Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition

Demiao LIN
chatdoc.com

Abstract

With the rapid development of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) has become a predominant method in the field of professional knowledge-based question answering. Presently, major foundation model companies have opened up Embedding and Chat API interfaces, and frameworks like LangChain have already integrated the RAG process. It appears that the key models and steps in RAG have been resolved, leading to the question: are professional knowledge QA systems now approaching perfection? This article discovers that current primary methods depend on the premise of accessing high-quality text corpora. However, since professional documents are mainly stored in PDFs, the low accuracy of PDF parsing significantly impacts the effectiveness of professional knowledge-based QA. We conducted an empirical RAG experiment across hundreds of questions from the corresponding real-world professional documents. The results show that, ChatDOC (chatdoc.com), a RAG system equipped with a panoptic and pinpoint PDF parser, retrieves more accurate and complete segments, and thus better answers. Empirical experiments show that ChatDOC is superior to baseline on nearly 47% of questions, ties for 38% of cases, and falls short on only 15% of cases. It shows that we may revolutionize RAG with enhanced PDF structure recognition.

1 Introduction

Large language models (LLM) are trained on data that predominantly come from publicly available internet sources, including web pages, books, news, and dialogue texts. It means that LLMs primarily rely on internet sources as their training data, which are vast, diverse, and easily accessible, supporting them to scale up their capabilities. However, in vertical applications, professional tasks require LLMs to utilize domain knowledge, which unfortunately is private, and not part of their pre-training data.

A popular approach to equip LLM with domain knowledge is Retrieval-Augmented Generation (RAG). RAG framework answers a question in four steps: the user proposes a query, the system retrieves the relevant content from private knowledge bases, combines it with the user query as context, and finally asks the LLM to generate an answer. This is illustrated in Figure 1 with a simple example. This process mirrors the typical cognitive process of encountering a problem, including consulting relevant references and subsequently deriving an answer. In this framework, the pivotal component is the accurate retrieval of pertinent information, which is critical for the efficacy of the RAG model.

However, the process of retrieval from PDF files is fraught with challenges. Common issues include inaccuracies in text extraction and disarray in the row-column relationships of tables inside PDF files. Thus, before RAG, we need to convert large documents into retrievable content. The conversion involves several steps, as shown in Figure 2:

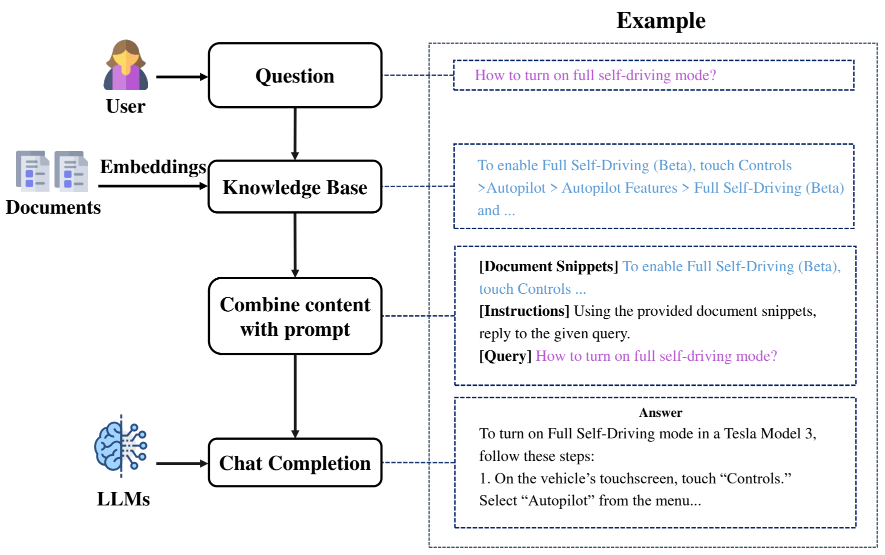

Example
User
Question
How to turn on full self-driving mode?
Documents
Embeddings
Knowledge Base
To enable Full Self-Driving (Beta), touch Controls >Autopilot > Autopilot Features > Full Self-Driving (Beta) and ...
Combine content with prompt
[Document Snippets] To enable Full Self-Driving (Beta), touch Controls ...
[Instructions] Using the provided document snippets, reply to the given query.
[Query] How to turn on full self-driving mode?
LLMs
Chat Completion
Answer
To turn on Full Self-Driving mode in a Tesla Model 3, follow these steps:
1. On the vehicle's touchscreen, touch "Controls."
Select "Autopilot" from the menu...

Figure 1. The workflow of Retrieval-Augmented Generation (RAG).

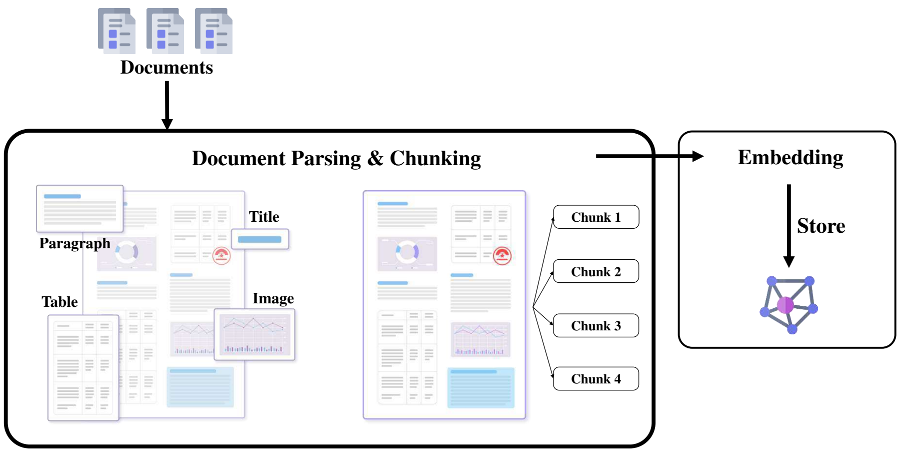


Figure 2. The process of converting PDFs into retrievable contents.

- Document Parsing & Chunking. It involves extracting paragraphs, tables, and other content blocks, then dividing the extracted content into chunks for subsequent retrieval.
- Embedding. It transforms text chunks into real-valued vectors and stores them in a database.

Since each of these steps can lead to information loss, the compounded losses can significantly impact the effectiveness of RAG's responses.

This paper primarily addresses the question of whether the quality of PDF parsing and chunking affects the outcomes of RAG. We will explore the challenges, methodologies, and real-world case studies pertaining to this issue. It will include an examination of two types of methods in this field, namely rule-based and deep learning-based methods, followed by empirical evaluations of their efficacy through practical examples.

2 PDF Parsing & Chunking

2.1 Challenges and Methods Overview

To humans, the cognitive process of perusing any document page is similar. When we read a page, characters are captured by our retinas. Then, in our brains, these characters are organized into

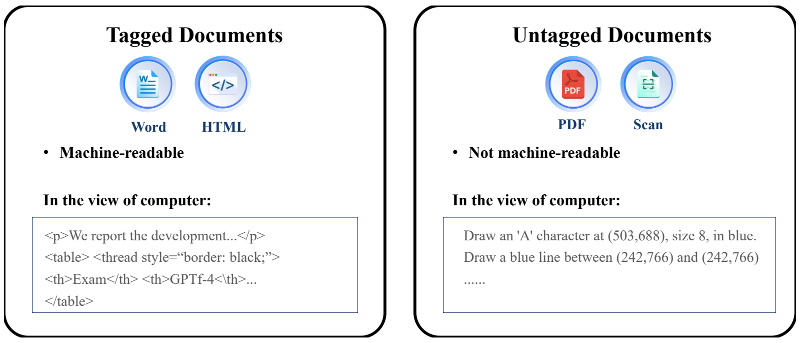

Tagged Documents
Word
HTML
• Machine-readable
In the view of computer:
<p>We report the development...</p>
<table> <thread style="border: black;">
<th>Exam</th> <th>GPTf-4<\th>...
</table>
Untagged Documents
PDF
Scan
• Not machine-readable
In the view of computer:
Draw an 'A' character at (503,688), size 8, in blue.
Draw a blue line between (242,766) and (242,766)
......

Figure 3. Two types of documents in the view of computers.

paragraphs, tables, and charts, and then understood or memorized. However, computers perceive information as binary codes. From their perspective, as illustrated in Figure 3, documents can be categorized into two distinct types:

- Tagged Documents: Examples include Microsoft Word and HTML documents, which contain special tags like <p>and <table> to organize the text into paragraphs, cells, and tables.
- Untagged Documents: Examples include PDFs, which store instructions on the placement of characters, lines, and other content elements on each document page. They focus on 'drawing' these basic content elements in a way that makes the document legible to human readers. They do not store any structural information of the document, like tables or paragraphs. Thus, untagged documents are only for human e-reading, but are unreadable by machines. This becomes evident when attempting to copy a table from a PDF into MS Word, where the original structure of the table is often completely lost.

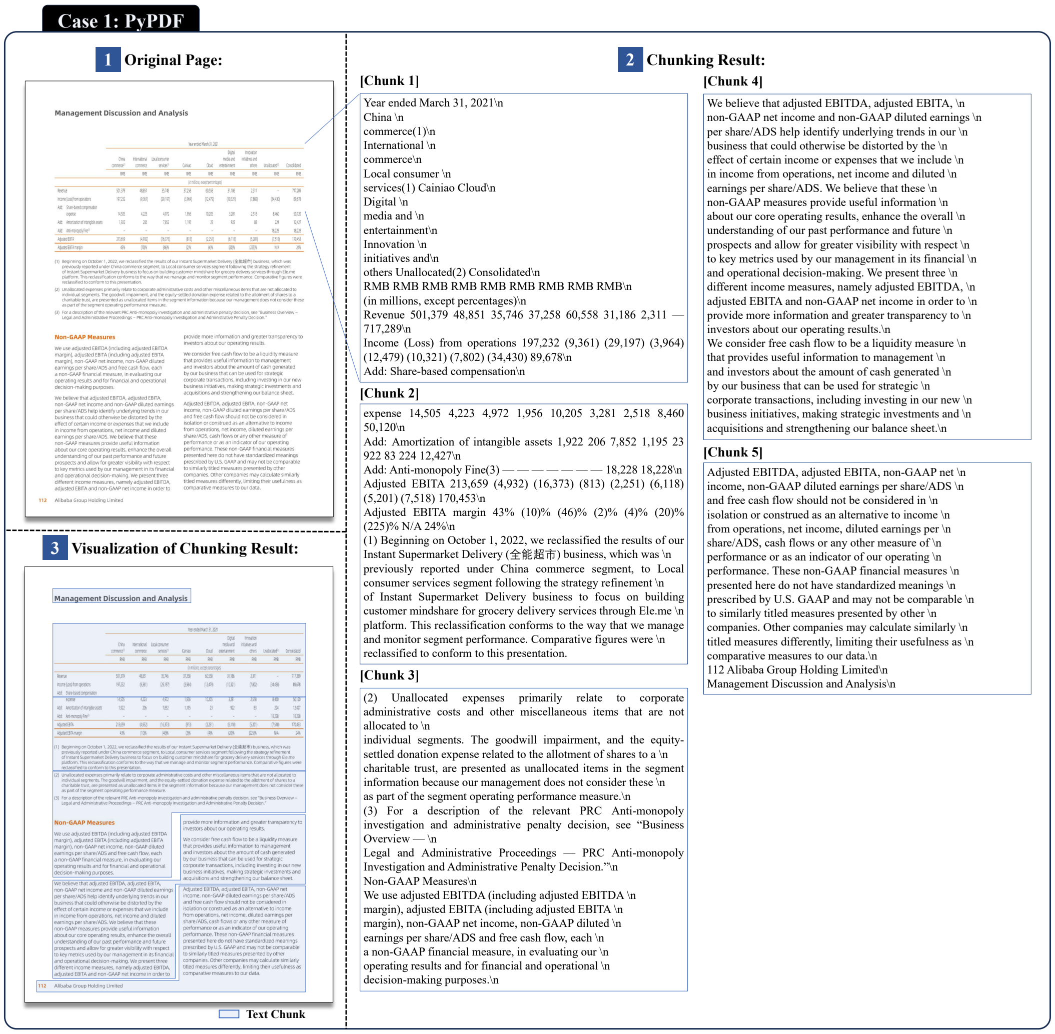
However, Large Language Models (LLMs) exhibit proficiency in processing serialized text. Consequently, to enable LLMs to effectively manage untagged documents, a parser that organizes scattered characters into coherent texts with their structures is necessary. Ideally, a PDF Parser should exhibit the following key features:

- **Document Structure Recognition**: It should adeptly divide pages into different types of content blocks like paragraphs, tables, and charts. This ensures that the divided text blocks are complete and independent semantic units.
- **Robustness in Complex Document Layout**: It should work well even for document pages with complex layouts, such as multi-column pages, border-less tables, and even tables with merged cells.

Currently, there are two main types of methods of PDF Parsing: rule based approaches and deep learning-based approaches. Among them, PyPDF, a widely-used rule-based parser, is a standard method in LangChain for PDF parsing. Conversely, our approach, ChatDOC PDF Parser (`https://pdfparser.io/`), is grounded in the deep learning models. Next, we illustrate the difference between them by introducing the methods and delving into some real-world cases.

2.2 Rule-based Method: PyPDF

We first introduce the parsing & chunking workflow based on PyPDF. First, PyPDF serializes characters in a PDF into a long sequence without document structure information. Then, this sequence undergoes segmentation into discrete chunks, utilizing some segmentation rule, such as the "`RecursiveCharacterTextSplitter`" function in LangChain. Specifically, this function divides the document based on a predefined list of separators, such as the newline character "\n". After this initial segmentation, adjacent chunks are merged only if the length of the combined chunks is not bigger than a predetermined limit of N characters. Hereafter, we use "PyPDF" to refer to the method of document parsing and chunking using PyPDF+`RecursiveCharacterTextSplitter`, provided there is no contextual ambiguity. The maximum length of a chunk is set to 300 tokens in the following. Next, we use a case to observe the inherent nature of PyPDF.



[Chunk 1]

Year ended March 31, 2021\n
China \n
commerce(1)\n
International \n
commerce\n
Local consumer \n
services(1) Cainiao Cloud\n
Digital \n
media and \n
entertainment\n
Innovation \n
initiatives and\n
others Unallocated(2) Consolidated\n
RMB RMB RMB RMB RMB RMB RMB RMB RMB\n
(in millions, except percentages)\n
Revenue 501,379 48,851 35,746 37,258 60,558 31,186 2,311 — 717,289\n
Income (Loss) from operations 197,232 (9,361) (29,197) (3,964) (12,479) (10,321) (7,802) (34,430) 89,678\n
Add: Share-based compensation\n

[Chunk 2]

expense 14,505 4,223 4,972 1,956 10,205 3,281 2,518 8,460 50,120\n
Add: Amortization of intangible assets 1,922 206 7,852 1,195 23 922 83 224 12,427\n
Add: Anti-monopoly Fine(3) — — — — — — — — 18,228 18,228\n
Adjusted EBITA 213,659 (4,932) (16,373) (813) (2,251) (6,118) (5,201) (7,518) 170,453\n
Adjusted EBITA margin 43% (10)% (46)% (2)% (4)% (20)% (225)% N/A 24%\n
(1) Beginning on October 1, 2022, we reclassified the results of our Instant Supermarket Delivery (全能超市) business, which was \n
previously reported under China commerce segment, to Local consumer services segment following the strategy refinement \n
of Instant Supermarket Delivery business to focus on building customer mindshare for grocery delivery services through Ele.me \n
platform. This reclassification conforms to the way that we manage and monitor segment performance. Comparative figures were \n
reclassified to conform to this presentation.

[Chunk 3]

(2) Unallocated expenses primarily relate to corporate administrative costs and other miscellaneous items that are not allocated to \n
individual segments. The goodwill impairment, and the equity-settled donation expense related to the allotment of shares to a \n
charitable trust, are presented as unallocated items in the segment information because our management does not consider these \n
as part of the segment operating performance measure.\n
(3) For a description of the relevant PRC Anti-monopoly investigation and administrative penalty decision, see "Business Overview — \n
Legal and Administrative Proceedings — PRC Anti-monopoly Investigation and Administrative Penalty Decision."\n
Non-GAAP Measures\n
We use adjusted EBITDA (including adjusted EBITDA \n
margin), adjusted EBITA (including adjusted EBITA \n
margin), non-GAAP net income, non-GAAP diluted \n
earnings per share/ADS and free cash flow, each \n
a non-GAAP financial measure, in evaluating our \n
operating results and for financial and operational \n
decision-making purposes.\n

[Chunk 4]

We believe that adjusted EBITDA, adjusted EBITA, \n
non-GAAP net income and non-GAAP diluted earnings \n
per share/ADS help identify underlying trends in our \n
business that could otherwise be distorted by the \n
effect of certain income or expenses that we include \n
in income from operations, net income and diluted \n
earnings per share/ADS. We believe that these \n
non-GAAP measures provide useful information \n
about our core operating results, enhance the overall \n
understanding of our past performance and future \n
prospects and allow for greater visibility with respect \n
to key metrics used by our management in its financial \n
and operational decision-making. We present three \n
different income measures, namely adjusted EBITDA, \n
adjusted EBITA and non-GAAP net income in order to \n
provide more information and greater transparency to \n
investors about our operating results.\n
We consider free cash flow to be a liquidity measure \n
that provides useful information to management \n
and investors about the amount of cash generated \n
by our business that can be used for strategic \n
corporate transactions, including investing in our new \n
business initiatives, making strategic investments and \n
acquisitions and strengthening our balance sheet.\n

[Chunk 5]

Adjusted EBITDA, adjusted EBITA, non-GAAP net \n
income, non-GAAP diluted earnings per share/ADS \n
and free cash flow should not be considered in \n
isolation or construed as an alternative to income \n
from operations, net income, diluted earnings per \n
share/ADS, cash flows or any other measure of \n
performance or as an indicator of our operating \n
performance. These non-GAAP financial measures \n
presented here do not have standardized meanings \n
prescribed by U.S. GAAP and may not be comparable \n
to similarly titled measures presented by other \n
companies. Other companies may calculate similarly \n
titled measures differently, limiting their usefulness as \n
comparative measures to our data.\n
112 Alibaba Group Holding Limited\n
Management Discussion and Analysis\n

Figure 4. Parsing and chunking results of PyPDF on Case 1 (original document: [1]). Zoom in to see the details.

Case 1 in Figure 4 is a page from a document that features a mix of a table and double-column text where their boundaries are difficult to distinguish. Rows in the middle of the table do not have horizontal lines, making it difficult to recognize the rows in the table. And paragraphs have both single-column layout (for notes below the table) and double-columns layout (for paragraphs in the lower part of the page).

The parsing and chunking result of PyPDF is shown in Figure 4. In the "3 Visualization" part, we can see that PyPDF correctly recognizes the one-column and two-column layout parts of the page. But there are three shortcomings of PyPDF:

1. It cannot recognize the boundary of paragraphs and tables. It wrongly splits the table into two parts and merges the second part with the subsequent paragraph as one chunk.

   PyPDF seems to be good at detecting the boundary of a paragraph, as it does not split one paragraph into multiple chunks. But it actually does not parse the boundary of a paragraph. In the "2 Chunking Result" part we can see that each visual text line in the page is parsed as a line ended with "\n" in the result, and there is no special format at the end of a paragraph. It chunks paragraphs correctly because we use a special separator ".\n" that regards a line ending with a period as likely to be the end of a paragraph. However, this heuristic may not hold in many cases.

2. It cannot recognize the structure within a table. In the "2 Chunking Result" part, in chunk1, the upper part of the table is represented as a sequence of short phrases, where a cell may be split

into multiple lines (e.g. the cell "China commerce(1)") and some adjacent cells may be arranged in one line (e.g. the third to the fifth cells in the second line, "services(1) Cainiao Cloud"). So, the structure of the table is completely destroyed. If this chunk is retrieved for RAG, LLM is unable to perceive any meaningful information from it. Similar situation for Chunk 2. Moreover, the headers of the table only exist in Chunk 1, so the lower part of the table in Chunk 2 becomes meaningless.

3. It cannot recognize the reading order of the content. The last line of Chunk 5, "Management Discussion and Analysis" is actually located at the top of the page, but is parsed as the last sentence in the result. This is because PyPDF parses the document by the storage order of the characters, instead of their reading order. This may cause chaotic results when faced with complex layouts.

The result on another case Case 2 features with a complex and cross-page table is shown in Figure 15 in the Appendix.

2.3 Deep Learning-based Method: ChatDOC PDF Parser

Next, we turn our attention to the method of deep learning-based parsing, exemplified by our ChatDOC PDF Parser. The ChatDOC PDF Parser (`https://pdfparser.io/`) has been trained on a corpus of over ten million document pages. Following the method in [2], it incorporates a sequence of sophisticated steps, including:

1. OCR for text positioning and recognition;
2. Physical document object detection;
3. Cross-column and cross-page trimming;
4. Reading order determination;
5. Table structure recognition;
6. Document logical structure recognition.

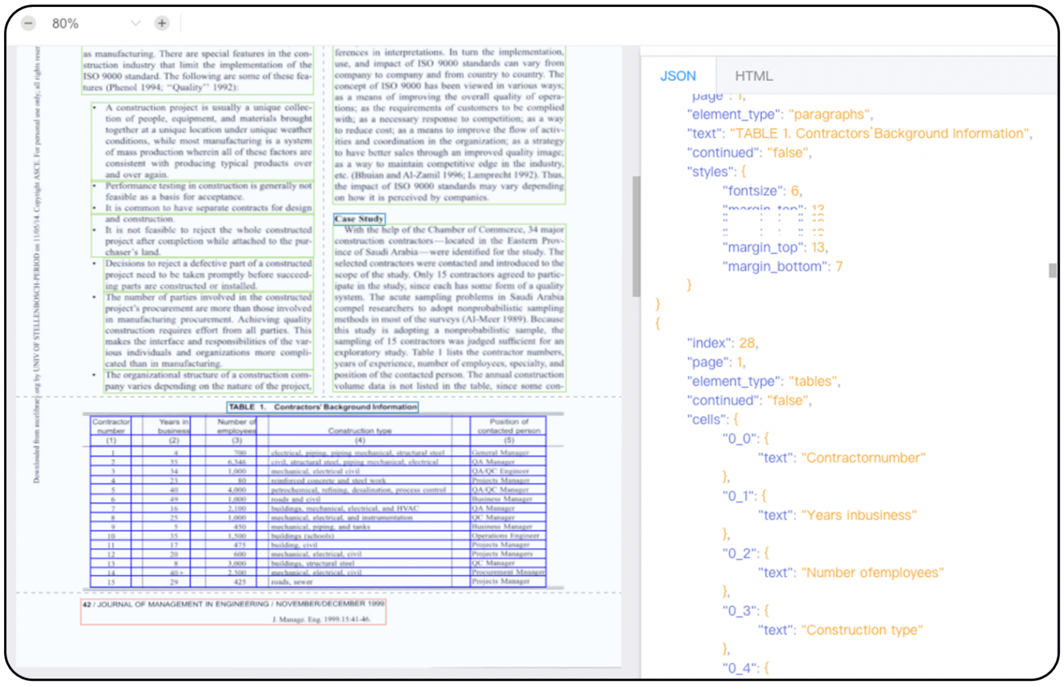
Readers might refer to [2] for the details of these steps. After parsing, we use the paragraphs and tables as basic blocks, and merge adjacent blocks until reaching the token limit to form a chunk.

ChatDOC PDF Parser is designed to consistently deliver parsing results in JSON or HTML formats, even for challenging PDF files. It parses a document into content blocks where each block refers to a table, paragraph, chart, or other type. For tables, it outputs the text in each table cell and also tells which cells are merged into a new one. Moreover, for documents with hierarchical headings, it outputs the hierarchical structure of the document. In summary, the parsed result is like a well-organized Word file. Figure 5 shows a scan-copy page and its parsing result. The left side displays the document and the recognized content blocks (with different colored rectangles). The right side shows the parsing result in JSON or HTML format. Readers might refer to [3] for the live demo of this parsing result.

Then, we check the result of ChatDOC PDF Parser on Case 1 in Figure 6. It successfully addresses the three shortcomings of PyPDF.

1. As shown in the "3 Visualization" part, it recognizes the mixed layout and correctly sets the whole table as a separate chunk. For paragraphs, as shown in chunk 2 in the "2 Chunking Result" part, text lines in the same paragraphs are merged together, making it easier to understand.
2. In the "2 Chunking Result" part, in Chunk 1, we can see the table is represented using the `markdown` format, which preserves the table's internal structure. Additionally, ChatDOC PDF Parser can recognize the merged cells inside a table. Since the `markdown` format cannot represent the merged cells, we put the whole text in the merged cell into each original cell in the `markdown` format. As you can see, in Chunk 1 the text "Year ended March 31, 2021" repeats 9 times, which stands for a merged cell with the original 9 ones.
3. Moreover, "Management Discussion and Analysis" and "112 Alibaba Group Holding Limited" is recognized as the page header and footer, and they are placed at the top and bottom of the parsing result which is consistent with reading order.

The result on another case of Case 2 featured with complex and cross-page table is shown in Figure 16 in the Appendix.



Downloaded from ascelibrary.org by UNIV OF STELLENBOSCH-PERIOD on 11/05/14. Copyright ASCE. For personal use only; all rights reser

as manufacturing. There are special features in the construction industry that limit the implementation of the ISO 9000 standard. The following are some of these features (Phenol 1994; "Quality" 1992):

- A construction project is usually a unique collection of people, equipment, and materials brought together at a unique location under unique weather conditions, while most manufacturing is a system of mass production wherein all of these factors are consistent with producing typical products over and over again.
- Performance testing in construction is generally not feasible as a basis for acceptance.
- It is common to have separate contracts for design and construction.
- It is not feasible to reject the whole constructed project after completion while attached to the purchaser's land.
- Decisions to reject a defective part of a constructed project need to be taken promptly before succeeding parts are constructed or installed.
- The number of parties involved in the constructed project's procurement are more than those involved in manufacturing procurement. Achieving quality construction requires effort from all parties. This makes the interface and responsibilities of the various individuals and organizations more complicated than in manufacturing.
- The organizational structure of a construction company varies depending on the nature of the project,

ferences in interpretations. In turn the implementation, use, and impact of ISO 9000 standards can vary from company to company and from country to country. The concept of ISO 9000 has been viewed in various ways; as a means of improving the overall quality of operations; as the requirements of customers to be complied with; as a necessary response to competition; as a way to reduce cost; as a means to improve the flow of activities and coordination in the organization; as a strategy to have better sales through an improved quality image; as a way to maintain competitive edge in the industry, etc. (Bhuian and Al-Zamil 1996; Lamprecht 1992). Thus, the impact of ISO 9000 standards may vary depending on how it is perceived by companies.

Case Study

With the help of the Chamber of Commerce, 34 major construction contractors—located in the Eastern Province of Saudi Arabia—were identified for the study. The selected contractors were contacted and introduced to the scope of the study. Only 15 contractors agreed to participate in the study, since each has some form of a quality system. The acute sampling problems in Saudi Arabia compel researchers to adopt nonprobabilistic sampling methods in most of the surveys (Al-Meer 1989). Because this study is adopting a nonprobabilistic sample, the sampling of 15 contractors was judged sufficient for an exploratory study. Table 1 lists the contractor numbers, years of experience, number of employees, specialty, and position of the contacted person. The annual construction volume data is not listed in the table, since some con-

TABLE 1. Contractors' Background Information

Contractor number	Years in business	Number of employees	Construction type	Position of contacted person
(1)	(2)	(3)	(4)	(5)
1	4	700	electrical, piping, piping mechanical, structural steel	General Manager
2	35	6,346	civil, structural steel, piping mechanical, electrical	QA Manager
3	34	1,000	mechanical, electrical civil	QA/QC Engineer
4	23	80	reinforced concrete and steel work	Projects Manager
5	40	4,000	petrochemical, refining, desalination, process control	QA/QC Manager
6	49	1,000	roads and civil	Business Manager
7	16	2,100	buildings, mechanical, electrical, and HVAC	QA Manager
8	25	1,000	mechanical, electrical, and instrumentation	QC Manager
9	5	450	mechanical, piping, and tanks	Business Manager
10	35	1,500	buildings (schools)	Operations Engineer
11	17	475	building, civil	Projects Manager
12	20	600	mechanical, electrical, civil	Projects Managers
13	8	3,000	buildings, structural steel	QC Manager
14	40+	2,500	mechanical, electrical, civil	Procurement Manager
15	29	425	roads, sewer	Projects Manager

42 / JOURNAL OF MANAGEMENT IN ENGINEERING / NOVEMBER/DECEMBER 1999

J. Manage. Eng. 1999.15:41-46.

```
        "page": 1,
        "element_type": "paragraphs",
        "text": "TABLE 1. Contractors’Background Information",
        "continued": "false",
        "styles": {
                "fontsize": 6,
                "margin_top": 13,
                "margin_bottom": 7
        }
}
{
        "index": 28,
        "page": 1,
        "element_type": "tables",
        "continued": "false",
        "cells": {
                "0_0": {
                        "text": "Contractornumber"
                },
                "0_1": {
                        "text": "Years inbusiness"
                },
                "0_2": {
                        "text": "Number ofemployees"
                },
                "0_3": {
                        "text": "Construction type"
                },
                "0_4": {
```

Figure 5. An example illustrating the results of the ChatDOC PDF Parser. Zoom in to see the details.

3 Experiments on the Impact of PDF Recognition on RAG

Back to the main topic of this paper, does the way a document is parsed and chunked affect the quality of answers provided by an RAG system? To answer this, we have carried out a systematic experiment to assess the impacts.

3.1 Quantitative Evaluation of RAG Answer Quality

3.1.1 Settings

We compared two RAG systems as listed in Table 1:

- ChatDOC: uses ChatDOC PDF Parser to parse the document and leverage the structure information for chunking.
- Baseline: uses PyPDF to parse the document and use `RecursiveCharacterTextSplitter` function for chunking.

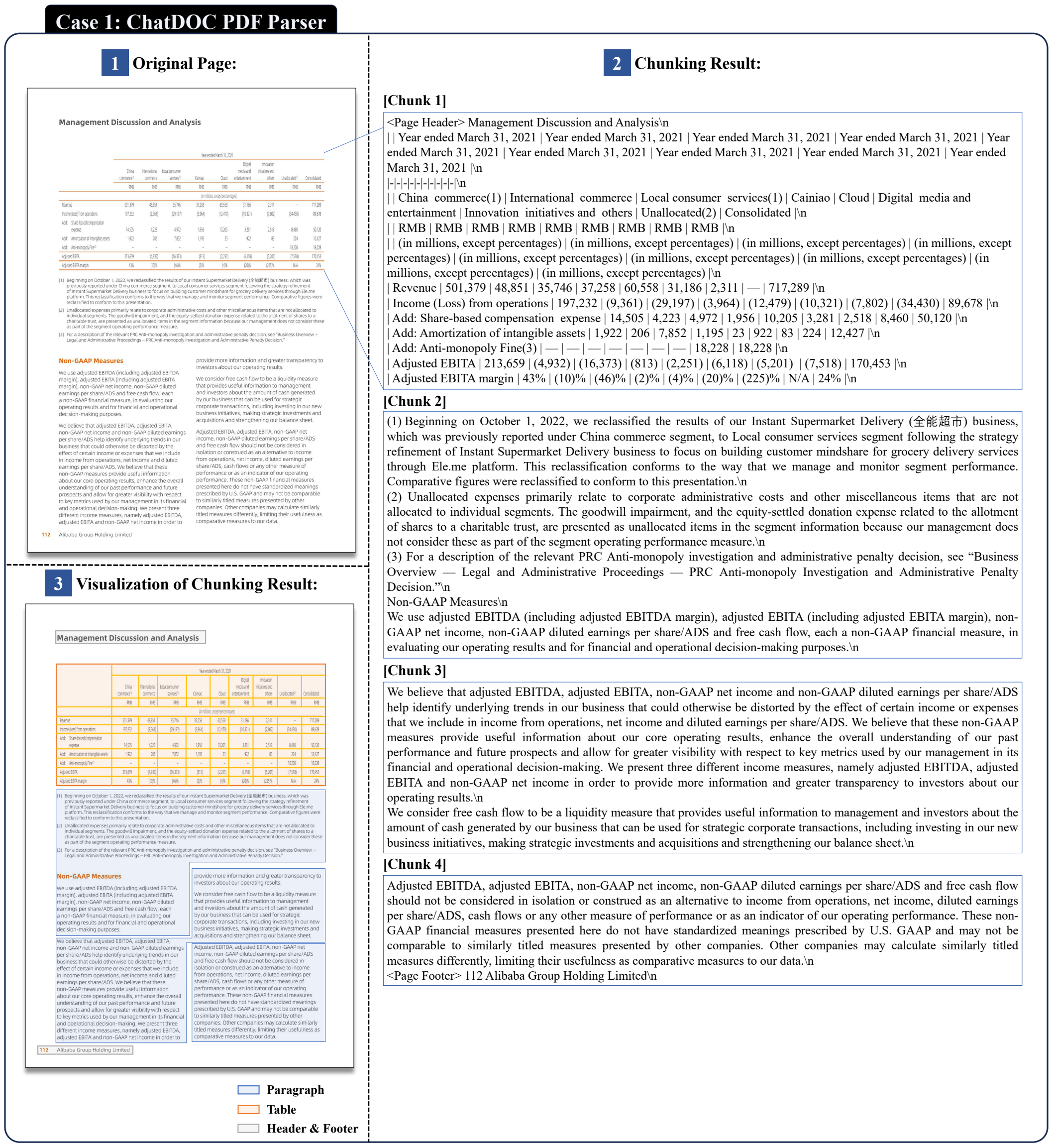
Other components, like embedding, retrieval, and QA, are the same for both systems.

3.1.2 Data Preparation

For our experiment, we assembled a dataset that closely mirrors real-world conditions, comprising 188 documents from various fields. Specifically, This collection includes 100 academic papers, 28 financial reports, and 60 documents from other categories such as textbooks, courseware, and legislative materials.

We then gathered 800 manually generated questions via crowd-sourcing. After careful screening, we removed low-quality questions and got 302 questions for evaluation. These questions were divided into two categories (as shown in Table 2):

- **Extractive questions** are those that can be answered with direct excerpts from the documents. Usually, they require pinpoint answers because they seek specific information. We found when



[Chunk 1]

```
<Page Header> Management Discussion and Analysis\n
| | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 | Year ended March 31, 2021 |\n
|-|-|-|-|-|-|-|-|-|-|\n
| | China commerce(1) | International commerce | Local consumer services(1) | Cainiao | Cloud | Digital media and entertainment | Innovation initiatives and others | Unallocated(2) | Consolidated |\n
| | RMB | RMB | RMB | RMB | RMB | RMB | RMB | RMB | RMB |\n
| | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) | (in millions, except percentages) |\n
| Revenue | 501,379 | 48,851 | 35,746 | 37,258 | 60,558 | 31,186 | 2,311 | — | 717,289 |\n
| Income (Loss) from operations | 197,232 | (9,361) | (29,197) | (3,964) | (12,479) | (10,321) | (7,802) | (34,430) | 89,678 |\n
| Add: Share-based compensation expense | 14,505 | 4,223 | 4,972 | 1,956 | 10,205 | 3,281 | 2,518 | 8,460 | 50,120 |\n
| Add: Amortization of intangible assets | 1,922 | 206 | 7,852 | 1,195 | 23 | 922 | 83 | 224 | 12,427 |\n
| Add: Anti-monopoly Fine(3) | — | — | — | — | — | — | — | 18,228 | 18,228 |\n
| Adjusted EBITA | 213,659 | (4,932) | (16,373) | (813) | (2,251) | (6,118) | (5,201) | (7,518) | 170,453 |\n
| Adjusted EBITA margin | 43% | (10)% | (46)% | (2)% | (4)% | (20)% | (225)% | N/A | 24% |\n
```

[Chunk 2]

(1) Beginning on October 1, 2022, we reclassified the results of our Instant Supermarket Delivery (全能超市) business, which was previously reported under China commerce segment, to Local consumer services segment following the strategy refinement of Instant Supermarket Delivery business to focus on building customer mindshare for grocery delivery services through Ele.me platform. This reclassification conforms to the way that we manage and monitor segment performance. Comparative figures were reclassified to conform to this presentation.\n
(2) Unallocated expenses primarily relate to corporate administrative costs and other miscellaneous items that are not allocated to individual segments. The goodwill impairment, and the equity-settled donation expense related to the allotment of shares to a charitable trust, are presented as unallocated items in the segment information because our management does not consider these as part of the segment operating performance measure.\n
(3) For a description of the relevant PRC Anti-monopoly investigation and administrative penalty decision, see "Business Overview — Legal and Administrative Proceedings — PRC Anti-monopoly Investigation and Administrative Penalty Decision."\n
Non-GAAP Measures\n
We use adjusted EBITDA (including adjusted EBITDA margin), adjusted EBITA (including adjusted EBITA margin), non-GAAP net income, non-GAAP diluted earnings per share/ADS and free cash flow, each a non-GAAP financial measure, in evaluating our operating results and for financial and operational decision-making purposes.\n

[Chunk 3]

We believe that adjusted EBITDA, adjusted EBITA, non-GAAP net income and non-GAAP diluted earnings per share/ADS help identify underlying trends in our business that could otherwise be distorted by the effect of certain income or expenses that we include in income from operations, net income and diluted earnings per share/ADS. We believe that these non-GAAP measures provide useful information about our core operating results, enhance the overall understanding of our past performance and future prospects and allow for greater visibility with respect to key metrics used by our management in its financial and operational decision-making. We present three different income measures, namely adjusted EBITDA, adjusted EBITA and non-GAAP net income in order to provide more information and greater transparency to investors about our operating results.\n
We consider free cash flow to be a liquidity measure that provides useful information to management and investors about the amount of cash generated by our business that can be used for strategic corporate transactions, including investing in our new business initiatives, making strategic investments and acquisitions and strengthening our balance sheet.\n

[Chunk 4]

Adjusted EBITDA, adjusted EBITA, non-GAAP net income, non-GAAP diluted earnings per share/ADS and free cash flow should not be considered in isolation or construed as an alternative to income from operations, net income, diluted earnings per share/ADS, cash flows or any other measure of performance or as an indicator of our operating performance. These non-GAAP financial measures presented here do not have standardized meanings prescribed by U.S. GAAP and may not be comparable to similarly titled measures presented by other companies. Other companies may calculate similarly titled measures differently, limiting their usefulness as comparative measures to our data.\n
<Page Footer> 112 Alibaba Group Holding Limited\n

Figure 6. Parsing and chunking results of ChatDOC PDF Parser on Case 1 (original document: [4]). Zoom in to see the details.

using LLM for evaluation, it may fail to distinguish the subtle but important differences between answers, so we relied on human assessment. We used a 0-10 scale to rate the results. An annotator is given the retrieved content and answer of both methods and rates the two methods at the same time. We show the retrieved content because it usually cannot evaluate the answer without document content, and show two methods together to promote detailed comparison, especially on partially correct results.

- **Comprehensive analysis questions** necessitate synthesizing information from multiple sources and aspects and making a summary. Since the answer is lengthy and requires a comprehensive understanding of the given document contents, we found it difficult and time-consuming for humans to evaluate. Hence, we used GPT-4 to evaluate the answer quality, scoring from 1-10. We also rate the result of two methods in one request. But we only give the retrieved content without an answer because the answer is lengthy (thus costly) compared with extractive questions and a better retrieved content can imply a better answer (since the used LLM is the same). A pair of results of two methods is scored 4 times to avoid bias [5], and their average value is used. Specifically, for a pair of content (A, B) to be compared for the same question, we feed both