iApp Technology · Technical Whitepaper WP-2026-12

Thai Business Document Extraction: Field Accuracy on a Synthetic Test Set of Eleven Document Types

Production service as of 29 October 2026 · Measured 28 October 2026; withholding-tax certificate re-measured 29 October 2026 after a fix · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. iApp offers eleven APIs that turn a photograph or scan of a Thai business document — bank and credit card statements, utility bills, a credit bureau report, company, civil registration and vehicle registration documents, a withholding-tax certificate, a resume and a job description — into named JSON fields. No public Thai test data with checked answers exists for these documents, so we generated 440 documents, 11 types of 40 each, filled with fictitious values, laid out like the real documents, and degraded as office scans or phone photographs. The answers are the generated values themselves. Over 5,092 fields, the service returned 97.9% exactly right after a stated normalisation (95% CI [97.5, 98.3]), and every field was right on 79.5% of documents. The weakest type was the vehicle registration book (93.4%), the strongest the bank statement (99.7%). The withholding-tax certificate was measured again on 29 October 2026, after we corrected how the service reads that form's tick boxes: its accuracy rose from 87.9% to 98.2%, and the pooled figure from 97.0% to 97.9% (Section 6). These figures measure reading under scan and photograph degradation on clean, known layouts; real documents with stamps, handwriting and unfamiliar layouts will score lower.

1 Introduction

Each of the eleven document APIs accepts an image or PDF and returns one JSON object whose fields are defined per document type: for a bank statement, the bank, account holder, account number, statement period, totals and every transaction row; for a vehicle registration book, the plate, province, make, model, chassis and engine numbers, owner and occupant. A customer uses these fields directly, so this paper measures the share of fields the service returns exactly right, per document type and per field.

Measuring this needs documents with known answers. Real examples of these documents carry personal data — national ID numbers, account numbers, credit histories — and cannot be collected and published as a test set. We therefore built a synthetic set in which every value is invented and known exactly. Section 8 states what a synthetic set does and does not show.

2 Test Data

Generation. A generator script, published with this paper, draws fictitious values for each document type and writes them into an HTML page laid out after the public blank forms and published samples of the real document: a bank statement with a header block and a transaction table, a utility bill with meter readings and a charges table, a company certificate in the registrar's letter format, a withholding-tax certificate in the two-party form with income-type rows, a vehicle registration page with pre-printed labels and dot-matrix values, and so on. Thai fonts released under the SIL Open Font License are used. The page is rendered at about 240 dots per inch.

Values. Names and addresses are random combinations of Thai given names, family names and place names; companies, banks, card issuers and lenders have invented names. Personal ID and company registration numbers carry a valid check digit but use a registration-office code that is not assigned to any province, so no generated number belongs to a real person or company. Amounts are internally consistent: running balances add up, bill totals equal the sum of the charges plus 7% VAT, and ages follow from dates of birth.

© 2026 iApp Technology Co., Ltd. · iapp.co.th · WP-2026-12

Dates use the formats of each document: Buddhist-era years with full or abbreviated Thai month names, day/month/year numerals with four- or two-digit years, and English dates on bilingual documents. Resumes are in Thai (one third) or English; job descriptions in Thai (two thirds) or English.

Safeguards. Every page carries a diagonal "specimen" watermark in Thai and English, and no official emblem or logo is drawn. No generated document is reproduced in this paper.

Degradation. Documents alternate between two capture profiles, 220 each:

- **Scan**: about 200 dpi, skew up to 1.5 degrees, half in greyscale, sensor noise, slight blur, JPEG quality 60–85.
- **Photo**: the page on a coloured desk surface, each corner displaced by up to 3.5% of the page length (perspective), rotation up to 2 degrees, uneven light with a soft shadow, a colour cast, noise, blur of 0.6–1.5 pixels, JPEG quality 60–81, 1,600–2,000 pixels on the long side.

3 Method

Each image was sent once, as a single-page JPEG with the documented default parameters (Thai output; for resumes and job descriptions, output in the document's own language), directly to the production service, bypassing the API gateway, with at most two requests in flight. A seeded sample of 44 documents was also sent through the public API (Section 7). The service was changed on 29 October 2026 for the withholding-tax certificate only (Section 6); its 40 documents, and the 4 of them in the public-API sample, were sent again the same way that day, and those answers replace the 28 October ones in every figure of this paper except where a figure is marked as before the fix.

Scoring. Every field that has a value on the page is compared with the generated value. A field is correct when the two are equal after this normalisation, fixed before the run:

- **Text**: Unicode NFKC, Thai digits as Arabic digits, case, spaces and the punctuation . , ; : ’ " ( ) ignored; the abbreviated titles for Miss, Master and Girl read as the full titles.
- **Addresses**: as text, with the administrative words (house number, village, lane, road, sub-district, district, province) ignored in full or abbreviated form.
- **Organisation names** of banks and lenders: the legal-form suffix ignored.
- **Identifiers**: spaces and hyphens ignored; masked digits must be masked.
- **Dates**: both sides read as a calendar date; any printed format and either era is accepted. The credit report's query time must also match.
- **Amounts**: equal within 0.005 (0.00005 for the per-unit fuel-adjustment rate); a discount printed with a minus sign may be returned with either sign.
- **Categories** such as account type, credit type, gender and fuel type: as text.

A missing or empty answer is wrong, and a failed request counts every field of its document as wrong (there were 0). Only fields printed on the page are scored, so a value the service invents for a field that is absent is not counted. The generated assessments of the resume API (scores, strengths, suggested skills) are not extraction and are not scored. Table rows (transactions, credit accounts, income rows, education and employment entries) are compared row by row in printed order and reported separately. For job descriptions, each printed responsibility and qualification counts as returned when one item of the answer equals it.

Statistics. Field accuracy is pooled over fields; its 95% confidence interval is a percentile bootstrap that resamples documents (2,000 resamples, seed 1337), because fields on one page are not independent.

4 Results

Table 1 gives the accuracy per document type and Figure 1 by capture profile; Tables 2 and 3 break the header fields down by kind and list the weakest. Every field is listed in the Appendix.

Capture profile. Scans scored 98.0% [97.4, 98.5] and photographs 97.8% [97.2, 98.4]; the difference is 0.1 points (95% CI [-0.7, 1.0]).

Tables and lists. Over 3,648 table-row fields the accuracy was 98.7%; the returned row count equalled the printed row count in 100.0% of 240 tables. Of 384 printed job-description list items, 98.4% were returned verbatim.

Errors. Of the wrong header fields, 8 were missing from the answer and 100 carried a wrong value.

Table 1: Field accuracy (%) per document type, 40 documents each. Fields: scored header fields. Doc. all: share of documents with every header field right. Rows: accuracy of table-row fields (count); [a] for job descriptions, printed list items returned verbatim.

Document type	Docs	Fields	Accuracy [95% CI]	Scan	Photo	Doc. all	Rows
Bank statement	40	360	99.7 [99.2, 100.0]	100.0	99.4	97.5	100.0 (1,530)
Company registration certificate	40	280	99.6 [98.9, 100.0]	100.0	99.3	97.5	–
Electricity bill	40	724	99.4 [98.9, 99.9]	99.2	99.7	90.0	–
Water bill	40	610	99.5 [98.9, 100.0]	99.0	100.0	92.5	–
Credit bureau report	40	320	95.6 [93.8, 97.5]	96.2	95.0	65.0	99.2 (860)
Civil registration certificate	40	680	98.7 [97.9, 99.4]	98.8	98.5	77.5	–
Credit card statement	40	720	97.1 [95.7, 98.3]	96.9	97.2	62.5	100.0 (568)
Withholding-tax certificate	40	455	98.2 [97.1, 99.3]	98.2	98.2	80.0	92.2 (141)
Vehicle registration book	40	623	93.4 [91.9, 95.1]	93.5	93.4	27.5	–
Resume (CV)	40	160	96.9 [94.4, 99.4]	98.8	95.0	87.5	94.9 (549)
Job description	40	160	99.4 [98.1, 100.0]	100.0	98.8	97.5	98.4 (384)[a]
All types	440	5,092	97.9 [97.5, 98.3]	98.0	97.8	79.5	98.7 (3,648)

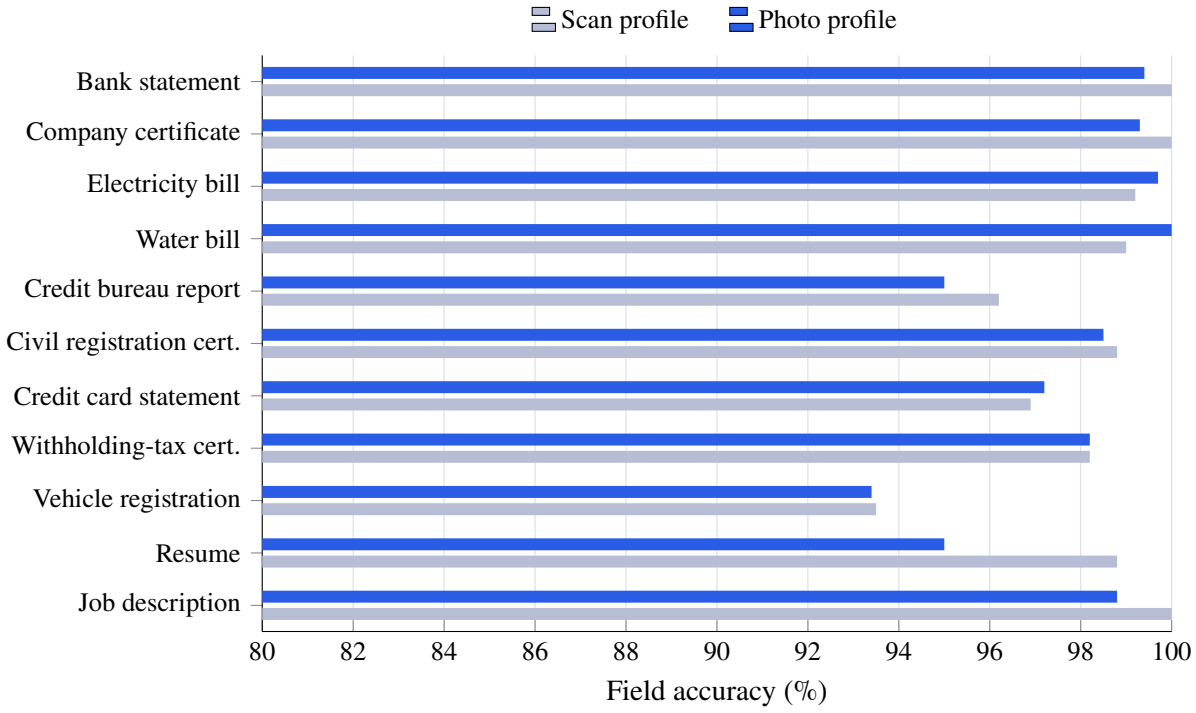


Figure 1: Field accuracy per document type under the two capture profiles (20 documents each).

Table 2: Header-field accuracy by kind of field, all types.

Kind of field	Fields	Accuracy (%)
Amounts and numbers	1,549	99.3
Identifiers	1,048	94.7
Names	775	98.7
Dates	720	99.9
Categories and other text	680	98.1
Addresses	240	94.2
Contact details	80	96.2

Table 3: The eight weakest header fields (at least 20 scored).

Document type	Field	n	Accuracy (%)
Vehicle registration book	`body_number`	40	32.5
Credit bureau report	`reference_number`	40	67.5
Credit card statement	`cardNumber`	40	80.0
Credit card statement	`reward.pointsRedeemed`	40	82.5
Civil registration certificate	`registrationOffice`	40	85.0
Electricity bill	`customer_address`	40	90.0
Vehicle registration book	`engine_number`	33	90.9
Civil registration certificate	`address`	40	92.5

5 Discussion

Printed amounts, dates and tables are read reliably. Amounts and other numbers were right on 99.3% of 1,549 fields and dates on 99.9% of 720; bank statements, company certificates and both utility bills had the highest accuracy (Table 1). The returned row count matched the printed one for 100.0% of the 240 tables, and 98.7% of table cells were right. Photographs cost little: the scan–photo difference of 0.1 points has a confidence interval that includes zero.

Long character strings in a dot-matrix typeface are the main reading weakness. The chassis number on the vehicle registration page was right on 13 of 40 documents. The errors are confusions of 0 with O and 8 with B: 13 answers contained the letter I, O or Q, which never occur in a chassis number. The typeface we used draws these characters alike, so the real page may fare better, but an integration should check chassis numbers against their alphabet and length. Masked card numbers were wrong on 8 of 40 statements, 6 of them by an extra O or zero inside the masked group, and an underscore in a credit-report reference number was returned as a space on 13 of 14 reports.

Tick boxes are still read less reliably than text. Even after the fix (Section 6), the ticked form type on the withholding-tax certificate was wrong on 3 of 40 documents, and when the ticked payer method was not the first option the answer named it in 3 of 5 cases. An integration that depends on these fields should cross-check them, for example the form type against whether the payee is a company or a person.

A row label is sometimes returned with the value. On the civil registration certificate, the registration office was returned with the printed row label in front of it on 6 documents; that is the office's full name, but it is not the value printed in the field, and we score it as wrong.

Uncommon spellings are sometimes regularised. Several name and address errors are not misreadings of a blurred character: the service returned the common spelling of a similar word where the page prints an unusual one, or exchanged two similar Thai consonants. Thai personal names are often spelled unusually, so names should be checked against another source where it matters.

The per-field confidence does not separate right from wrong. The service returns a confidence for each field. It averaged 0.956 on the 108 wrong header fields and 0.966 on the right ones, and 96.3% of wrong fields carried a confidence of 0.90 or more. On this evidence, confidence alone is not a reliable trigger for human review; arithmetic checks (totals, VAT, balances) and check digits on identifiers are.

Document-level accuracy is lower than field accuracy because one wrong field fails a document: 79.5% of documents had every header field right. For the vehicle registration book the chassis number dominates it. Before the fix, the payer-method wording alone set it to zero for the withholding-tax certificate; after it, 32 of 40 certificates had every header field right.

6 Fix to the Withholding-Tax Certificate

The 28 October run found the withholding-tax certificate to be the weakest type, at 87.9% of fields and no document fully right, for three reasons in the service itself rather than in reading the page:

- **The tick boxes were not looked for.** The form has a row of seven boxes for the form type and a row of four for the payer method. The service's instructions for this type named the two fields but neither their options nor that the answer is the ticked box. The machine-read text the service also uses does not show which box is ticked (box symbols come out as stray letters), so the answer for the payer method was the first printed option on every document, ticked or not, and the form type was wrong on 12 of 40.
- **The documented values were not returned.** The payer method is documented as one of four fixed values; the service returned the form's shorter printed wording of the first one on 39 of 40 documents.
- **A hidden character.** The service's own definition of the first documented value contained an invisible zero-width space, so even an answer copied from it would not equal the documented value.

The corrected service names each field's printed options and the values to return, and says that the answer is the box that carries a mark in the image. A final step maps the printed wordings of the payer options and spacing variants of the form names onto the documented values, and the hidden character is removed. Nothing else changed: the other ten document types have separate definitions that were not touched. Table 4 compares the two runs.

Table 4: Withholding-tax certificate before (28 October 2026) and after (29 October 2026) the fix, 40 documents each; the pooled rows cover all 11 types.

Measure	Before	After
Field accuracy, withholding-tax certificate (%)	87.9	98.2
Documents with every field right	0	32
Form type right (%)	70.0	92.5
Payer method right (%)	0.0	95.0
Payer method in the form's shorter wording	39	0
Ticked payer method other than the first, named	0 of 5	3 of 5
Field accuracy, all 11 types (%)	97.0	97.9
Documents with every field right, all types (%)	72.3	79.5

7 Production Performance

44 documents (4 per type, seed 2026) were sent through the public endpoints at `api.iapp.co.th` with an ordinary API key, from a client in Bangkok, at most two at a time; the 4 withholding-tax certificates among

them were sent again on 29 October, after the fix. Status codes: 200: 44. On those documents the public API scored 97.6% of 509 fields against 97.6% for the direct run, and the two runs agreed on whether a field was right for 100.0% of fields and returned an identical value for 97.6%. End to end, a page took 12.4 s at the median and 19.9 s at the 95th percentile. In the direct run the service's own processing time was 11.1 s per page at the median; documents with long tables, such as bank statements, take longest. Each page costs 1 IC.

8 Limitations

- **Synthetic documents.** The layouts are our approximations of the real documents, rendered from clean fonts. Real documents vary between issuers and over time, carry stamps, signatures, folds, handwriting and low-contrast printing, and arrive as multi-page PDFs. These results show how accurately the service reads printed values through scan and photo degradation on known layouts; they do not show its accuracy on a customer's real documents, which will be lower.
- **Only printed fields are scored.** The test does not measure how often the service returns a value for a field that is absent from the page.
- **Exact match is strict** for long free text: an address returned with one abbreviation expanded differently is wrong, although usable.
- **One run per type.** The service answers each request afresh; repeated runs can differ slightly (Section 7). The withholding-tax certificate was measured on 29 October, after a change to the service; the other ten types on 28 October.
- **The fix was checked on this test set.** Before the 29 October run, the corrected service was tried on these same 40 certificates. Its instructions name only the options printed on the official form, nothing specific to our generated documents, but the after-fix figures for the tick-box fields may still be somewhat optimistic.
- **Forty documents per type.** Confidence intervals per type are wide; the pooled figure is the more reliable one.

9 Reproducibility

The generator (`gen_docs.py`, `synthdoc.py`), the scoring rules (`scoring.py`), the per-document answers and the generated values are published with this paper; a fixed seed regenerates the same values. The fonts are open-licence fonts fetched from a public font repository by `get_data.sh`. Rendering depends on the browser build, so regenerated images are equivalent but not byte-identical. 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

A Accuracy per Field

Accuracy (%) of every scored field, with the number scored in parentheses where it is not 40. Fields marked `[]` are table-row cells.

Bank statement. Below 100%: `account_number` 97.5 (40). At 100.0%: `bank_name`, `account_holder_name`, `account_type`, `branch`, `period_from`, `period_to`, `total_withdrawal_amount`, `total_deposit_amount`, `transactions[].date` (510), `transactions[].withdrawal` (241), `transactions[].balance` (510), `transactions[].deposit` (269).

Company registration certificate. Below 100%: `registration_office` 97.5 (40). At 100.0%: `form_code`, `company_name`, `registration_number`, `registration_date`, `issue_date`, `registrar_name`.

Electricity bill. Below 100%: `customer_address` 90.0 (40). At 100.0%: `provider`, `customer_name`, `account_number`, `invoice_number`, `tariff_type`, `reading_date`, `last_reading`, `previous_reading`, `units_consumed`, `energy_charge`, `service_charge`, `ft_rate_per_unit`, `ft_charge`, `subtotal_before_vat`, `vat_amount`, `total_current_bill`, `due_date`, `discount` (4).

Water bill. Below 100%: `customer_address` 92.5 (40). At 100.0%: `provider`, `customer_name`, `account_number`, `invoice_number`, `invoice_date`, `current_reading`, `previous_reading`, `consumption_cubic_meter`, `water_usage_charge`, `monthly_service_fee`, `discount` (10), `subtotal_before_vat`, `vat_amount`, `total_amount`, `due_date`.

Credit bureau report. Below 100%: `reference_number` 67.5 (40); `full_name` 97.5 (40); `accounts[].lender_name` 95.9 (172). At 100.0%: `report_number`, `query_date`, `id_card_number`, `date_of_birth`, `total_accounts`, `accounts_under_dispute`, `accounts[].credit_type` (172), `accounts[].account_status` (172), `accounts[].credit_limit` (172), `accounts[].outstanding_balance` (172).

Civil registration certificate. Below 100%: `address` 92.5 (40); `registrationOffice` 85.0 (40). At 100.0%: `nationalIdNumber`, `houseRegistrationNumber`, `name`, `gender`, `nationality`, `dateOfBirth`, `age`, `residentialStatus`, `mothersName`, `mothersNationality`, `fathersName`, `fathersNationality`, `moveInDate`, `footer.date`, `footer.registrar`.

Credit card statement. Below 100%: `cardNumber` 80.0 (40); `cardHolderName` 95.0 (40); `interestRate` 95.0 (40); `reward.pointsRedeemed` 82.5 (40); `reward.pointsOutstanding` 95.0 (40). At 100.0%: `statementDate`, `dueDate`, `bankName`, `bankTaxID`, `bankPhone`, `creditLimit`, `closingDate`, `newBalancePayment`, `minimumPaymentDue`, `previousBalance`, `totalPayments`, `totalInterestCharges`, `reward.rewardPointsEarned`, `transactions[].transactionDate` (284), `transactions[].amountTHB` (284).

Withholding-tax certificate. Below 100%: `documentForm` 92.5 (40); `taxpayer.name` 95.0 (40); `taxpayer.address` 97.5 (40); `taxPayerMethod` 95.0 (40); `paymentTypes[].paymentDate` 81.1 (37); `paymentTypes[].amount` 96.2 (52); `paymentTypes[].withholdingTax` 96.2 (52). At 100.0%: `bookNumber`, `documentNumber`, `taxDeductor.name`, `taxDeductor.taxId`, `taxpayer.taxId`, `totalAmount`, `totalTax`, `socialSecurity.amount` (15).

Vehicle registration book. Below 100%: `license_plate` 95.0 (40); `province` 97.5 (40); `model` 97.5 (40); `body_number` 32.5 (40); `engine_number` 90.9 (33); `owner_name` 92.5 (40); `owner_id` 97.5 (40); `owner_address` 95.0 (40); `occupant_name` 93.3 (15). At 100.0%: `registration_date`, `vehicle_type`, `brand`, `year`, `color`, `fuel_type`, `seating_capacity`, `occupant_id` (15).

Resume (CV). Below 100%: `personalInformation.name` 97.5 (40); `personalInformation.birthDate` 97.5 (40); `personalInformation.email` 92.5 (40); `workExperience[].company` 98.0 (99); `workExperience[].startDate` 87.9 (99); `workExperience[].endDate` 85.9 (99). At 100.0%: `personalInformation.phoneNumbers`, `education[].university` (51), `education[].year` (51), `education[].gradePointAverage` (51), `workExperience[].title` (99).

Job description. Below 100%: `location` 97.5 (40). At 100.0%: `jobTitle`, `companyName`, `salaryRange`.