



General Thai Document OCR: Accuracy Benchmark on Thai Government Documents

Service version v2.1.13-20260822 · Benchmark executed 22 August 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. We evaluated the iApp General Thai Document OCR API (v2.1.13-20260822) against a leading global cloud OCR service on 242 scanned pages from 20 Thai Administrative Court judgments drawn from a publicly available dataset. Both systems received byte-identical page images. The iApp service achieved a mean character error rate of 5.18% versus 6.09% for the cloud baseline — a 15% relative reduction — and produced the lower character error rate on 20 of 20 documents. In production, the service processes a page in 0.53 s (median) and sustains 4.28 pages per second.

1 Introduction

The General Thai Document OCR API converts printed Thai and English documents — images, PDF files, and Microsoft Office files — into plain text, layout-structured JSON, or generated Microsoft Word files. It is operated by iApp Technology Co., Ltd. and is available as a public cloud API at api.iapp.co.th and as a fully self-contained on-premise deployment.

Published OCR accuracy figures are frequently difficult to compare: vendors evaluate on private data, with unspecified normalization, against ground truth of unknown provenance. This whitepaper documents a benchmark designed to be verifiable by third parties. The dataset is public, the evaluation protocol is specified in full, including text normalization and metric definitions, and the baseline is a widely deployed commercial cloud OCR service, referred to here as *Cloud OCR A*.

2 Benchmark Dataset

The evaluation uses the *iApp Thai Administrative Court Judgment OCR* dataset, publicly available on Hugging Face ([kobkrit/iApp-Thai-Admin-Court-Judgment-OCR](https://huggingface.co/kobkrit/iApp-Thai-Admin-Court-Judgment-OCR)). It contains 1,079 scanned Thai Administrative Court judgments as PDF files, each paired with the word-processor source document from which the judgment was typeset. The scans are authentic government documents: multi-page, formally typeset Thai with official headers, seals, and signatures — representative of the documents Thai organizations process in practice.

From the 2560 B.E. (2017) partition, 20 documents (242 pages in total) were sampled with a fixed random seed (1337), so the sample is reproducible exactly.

3 Methodology

3.1 Protocol

Each PDF page is rendered once to a JPEG image at 200 dpi (quality 95). Both systems receive the byte-identical image bytes for every page; neither system receives any encoding, resolution, or preprocessing advantage. The cloud baseline is invoked through its standard document-text API with Thai and English language hints, matching its recommended configuration for this document class. Per-document output is the concatenation of its page outputs.

3.2 Ground truth

The ground truth for each document is the text extracted from its word-processor source file. Because the source document is the typeset original rather than a manual transcript of the scan, both systems carry the same fixed penalty from headers, page furniture, and typesetting differences between source and scan. Absolute error rates reported here are therefore *conservative* — the achievable error on clean transcripts is lower — and the meaningful figure is the gap between the two systems, which are scored against identical references.

3.3 Metrics

All text is Unicode-repaired and NFC-normalized, and whitespace is collapsed, before scoring.

- **CER** — character-level Levenshtein distance divided by reference length.
- **WER** — word-level Levenshtein distance over tokens produced by PyThaiNLP `newmm` word segmentation, applied identically to reference and hypothesis. Thai is written without spaces between words, so a specified segmenter is required for a well-defined WER.
- **ROUGE-L** — longest-common-subsequence F1 over the same `newmm` tokens.

4 Results

4.1 Aggregate

Metric	iApp Document OCR	Cloud OCR A
Character error rate (CER), mean	5.18%	6.09%
Character error rate (CER), median	4.32%	5.04%
Word error rate (WER), mean	7.03%	7.52%
ROUGE-L (F1), mean	0.9609	0.9576

Table 1: Aggregate results over 20 documents (242 pages). Lower is better for CER and WER; higher is better for ROUGE-L. Best result in bold.

The iApp service reduced the mean character error rate by 15% relative to Cloud OCR A and was more accurate on every document in the sample (20/20). Equivalently, measured character accuracy is 94.8% and word accuracy 93.0% under this conservative protocol.

4.2 Per-document results

One document (no. 12) shows a substantially higher error rate for both systems, consistent with a lower-quality scan or greater divergence between the source document and the scan; the ranking between the systems is unchanged on that document.

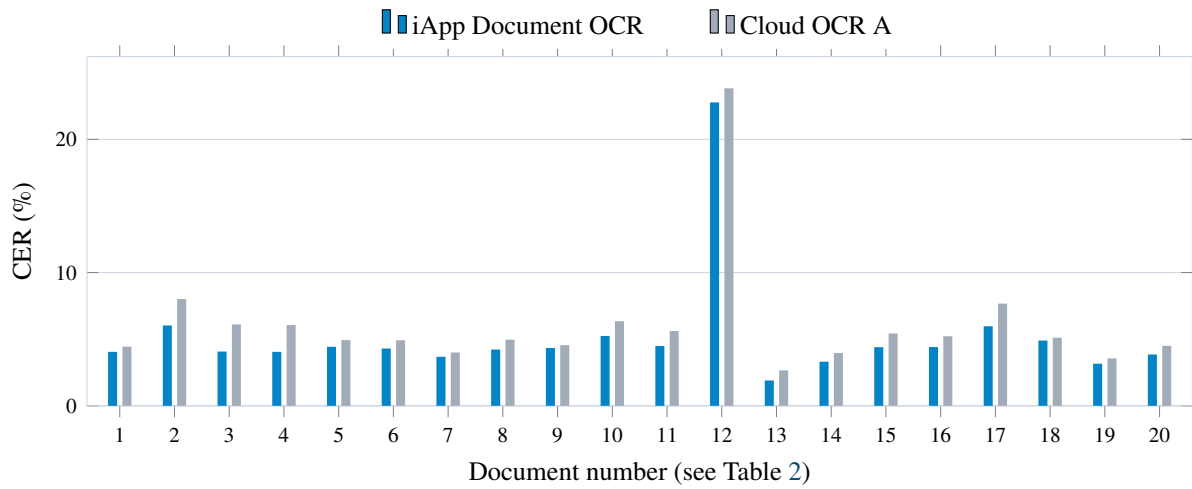


Figure 1: Character error rate per document. The iApp service (blue) achieved the lower CER on all 20 documents.

Doc	Pages	Ref. chars	iApp CER (%)	Cloud OCR A CER (%)	Δ (pp)
1	9	16,832	4.05	4.44	+0.40
2	4	5,054	6.03	8.01	+1.98
3	10	18,222	4.07	6.11	+2.05
4	7	11,790	4.05	6.06	+2.01
5	9	16,555	4.43	4.93	+0.50
6	14	26,350	4.30	4.92	+0.62
7	27	55,004	3.68	4.00	+0.32
8	8	15,059	4.22	4.96	+0.74
9	11	21,266	4.34	4.55	+0.22
10	5	8,488	5.24	6.35	+1.11
11	7	12,927	4.49	5.61	+1.11
12	5	9,977	22.77	23.83	+1.06
13	13	25,451	1.90	2.66	+0.77
14	21	45,080	3.31	3.96	+0.65
15	8	14,528	4.40	5.43	+1.03
16	5	9,028	4.41	5.22	+0.81
17	5	6,984	5.97	7.67	+1.70
18	9	16,170	4.90	5.11	+0.22
19	45	100,301	3.16	3.56	+0.39
20	20	42,833	3.85	4.50	+0.66

Table 2: Per-document character error rate. Δ is the absolute improvement of the iApp service over Cloud OCR A in percentage points; the better result per row is in bold. Document file names are listed in Appendix A.

5 Production Performance

Accuracy figures above are engine-level. The figures below were measured on the iApp production deployment of v2.1.13-20260822 in August 2026, through the same code path that serves customer traffic.

Metric	Value
Median end-to-end processing time (plain-text endpoint)	0.53 s per page
Sustained throughput	4.28 pages/s ($\approx$ 15,400 pages/hour)
Maximum input file size	30 MB
Supported inputs	PNG, JPEG, PDF, DOC(X), XLS(X), PPT(X)

Table 3: Production service performance, measured August 2026.

6 Deployment and Data Security

The public cloud API is GDPR- and PDPA-compliant: uploaded documents are used only to produce the requested output, generated download links expire after 10 minutes, and generated files are deleted within one hour. For organizations with stricter requirements, the service is available as a fully self-contained on-premise deployment in which no document data leaves the customer’s infrastructure.

7 Reproducibility

The dataset is public, the sample is fixed by seed, and Sections 3.1–3.3 specify the rendering, invocation, normalization, and metric conventions in full. Third parties can reproduce the evaluation of either system from this document alone. Enquiries regarding the benchmark, including the identity of the baseline service, may be addressed to sale@iapp.co.th.

About iApp Technology. iApp Technology Co., Ltd. is a Thai artificial-intelligence company providing OCR, document understanding, speech, and large-language-model services for the Thai language.

Web: iapp.co.th · E-mail: sale@iapp.co.th · Tel: +66 86 322 5858

Appendix A: Document File Names

Document numbers in Table 2 and Figure 1 correspond to the following file names in the public dataset (PDF/DOC pairs, 2560 B.E. partition).

Doc	Dataset file name
1	08013-600140-2F-621018-0000648823
2	01013-600315-2F-600921-0000606624
3	09013-600261-2F-620513-0000639495
4	03013-600204-2F-610213-0000615245
5	09013-600227-2F-620110-0000634060
6	09013-600096-2F-610427-0000620180
7	03012-600477-1F-620502-0000639287
8	01013-600427-2F-610201-0000614674
9	01013-600276-2F-601208-0000611364
10	09013-600008-2F-600627-0000606896
11	02013-600008-2F-600803-0000603597
12	01013-600216-2F-610131-0000614723
13	01013-600163-2F-600810-0000603887
14	01013-600125-2F-611122-0000631729
15	08013-600022-2F-600816-0000604217
16	01013-600011-2F-600823-0000604607
17	03013-600092-2F-600816-0000604139
18	01012-600387-2F-600824-0000604662
19	08012-600443-1F-620221-0000636050
20	01013-600245-2F-610809-0000626076