Benchmark Document for PDF Extraction

A synthetic test document with headers, tables, formulas, and images

1. Introduction

This document is designed to test PDF extraction engines. It contains various elements commonly found in technical documents: headings at multiple levels, paragraphs of text, data tables, mathematical formulas, and figures. The goal is to evaluate how well each extraction engine preserves the structure and content of the original document.

2. Mathematical Formulas

This section contains mathematical formulas in LaTeX notation:

2.1 Famous Equations

- Einstein's mass-energy equivalence:
  ```
  E = mc^2
  ```
- Euler's identity:
  ```
  e^(i*pi) + 1 = 0
  ```
- Pythagorean theorem:
  ```
  a^2 + b^2 = c^2
  ```
- Quadratic formula:
  ```
  x = (-b +/- sqrt(b^2 - 4ac)) / 2a
  ```

2.2 LaTeX Block Formula

The Schrodinger equation in Dirac notation:

```
$$i*hbar * d/dt |psi(t)> = H |psi(t)>$$
```

3. Data Table

The following table shows benchmark results for different algorithms:

Algorithm	Time (ms)	Memory (MB)	Accuracy (%)
Quick Sort	45.2	12.5	100.0
Merge Sort	52.8	24.0	100.0
Bubble Sort	1250.3	8.2	100.0
Neural Net	89.5	256.0	98.7

4. Figure

Figure 1 below shows a plot of trigonometric functions:

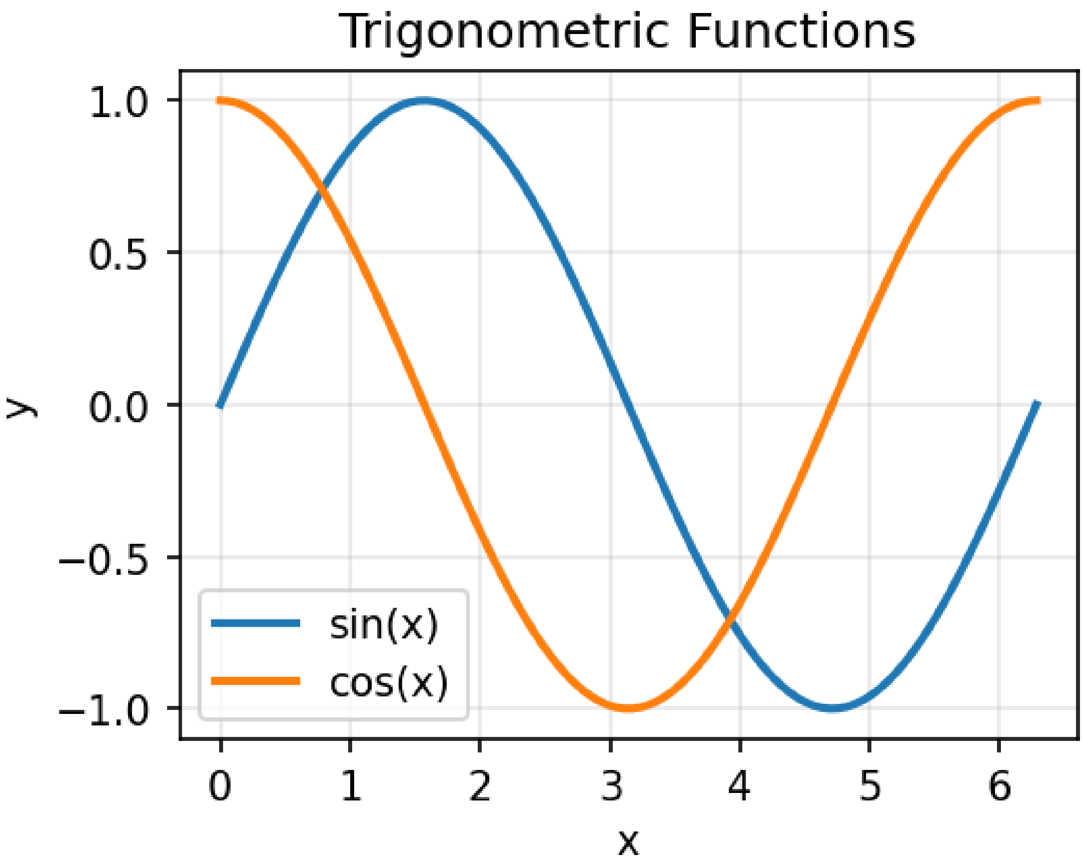

*Figure 1: Plot of sin(x) and cos(x) functions over [0, 2*pi]*

5. Conclusion

This benchmark document provides a standardized test case for evaluating PDF extraction engines. A successful extraction should preserve:

1. Document structure (headings, sections)
2. Table formatting and data
3. Mathematical formulas
4. Figure references
5. Text formatting (bold, italic)

The extracted content can be compared against this source to measure extraction quality.

References

[1] Smith, J. (2024). PDF Extraction Methods. Journal of Document Processing, 15(3), 45-67.

[2] Johnson, A. & Lee, B. (2023). Benchmarking Document AI Systems. arXiv:2301.12345.

[3] Williams, C. (2024). OCR in the Age of Large Language Models. ACM Computing Surveys.