

"95% accurate" still leaves more than half your invoices needing a person.

Capture tools quote accuracy per field. AP teams get whole invoices, and an invoice is only usable when every field the ERP needs is right. At 95% per field, 56 of 100 invoices still need a person.

01 The numbers behind the problem

38 in 100

invoices go through without a person touching them.

Ardent Partners research, The State of AP 2026 (n=194).

1 in 5

invoices gets stuck as an exception.

Ardent Partners research, The State of AP 2026: 19.9%.

92.7%

of fields right on scanned invoices, for the best of eight general-purpose AI models.

Berghaus et al., Fraunhofer IAIS and Lamarr Institute, 2025 preprint (arXiv 2509.04469).

4 days

to resolve a single invoice error.

APQC Open Standards Benchmarking, median of 2,861 organizations.

56 of 100 is a Nuntiq calculation and an illustration: 26 header fields and 5 lines of 6 fields make 5,600 fields per 100 invoices, so 280 errors at 95%; errors cluster at about 5 per bad invoice. Line count and clustering are assumptions.

02 Flagged, not guessed: how AP Receiving checks a capture

1 Add it up

Lines, charges and discounts are added up and compared with the totals printed on the invoice. Deductions such as retainage or withholding tax are checked against the amount due.

2 Read it again

When the numbers do not match, a second AI model reads the invoice again, before anyone on the team has to look at it.

3 Flag it for a person

If it still does not add up, the invoice goes to the team flagged, with the reason and the original beside the captured data. Validation rules then check every invoice before release.

03 14 AI models, the same 210 invoices

Model	Complete	Header	Lines	Tax
DeepSeek V4.1 Flash	100.0%	97.6%	99.8%	100.0%
Claude Haiku 5.5	100.0%	98.0%	99.8%	100.0%
GLM 5.3 Flash	99.4%	95.1%	99.8%	99.4%
GPT-5.6 Luna	99.8%	98.2%	99.8%	99.6%
GPT-6 Luna	99.5%	98.3%	99.7%	99.6%
Gemini 3.5 Flash-Lite	99.7%	97.7%	99.1%	99.8%
Gemini 3.1 Flash-Lite	98.4%	87.8%	94.2%	99.1%
Qwen3 Next 80B	97.4%	91.3%	96.0%	98.7%
GLM 4.7 Flash	89.0%	74.9%	87.5%	83.0%
Ministral 3 14B	92.4%	87.2%	66.1%	90.4%
gpt-oss-120b	85.7%	85.6%	61.7%	94.2%
gpt-oss-20b	79.0%	88.7%	55.0%	79.8%
Nemotron 3 Super	81.2%	71.7%	57.7%	38.5%
Gemini 2.5 Flash-Lite	49.7%	44.5%	76.1%	44.2%

Complete: every invoice line captured. Each model at its best setting. Rounded to one decimal, so 100.0% can include a single miss.

04 The Nuntiq process vs. a model alone

Setup	Needs a person
Nuntiq process	10 of 420 (2.4%)
Claude Haiku 5.5 alone	14 of 420 (3.3%)
DeepSeek V4.1 Flash alone	15 of 420 (3.6%)
GPT-6 Luna alone	29 of 420 (6.9%)

The Nuntiq process captured 100.0% of invoice lines (rounded), and no setup left missed lines unnoticed (0 of 210). The figures above use two reads per invoice (420 reads). A model on its own has nothing to fall back on: every capture that falls short needs a person.

How we tested. 210 test invoices generated by Nuntiq in nine languages: digital PDFs, scans, faxes and phone photos, 1 to 15 pages, including credit notes and multi-invoice PDFs. Every model got the same instructions and read each invoice two or three times. This is Nuntiq's own benchmark on its own test set.

SEE EVERY ANSWER

Eight invoices built to break capture, at nomorebaddata.com

Scans, faxes, phone photos and a 424-line time sheet, with every captured value next to the page.

MEDIA CONTACT

ewout.vocking@nuntiq.com