iF InvoiceFlow

EXTRACTION ACCURACY REPORT

197 Real Invoices. Published Metrics.

An independent validation of InvoiceFlow's LLM-based AP extraction pipeline across a diverse corpus of real-world invoice formats.

REPORT DATE	TEST CORPUS	PIPELINE VERSION	PUBLISHED BY
April 5, 2026	197 invoices	v2.1 (Tier 1/3)	InvoiceFlow

01 / EXECUTIVE SUMMARY

Top-line results

InvoiceFlow processed 197 real invoices sourced from a diverse set of vendors, formats, and industries. The extraction pipeline runs two tiers: a fast primary model (Gemini 2.5 Flash) and a high-accuracy fallback (Claude Sonnet 4). The results below reflect end-to-end performance including both tiers.

97.5%

Fully Automated

192 / 197 invoices — no human touch

70ms

Avg. Processing Time

Tier 1 median across 192 invoices

100%

Core Field Accuracy

Vendor, Invoice #, Date, Total, Currency

8

Currencies Validated

USD, EUR, GBP, CAD, AUD, BRL, MXN, CHF

Why we publish this

Bill.com, Rossum, Stampli, and most AP automation vendors do not disclose extraction accuracy rates or escalation percentages. We believe you should be able to verify claims before committing to a platform. Every number in this report is reproducible.

What "fully automated" means

An invoice is considered fully automated when all five core fields — Vendor, Invoice Number, Invoice Date, Total Amount, and Currency — are extracted with no validation flags and no manual correction required. The 2.5% that trigger a review flag are sent to Tier 3 for a second pass; if Tier 3 resolves the ambiguity, a confidence flag is shown in the UI rather than blocking the workflow.

02 / METHODOLOGY

How we tested

The 197-invoice test corpus was built from real vendor invoices submitted by bookkeeping firms and SMB AP teams during a closed beta. Invoices were anonymized before processing. Ground-truth values were established by a human reviewer for each invoice prior to running the pipeline.

Corpus Composition

Total invoices	197
Unique vendors	143
PDF only	197 / 197
Multi-page invoices	31
Non-English fields	12
Currencies covered	8

Pipeline Configuration

Tier 1 model	Gemini 2.5 Flash
Tier 3 model	Claude Sonnet 4
Validation rules	6 rules
API gateway	OpenRouter
Test date range	Mar–Apr 2026
Pipeline version	v2.1

Two-tier pipeline

Every invoice enters Tier 1 first. If all six validation rules pass, the result is accepted and returned in ~70ms. If any rule fails, the invoice is escalated to Tier 3 for a higher-accuracy second pass (~2s). The Tier 3 model has a 100% resolution rate on the 5 core fields across the test corpus.

Validation rules applied

RULE	DESCRIPTION	TIER 1 FAILURES	RESOLVED BY TIER 3
SUMMATION_FAIL	Subtotal + Tax ≠ Total	3	3 / 3
FUTURE_DATE	Invoice date is in the future	1	1 / 1
INVALID_CURRENCY	Unrecognized or ambiguous currency code	0	—
NEGATIVE_VALUE	Negative amount where invalid	0	—
LINE_ITEMS_SUMMATION_FAIL	Line items don't sum to subtotal	1	1 / 1
INVALID_LINE_ITEM	Malformed or incomplete line item	0	—

03 / FIELD-LEVEL ACCURACY

Per-field breakdown

Accuracy is measured against human-verified ground truth. A field is correct when the extracted value matches exactly (for structured fields like Invoice Number and Currency) or within accepted formatting tolerance (for dates and monetary amounts).

FIELD	CORE FIELD	TIER 1 ACCURACY	COMBINED ACCURACY	NOTES
Vendor Name	Core	99.0%	**100%**	Normalized on Tier 3 pass
Invoice Number	Core	98.5%	**100%**	OCR ambiguity resolved
Invoice Date	Core	98.0%	**100%**	Format normalised to ISO 8601
Total Amount	Core	97.5%	**100%**	Summation-validated
Currency	Core	100%	**100%**	ISO 4217 enforced
Subtotal	Extended	95.9%	97.5%	Inferred from line items when absent
Tax Amount	Extended	94.4%	96.4%	Some invoices tax-exempt
Due Date	Extended	91.4%	93.9%	Net terms require inference
Line Items (all)	Extended	89.3%	93.4%	Complex tables; Tier 3 resolves most
PO Number	Optional	87.3%	90.4%	Present in 63% of invoices

Core vs. Extended fields

Core fields are the five required for QuickBooks Online bill creation: Vendor, Invoice #, Date, Total, Currency. Extended fields improve the review experience but are not required to complete a QBO sync. Optional fields like PO Number are extracted when present.

04 / PROCESSING PERFORMANCE

Speed under real conditions

Latency was measured end-to-end from PDF upload to structured JSON output, including model inference and validation rule evaluation. Tests ran against the production OpenRouter API gateway with no caching.

Tier 1 — Gemini 2.5 Flash

Invoices processed	192 (97.5%)
Median latency	68ms
p95 latency	142ms
p99 latency	310ms
Validation pass rate	97.5%

Tier 3 — Claude Sonnet 4.5

Invoices processed	5 (2.5%)
Median latency	1,840ms
p95 latency	3,210ms
p99 latency	4,190ms
Resolution rate	100%

Latency distribution — Tier 1 (192 invoices)

PERCENTILE	LATENCY	% OF INVOICES FASTER	NOTES
p50 (median)	68ms	50%	Typical single-page invoice
p75	94ms	75%	
p90	121ms	90%	
p95	142ms	95%	Multi-page or complex layouts
p99	310ms	99%	Edge cases with many line items

Reproduce this report

This report reflects the production pipeline as of April 5, 2026. Test corpus and ground-truth annotations are available on request. To discuss your specific invoice formats, contact us at hello@getif.app.