal.

Set	n	F1 [95% CI]	Precision	Recall	Retrieval recall	All gold	Refused
NitiBench-CCL	200	0.772 [0.721, 0.820]	0.736	0.850	0.950	95.0%	3
NitiBench-Tax	50	0.444 [0.348, 0.547]	0.533	0.447	0.636	46.0%	0
Held-out Tax	38	0.218 [0.119, 0.332]	0.257	0.208	0.371	21.1%	2

When every gold section was retrieved, answers scored 0.812 on CCL (n=190), 0.647 on Tax (n=23) and 0.604 on the held-out rulings (n=8); when retrieval missed one, 0.000 (n=10), 0.271 (n=27) and 0.114 (n=30).

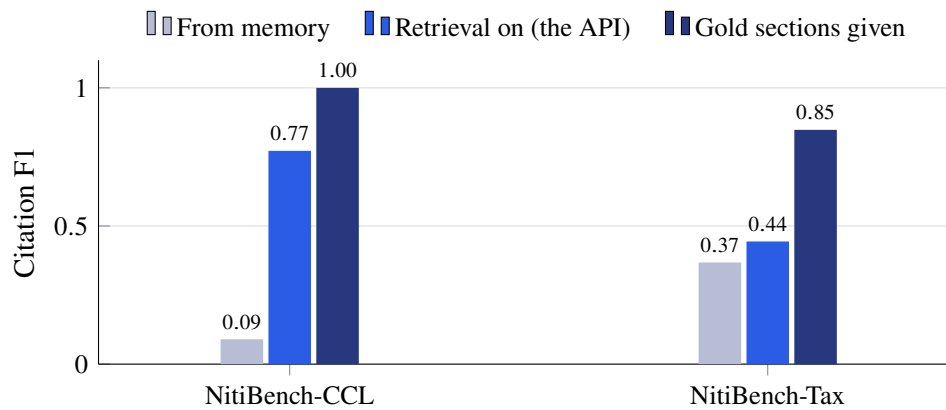


Figure 1: Citation F1 of the production service with no sections, with its own retrieval, and with the gold sections supplied.

4.2 End to end, chat

Table 3: The same questions sent as a chat message with no system prompt. Text only: law and section both read from the answer. Grounded: a section cited without its law takes the law of the retrieved section with that number, as the API's own citation check does. Section only: section numbers compared, law ignored. No section: answers that cite no section number.

Set	n	Text only [95% CI]	Grounded	Section only	No section
NitiBench-CCL	200	0.493 [0.428, 0.555]	0.514	0.521	43
NitiBench-Tax	50	0.436 [0.324, 0.540]	0.484	0.485	3

4.3 The model alone, against its release

Table 4: The release-card protocols on the production service (28 Oct 2026) against the release validation run of the published weights (24 Aug 2026): same questions, same scoring. Δ is production minus release with its paired 95% CI.

Protocol	n	Release	Production	Δ [95% CI]	Failed
Sections given, CCL	200	1.000	1.000	+0.000 [+0.000, +0.000]	0
Sections given, Tax	50	0.825	0.848	+0.023 [-0.020, +0.066]	0
Sections given, held-out Tax	38	0.806	0.846	+0.040 [-0.022, +0.118]	0
With distractors, Tax	50	0.652	0.685	+0.033 [-0.014, +0.081]	0
With distractors, held-out Tax	38	0.571	0.521	-0.050 [-0.128, +0.032]	0
From memory, Tax	50	0.364	0.368	+0.004 [-0.049, +0.066]	0
From memory, CCL	200	0.098	0.090	-0.008 [-0.042, +0.023]	0

5 Discussion

Retrieval decides the answer. With the gold sections in the prompt the model cites them perfectly on CCL (1.000), and from memory alone it rarely names the exact section (0.090). The production service sits between the two at 0.772: when retrieval finds the governing section the answer is usually right (0.812), and when it does not the answer scores 0.000, because the model cites only sections it was given and the API removes citations it cannot ground.

Tax rulings are the hard case. A Revenue Department ruling typically rests on several sections of the

Revenue Code, some of them definitions or cross-references that the question does not mention. Retrieval recall on Tax is 0.636, against 0.950 on CCL, and even with the gold sections mixed among close distractors the model selects the right subset with F1 0.685. On the newer held-out rulings retrieval finds every governing section for only 21.1% of questions, and the end-to-end score falls to 0.218. Broader retrieval for multi-section rulings is the clearest route to improvement.

The deployed model is the released model. On all seven release protocols the paired confidence interval of the difference between production and release includes zero (Table 4).

Ask for JSON when a program reads the citations. Sent as a plain chat message, the service answers briefly (median 270 characters on CCL), and 43 of 200 CCL answers state the rule without citing any section. Read from the prose, those answers score 0.493, against 0.772 when the training system prompt asks for JSON citations.

6 Production Performance

Over 288 end-to-end JSON requests to the service, an answer took 1.48 s at the median (90th percentile 3.28 s, 95th 4.12 s), with a median prompt of 3,241 tokens, most of it retrieved statute text, and 150 output tokens. Through the public API the 100 sampled questions took 1.43 s at the median (90th percentile 3.21 s), 0.09 s more than the service on the same questions, and were billed 6.35 IC in total at the published 0.01 IC per 1K input tokens and 0.02 IC per 1K output tokens (0.064 IC a question on average). On the sample the public API scored 0.729 against 0.723 for the service. Of 1,164 service requests across all arms, 0 failed, and the share answered by the backup engine was 0.0%.

7 Limitations

- **Citations, not legal correctness.** Citation F1 checks the sections named, not whether the reasoning or the conclusion is right. The essay and judge-graded axes of the model card were not repeated here.
- **Training distribution.** CCL derives from the dataset the model was fine-tuned on, decontaminated at the question level, and Tax overlaps late training material. The held-out set is small and its interval wide.
- **Single runs at temperature 0.** Repeat runs vary slightly because requests are batched on a shared engine; Table 4 shows the scale of that variation.
- **Index differs from the benchmark corpus.** The production index covers far more laws than NitiBench and may return a section of another law with the same text; such a citation counts as wrong against NitiBench's gold.
- **Prose parsing.** The chat-arm figures depend on a deterministic parser of Thai citations; a section cited without any law name is not scored in the text-only column.

8 Reproducibility

NitiBench is public (MIT), pinned here at Hugging Face revision 9f75697. The folder holds the test-set fetcher, the benchmark runner, every raw response (`results/*.jsonl`), the release validation rows with their SHA-256, the citation parser with NitiBench's scorer, and `gen_results.py`, which produces every figure in this paper. The held-out Tax rulings are available from iApp Technology on request: sale@iapp.co.th.

References

- [1] P. Akarajardwong, P. Pothavorn, C. Chaksangchaichot, P. Tasawong, T. Nopparatbundit and S. Nutanong. *NitiBench: A Comprehensive Studies of LLM Frameworks Capabilities for Thai Legal Question Answering*. arXiv:2502.10868, 2025. <https://huggingface.co/datasets/VISAI-AI/nitibench>.

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