

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: Markdown (Microsoft), Docling (IBM), and Marker (marker-pdf with Surya OCR).

Each backend has distinct strengths: Markdown 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	$\geq 90\%$
Hierarchy Accuracy	Correct parent-child edges	$\geq 85\%$
Content Fidelity	Word overlap with source	$\geq 95\%$
Latency (PDF)	Time per document	$< 5s$