

Section 1: Executive Summary

This report provides an overview of PDF generation and chunk extraction for AI knowledge search systems. The document is structured into several sections to simulate a real-world report.

1.1 Background

PDFs are widely used for document exchange. Accurate extraction of structured information, such as headers and sections, is crucial for downstream AI models.

Section 2: Data Extraction Results

The following table summarizes extraction statistics from several sample documents.

Document	Pages	Chunks	Sections
report1.pdf	5	24	3
whitepaper.pdf	12	67	8
manual.pdf	20	182	10

Section 3: Conclusions & Future Work

Improving the realism of test PDFs, with headers, sections and multi-page layouts, enhances the evaluation of extraction and search capabilities.

Future improvements include:

- Adding figures and tables
- Simulating footnotes and references
- Including section numbering and TOC

