# **Comparative OCR Benchmark: Tesseract 5.5.0, EasyOCR 1.7.2, and Mistral OCR**

## **Introduction**

Optical Character Recognition (OCR) is a cornerstone technology for digitizing and extracting text from scanned documents, PDFs, and images. As organizations migrate from paper and static files to structured, searchable content, OCR’s accuracy and reliability critically impact productivity, regulatory compliance, and accessibility. In this report, we benchmark three leading open-source OCR solutions—Tesseract 5.5.0, EasyOCR 1.7.2, and Mistral-OCR—across a set of challenging government and legal documents typical for the Saudi financial sector. The aim: provide a transparent, quantitative comparison of their Arabic OCR capabilities (including mixed-language context) and highlight their suitability for real-world enterprise and compliance digitization projects.

## **Technology Overview**

### **Tesseract: The Industry Standard for Classical OCR**

Tesseract, developed by HP and maintained by Google, is the most widely used open-source OCR engine. Its recognition pipeline uses multi-stage image processing: binarization, connected component analysis (blob detection), feature extraction, and then statistical or LSTM character classification models. Tesseract’s strength lies in its fast and reliable recognition for clean, typewritten documents, especially for Latin scripts. Its support for Arabic and other complex scripts relies predominantly on automated character segmentation and language-specific pre-trained language models. Tesseract remains the mainstream baseline, especially in government and enterprise solutions.

### **EasyOCR: Deep Learning-Based Neural OCR**

EasyOCR is a modern, open-source OCR engine that utilizes deep learning—specifically, Convolutional Neural Networks (CNNs) for text detection, followed by recurrent neural networks (RNNs) for text recognition. The model is trained on very large, script-specific datasets and excels in handling noisy, heterogeneous input such as photographs, billboards, and multi-lingual documents. EasyOCR’s flexible, end-to-end architecture improves robustness to skew, font variation, and image noise and is especially strong on scripts, like Arabic, which exhibit contextual glyph variation.

### **Mistral OCR: Transformer-based, Foundation Model for Document OCR**

Mistral OCR represents a new generation of “foundation model” OCR systems informed by the advances in large language models (LLMs) and computer vision transformers. The system leverages vision transformers (ViT) combined with sequence modeling and self-attention, excelling in pattern recognition for varied fonts, degraded scans, and dense, multi-column layouts. It is trained on billions of tokens across English and Arabic, including legacy and modern text. Mistral’s architecture is especially tuned for script-complex and layout-challenged documents, as is typically the case in financial regulation and government forms.

## **Benchmark Dataset Description**

Three representative documents were chosen for rigorous, line-level accuracy assessment:

* **Document 1:** The bilingual (Arabic/English) “Banking Control Law”, a government regulation in parallel columns, exported directly from a document processing system (high-quality digital PDF).
* **Document 2:** “Regulations for Consumer Financing”, a bilingual (Arabic/English) circular in table-heavy, multi-section format. This is an image-based scan (degraded, slightly skewed, with bleed-through), rebuilt using OCRmyPDF.
* **Document 3:** “مبادئ الالتزام للبنوك والمصارف التجارية العاملة في المملكة العربية السعودية” (Principles of Compliance for Commercial Banks in Saudi Arabia), a full-Arabic, text-dense compliance code, scanned in monochrome (font bleed, moderate noise, complex headers).

Each document was evaluated over both English and Arabic text (where present), but this report focuses on Arabic content performance.

## **Document-by-Document Model Comparison**

### **Document 1: Banking Control Law (Bilingual, Born-Digital PDF)**

|  |  |  |
| --- | --- | --- |
| **Model** | **Char. Accuracy (%)** | **Word Accuracy (%)** |
| Tesseract | 82.7 | 53.9 |
| EasyOCR | 86.3 | 59.7 |
| Mistral-OCR | 93.2 | 74.5 |

**Observations:**

* Tesseract, while consistent on digital-born PDF, exhibited numerous joining-form and kashida (tatweel) confusion issues in the Arabic script, breaking words and lines unpredictably.
* EasyOCR, thanks to its neural network backbone, mitigated most segmentation noise, achieving a 4–6 point improvement in character-level and a 6-point improvement in word-level accuracy over Tesseract.
* Mistral-OCR performed at near-commercial-transcription levels (>93% CA), robustly segmenting complex lines and preserving word unity.

### **Document 2: Consumer Financing Regulations (Bilingual, Image-Scan)**

|  |  |  |
| --- | --- | --- |
| **Model** | **Char. Accuracy (%)** | **Word Accuracy (%)** |
| Tesseract | 77.9 | 46.1 |
| EasyOCR | 82.1 | 52.9 |
| Mistral-OCR | 89.7 | 68.2 |

**Observations:**

* The image-scanned, degraded layout revealed Tesseract’s limitations: tatweel-as-blank errors fragmented long Arabic tokens, and noise in the tables caused severe word-level loss.
* EasyOCR proved more robust to scanning artifacts and layout noise, pulling ahead especially in preserving line and word integrity within noisy table contexts.
* Mistral-OCR provided a standout performance in this adverse setting, with strong error correction and word segmentation.

### **Document 3: Principles of Compliance (Full Arabic, Image-Scan)**

|  |  |  |
| --- | --- | --- |
| **Model** | **Char. Accuracy (%)** | **Word Accuracy (%)** |
| Tesseract | 68.4 | 34.7 |
| EasyOCR | 73.1 | 41.8 |
| Mistral-OCR | 87.9 | 62.5 |

**Observations:**

* The most demanding test: heavy, stylized Arabic text, moderate-to-severe scan artifacts. Tesseract fell below 70% character accuracy, with word accuracy below 35% (barely readable).
* EasyOCR, while better, still struggled with excessive duplicated letters and tatweel-segmentation, especially in headers and dense bullet points.
* Mistral-OCR proved robust, maintaining clarity, word unity, and layout fidelity even under challenging scan conditions.

## **Cross-Document Performance Summary**

|  |  |  |
| --- | --- | --- |
| **Model** | **Avg. Char. Accuracy (%)** | **Avg. Word Accuracy (%)** |
| Tesseract | 76.3 | 44.9 |
| EasyOCR | 80.5 | 51.5 |
| Mistral-OCR | 90.3 | 68.4 |

**Key Takeaways:**

* Mistral-OCR advances accuracy by 10–14 points (CA) and 17–23 points (WA) over the previous generation (EasyOCR/Tesseract) for modern and legacy Arabic, in both clean and scanned/mixed layouts.
* EasyOCR represents a usable improvement over Tesseract in all realistic conditions and is a sound choice where lightweight, on-premise neural OCR is wanted.
* Tesseract’s stock model is only suitable for “clean” digital-born PDFs or as a fallback baseline for Latin-script forms; word accuracy on Arabic, especially in scanned or degraded documents, is generally below enterprise and compliance standards.

## **Conclusion and Industry Fit Assessment**

This benchmark clearly demonstrates that, while neural-network and transformer-based OCR models—represented here by EasyOCR and Mistral-OCR—deliver substantial gains over traditional models like Tesseract, *none* of the evaluated open-source approaches currently reach the word-accuracy thresholds historically accepted as sufficient for enterprise, regulatory, or archival use in production Arabic document digitization.

**Industry Minimum Thresholds for “Acceptable” OCR (Word Accuracy):**

* **English (clean):** 97–98%
* **Arabic (clean):** 85–90%
* **Arabic (scanned/legacy):** 80–85%
* **Mixed (scan):** 80–85%

Mistral-OCR, the highest performer in our benchmark, averaged only **68.4% word accuracy** across three realistic, representative legal and regulatory documents—including born-digital bilingual files and scanned government circulars with complex layouts. This is *far below* the 80–85% generally cited as a minimum for practical search, compliance review, or downstream automation. EasyOCR and Tesseract performed at still lower levels (51.5% and 44.9% average word accuracy, respectively).