

PDF OCR Production Test Report

Reviewed by PDF Toolbox | Tested August 6, 2026 (UTC) | Synthetic data only

PASS

One live production job created a searchable four-page English PDF from clean, degraded, sideways, and mixed-content controls. Separate signed and encrypted controls were refused without partial output, and the scheduled retention check passed.

Scope and method

The test used the live PDF Toolbox upload, PDF OCR job, queue-status, and download endpoints. The synthetic source contained four Letter pages: a clean scan, a faint and blurred scan, a scan with text rotated 90 degrees inside the page image, and a mixed page containing both raster text and native PDF text.

The source also contained selected document metadata and one synthetic embedded attachment. The downloaded result was checked by PDF text extraction, rendered-page inspection, structure and metadata inspection, MIME and PDF-header checks, and SHA-256 comparison. Separate synthetic signed and encrypted PDFs tested refusal behavior and partial-output cleanup.

Check	Observed result	Status
Clean scan	The expected sentence, invoice amount, reference ID, and CLEAN / ALPHA / 4827 controls were searchable.	PASS
Degraded scan	FAINT / BRAVO / 5938 were searchable. The source code 00I1-L5S8-B6G9 was read as 0011-L5S8-B6G9, demonstrating OCR error risk.	PASS WITH ERROR
Sideways scan	ROTATED / CHARLIE / 6049 became searchable. The visible page content remained sideways in the tested output.	PASS WITH LIMIT
Mixed page	Native text ECHO 8261 remained searchable; raster text DELTA 7150 on the same page was not OCR-processed under skip-text behavior.	EXPECTED LIMIT
Output format	The result was ordinary PDF 1.7 with no PDF/A part or conformance marker.	PASS
Rendering	All four output pages rendered successfully.	PASS

Original production output evidence

Each image below was rendered directly from the accepted production output PDF. Page 4 is not published as an image because it is the mixed native-text control documented in the results and downloadable files.

FIX 3 OCR AUDIT - CLEAN SCAN

CLEAN ENGLISH CONTROL ALPHA 4827
Searchable output should include this sentence.
Invoice total: \$1,234.56
Reference ID: OCR-CLEAN-2026

FIX 3 OCR AUDIT - DEGRADED SCAN

FAINT CONTROL BRAVO 5938
Low contrast and blur can cause recognition errors.
Code: 001-1558-8609

FIX 3 OCR AUDIT - ROTATED SCAN
ROTATED CONTROL CHARLIE 6049
The source image is rotated 90 degrees.

1. Clean scan output - expected text recognized

2. Degraded scan output - recognition error documented

3. Sideways scan output - searchable text, visible orientation retained

What the images do and do not show

These images show page appearance in the tested output. Searchability was verified separately by text extraction. A rendered page alone cannot prove text accuracy, reading order, attachment preservation, metadata preservation, signature state, or retention behavior.

Preservation observations

The tested attachment fix3-attachment-control.txt remained embedded. The selected Title, Author, Subject, and Keywords document fields remained. This is evidence for the published synthetic file only and is not a guarantee that every metadata field, attachment, form, annotation, link, accessibility feature, layer, or other PDF structure will be preserved.

Runtime, refusal, and exact production record

Recorded production runtime

Component	Recorded version
Core OCR	Python 3.10.12; OCRmyPDF 16.13.0
Recognition	Tesseract 4.1.1; Leptonica 1.82.0
PDF utilities	Ghostscript 9.55.0; unpaper 6.1; pikepdf 10.11.0
Shared Python runtime	PyMuPDF 1.26.7; pdf2docx 0.5.12

Signed and encrypted controls

Control	Production result	Partial output
Digitally signed PDF - job 695	Refused because OCR changes the document and would invalidate signatures. The user-facing message requests an unsigned copy.	Absent
Encrypted PDF - job 696	Refused because the PDF was encrypted or password-protected. The user-facing message requests an unencrypted copy.	Absent

Exact production record

Production	https://pdftoolboxonline.com Commit: a320ccfe4b6cff97e9661dca36abc06fc935c5fa
Evidence run	/root/pdf-toolbox-validation/fix3-ocr-20260806-030448 Valid session / job: UYue_GcXF_Y6kS1Ojh9t5 / 694
Source	pdf-ocr-synthetic-source.pdf - 422070 bytes SHA-256: f3483d734712cc3be10ca8031f849d5fdd8942cf5e0d82adbb734c1d95f5d940
Output	pdf-ocr-production-output.pdf - 581827 bytes SHA-256: fdad27ee02dedeefa3afd3775354bc674708d4ad4d346637076394c74f8811fb

Interpretation, retention, and user guidance

What the production result supports

For this synthetic English PDF, the production tool created searchable text for clean, degraded, and sideways scanned pages; preserved native text on a mixed page; produced standard PDF 1.7 output; retained the tested attachment and selected metadata; and refused signed and encrypted controls without partial output.

What the result does not guarantee

1	Correct recognition of every character, word, number, code, or reading order.
2	OCR of raster text on pages that already contain native text.
3	Correction of visible page orientation or exact preservation of page appearance.
4	Preservation of every PDF feature, metadata field, attachment, form, link, annotation, layer, or accessibility structure.
5	Suitability as legal, financial, archival, accessibility, or transcription certification.

User verification checklist

1	Keep the original PDF as the authoritative source.
2	Compare important names, dates, amounts, reference numbers, and codes.
3	Inspect every output page for visual, orientation, and layout changes.
4	Test search, selection, copy-paste, and reading order where they matter.
5	Check links, forms, annotations, accessibility, metadata, and attachments when relevant.

Scheduled retention test

The acceptance test created a valid synthetic session with upload, output, and preview directories older than the 60-minute eligibility threshold. The scheduled janitor deleted all three. It runs every five minutes, so deletion is not guaranteed at the exact 60-minute mark. The distributed source and output are retained synthetic evidence copies, not the temporary session copies.

Conclusion

The production tool can add searchable English text to eligible scanned PDF pages, with accuracy and preservation limitations that require review. It is not an exact transcription service, a visual-preservation guarantee, a PDF/A converter, or a certification workflow.