iApp Technology · Technical Whitepaper WP-2026-29

Thai Business Document Extraction, Part 2: Field Accuracy on Ten More Document Types

Production service as of 18 November 2026 · Measured 18 November 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. On 18 November 2026 iApp published ten more document APIs that turn a photograph or scan of a Thai document into named JSON fields: a company certificate read in full, the partnership and company registration certificate, death, marriage and divorce certificates, the house registration, employment and salary certificates, pay slips, loan agreements, loan statements and MRTA payment vouchers. We measured them exactly as the first eleven types were measured in WP-2026-12: 400 synthetic documents, 40 per type, filled with fictitious values, laid out like the real documents and degraded as office scans or phone photographs, with the generated values as the answers. Over 5,712 fields the service returned 96.9% exactly right after a stated normalisation (95% CI [96.4, 97.4]), and every field was right on 68.8% of documents. The weakest type was the death certificate (94.6%), the strongest the MRTA payment voucher (99.8%). 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 API accepts an image or PDF and returns one JSON object whose fields are defined per document type. The ten types had been served for some time without public pages; before publishing them we measured them with the method, toolkit and scoring rules of WP-2026-12, so that the accuracy on each page means the same as on the eleven earlier ones. This paper states only what differs and the results; the method is described in full in WP-2026-12.

2 Test Data

Generation. A generator script published with this paper (`gen_docs_more.py`) uses the same toolkit as WP-2026-12: fictitious Thai names, addresses and companies; personal and juristic ID numbers with a valid check digit but an unassigned registration-office code; invented banks, lenders and insurers; a diagonal "specimen" watermark on every page and no official emblem. Each type has two layouts, drawn at random:

- **Company certificate (full)**: the registrar's letter in numbered paragraphs, or a table of the same items; one to six directors, signing authority, capital in figures and words.
- **Registration certificate**: the certificate for a limited company, limited partnership or registered ordinary partnership, plain or in a double border.
- **Death certificate**: a two-column table of the certificate's items, or a boxed form; hospital or home as the place of death.
- **Employment certificate**: a letter on company letterhead, in prose or with a table; with a salary (salary certificate, 60%) or without, current or former employee.
- **House registration**: the house page and one person's page, stacked or side by side.
- **Loan agreement**: a contract in numbered clauses or in four parts, with interest rates by period and a repayment schedule.

© 2026 iApp Technology Co., Ltd. · iapp.co.th · WP-2026-29

- **Loan statement**: account summary, payment records and arrears.
- **Marriage or divorce certificate**: 35% divorce; plain or bordered.
- **MRTA payment voucher**: a premium remittance to an insurer, or a fee with VAT and withholding tax; an approval block whose first row carries the preparer's code.
- **Pay slip**: earnings and deductions side by side, year-to-date totals, net pay; only the lines that apply are printed.

Amounts are internally consistent (totals add up; VAT is 7%; net pay is earnings less deductions). Dates use Buddhist-era years with full or abbreviated Thai month names, or day/month/year numerals.

Degradation. As in WP-2026-12, documents alternate between the scan and the photo profile, 20 of each per type.

3 Method

Each image was sent once, as a single-page JPEG with the documented default parameters (Thai output), directly to the production service, bypassing the API gateway, with at most two requests in flight; a seeded sample of 4 documents per type was also sent through the public API (Section 5).

Scoring. The normalisation of WP-2026-12 applies unchanged. Five kinds were added before the run, each for a field whose schema names a value differently from how the page prints it; each is a normalisation of both sides, never a looser match:

- **Given names**: a leading Thai title (Mr, Mrs, Miss, Master, Girl, and their abbreviations) is ignored, because the schemas split given and family names while the page prints the title with the given name.
- **Registration office**: a leading "registration office" label is ignored.
- **Position and department**: a leading "position", "division" or "department" label is ignored.
- **Province**: a leading "province" label is ignored.
- **Time**: hours and minutes must match. For loan-agreement table rows, the interest-rate period and rate must contain the same numbers.

Addresses also ignore a leading "house number" label, an address word like the "number" label that WP-2026-12 already ignores; it was added after a first look at the loan-agreement answers showed that an address printed after "resides at house number" was read correctly with the label kept. Full names, such as a registrar's, keep the title: a name returned without its printed title is wrong. Fields printed with the same value on every document of a type (the document title, the issuing department, the house type) are not scored. A failed request counts every field of its document as wrong (there were 0). Statistics are as in WP-2026-12.

4 Results

Table 1: Field accuracy (%) per document type, 40 documents each. Fields: scored top-level fields. All right: share of documents with every scored field right.

Document type	Docs	Fields	Accuracy [95% CI]	Scan	Photo	All right
Company certification (long form)	40	400	95.0 [93.0, 96.8]	94.5	95.5	57.5
Partnership and company registration certificate	40	280	99.3 [98.2, 100.0]	100.0	98.6	95.0
Death certificate	40	1,080	94.6 [93.6, 95.6]	94.4	94.8	12.5
Employment and salary certificate	40	472	95.1 [93.6, 96.6]	94.5	95.7	47.5
House registration	40	895	95.6 [94.1, 97.0]	98.2	93.1	35.0
Loan agreement	40	440	98.9 [98.0, 99.8]	97.7	100.0	87.5
Loan statement	40	640	98.6 [96.2, 100.0]	97.8	99.4	92.5
Marriage and divorce certificate	40	320	98.8 [97.5, 99.7]	99.4	98.1	90.0
MRTA payment voucher	40	412	99.8 [99.2, 100.0]	99.5	100.0	97.5
Salary statement (pay slip)	40	773	98.2 [97.2, 99.1]	97.7	98.7	72.5
All ten types	400	5,712	96.9 [96.4, 97.4]	97.0	96.9	68.8

Tables and lists. Over 610 table-row values (loan-agreement interest rates and schedule, loan-statement payment records) the accuracy was 96.9%. Of 144 printed directors' names on the company certificates, 96.5% were returned.

Errors. Of the wrong fields, 16 were missing from the answer and 159 carried a wrong value.

What went wrong. The errors fall into five kinds.

- **Titles dropped from full names.** On the death certificate the registrar's printed name, such as "(Mrs Arunee Niyomtham)" in Thai, came back without its title on 23 of 40 documents. Under the rules of WP-2026-12 a full name keeps its title, so these count as wrong; a reader who strips titles before matching loses nothing. This is most of the gap between the death certificate and the other types.
- **Short reference numbers misread.** The certificate number at the top left of the full company certificate (wrong on 37.5% of documents, set in a small font) and the letter number of the employment certificate (wrong on 32.5%) were the weakest fields: a two-letter Thai prefix was misread, or returned with a dot inserted between its two letters.
- **Text wrapped in a table cell.** On the death certificate, a hospital name or address that wrapped onto a second line in its cell was often cut at the line end.
- **A value placed under another key.** On 11 of 40 house registrations the house code was left empty under the house's part of the answer; on 8 of them it was returned under the person's part instead; on pay slips, less common deductions were sometimes filed under a neighbouring key (other deductions as absence, a work deposit as other deductions); on 2 Bangkok marriage certificates the province was returned as the registration office.
- **Single characters.** The rest are one-character misreads in names and addresses, a zero read as the letter O in account numbers, and names or places returned with the spelling of a more common word instead of the printed one. One loan statement came back with a different district and province in its address.

Amounts, totals and dates were the strongest fields: every pay-slip total and net pay, and every voucher amount, was right.

Confidence. The per-field confidence averaged 0.897 on wrong fields and 0.953 on right ones, and 94.1% of wrong fields scored 0.90 or more: as in WP-2026-12, it does not mark errors.

5 Production Performance

40 documents (4 per type, seed 2026) were sent through the public endpoints at `api.iapp.co.th` with an ordinary API key, at most two at a time. Status codes: 40×200. On those documents the public API scored 97.3% of 566 fields against 97.2% for the direct run; the two runs agreed on whether a field was right for 99.5% of fields and returned an identical value for 99.1%. End to end, a page took 8.0 s at the median and 11.4 s at the 95th percentile; in the direct run the service's own processing time was 7.7 s per page at the median. Each page costs 1 IC.

6 Limitations

- **Synthetic documents.** The layouts are our approximations of the real documents. Real documents vary between issuers, carry stamps, signatures and handwriting, and arrive as multi-page PDFs; loan agreements in particular run to many pages, of which only the first is modelled here. Accuracy on a customer's documents will be lower.
- **Only printed fields are scored**, so the test does not measure values invented for absent fields.
- **Exact match is strict.** The place names come from an open word list in which a few are misspelt; where the service returned the correct spelling of a real district instead of the printed misspelling, the field counts as wrong.
- **One run, forty documents per type.** Per-type intervals are wide; the pooled figure is the more reliable one.

7 Reproducibility

The generator (`gen_docs_more.py` with `synthdoc.py`), the fields scored (`spec_more.py`), the scoring (`scoring.py`, `score_more.py`), the runner (`bench_more.py`), the per-document answers and the generated values are published with this paper; a fixed seed regenerates the same values. 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.

Company certification (long form): `document_number` 62.5; `certification_detail.registration_date` 97.5; `certification_detail.business_registration_number` 100.0; `company_info.company_name` 100.0; `company_info.board_members_count` 100.0; `company_info.registered_capital_number` 100.0; `company_info.registered_capital_text` 100.0; `company_info.headquarters_address` 90.0; `issued_date` 100.0; `registrar_signature` 100.0.

Partnership and company registration certificate: `registration_number` 100.0; `document_type` 100.0; `company_name` 100.0; `registration_office` 100.0; `registration_date` 100.0; `issue_date` 100.0; `registrar_name` 95.0.

Death certificate: `certificateNumber` 100.0; `registrarOffice` 100.0; `deceased.firstName` 92.5; `deceased.lastName` 97.5; `deceased.nationalId` 100.0; `deceased.sex` 100.0; `deceased.age` 100.0; `deceased.nationality` 100.0; `deceased.address` 97.5; `deceased.dateOfDeath` 100.0; `deceased.timeOfDeath` 100.0; `deceased.causeOfDeath`

100.0; deceased.placeOfDeath.name 65.0; deceased.placeOfDeath.address 82.5; deceased.mother.firstName 97.5; deceased.mother.lastName 97.5; deceased.father.firstName 90.0; deceased.father.lastName 100.0; informant.firstName 100.0; informant.lastName 97.5; informant.nationalId 100.0; informant.relationship 100.0; informant.address 100.0; burial.type 97.5; burial.place 97.5; dateOfRegistration 100.0; registrarName 42.5.

Employment and salary certificate: certificateType 100.0; companyName 100.0; companyAddress 92.5; phoneNumber 97.5; documentNumber 67.5; employeeName 92.5; employeePosition 100.0; department 100.0; startDate 100.0; endDate 100.0 (9); salary 100.0 (23); managerName 92.5; issueDate 100.0.

House registration: house_detail.house_code 72.5; house_detail.registration_office 100.0; house_detail.address 90.0; house_detail.village_name 100.0 (15); house_detail.house_structure 95.0; house_detail.house_number_issue_date 95.0; house_detail.registrar_name 87.5; house_detail.printed_date 97.5; person_detail.person_order 100.0; person_detail.name 97.5; person_detail.nationality 100.0; person_detail.gender 100.0; person_detail.citizen_id 100.0; person_detail.status 100.0; person_detail.birth_date 100.0; person_detail.mother_name 95.0; person_detail.mother_citizen_id 100.0; person_detail.mother_nationality 100.0; person_detail.father_name 90.0; person_detail.father_citizen_id 100.0; person_detail.father_nationality 100.0; person_detail.moved_from 90.0; person_detail.registrar_name 92.5.

Loan agreement: contractType 100.0; bank 100.0; borrowerInformation.fullName 100.0; borrowerInformation.age 100.0; borrowerInformation.ethnicity 100.0; borrowerInformation.nationality 100.0; borrowerInformation.fullAddress 87.5; loans.0.loanType 100.0; loans.0.loanDetails.amount 100.0; loans.0.loanDetails.period 100.0; contractDate 100.0; []period 100.0 (96); []rate 80.2 (96); []period 100.0 (73); []monthlyAmount 100.0 (73).

Loan statement: loanIssuer 97.5; loanAccountNumber 100.0; customerName 97.5; documentDeliveryAddress 95.0; creditLimit 100.0; remainingPrincipalBalance 100.0; monthlyInstallmentAmount 100.0; loanTerm 100.0; paymentDueDate 100.0; contractEndDate 100.0; deductionAccountNumber 97.5; latePaymentAmounts.principalAmount 97.5; latePaymentAmounts.interestAmount 97.5; latePaymentAmounts.otherCharges 97.5; latePaymentAmounts.total 97.5; notYetDueAmountForLoanTerm 100.0; []transactionDate 100.0 (68); []principalAmount 100.0 (68); []interestAmount 100.0 (68); []totalAmount 100.0 (68).

Marriage and divorce certificate: certificateType 100.0; certificateNumber 100.0; issueDate 100.0; placeOfIssue 95.0; provinceOfIssue 100.0; spouse1Name 100.0; spouse2Name 97.5; registrarName 97.5.

MRTA payment voucher: paymentVoucherNumber 97.5; preparedByName 100.0; preparedByDate 100.0; pycNumber 100.0; documentDate 100.0; payTo 100.0; amount 100.0; amountBeforeVat 100.0 (13); vat 100.0 (13); amountWithVat 100.0 (13); withholdingTax 100.0 (13); netPayment 100.0; netPaymentInWords 100.0.

Salary statement (pay slip): employerName 100.0; employeeName 92.5; employeeCode 97.5; position 100.0; payrollPeriod.startDate 100.0; payrollPeriod.endDate 100.0; paymentDate 100.0; bankAccount 92.5; earnings.salaryWage 100.0; earnings.overtime 100.0 (21); earnings.commission 100.0 (10); earnings.allowancesCostOfLiving 100.0 (17); earnings.bonus 100.0 (7); earnings.others 100.0 (12); earnings.totalEarnings 100.0; deductions.socialSecurityFund 100.0; deductions.withholdingTax 100.0 (37); deductions.studentLoanFund 100.0 (8); deductions.deposit 75.0 (8); deductions.absentLeaveLate 100.0 (9); deductions.others 57.1 (7); deductions.totalDeductions 100.0; ytd.earnings 100.0; ytd.withholdingTax 97.3 (37); ytd.accumulatedSsf 97.5; netPay 100.0.