ChatDOC：RAG

Revolutionizing Retrieval–Augmented Generation with Enhanced PDF Structure Recognition

本文探讨了基于检索增强生成（RAG）的专业知识问答系统是否已经接近完美，并发现当前主要方法依赖于高质量文本语料库的前提条件。然而，由于专业文档主要以PDF格式存储，低准确率的PDF解析显著影响专业知识问答的有效性。作者进行了实证RAG实验，使用了一个具有全景和精准PDF解析器的ChatDOC系统，在数百个来自相应真实世界专业文档的问题中检索出更准确、完整的段落，从而更好地回答问题。实验结果表明，ChatDOC在近47%的问题上优于基线，有38%的情况与基线打平，只有15%的情况下表现不如基线。这表明我们可以通过增强PDF结构识别来革命化RAG。

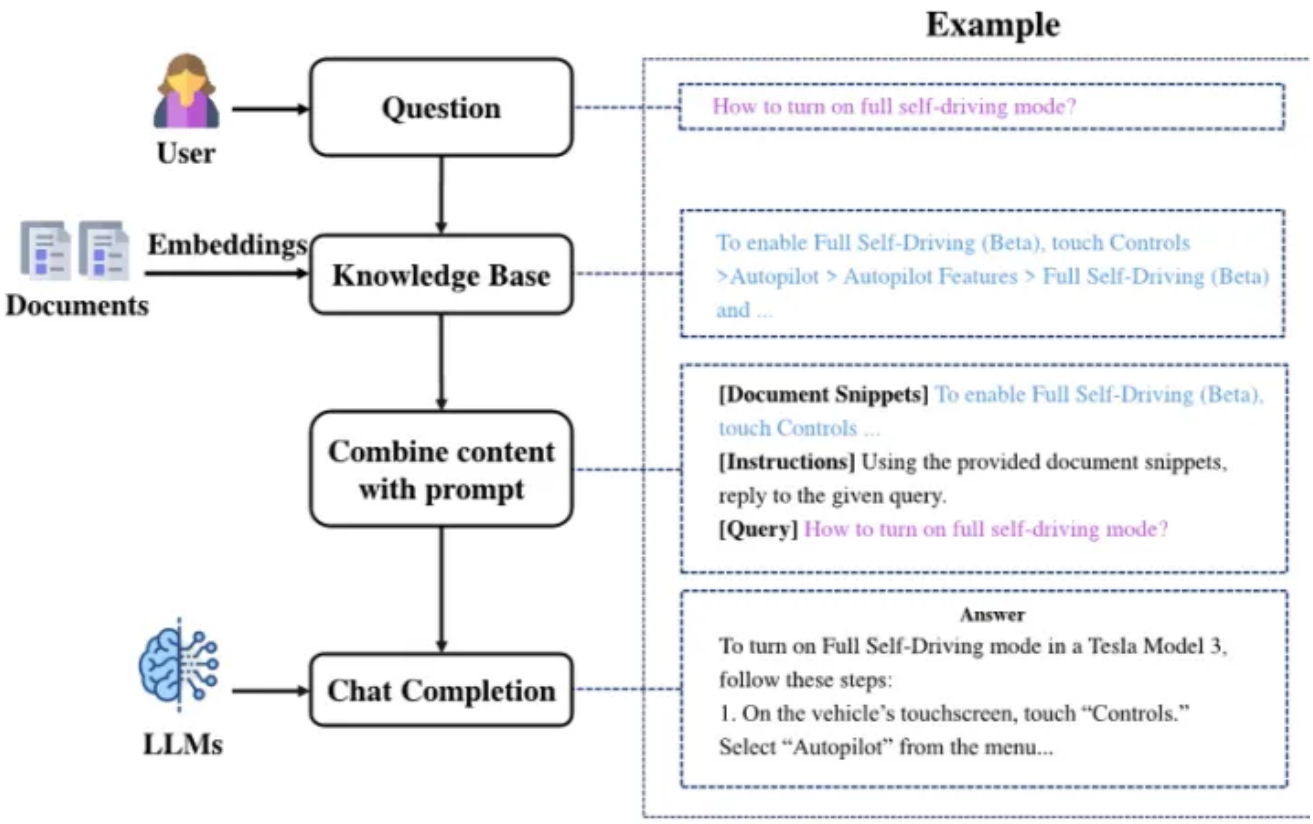


Figure 1. The workflow of Retrieval-Augmented Generation (RAG).

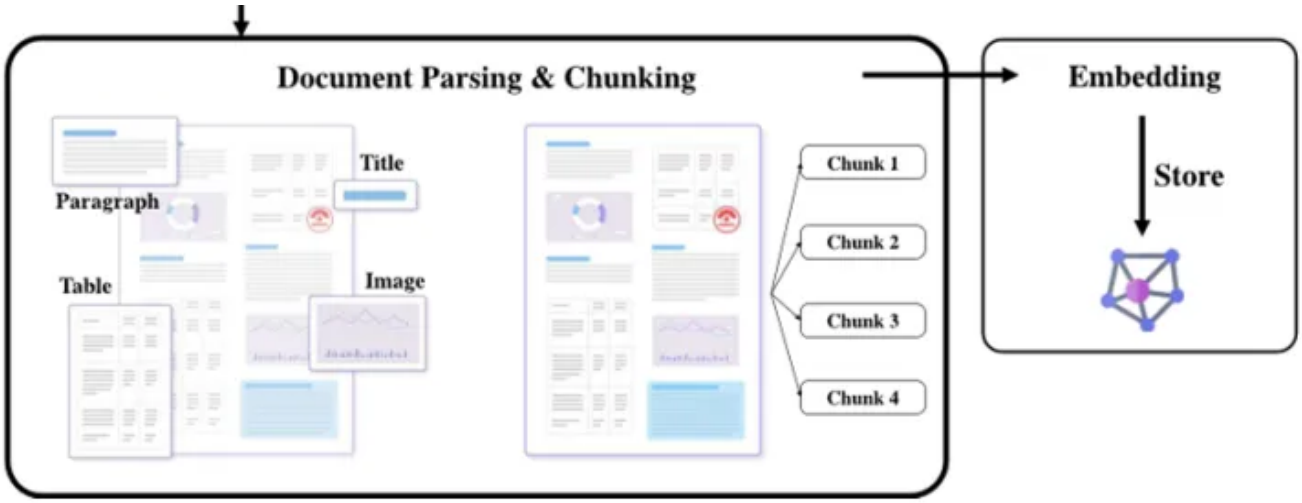


Figure 2. The process of converting PDFs into retrievable contents.

论文方法

方法描述

该论文介绍了两种PDF解析方法：基于规则的方法（PyPDF）和基于深度学习的方法（ChatDOC PDF Parser）。其中，PyPDF是一种广泛使用的基于规则的解析器，而ChatDOC PDF Parser则采用了深度学习模型来处理未标记文档。

方法改进

与PyPDF相比，ChatDOC PDF Parser具有以下优势：

1. 能够正确识别段落和表格之间的边界。
2. 能够正确识别表格内部结构，并使用Markdown格式保留表格的内部结构。
3. 能够正确识别文本的阅读顺序，避免因复杂布局而导致的结果混乱。

解决的问题

由于计算机只能理解二进制代码，无法感知信息的结构，因此需要将散乱的字符组织成有意义的文本块，并确定其结构。为此，需要一种能够有效地管理未标记文档的解析器。ChatDOC PDF Parser通过一系列复杂的步骤，包括OCR、物理对象检测、跨列和跨页修剪、阅读顺序确定、表格结构识别和文档逻辑结构识别等，能够准确地解析PDF文件并将其转换为JSON或HTML格式的内容块。这种方法能够更好地处理复杂布局和合并单元格等问题，从而提高了PDF解析的准确性。

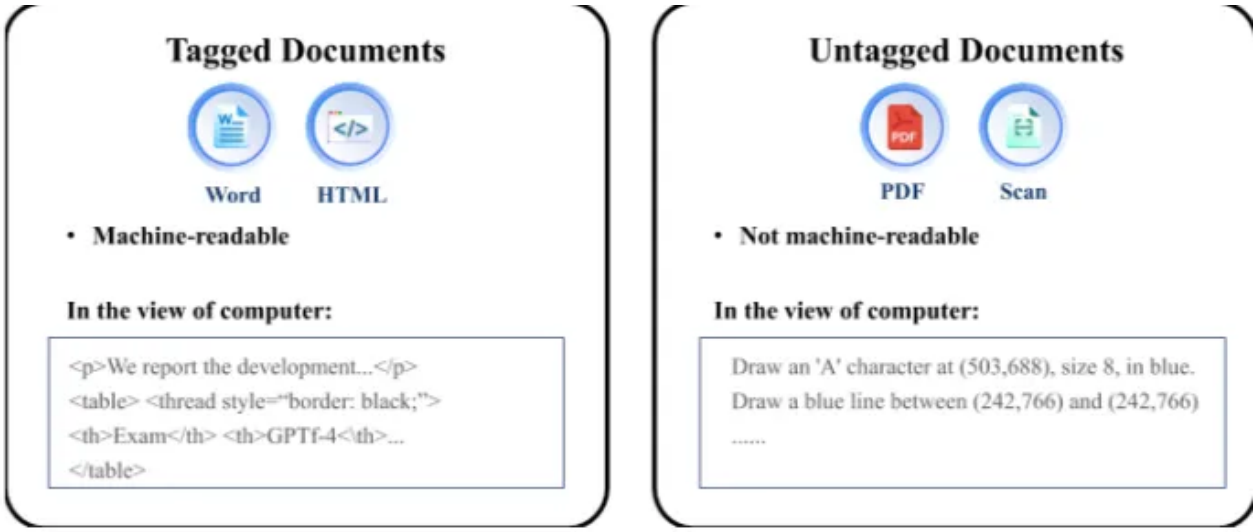


Figure 3. Two types of documents in the view of computers.

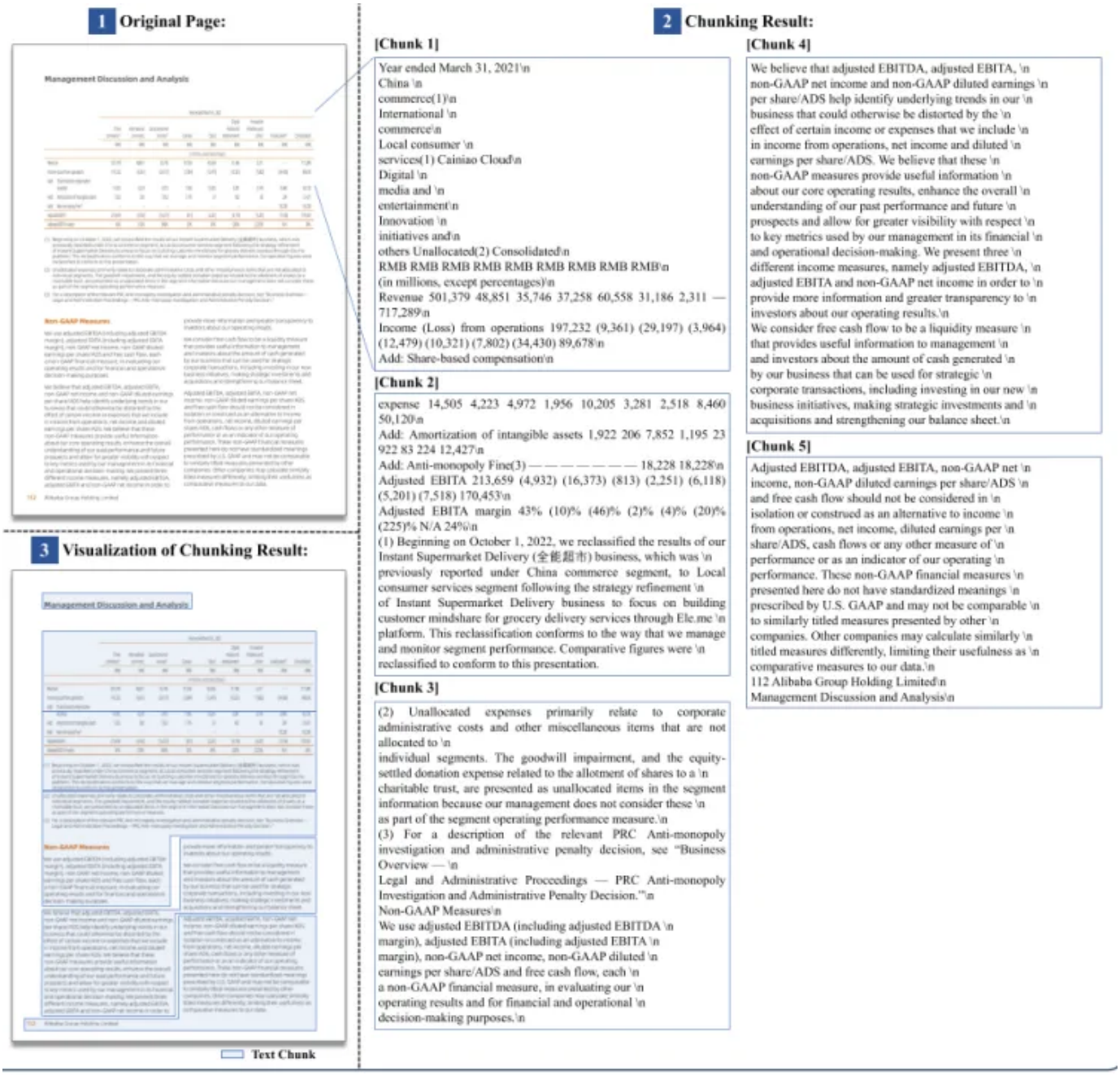


Figure 4. Parsing and chunking results of PyPDF on Case 1 (original document: [1]). Zoom in to see the details.

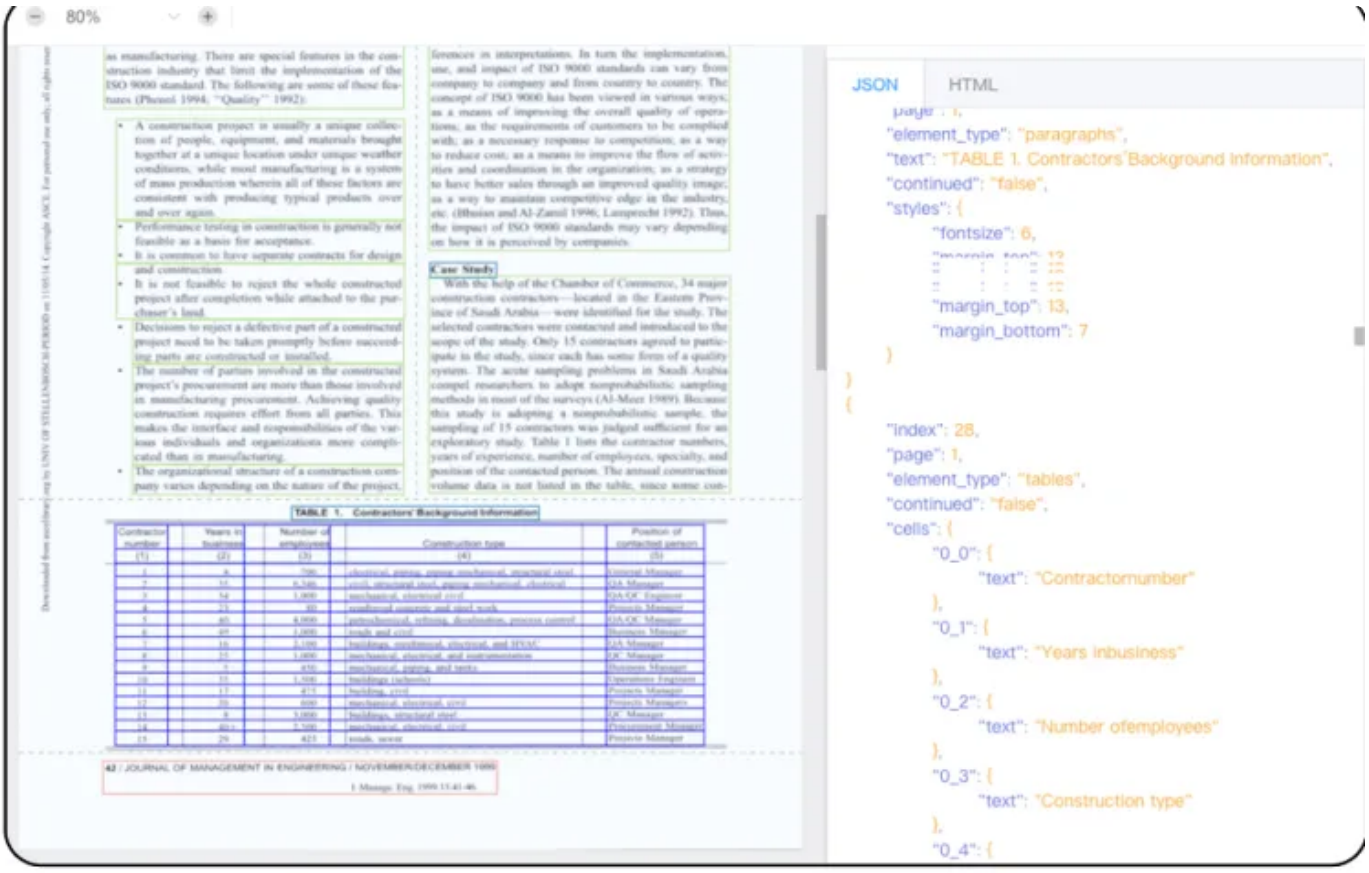


Figure 5. An example illustrating the results of the ChatDOC PDF Parser. Zoom in to see the details.

1 Original Page:

2 Chunking Result:

[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

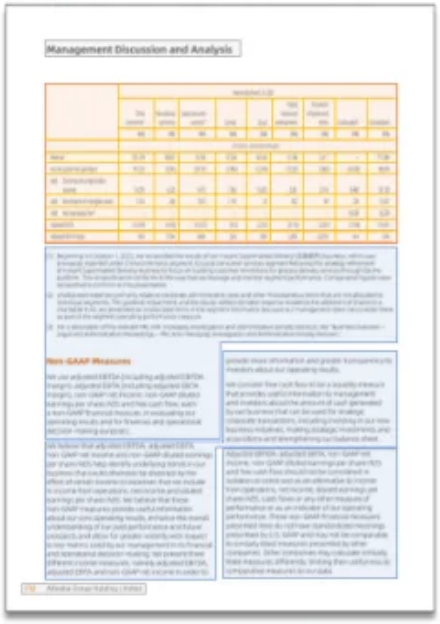
[Chunk 3]

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

3 Visualization of Chunking Result:



Paragraph
Table
Header & Footer

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

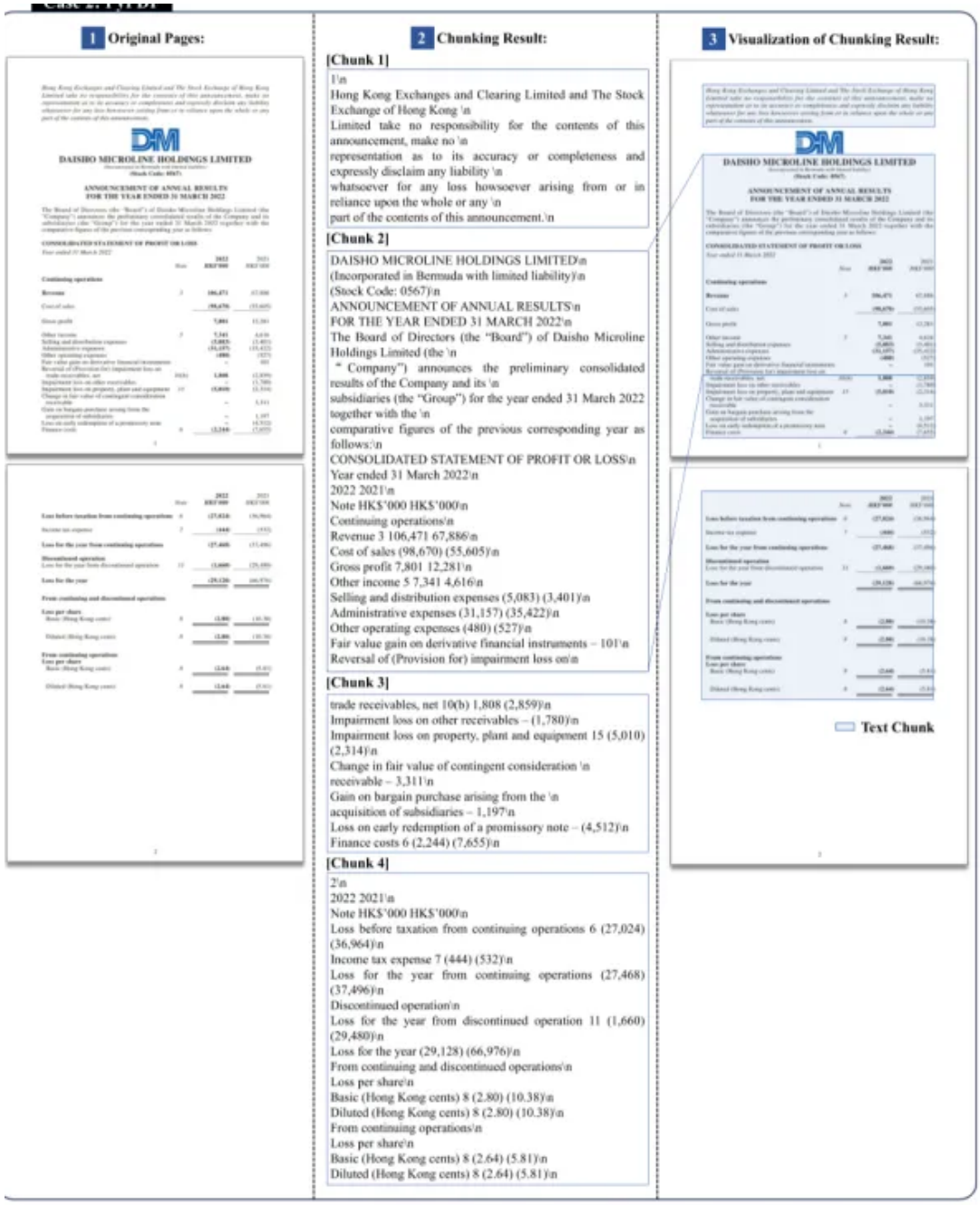


Figure 15. Parsing and chunking results of PyPDF on Case 2 (original document: [4]). Zoom in to see the details.

论文实验

本文主要介绍了针对RAG系统中文档解析和分块对答案质量的影响所做的实验，并通过一系列的对比实验来验证ChatDOC系统的优越性。

在实验中，作者比较了两个RAG系统，分别是使用ChatDOC PDF Parser进行文档解析并利用结构信息进行分块的ChatDOC系统，以及使用PyPDF进行文档解析并使用RecursiveCharacterTextSplitter函数进行分块的Baseline系统。其他组件如嵌入、检索和QA部分则保持一致。

实验分为两部分：

1. 对于提取式问题（extractive questions），作者手动收集了800个问题并通过众包筛选出302个高质量的问题用于评估。这些问题被分成两类：一类是直接从文档中提取答案的提取式问题，另一类需要综合多个来源和方面信息做出总结的分析式问题。对于提取式问题，作者使用人类评分来评估答案的质量，使用0–10分制进行打分。而对于分析式问题，则使用GPT–4来评估答案的质量，得分为1–10分。最终结果表明，ChatDOC系统在大多数情况下表现优于Baseline系统。
2. 对于案例研究（case studies），作者展示了几个具体例子以展示ChatDOC系统的优越性。这些例子包括：在一个关于特斯拉用户手册查询的例子中，ChatDOC系统能够更好地识别表格结构并提供更准确的答案；在一个关于论文的研究问题中，ChatDOC系统能够全面地检索到整个表格并准确回答问题。

综上所述，本文通过对不同类型的实验进行了详细的对比分析，证明了ChatDOC系统相对于Baseline系统具有更好的性能和优势。

Steps ↓	ChatDOC (PDFlux-LLM)	Baseline (PyPDF-LLM)
PDF Parsing	**PDFlux** (Deep Learning-based)	**PyPDF** (Rule-based, default method in LangChain)
Chunking	≈300 tokens per chunk + chunking via paragraphs, tables etc.	≈300 tokens per chunk + separator
Embedding	text-embedding-ada-002	
Retrieval	≤3000 tokens	
QA	GPT3.5-Turbo	

Table 1. Settings of two RAG systems: ChatDOC and Baseline.

	Extractive Questions	Comprehensive Analysis Questions
Number	86	216
Question Examples	*1. Locate the content of section ten, what is the merged operating cost in the income statement?* *2. What is the specific content of table 1.* *3. Extract financial data and profit forecast tables.* *4. Find the long-term loan table.*	*1. Summarize and analyze the profit forecast and valuation in the research report.* *2. Fully report the research approach of this text.* *3. Analyze the long-term debt-paying ability based on this report.* *4. How is the feasibility analysis done in this article?* *5. Give a simple example to explain the encoding steps and algorithm in the paper.*
Evaluation	Human Evaluation	GPT 4 evaluation

Table 2. The questions in the dataset are categorized into extractive questions and comprehensive analysis questions.

	Total	ChatDOC wins	Tie	Baseline wins
Extractive Questions	86	**42 (49%)**	36 (42%)	8 (9%)
Comprehensive Questions	216	**101 (47%)**	79 (37%)	36 (17%)
Summary	302	**143 (47%)**	115 (38%)	44 (15%)

Table 3. The comparison result between ChatDOC and Baseline.

论文总结

文章优点

- 论文提出了一种基于大型语言模型（LLM）和增强型PDF结构识别框架的文本检索系统ChatDOC。
- ChatDOC能够有效地处理表格，并且在多个文件中进行多轮对话，支持多种文件类型。
- 论文通过实验证明了ChatDOC相对于其他PDF解析器具有更高的可靠性和准确性。

方法创新点

- 提出了一种基于深度学习的PDF解析方法，可以有效地提取和整合文档中的结构化信息。
- 使用嵌入模型将文本块转换为实值向量，并将其存储在数据库中，以提高系统的效率和准确性。
- 在ChatDOC中应用了该PDF解析框架，使其成为一款高效的AI文件阅读助手。

未来展望

- 将会比较更多基于深度学习的文档解析方法，以更全面地了解RAG质量与文档解析质量之间的关系。
- 进一步优化PDF解析框架，提高其准确性和可靠性，以进一步提升ChatDOC的表现。