OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations

Linke Ouyang[1*] Yuan Qu[1*] Hongbin Zhou[1*] Jiawei Zhu[1*] Rui Zhang[1*] Qunshu Lin[2*]
Bin Wang[1*†] Zhiyuan Zhao[1] Man Jiang[1] Xiaomeng Zhao[1] Jin Shi[1] Fan Wu[1] Pei Chu[1] Minghao Liu[3]
Zhenxiang Li[1] Chao Xu[1] Bo Zhang[1] Botian Shi[1] Zhongying Tu[1] Conghui He[1‡]

[1]Shanghai AI Laboratory [2]Abaka AI [3]2077AI

Abstract

Document content extraction is a critical task in computer vision, underpinning the data needs of large language models (LLMs) and retrieval-augmented generation (RAG) systems. Despite recent progress, current document parsing methods have not been fairly and comprehensively evaluated due to the narrow coverage of document types and the simplified, unrealistic evaluation procedures in existing benchmarks. To address these gaps, we introduce OmniDocBench, a novel benchmark featuring high-quality annotations across nine document sources, including academic papers, textbooks, and more challenging cases such as handwritten notes and densely typeset newspapers. OmniDocBench supports flexible, multi-level evaluations—ranging from an end-to-end assessment to the task-specific and attribute-based analysis—using 19 layout categories and 15 attribute labels. We conduct a thorough evaluation of both pipeline-based methods and end-to-end vision-language models, revealing their strengths and weaknesses across different document types. OmniDocBench sets a new standard for the fair, diverse, and fine-grained evaluation in document parsing. Dataset and code are available at https://github.com/opendatalab/OmniDocBench.

1. Introduction

As large language models [1, 28, 39, 44] increasingly rely on high-quality, knowledge-rich data, the importance of accurate document parsing has grown substantially. Document parsing, a core task in computer vision and document intelligence, aims to extract structured, machine-readable content from unstructured documents such as PDFs. This task is particularly critical for ingesting academic papers, technical reports, textbooks, and other rich textual sources

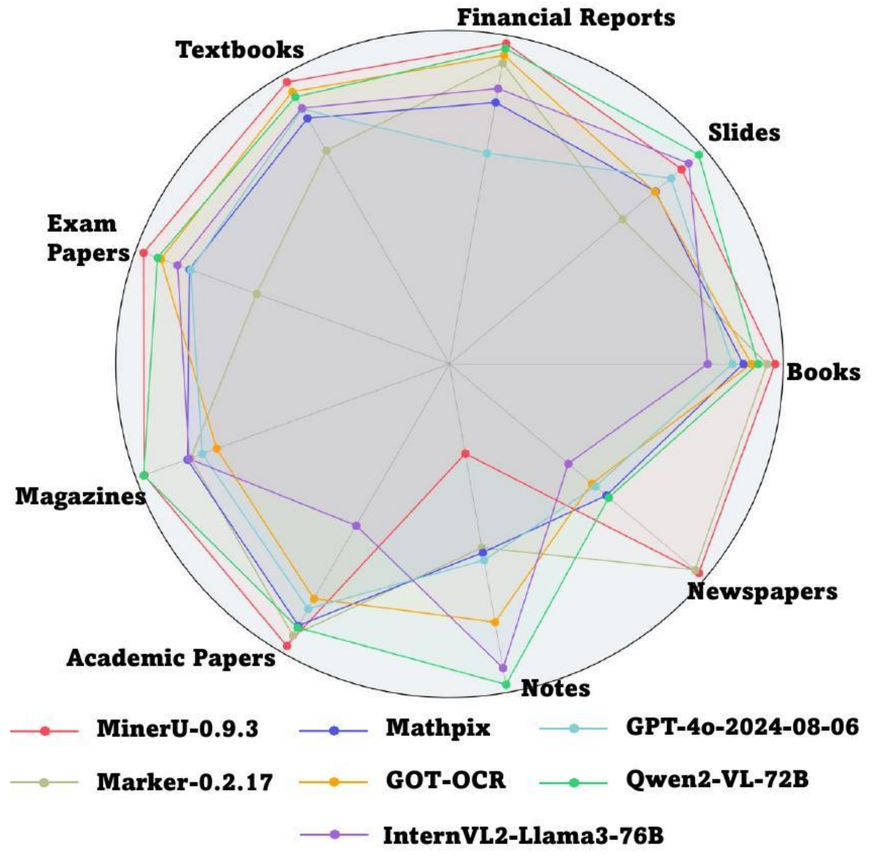

Figure 1. Results of End-to-End Text Recognition on OmniDocBench across 9 PDF page types.

into large language models, thereby enhancing their factual accuracy and knowledge grounding [19, 42, 45, 47, 52]. Moreover, with the emergence of retrieval-augmented generation (RAG) systems [12, 22], which retrieve and generate answers conditionally with external documents, the demand for precise document understanding has further intensified.

To address this challenging task, two main paradigms have emerged: 1) Pipeline-based approaches that decompose the task into layout analysis, OCR, formula/table recognition, and reading order estimation [34, 42]; and 2) End-to-end vision-language models (VLMs) that directly output structured representations (e.g., Markdown) [3, 7, 8, 29, 45, 46, 48]. Although both approaches have demonstrated promising results, conducting a broad comparison of their effectiveness remains challenging due to the absence of a comprehensive and unified evaluation benchmark.

As shown in Table 1, for pipeline-based document parsing systems, dedicated benchmarks [10, 26, 54] have been

* The authors contributed equally.
† Project lead.
‡ Corresponding author (heconghui@pjlab.org.cn).