

Gleann Document Extraction Benchmark Report

A comprehensive evaluation of document extraction backends for knowledge graph construction.

1. Introduction

Document extraction is a critical component of any knowledge management system. The quality of extraction directly impacts downstream tasks such as semantic search, question answering, and knowledge graph construction. This report evaluates three extraction backends: MarkItDown (Microsoft), Docling (IBM), and Marker (marker-pdf with Surya OCR).

Each backend has distinct strengths: MarkItDown excels at broad format support with minimal dependencies, Docling provides high-fidelity PDF parsing with table detection, and Marker leverages deep learning OCR for complex document layouts including scanned documents.

2. Methodology

2.1 Test Corpus

The test corpus consists of documents with known structure: headings at multiple levels, paragraphs with technical content, tables with numerical data, and inline formatting. Each document has a ground-truth section map used for accuracy measurement.

2.2 Metrics Definition

We measure section detection rate, hierarchy accuracy, content fidelity, and latency per document conversion.

Metric	Description	Target
Section Detection	Known headings found	>= 90%
Hierarchy Accuracy	Correct parent-child edges	>= 85%
Content Fidelity	Word overlap with source	>= 95%
Latency (PDF)	Time per document	< 5s

3. Results

3.1 PDF Extraction

All three backends successfully extracted content from standard PDF files. Marker showed the best section hierarchy detection due to its visual layout analysis. Docling excelled at table extraction. MarkItDown provided the fastest processing time but with minimal structural information.

3.2 DOCX Extraction

For DOCX files, MarkItDown and Marker both correctly preserved heading hierarchy from Word styles.

3.3 Image-based Documents

Only Marker via Surya OCR could extract text from scanned document images. MarkItDown and Docling return empty or minimal results for image inputs without embedded text layers.

4. Conclusion

No single backend dominates across all dimensions. For production deployments, we recommend Marker for PDF-heavy workflows requiring OCR, Docling for table-rich PDF documents, and MarkItDown as a lightweight fallback for Office formats. The gleann plugin system allows users to install the backend that best matches their corpus.

4.1 Future Work

Integration of LLM-assisted post-processing via Ollama could improve extraction quality for all backends. Additionally, ensemble approaches combining multiple backend outputs may yield higher accuracy than any single backend.