Data Extraction Project Test

Goal: Verify PyMuPDF's ability to extract content accurately, identify structural elements, and handle various file conditions, which is crucial for the AI data ingestion phase.

PHASE I: Basic Extraction Check

1. **This line is digital text, created directly in Word. It should be easily extracted.**
2. **The number of pages in the output JSON must match the actual page count.**

PHASE II: Header and Error Test

- **Header Test: The title "PHASE II" is a primary header. The logic must mark it as such (based on font size or bold flag).**
- **Size Limit Simulation: The system must check if the file size exceeds 20MB.**

Key Sentence for Verification: The crucial test case for the PyMuPDF library is its ability to handle both digital and image-based text reliably.