Reading Brazil: A Local Cost-versus-Accuracy Benchmark of Open OCR Engines on Brazilian Documents

Cristhian Kapelinski[1], **Aline Lunkes**[1], **Daniel Welfer**[2], **Dionatan Schmidt**[1], **Beatriz Machado**[1], **Diego Kreutz**[1]

[1]AI Horizon Labs, Federal University of Pampa (UNIPAMPA)

[2] Federal University of Santa Maria (UFSM)

{cristhianavila,beatrizmachado}.aluno@unipampa.edu.br

{dionatanschmidt,alinelunkes,diegokreutz}@unipampa.edu.br

welfer@gmail.com

Abstract. *Privacy rules push Brazilian institutions to read their forms and identity documents on their own hardware. No published benchmark tells them which open OCR engine does that well at the lowest cost: existing ones are English, Chinese, or non-Brazilian Portuguese, and none measures the cost of running locally. We pair a Brazilian-Portuguese accuracy measurement with a cost measurement on one consumer machine, scoring 14 open engines on forms, identity documents, clean prose, and degraded scans. No class of engine wins outright. Only two hold up on all four document types: the OCR engine Surya and the vision-language model Qwen2.5-VL. On identity photos the divide is old engines against recent ones, not OCR engines against vision-language models, and Surya recovers 0.921 of the fields, more than any vision-language model. Engine choice should follow the document type. We also rebuild the forms ground truth from the images and validate it by hand, finding no error in 250 sampled fields.*

1. Introduction

Brazilian public services run on paper, at scale. The federal driving-license application alone has over 54 million registered users.[1] Before any of these documents can be filed, queried, or audited, a machine has to read it.

The documents differ widely in layout. The Brazilian forms we score cover surveillance, social-registry and bank, academic, health, and veterinary templates [Xu et al. 2022], and citizens carry identity documents of their own, the driving license (CNH) and the identity card (RG). The software that reads them has changed kind in the last two years. Leading document readers are now end-to-end vision-language models (VLMs): one network takes a page image and returns its transcription in a single step, replacing the older detect-then-recognize pipelines [Zhong et al. 2026].

A Brazilian institution often cannot send such documents to a cloud service, because the General Data Protection Law restricts how personal data may be processed.[2] It has to read them on hardware it already owns. The question is therefore not which engine tops an English leaderboard, but which open engine reads its documents well at the lowest local cost. A recent Brazilian study shows that this concern is already live [Ribeiro et al. 2025].

1. Senatran, driver-licensing statistics, 2026. 2. Brazil, General Data Protection Law (LGPD), Lei 13.709/2018.

That question has no published answer for transcription on Brazilian documents. Optical character recognition (OCR) turns a document image into text. The open OCR models of 2025 and 2026 are ranked mainly on two benchmarks. OmniDocBench has English and Chinese text pages [Ouyang et al. 2025]; olmOCR-Bench is English by construction and drops non-English output [Poznanski et al. 2025]. Neither contains Portuguese, so neither answers the Brazilian question. To answer it we measure two quantities. Field-value recall (FVR) is the fraction of a form's known field values that the engine's transcription contains; higher is better, and 1.0 means every value was recovered. Normalized edit distance (NED) says how few character edits separate the transcription from the reference text, reported here as a percentage where 100 means the two are identical. We lead with FVR because a field value missing from a record costs the institution more than an extra token that should not be there, which a later key-value check can filter out.

The benchmarks that do include Portuguese widen the language coverage, but none of them answers this question. MDPBench and CC-OCR cover Portuguese, but they do not separate Brazil from Portugal and report no cost [Li et al. 2026, Yang et al. 2025]. A recent benchmark targets European Portuguese [Cardeira et al. 2026], a different variety and a different document mix. The one genuinely Brazilian full-page resource is a set of handwritten student essays [Neto et al. 2024], not printed administrative documents. The closest cost-versus-accuracy work is English or French, and runs on datacenter hardware [Poznanski et al. 2025, Rigal et al. 2026]. No public benchmark pairs a Brazilian accuracy axis with a local cost axis across these recent open engines, and none puts the new vision-language engines beside the OCR engines a Brazilian institution would normally reach for.

We close that gap with a benchmark and measurement study. We evaluate 14 open OCR engines, all locally and with no external service. 8 are end-to-end vision-language models, from 0.9B to 9B parameters. The other 6 are purpose-built OCR engines: 5 classical detect-then-recognize pipelines, plus Surya, which replaces that pipeline with a single transformer. For every engine we record both how well it reads and what it costs to run, the median time per page. We cover four document types: Brazilian forms and identity documents, scored by FVR, and clean Portuguese prose and degraded scans, scored by NED. English forms serve as a control. Only two engines hold up on all four, the OCR engine Surya and the strongest vision-language model Qwen2.5-VL, so engine choice should follow the document type rather than a single leaderboard rank.

We make four contributions, two measurements and two resources. The first is the measurement itself: Brazilian documents read by both classes of engine, side by side. The second is the cost measurement, processing time per document on one consumer machine rather than a datacenter figure. The third is an error analysis that says why engines fail and not only which ones do: the older classical pipelines collapse on identity-document photos and on impulse noise and rotation, the recent engines of both classes lose accuracy gently, and two engines stay accurate everywhere. The fourth is the one that took the most effort, and the one other work can reuse: a corrected Brazilian forms gold. We rebuilt it from the document images, because the source benchmark's key-value gold turned out to be contaminated. Two annotators validated it independently. It is released with the scoring harness that recomputes the paper's measurements.

2. Related Work

Table 1 positions the directly comparable benchmarks. None of these benchmarks measures cost on local consumer hardware (the Size column counts evaluated documents or pages, n/r where no fixed count is reported). The Gold src column names what produced each gold. Ours describes the Brazilian forms gold, the one we build from the images instead of inheriting: a large multimodal model reads them and two annotators validate a sample by hand (Section 3). Our other sets reuse their sources' gold (Table 2).

Table 1. Positioning against comparable OCR benchmarks.

Benchmark	Scope (doc type)	Langs	PT-BR gold	Gold src	Size (docs)	Cost axis
OmniDocBench [Ouyang et al. 2025]	Document parsing	EN/ZH	No	Model, human	981	none
olmOCR-Bench [Poznanski et al. 2025]	End-to-end OCR	EN	No	Mixed	1,403	dollars
MDPBench [Li et al. 2026]	Multilingual	Multi.	Generic PT	Model, human	3,400	none
CC-OCR [Yang et al. 2025]	VLM literacy	Multi.	Generic PT	Model, human	7,058	none
GlotOCR-Bench [Kargaran et al. 2026]	Script coverage	Multi.	Latin tier	Synthetic	16,375	none
French PDF2MD [Rigal et al. 2026]	PDF-to-Markdown	FR	No	Model	n/r	latency
[Ribeiro et al. 2025]	Feature extraction	PT-BR	No (labels)	Human	50	latency
This work	**Forms, IDs, prose, degraded**	**PT-BR/EN**	**Yes**	**Model, human**	**50/50/200/224**	**latency**

Benchmarks the models headline on. OmniDocBench is a widely used document-parsing benchmark that recent OCR models report on. Its annotations are model pre-labels corrected by annotators and checked by experts [Ouyang et al. 2025]. In the release we cite, its text pages are English and Chinese, so the engines that top it are untested on Brazilian documents. olmOCR-Bench pairs an open 7B vision-language model with thousands of binary unit tests and plots cost against accuracy [Poznanski et al. 2025]. We follow its way of reporting bootstrap confidence intervals. It is English by design, and its cost is a datacenter dollar figure rather than a measurement on a consumer GPU.

Multilingual benchmarks with Portuguese. MDPBench is the first multilingual document-parsing benchmark spanning digital and photographed pages across seventeen languages including Portuguese [Li et al. 2026]; its ground truth is produced by an ensemble of recognition models with manual correction, it does not distinguish Portugal from Brazil, and it reports no cost. CC-OCR is a peer-reviewed OCR benchmark covering Portuguese among ten languages, annotated by a model and corrected by hand [Yang et al. 2025]. GlotOCR-Bench measures recognition across 158 writing systems on the same recent engines, and finds Latin scripts strong while many low-resource scripts sit near zero [Kargaran et al. 2026]. It uses rendered synthetic text lines, treats Portuguese as part of the already-solved Latin group, and reports accuracy only. These works establish multilingual coverage but leave a Brazilian slice with a cost trade-off and classical baselines outside their scope.

Cost-versus-accuracy and Brazilian prior art. The French PDF-to-Markdown benchmark is the closest work in method [Rigal et al. 2026]. It evaluates the same generation of open engines, fifteen models, several of them in 8-bit floating point (fp8), on a single non-English language. It plots accuracy against throughput on an enterprise 80 GB GPU

rather than a consumer card. Its pages are picked where models disagree rather than sampled at random, and the language is French, not Portuguese. It flags that fp8 and hardware compatibility affect its throughput numbers, a caveat our fp8 deployment of two engines shares. On the Brazilian side, BRIDP is a peer-reviewed benchmark of synthetic identity documents, covering CNH, the taxpayer registry number (CPF), and RG. It validates a detector plus classical OCR, scoring each field by edit distance [Fernandes et al. 2024]; it tackles the harder problem of building the pipeline and reports no cost, so it sits beside an end-to-end OCR measurement rather than competing with it. A Brazilian SBC study evaluates classical OCR on Portuguese text in a misinformation setting [Santos et al. 2023b], and the closest work at this venue benchmarks vision-language models on a Brazilian tax-assessment task, weighing accuracy against wall-clock time [Ribeiro et al. 2025]; the latter is property-feature extraction rather than transcription, and its cost axis is time only.

We use established resources as test material: a Portuguese text-recognition suite [Santos et al. 2023a], the multilingual form benchmark whose Portuguese split we confirm to be Brazilian [Xu et al. 2022], an English forms benchmark as the language control [Jaume et al. 2019], and a Brazilian identity-document dataset, which is why our synthetic identity gold holds no real personal data [Fernandes et al. 2024].

One criterion selects the 8 vision-language engines we measure: an open-weight end-to-end OCR model released in 2025 or 2026 that runs on our reference runtime (vLLM) and fits a single consumer GPU. What we test is therefore the set of engines a Brazilian institution could actually deploy on its own hardware, not a sample from a leaderboard. They span 0.9B to 9B parameters and are themselves recent vendor technical reports: PaddleOCR-VL, a 0.9B model;[3] HunyuanOCR, a 1B OCR specialist;[4] MinerU2.5, a 1.2B page-parsing model;[5] DeepSeek-OCR, a 3B mixture-of-experts OCR model;[6] dots.ocr, a 3B model (a 1.2B vision encoder with a 1.7B language model);[7] Qwen3-VL, a 4B general VLM;[8] Qwen2.5-VL, a 7B general VLM;[9] and GLM-OCR, a 9B model.[10]

Against these we score six purpose-built OCR engines: the classical detect-then-recognize pipelines Tesseract [Smith 2007], EasyOCR, docTR, RapidOCR, and PaddleOCR (PP-OCR), and the modern transformer-based Surya, so the comparison spans both engine classes rather than the vision-language models alone. A live trend toward end-to-end OCR vision-language models is surveyed by a method paper [Zhong et al. 2026]. One further open engine did not enter the roster: a contemporaneous 1B specialist [Taghadouini et al. 2026] that failed to serve on the reference runtime.

3. Documents and Gold

Table 2 lists every scored set with its role, size, language scope, gold source, and license; this section describes each before any number is reported.

Brazilian forms. The Portuguese split of a multilingual form-understanding benchmark [Xu et al. 2022] supplies the forms. Its images are Brazilian administrative form templates. The benchmark's annotators filled them with made-up data for privacy and then scanned them, so they carry no real personal data. We score the 50 forms that have

3. PaddleOCR-VL technical report, arxiv.org/abs/2510.14528. 4. HunyuanOCR technical report, arxiv.org/abs/2511.19575. 5. MinerU2.5 technical report, arxiv.org/abs/2509.22186. 6. DeepSeek-OCR technical report, arxiv.org/abs/2510.18234. 7. dots.ocr, rednote-hilab, 2025, github.com/rednote-hilab/dots.ocr. 8. Qwen3-VL technical report, arxiv.org/abs/2511.21631. 9. Qwen2.5-VL technical report, arxiv.org/abs/2502.13923. 10. GLM-OCR technical report, arxiv.org/abs/2603.10910.

Table 2. Scored documents and their gold.

Set	Role	n	Lang	Gold source	License
BR forms [Xu et al. 2022]	Forms FVR axis	50	PT-BR	Model+human	CC BY-NC-SA 4.0
BR IDs [Fernandes et al. 2024]	IDs FVR axis	50	PT-BR	Synthetic	Unstated
EN forms [Jaume et al. 2019]	Language control	30	EN	Human-annotated	Research-only
RIB (clean) [Santos et al. 2023a]	Clean-text NED	200	PT	Reference transcription	CC BY 4.0
HYB (degraded) [Santos et al. 2023a]	Degradation stress	224	PT	Reference transcription	CC BY 4.0

non-empty field gold, that is, a known correct value for at least one field: surveillance, social-registry and bank, academic, health, and veterinary. The original task is entity tagging rather than full-page transcription, so we score field-value recall instead, matching up to punctuation and spacing.

Reading the images, we found the benchmark's key-value gold contaminated. About a fifth of its field values are printed checkbox instructions rather than filled-in data, often cut off mid-word. On table-heavy forms, column headers and labels are linked as if they were values. At least one entry is simply wrong: a CPF the gold records as `674...` reads `574...` in the document. We therefore rebuilt the field gold from the images, using a large multimodal model (Claude Opus) that is not among the evaluated engines. A value counts as gold only when it is legibly filled into a form field; printed instructions and mislinked labels are discarded. Dates and numbers vary in punctuation, so we match tolerantly, and a value read correctly is not penalized for a formatting difference alone.

Two annotators then checked the result independently. Each compared a random sample of 250 of the 1428 scored fields against the images, and neither found an error, which puts the gold's field-level accuracy at 1.00 (Wilson 95% interval [0.985, 1.000]) with full agreement between them. What their discussion surfaced was a choice of representation, not a mistake. A marked option, shown on the page as "A (X)", can be written down two equally valid ways: the full option text as the answer, or the option label as the field and the "X" as the answer. One consent form takes the second form, recording seven checklist items each as a "V". Both annotators agreed the value itself is right; the rebuilt gold simply does not use one encoding everywhere.

Brazilian identity documents. The identity set is 50 synthetic CNH and RG images [Fernandes et al. 2024]. Their values are made up, so no real personal data is involved, and the gold comes from the dataset's own structured fields rather than from reading the image. That gold is clean and well formed, unlike the noisy key-value linking of the forms. A low score on identity documents is therefore a genuine reading failure, not an artifact of how the answers were recorded.

English control. An English benchmark of noisy scanned forms [Jaume et al. 2019] serves as the control. If an engine scores low on the Brazilian forms, this tells us whether the engine is simply weak or whether Portuguese is the problem. We score all 14 engines on 30 English forms with non-empty field gold, using the same recall measure and the same matching as the Brazilian forms, so both engine classes appear in the control.

Clean Portuguese text (RIB). The clean reference partition (RIB) of a Portuguese text-recognition suite [Santos et al. 2023a] provides 200 clean literary pages with trusted human transcription as a clean-text NED reference. These pages are nineteenth-century literature with pre-1911 orthography, so we frame them as a Portuguese clean-text reference rather than as Brazilian administrative documents.

Degradation stress (HYB). The same suite supplies its pages under synthetic degradations as the hybrid partition (HYB) [Santos et al. 2023a], 224 pages built as eight degradation types over 28 base pages each. The degradations ship with the dataset and were not applied by us; six of the eight are DocCreator models [Journet et al. 2017]: bleed-through (Bleed), optical blur (Blur), character or ink degradation (CharDeg), holes and occlusion (Hole), Gaussian noise (Gauss), salt-and-pepper impulse noise (S&P), phantom-character ghosting (Phantom), and page rotation (Rot). We score all 14 engines on this set, spanning vision-language models and OCR engines, so the degradation result is a cross-engine comparison. We release the gold subsets whose licenses permit redistribution; the rest is license-restricted and fetched from its original source.

4. Measurement Protocol

No single number works across all four document types, so each task gets the metric that fits it. Forms, identity documents, and the English control are scored by FVR. Clean Portuguese pages and degraded pages are scored by NED, in percent, and the degraded set is reported both overall and per degradation type.

Every rate comes with its denominator and a 95% confidence interval, following olmOCR-Bench [Poznanski et al. 2025]. For NED we bootstrap over per-document means (10,000 resamples, seed 20260609). For FVR we use a Wilson interval over the field-level trials, micro-averaged across fields, so the 50 identity images give 354 scored fields. Where we compare two engines head to head we add a paired bootstrap over the documents they share, at the same resamples and seed. These pairwise tests carry no multiple-comparison correction, so an interval that excludes zero supports that one comparison and not a claim about the whole table.

All inference runs on one reference machine: an AMD Ryzen 7 9700X (8 cores, 16 threads), 64 GB of RAM, and a single 16 GB consumer GPU (NVIDIA GeForce RTX 5080), under the vLLM runtime with greedy decoding (temperature 0) at each engine's vendor defaults. The setup reflects what a deployer gets out of the box rather than a per-engine tuning an institution would be unlikely to replicate; the same machine times the classical pipelines on its CPU. Six vision-language engines run at their native precision; two, GLM-OCR and Qwen2.5-VL, run in fp8 to fit the card, which we treat as part of the measured deployment picture rather than a separate condition. Each vision-language engine receives the page image and the same transcription instruction in a single call, so the comparison is on identical inputs.

The cost we measure is the median per-page processing time per document type, each engine loading once and timing only the per-page OCR call. Model size and device appear in Table 5 as descriptors of each engine, not as part of the cost axis. We do not report peak memory, which is utilization-bound here because the runtime pre-allocates its key-value pool to a configured fraction of GPU memory; every vision-language engine was served within the card.

5. Results

Table 3 reports accuracy per task with confidence intervals, grouped by engine type, and Table 5 reports the cost. Figure 1 plots accuracy against latency, forms on the left and identity documents on the right. Only two engines hold up on all four document types, the

OCR engine Surya and the vision-language model Qwen2.5-VL. The rest change places depending on what they are asked to read.

Table 3. Per-task accuracy by engine, grouped by type, with 95% confidence intervals; bold is the best in each column.

Engine	BR forms FVR (n=50)	BR IDs FVR (n=50)	RIB NED% (n=200)	HYB NED% (n=224)
Vision-language models				
PaddleOCR-VL (0.9B)	0.778 [0.756, 0.799]	0.636 [0.584, 0.684]	95.9 [93.81, 97.65]	91.14 [87.9, 94.02]
HunyuanOCR (1B)	0.832 [0.812, 0.85]	0.78 [0.734, 0.82]	97.2 [96.57, 97.66]	88.57 [85.01, 91.89]
MinerU (1.2B)	0.905 [0.888, 0.919]	0.463 [0.412, 0.515]	92.59 [91.90, 93.16]	89.67 [88.09, 91.08]
DeepSeek-OCR (3B)	0.855 [0.836, 0.872]	0.864 [0.825, 0.896]	97.11 [96.47, 97.58]	90.59 [87.18, 93.64]
dots.ocr (3B)	0.931 [0.916, 0.943]	0.771 [0.725, 0.812]	95.3 [93.84, 96.45]	96.01 [94.3, 97.33]
Qwen3-VL (4B)	0.888 [0.871, 0.903]	0.703 [0.654, 0.749]	91.71 [90.8, 92.57]	87.59 [85.0, 89.87]
Qwen2.5-VL (7B)	**0.97** [0.96, 0.978]	0.842 [0.8, 0.876]	**97.46** [96.84, 97.91]	**97.53** [96.49, 98.16]
GLM-OCR (9B)	0.965 [0.954, 0.973]	0.828 [0.785, 0.864]	93.87 [91.98, 95.5]	95.99 [94.86, 96.77]
OCR engines				
Tesseract	0.918 [0.903, 0.931]	0.24 [0.199, 0.287]	93.68 [92.58, 94.66]	77.25 [72.36, 82.03]
EasyOCR	0.919 [0.904, 0.932]	0.486 [0.434, 0.538]	80.81 [79.81, 81.76]	73.74 [71.13, 76.22]
docTR	0.898 [0.881, 0.912]	0.822 [0.779, 0.858]	90.43 [89.46, 91.33]	79.96 [77.48, 82.27]
RapidOCR	0.863 [0.844, 0.88]	0.87 [0.831, 0.901]	92.03 [91.37, 92.57]	83.34 [81.32, 85.22]
PaddleOCR	0.9 [0.883, 0.914]	0.822 [0.779, 0.858]	94.98 [94.31, 95.51]	77.43 [74.77, 79.95]
Surya	0.965 [0.954, 0.973]	**0.921** [0.888, 0.945]	96.67 [95.34, 97.62]	93.89 [91.99, 95.44]

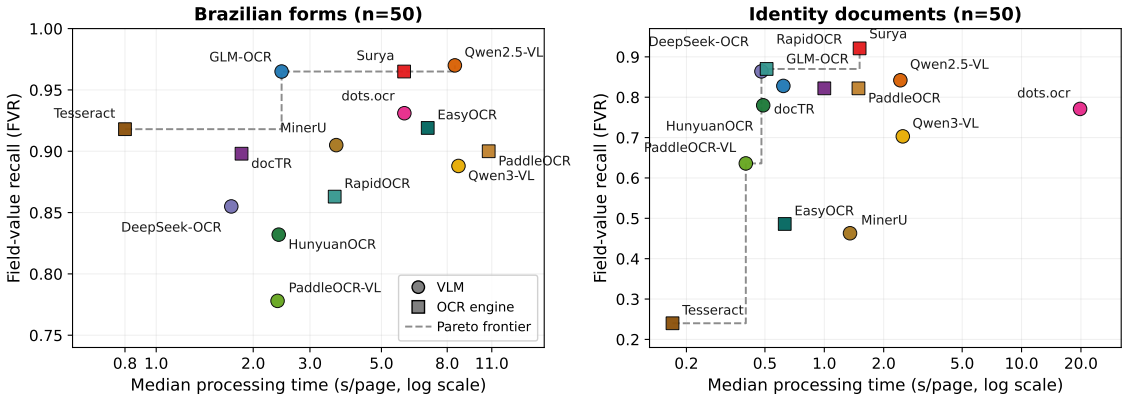


Figure 1. Field-value recall against latency, Brazilian forms (left) and identity documents (right); time on a logarithmic scale. Marker shape is engine type; the dashed line is the cost-accuracy frontier, the engines no other beats on both axes.

Brazilian forms. Both classes of engine read Brazilian form fields well (Figure 1, left). The best is Qwen2.5-VL at 0.97 [0.96, 0.978], with GLM-OCR and Surya next at 0.965 and 0.965, and their confidence intervals overlap (Table 3). A paired bootstrap over the 50 forms does not separate Qwen2.5-VL from the next engine (0.005 FVR, 95% interval [-0.010, 0.019], which includes zero), so the leaders are tied here.

This input is typed and easy, and the classical pipelines keep up: EasyOCR 0.919,

Tesseract 0.918, and PaddleOCR 0.9 all land among the mid-pack vision-language models. Only the strongest vision-language model and Surya pull clearly ahead, and the ordering does not follow parameter count. Two of the smallest vision-language models report lower raw scores, PaddleOCR-VL at 0.778 and HunyuanOCR at 0.832, because they collapse on the densest table-heavy forms. The collapse hits 1 of the 50 forms for PaddleOCR-VL and 3 for HunyuanOCR. Dropping those, their recall rises to 0.878 and 0.884, in line with the others. We report both numbers: the raw recall is what a deployment would see, and the other says how well the engine reads once the collapse is set aside.

Identity documents and the English control. Identity-document photos are harder than forms and reorder the roster (Figure 1, right), and the split is legacy versus modern rather than vision-language versus OCR engine. The best engine is the OCR engine Surya at 0.921 [0.888, 0.945], the highest on this axis and ahead of every vision-language model; its Wilson interval overlaps that of the best vision-language model on this axis, DeepSeek-OCR, but a paired bootstrap over the 50 documents confirms the lead (Surya minus DeepSeek-OCR is 0.057 FVR, 95% interval [0.023, 0.089], excluding zero), and the same test puts it 0.051 [0.023, 0.079] above the best classical pipeline, RapidOCR. That pipeline, RapidOCR at 0.87, in turn edges the best vision-language model on the point estimate though their confidence intervals overlap (Table 3). The best vision-language model on this axis is DeepSeek-OCR at 0.864, ahead of Qwen2.5-VL at 0.842 and GLM-OCR at 0.828, while the general vision-language models Qwen3-VL and dots.ocr fall to 0.703 and 0.771. The collapse is concentrated in the legacy classical engines: EasyOCR drops to 0.486 and Tesseract to 0.24, far below the modern transcription engines, which hold on these low-resolution photos. One modern engine falls in with the legacy floor: MinerU drops to 0.463, below even EasyOCR, because its page-parsing layout pass returns no text on 19 of the 50 cards, mostly the sparse driving licenses, which then score zero, the document parser meeting a single-field photo rather than a reading failure. The deployer lesson is to pick a modern engine matched to the document type and not to assume that classical means weak. Even the best engines leave roughly one identity field in twelve unread, so this axis still needs human verification before any automated filing.

All 14 engines run on the English control. On the 30 English forms they range from 0.388 (EasyOCR) to 0.906 (dots.ocr), close to the range they reach on the Brazilian forms. Reading forms is therefore what limits recall, not the language. The old-versus-recent divide survives the change of language: Surya (0.850) and the strong classical pipelines docTR (0.839) and PaddleOCR (0.830) sit mid-pack and beat several vision-language models, while Tesseract (0.475) and EasyOCR (0.388) trail, falling furthest from their Brazilian scores. MinerU is a useful case here: it reads these dense English forms at 0.839, level with docTR, yet falls to 0.463 on the sparse identity cards. That supports reading its identity result as a layout parser meeting the wrong kind of page, not an inability to read. Among the recent engines the largest drops are GLM-OCR (0.965 versus 0.607) and PaddleOCR-VL (0.778 versus 0.522). One caveat bounds all of this: the two sets are different corpora (FUNSD and XFUND), so the control limits how large a Portuguese penalty could be without isolating language document by document.

Clean Portuguese text. On the 200 clean literary pages, scored against trusted human transcription, NED runs from 80.81 to 97.46 across the roster, and clean printed Portuguese is near-solved across model classes. The vision-language model Qwen2.5-VL leads at

97.46 [96.84, 97.91], with HunyuanOCR at 97.2, DeepSeek-OCR at 97.11, and Surya at 96.67 close behind (Table 3). The classical pipelines hold here too: PaddleOCR reaches 94.98 and Tesseract 93.68, both within a few points of the best, and the one low outlier is EasyOCR at 80.81, well below the rest, which otherwise cluster between 90 and 97.5. Size does not predict clean-text quality: the 9B GLM-OCR at 93.87 and the 4B Qwen3-VL at 91.71 trail the 1B HunyuanOCR and the 3B DeepSeek-OCR, so on clean printed Portuguese, as on forms, accuracy does not track size.

Degradation behavior. We stress all 14 engines on the degradation set, 224 pages across eight degradation types, reporting NED overall and per type, and this is where the vision-language models and Surya separate cleanly from the classical pipelines. Qwen2.5-VL stays highest at 97.53 [96.49, 98.16] and Surya at 93.89, with dots.ocr at 96.01 and GLM-OCR at 95.99 also holding above 95; a paired bootstrap over the 224 degraded pages puts Qwen2.5-VL 3.6 NED above Surya (95% interval [1.9, 5.6], excluding zero), the steadier of the two leaders. The classical pipelines fall to a 73.74-to-83.34 band: RapidOCR 83.34, docTR 79.96, PaddleOCR 77.43, Tesseract 77.25, and EasyOCR 73.74 (Table 3).

Table 4 localizes the classical collapse, and the two hard degradations differ. Impulse noise (salt and pepper) breaks every classical pipeline, from Tesseract at NED 0.8 up to RapidOCR at 70.2. Rotation splits them: Tesseract (48.2), EasyOCR (54.1), and docTR (49.3) fail, while RapidOCR (84.2) and PaddleOCR (79.1) hold. Qwen2.5-VL and Surya stay high on both (97.6 and 98.6, 95.5 and 95.0). Tesseract's empty outputs on this set fall entirely on impulse-noise or rotation pages, a genuine failure rather than a scoring gap. Across the eight degradation types the strongest engines lose accuracy gently: Qwen2.5-VL, dots.ocr, GLM-OCR, and Surya never drop below the low 80s, and none of the four fails on impulse noise or rotation. Occlusion is where the smaller vision-language models give way instead: on holes DeepSeek-OCR falls to 55.8 and PaddleOCR-VL to 61.4, below even Tesseract at 81.3, though both hold on impulse noise and rotation. Impulse noise is the sharpest divide in the set: every classical pipeline lands at 70.2 or below, every recent engine at 89.6 or above.

Table 4. Per-degradation NED (%) on the HYB set for all 14 engines, grouped by type; bold is the best in each column.

Engine	Bleed	Blur	CharDeg	Hole	Gauss	S&P	Phantom	Rot
Vision-language models								
PaddleOCR-VL	88.1	97.7	98.1	61.4	98.6	94.0	99.0	92.1
HunyuanOCR	92.3	94.3	69.7	66.3	98.0	96.4	94.2	97.3
MinerU	94.0	94.1	92.2	86.2	94.2	91.0	94.5	71.1
DeepSeek-OCR	97.4	98.2	96.1	55.8	98.2	89.8	98.4	90.8
dots.ocr	98.1	97.9	97.8	82.3	98.3	96.7	98.5	98.4
Qwen3-VL	87.8	88.7	94.0	84.9	90.7	89.6	92.5	72.5
Qwen2.5-VL	98.4	98.7	**98.3**	**91.3**	98.4	**97.6**	98.8	**98.6**
GLM-OCR	97.4	97.1	97.1	89.2	97.2	95.3	98.1	96.6
OCR engines								
Tesseract	**98.7**	98.3	92.9	81.3	**99.1**	0.8	98.8	48.2
EasyOCR	73.7	82.9	82.9	77.2	86.4	45.2	87.6	54.1
docTR	84.0	93.3	87.4	72.6	93.9	66.4	92.8	49.3
RapidOCR	81.0	87.2	85.1	79.1	91.3	70.2	88.6	84.2
PaddleOCR	75.1	81.9	81.3	73.5	82.8	61.3	84.5	79.1
Surya	84.6	**98.8**	93.2	89.9	95.2	95.5	**99.1**	95.0

Local cost axis. On the same images and the same machine, processing time separates the

engines along a different axis than accuracy (Table 5). The two general vision-language models are the slow end: Qwen3-VL and Qwen2.5-VL take 8.67 and 8.45 s per form, against 2.4 s for HunyuanOCR and 1.71 s for DeepSeek-OCR. Size is not the predictor: GLM-OCR, the largest engine, is among the fastest at 2.45 s per form, while dots.ocr at 3B is the highest-latency engine on identity documents at 19.8 s per page, well above the rest, because it emits dense layout markup even for a small card image. Surya, an OCR engine and the cross-axis accuracy leader, runs at 5.88 s per form and 1.51 s per identity document, a moderate cost for its accuracy. The five classical pipelines, timed on the CPU, range from Tesseract at 0.80 s per form to PaddleOCR at 10.76 s, the slowest engine on forms, yet on the smaller identity images all five run in about 0.17 to 1.49 s, so on small inputs a CPU pipeline can be the cheapest option. We read these recorded medians as broad time tiers, the slow general vision-language models against the faster specialists and fp8 engines, rather than precise ratios: the slow general vision-language models split on what the cost buys, Qwen3-VL slow with no accuracy payoff on forms, the equally slow Qwen2.5-VL buying the top forms recall.

Table 5. Median per-page processing time (s) by document type, with model size and device. "–" marks cells that do not apply.

Vision-language models

Engine	Params	Device	Forms	IDs
PaddleOCR-VL	0.9B	GPU	2.38	0.4
HunyuanOCR	1B	GPU	2.4	0.49
MinerU	1.2B	GPU	3.62	1.35
DeepSeek-OCR	3B	GPU	1.71	0.48
dots.ocr	3B	GPU	5.89	19.8
Qwen3-VL	4B	GPU	8.67	2.5
Qwen2.5-VL	7B	GPU	8.45	2.43
GLM-OCR	9B	GPU	2.45	0.62

OCR engines

Engine	Params	Device	Forms	IDs
Tesseract	–	CPU	0.80	0.17
EasyOCR	–	CPU	6.96	0.63
docTR	–	CPU	1.84	1.00
RapidOCR	–	CPU	3.58	0.51
PaddleOCR	–	CPU	10.76	1.49
Surya	–	GPU	5.88	1.51

6. Discussion and Limitations

For a deployer choosing an engine, these results support a decision, not a ranking: no single model class wins outright, and the forms and identity evidence rests on 50 documents each. Section 7 gives the recommendations. Two engines, the OCR engine Surya and the vision-language model Qwen2.5-VL, hold up on all four document types, while most general vision-language models vary a lot from one to the next (GLM-OCR reads forms at 0.965 but clean prose at 93.87; dots.ocr reads forms at 0.931 but identity photos at 0.771). Accuracy does not track parameter count on our documents; that describes these samples, not a general law.

Why engines fail, and not only which ones, is the actionable part for a deployer. Legacy classical pipelines, Tesseract and EasyOCR, collapse on identity-document photos and on impulse noise and rotation, where their detect-then-recognize front end loses the line structure it depends on, while the modern classical pipelines, RapidOCR, PaddleOCR, and docTR, hold on clean and identity inputs but still fall on heavy degradation. The two small specialist vision-language models, PaddleOCR-VL and HunyuanOCR, collapse specifically on dense table-heavy forms, which lowers their raw forms recall and is why we report the excluding-degenerate numbers. Surya and Qwen2.5-VL lose accuracy gently on every document type, which is what makes them the all-rounders and gives the deployer a practical guide: legacy front ends fail on photos and noise, the small OCR vision-language models fail on dense layout, and the all-rounders fail nowhere badly.

Three limitations bound this reading, and we take them in turn.

Sample size. The forms and identity samples hold 50 and 50 documents, so the confidence intervals are wide. We run no equivalence test against a pre-stated margin, so we cannot resolve the size question at this sample size; that is not the same as showing there is no difference. A sample of $n{=}100$ would tighten each half-width by about a third and would likely separate several pairs that overlap here. Every ordering we report therefore characterizes these samples, not the full diversity of Brazilian documents.

The reconstructed gold. A large multimodal model reconstructed the forms gold. It is not among the evaluated engines, but it belongs to the same model class as the vision-language engines, so the fields such models read well are the ones most likely to have entered the gold. That selection bias would favor the vision-language class. The two-annotator review bounds outright errors, with none found in 250 sampled fields, but it does not bound this bias; a gold built by human annotators from scratch would settle it, and we leave that to future work. Our annotation decisions become part of the benchmark as well: we keep only filled values, match up to punctuation, and do not standardize how a marked checkbox is encoded. Any engine later scored against this gold inherits those choices.

The cost axis. We measure per-page processing time on one machine, one runtime, and each engine's vendor defaults. Two engines run in fp8 to fit the card, so for those two the engine and its numerical precision are confounded. The classical pipelines are timed on CPU and the vision-language engines on GPU. These are the costs of one representative on-premises deployment, not inference costs that transfer to other hardware. We report no peak-memory figure. The runtime reserves most of the card up front, so what we would measure is that setting, not the engine.

Future work covers three gaps: the open engines we could not run in this pass [Taghadouini et al. 2026], a larger Brazilian gold set validated by human annotators, and a born-digital Brazilian full-page corpus, which does not yet exist.

These engines read citizens' forms and identity documents, where a misread field can corrupt an administrative record and no engine is yet turnkey on identity photos. The local-hardware regime this benchmark measures is also the one that keeps personal documents off external services, so we evaluate only on-premises open engines and use synthetic identity images with fabricated values, processing no real personal data.

7. Conclusion

We measured the recent generation of open OCR engines on Brazilian documents under a local cost budget, placing vision-language models and purpose-built OCR engines side by side, and the result is a decision guide rather than a single winner. For typed administrative forms and clean printed prose, a modern classical pipeline or a small OCR model reads well at low latency and model size, so no large vision-language model is needed. For identity-document photos, choose a modern engine of any class and add human verification: Surya (0.921 FVR) leads every vision-language model and the best modern classical pipeline matches them, while the legacy classical engines collapse and should be avoided. For degraded or noisy scans, routine in the public service, prefer Surya or a strong vision-language model, since every classical pipeline falls sharply on impulse noise and the legacy pipelines also on rotation. For one engine across all of these, Surya and

Qwen2.5-VL are the only consistent all-rounders over all four document types; whether accuracy tracks model size our samples cannot resolve either way. The contribution is the Brazilian measurement and the corrected gold and scoring harness, released for an institution to run on its own documents.[11][12]

References

Cardeira, J., Glória-Silva, D., da Luz, M. L., Ferreira, R., Tavares, D., Semedo, D., and Magalhães, J. (2026). PorTEXTO: A european portuguese benchmark for visual text extraction. *arXiv:2606.19096*.

Fernandes, L. et al. (2024). BRIDP: Dataset and validation method for BRazilian identity document parsing. In *ISDA*.

Jaume, G. et al. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR-OST*.

Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., and Billy, A. (2017). DocCreator: A new software for creating synthetic ground-truthed document images. *Journal of Imaging*, 3(4):62.

Kargaran, A. H. et al. (2026). GlotOCR bench: OCR models still struggle beyond a handful of unicode scripts. *arXiv:2604.12978*.

Li, Z. et al. (2026). MDPBench: A benchmark for multilingual document parsing in real-world scenarios. *arXiv:2603.28130*.

Neto, A. F. d. S. et al. (2024). BRESSAY: A Brazilian Portuguese dataset for offline handwritten text recognition. In *ICDAR*.

Ouyang, L. et al. (2025). OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *CVPR*.

Poznanski, J. et al. (2025). olmOCR: Unlocking trillions of tokens in PDFs with vision language models. *arXiv:2502.18443*.

Ribeiro, G. R. et al. (2025). Vision-language models for automated property feature extraction in tax assessment: A case study. In *ENIAC*.

Rigal, B. et al. (2026). Benchmarking vision-language models for French PDF-to-Markdown conversion. *arXiv:2602.11960*.

Santos, M. K. et al. (2023a). ESTER-Pt: An evaluation suite for TExt recognition in Portuguese. In *ICDAR*.

Santos, Y. et al. (2023b). Evaluation of optical character recognition (OCR) systems dealing with misinformation in Portuguese. In *SIBGRAPI*.

Smith, R. (2007). An overview of the Tesseract OCR engine. In *ICDAR*.

Taghadouini, S. et al. (2026). LightOnOCR: A 1B end-to-end multilingual vision-language model for state-of-the-art OCR. *arXiv:2601.14251*.

Xu, Y. et al. (2022). XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of ACL*.

Yang, Z. et al. (2025). CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. In *ICCV*.

Zhong, Y. et al. (2026). OCRVerse: Towards holistic OCR in end-to-end vision-language models. *arXiv:2601.21639*.

11. `https://gitlab.com/cristhianavila.aluno/ocr-ptbr-benchmark`

12. **Acknowledgments.** This work was partially supported by Capes, CNPq (Grant No. 409743/2025-9), and FAPERGS (Grant Nos. 24/2551-0001368-7 and 24/2551-0000726-1). We thank the anonymous reviewers of ENIAC 2026 for their suggestions and comments. The authors also acknowledge the use of Generative AI as a supporting tool for reviewing and improving the code and manuscript, and the use of a large multimodal model to reconstruct the forms gold, as described in Section 3; they verified every number and take responsibility for all content.