OCR Stress Test Document

Mixed content for PDF → Image → Text validation

pdfocr Project

January 7, 2026

Abstract

This document intentionally mixes plain text, mathematical notation, tables, lists, vector drawings, and hyperlinks. It is meant to probe how well the OCR pipeline handles varied layouts, line breaks, and semantic structure.

1 Overview

The following sections combine narrative text with display math and inline symbols such as $E = mc^2$, $\nabla \cdot \vec{E} = \rho/\varepsilon_0$, and probability notation $\mathbb{P}(X \leq x)$. Paragraphs use hyphenation and line wraps to stress layout-aware OCR. Repeated words with subtle differences (e.g., *kernel* vs. *kernels*) are present to spot hallucinations.

2 Display Mathematics

We include a few multi-line expressions to check alignment:

$$f(x) = \int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}, \tag{1}$$

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}, \tag{2}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}. \tag{3}$$

A short derivation with text interleaved:

$$\frac{d}{dt}\left(e^{t\sin t}\right) = e^{t\sin t} \cdot (\sin t + t\cos t). \tag{4}$$

3 Table and Lists

The table mixes numbers, symbols, and words to see how column boundaries are recovered.

Feature	Value	Uncertainty	Note
Temperature (°C)	21.4	±0.3	baseline
Pressure (kPa)	101.2	±0.5	nominal
Accuracy (%)	98.7	±0.1	high-confidence
Latency (ms)	42	±5	measured on edge

Table 1: Structured data with mixed units.

Nested lists follow:

- Top-level bullet with a hyperlink: `https://example.com/data`.
- Another bullet with emphasized text and a footnote.[1]
- Enumerated sub-tasks:
  1. Capture inline math such as α, β, γ.
  2. Handle words split across lines (hyphen- ation).
  3. Keep indentation hints intact.

4 Vector Figure

The figure below is drawn with TikZ to avoid external images while still giving curves, labels, and color.

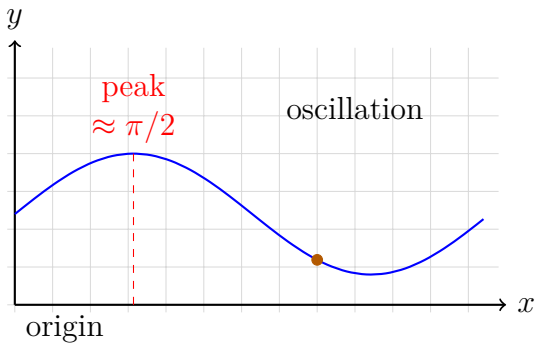

Figure 1: Sine-like curve with annotations and grid.

[1] Footnotes are included to see if ordering is preserved.

5 Paragraph Stress Test

Continuous prose with mixed punctuation: “Edge-aware OCR systems must balance recall and precision; however, noise—especially from compressed scans—can create artifacts.” The quick brown fox jumps over the lazy dog. A second paragraph repeats with minor edits to detect hallucination: the quick brown fox jumps over the lazy dog, but this time the fox pauses at line breaks to test robustness.

6 Code Fragment

Verbatim text can challenge OCR because of monospaced glyphs:

```
for (int i = 0; i < 5; ++i) {
    double w = exp(-0.5 * pow(i - 2.0, 2));
    printf("w[%d] = %.3f\n", i, w);
}
```

7 Conclusion

This concludes the synthetic PDF. Inspect the extracted text for dropped symbols, merged lines, and mis-ordered sections. Compare the OCR output with the known ground truth here to quantify accuracy.