



Thai Receipt OCR: Field Accuracy on Synthetic Thai Receipts and Tax Invoices

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

Executive summary. The iApp Receipt OCR API reads a Thai receipt or tax invoice and returns the issuer, tax ID, document number, date, customer, line items and totals as JSON. We generated 60 receipts with fictitious values in three common layouts — full A4 tax invoices, point-of-sale slips with VAT included, and small-business cash receipts without VAT — and degraded half as office scans and half as phone photographs. Over 470 printed fields, the service returned 94.0% exactly right after a stated normalisation (95% CI [92.0, 95.9]); on the six key fields (issuer, issuer tax ID, document number, date, VAT, grand total) it returned 92.0% [89.8, 94.2]. Every field was right on 56.7% of receipts; most errors are one field on point-of-sale slips, where the service names the shop rather than the registered company (98.7% if the shop name is accepted). The set is synthetic: it measures reading through scan and photo degradation on clean layouts, and real receipts — faded thermal paper, handwriting, stamps, creases — will score lower.

1 Introduction

Receipts and tax invoices are the most common input to expense, accounting and tax-reclaim systems in Thailand. The Receipt OCR API returns, for each page, the document type, number and date, the issuer's name, address, tax ID and phone, the customer's name, address and tax ID, the line items, and the totals: amount before discount, discount, VAT, grand total and the amount in words. This paper measures how often those fields are returned exactly right.

A public benchmark needs documents whose answers are known and which may be published. Real receipts carry the names, tax IDs and purchases of real businesses and people, so we built a synthetic set in which every value is invented. Section 7 states what such a set does and does not show.

2 Test Data

A generator script, published with this paper, writes fictitious values into three layouts:

- **Full tax invoice (A4)**, 24 documents: issuer block with address and tax ID, customer block with tax ID, a numbered item table, total, optional discount, VAT at 7% added, grand total and the amount in Thai words.
- **Point-of-sale slip**, 24 documents: an abbreviated tax invoice from a till, 360 pixels wide, with store and company names, a branch number, tax, till and receipt numbers, a two-digit Buddhist-era date and time, VAT included in the total, cash and change.
- **Cash receipt without VAT (A5)**, 12 documents: a small shop's receipt book with book and receipt numbers, buyer, items and amount in words filled in a handwriting-style italic.

Company and shop names, people and addresses are invented; tax IDs carry a valid check digit but use a registration-office code that is not assigned, so none belongs to a real taxpayer. Totals, VAT and amounts in words are internally consistent. Every page carries a diagonal “specimen” watermark, and none is reproduced in this paper.

Receipts alternate between two capture profiles, 30 each. **Scan**: about 200 dpi, skew up to 1.5 degrees, half in greyscale, noise, slight blur, JPEG quality 60–85. **Photo**: the receipt on a coloured surface, each corner displaced by up to 2.5% of its width, rotation up to 2 degrees, uneven light, a colour cast, noise, blur of 0.6–1.5 pixels, JPEG quality 60–81.

3 Method

Each image was sent once with the documented default request (the file only) directly to the production service, bypassing the API gateway, at most two at a time; a seeded sample of 15 receipts also went through the public API (Section 6). Every printed field with a generated value is compared with the answer after this normalisation, fixed before the run:

- **Names**: Unicode NFKC, Thai digits as Arabic digits, case, spaces and the punctuation . , ; : ' " () ignored; a trailing head-office or branch designation on the issuer name ignored, since the response has no branch field.
- **Tax IDs and document numbers**: spaces and hyphens ignored; all other characters, including prefixes such as “R#”, must match.
- **Dates**: both sides read as a calendar date in either era.
- **Amounts**: equal within 0.005. **Amount in words**: as text.

A missing answer is wrong. Only printed fields are scored; separately, we count how often the service returned a VAT amount for a receipt that prints none. Line items are not scored. Confidence intervals are percentile bootstraps over receipts (2,000 resamples, seed 1337).

4 Results

Table 1: Field accuracy (%) per layout. Doc. all: receipts with every printed field right.

Layout	Receipts	Fields	Accuracy [95% CI]	Scan	Photo	Doc. all
Full tax invoice (A4)	24	242	97.9 [95.8, 99.6]	98.4	97.5	83.3
Point-of-sale slip	24	144	84.0 [81.9, 86.1]	84.7	83.3	8.3
Cash receipt, no VAT (A5)	12	84	100.0 [100.0, 100.0]	100.0	100.0	100.0
All receipts	60	470	94.0 [92.0, 95.9]	94.5	93.6	56.7

Table 2: Accuracy (%) per field and layout, with the number scored. ^kKey field.

Field	A4 tax invoice	Point-of-sale slip	Cash receipt	All
Issuer name ^k	100.0 (24)	8.3 (24)	100.0 (12)	63.3 (60)
Issuer tax ID ^k	100.0 (24)	100.0 (24)	–	100.0 (48)
Book number	–	–	100.0 (12)	100.0 (12)
Document number ^k	83.3 (24)	95.8 (24)	100.0 (12)	91.7 (60)
Date ^k	100.0 (24)	100.0 (24)	100.0 (12)	100.0 (60)
Customer name	95.8 (24)	–	100.0 (12)	97.2 (36)
Customer tax ID	100.0 (24)	–	–	100.0 (24)
Total before VAT	100.0 (24)	–	–	100.0 (24)
Discount	100.0 (2)	–	–	100.0 (2)
VAT amount ^k	100.0 (24)	100.0 (24)	–	100.0 (48)
Grand total ^k	100.0 (24)	100.0 (24)	100.0 (12)	100.0 (60)
Amount in words	100.0 (24)	–	100.0 (12)	100.0 (36)

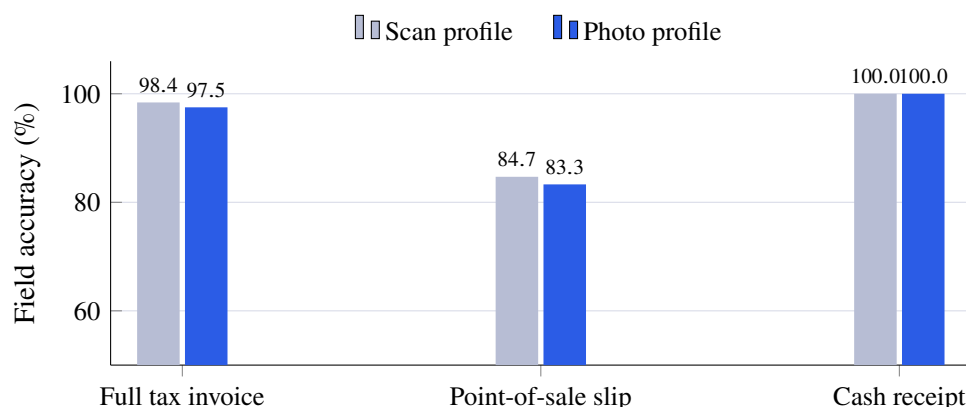


Figure 1: Field accuracy per layout and capture profile.

Scans scored 94.5% [91.5, 97.2] and photographs 93.6% [90.6, 96.2]. Every key field was right on 56.7% of receipts [45.0, 68.3].

Table 3: The most frequent errors (field, kind of error, count).

Field	Error	Count
Issuer name	wrong	22
Document number	wrong	5
Customer name	wrong	1

5 Discussion

Full tax invoices and cash receipts are read almost without error. The A4 tax invoices scored 97.9% and the cash receipts without VAT 100.0%. Most remaining errors on the A4 invoices are in document numbers: across all layouts 5 document numbers were wrong, 4 of them because a zero was returned as the letter O. On the 12 receipts that print no VAT, the service returned a VAT amount for 0: it did not compute a VAT figure the page does not show.

On point-of-sale slips the issuer field names the shop, not the company. The slips print the shop’s trading name above the registered company name and its tax ID. The response defines the issuer as “hospital, company or store”, and the service returned the trading name — alone on 8 slips, followed by the company name on the rest — on 22 of 24 slips; it returned the company name alone on 2. We score the trading name as wrong, because the registered taxpayer is what an accounting or tax-reclaim system needs; that one field accounts for the low score of the slips (84.0%). Counting the trading name as right, accuracy would be 99.3% on the slips and 98.7% [97.5, 99.8] over all receipts, with every field right on 91.7% of receipts. An integration that needs the registered taxpayer should key on the issuer tax ID, which the slips print and the service read correctly (Table 2).

Photographs cost little on this set: 94.5% for scans and 93.6% for photographs, with overlapping intervals.

6 Production Performance

15 receipts (5 per layout, seed 2026) were sent through <https://api.iapp.co.th/v3/store/ocr/receipt> with an ordinary API key from a client in Bangkok, at most two at a time. Status codes: 200: 15. On those receipts the public API scored 95.7% of 115 fields against 95.7% for the direct run, and the two

runs agreed on whether a field was right for 100.0% of fields. End to end, a receipt took 8.9 s at the median and 10.6 s at the 95th percentile. Each page costs 1 IC.

7 Limitations

- **Synthetic receipts.** Layouts are our own, rendered from clean fonts. Real receipts are printed by thousands of different systems, fade, crease and curl, carry stamps and handwritten corrections, and are often photographed in poor light. These results show how accurately the service reads printed values through scan and photo degradation; they do not show its accuracy on a customer's real receipts, which will be lower. iApp also measures this service on real receipts internally; those documents contain personal and commercial data, cannot be published, and are not reported here.
- **Line items are not scored**, and neither are fields absent from the page, apart from the VAT check above.
- **Sixty receipts.** Per-layout intervals are wide; one run of a service that answers each request afresh.

8 Reproducibility

The generator (`gen_receipts.py`, `synthdoc.py`), the scoring rules (`scoring.py`), the generated values and every answer are published with this paper; a fixed seed regenerates the same values, and the open-licence fonts are fetched from a public font repository by `get_data.sh`. Contact 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