Document Format Extraction Test

Introduction

This document contains various types of content to test text extraction quality across different formats.

Text Formatting

This paragraph contains **bold text**, *italic text*, and `inline code`.

Code Block

```
def hello_world():
    print("Hello, World!")
    return 42
```

Lists

Unordered list:
- Item 1
- Item 2
  - Nested item A
  - Nested item B
- Item 3

Ordered list:
1. First item
2. Second item
3. Third item

Table

Format	Extraction Quality	Processing Speed
PDF	Good	Medium
EPUB	Excellent	Fast
HTML	Very Good	Fast

Special Characters

Unicode test: café, naïve, résumé, ■■, ■■■, ■■■■■■■
Mathematical notation: $E = mc^2$, $\sum$, $\int$, $\sqrt{}$, $\leq$, $\geq$

Conclusion

This document tests various formatting elements to evaluate extraction quality.