

ChatDOC: RAG

[论文方法](#)

[方法描述](#)

[方法改进](#)

[解决的问题](#)

[论文实验](#)

[论文总结](#)

[文章优点](#)

[方法创新点](#)

[未来展望](#)

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。

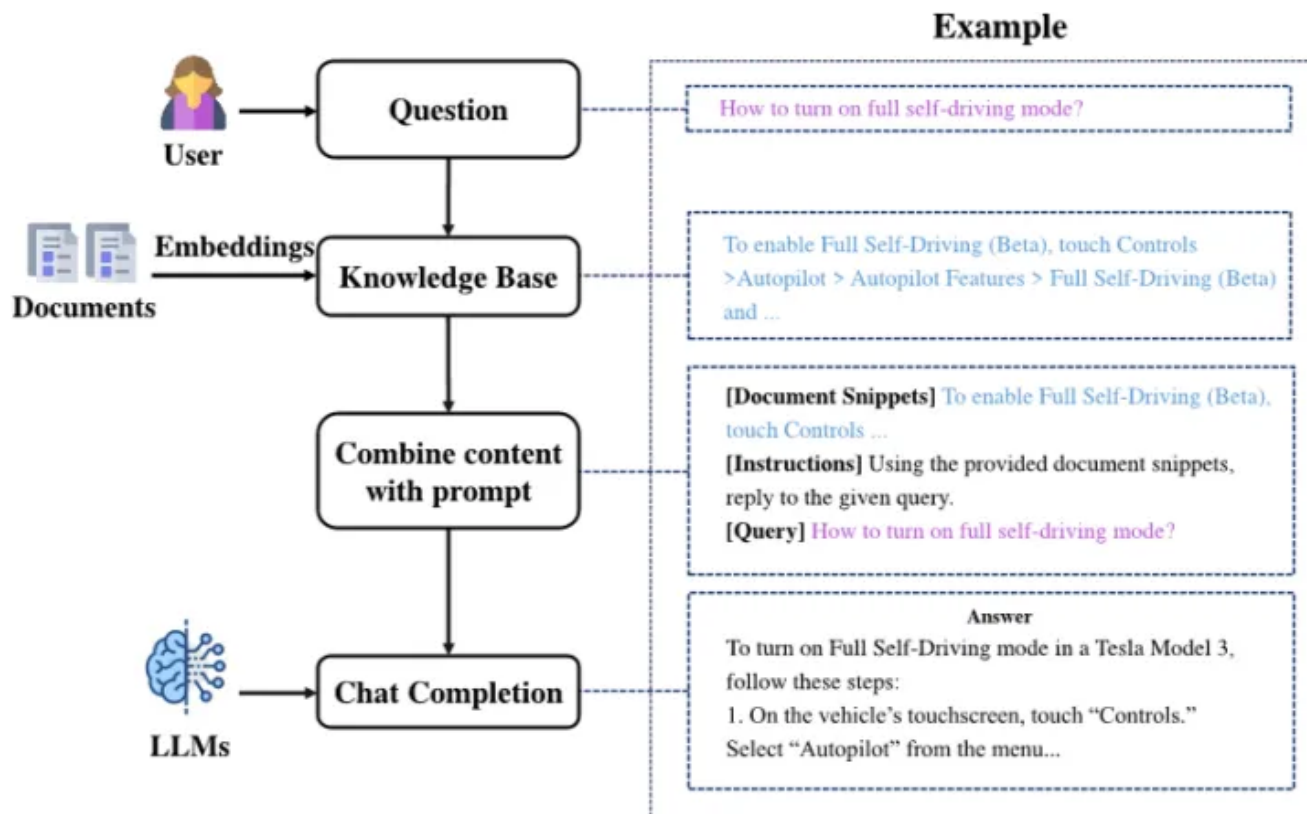


Figure 1. The workflow of Retrieval-Augmented Generation (RAG).

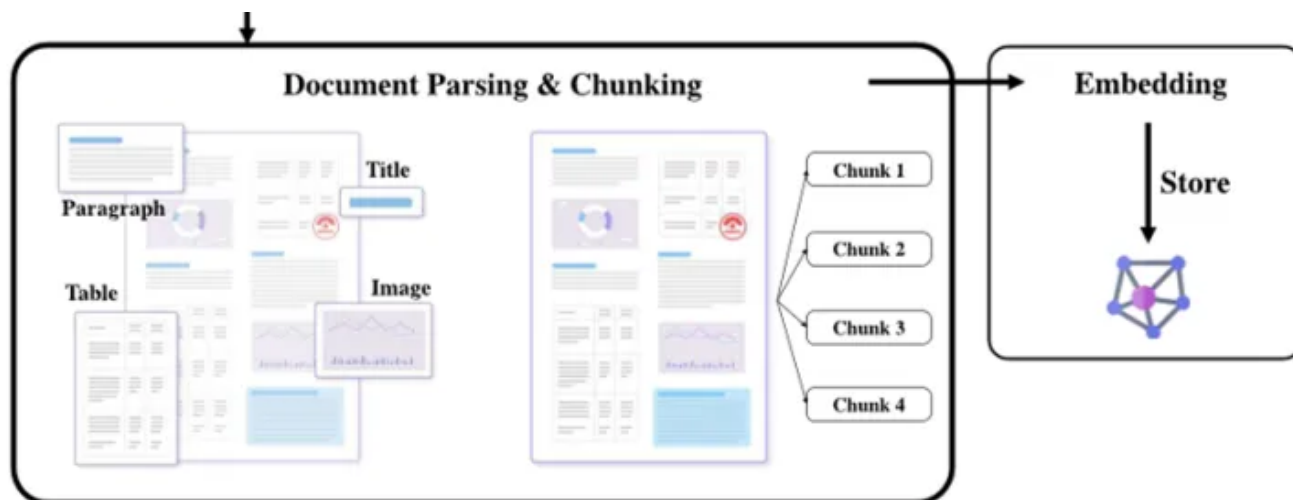


Figure 2. The process of converting PDFs into retrievable contents.

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。

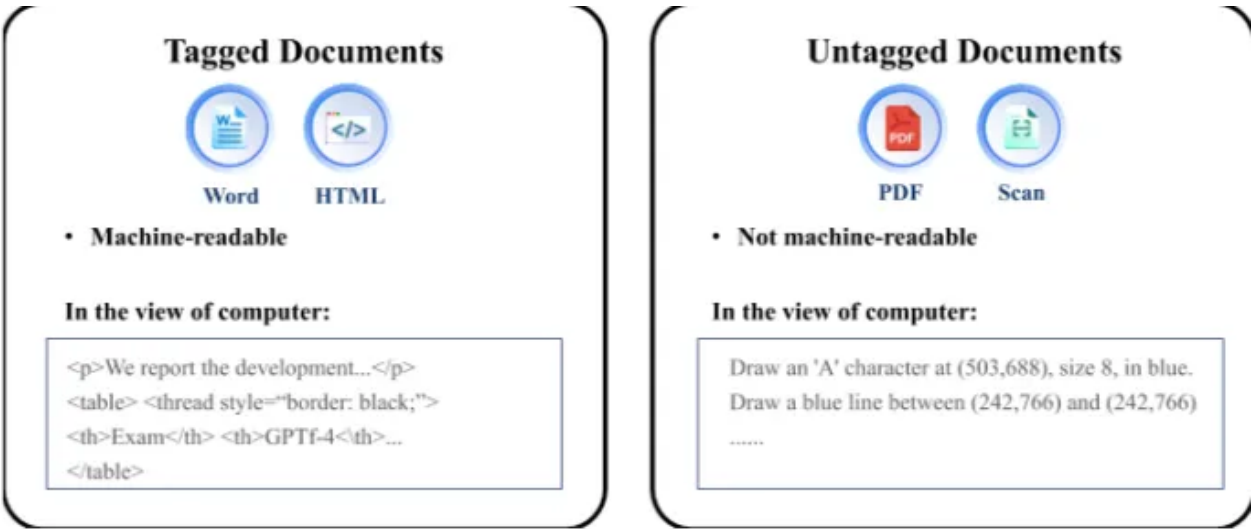


Figure 3. Two types of documents in the view of computers.

2 Chunking Result:

[Chunk 4]

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

[Chunk 5]

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.

112 Alibaba Group Holding Limited
Management Discussion and Analysis

[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

[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

Text Chunk

4

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.

ference 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. (Husain and Al-Zaidi 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 (1)	Years in business (2)	Number of employees (3)	Construction type (4)	Position of contacted person (5)
1	8	750	electrical, piping, piping, mechanical, structural steel	General Manager
2	12	5,500	civil, structural steel, piping, mechanical, electrical	QA Manager
3	24	1,000	mechanical, electrical, civil	QA/CV Engineer
4	25	30	structural concrete and steel work	Projects Manager
5	40	4,000	petrochemical, refining, distribution, process control	QA/CV Manager
6	48	3,400	roads and civil	Operations Manager
7	16	3,100	buildings, structural, electrical, and HVAC	QA Manager
8	22	1,000	mechanical, electrical, and instrumentation	QA Manager
9	8	450	mechanical, piping, and tanks	Business Manager
10	10	1,000	buildings, industrial	Operations Engineer
11	17	475	buildings, civil	Projects Manager
12	20	600	mechanical, electrical, civil	Projects Manager
13	8	3,000	buildings, structural steel	QA Manager
14	20	2,500	mechanical, electrical, civil	Construction Manager
15	26	400	roads, marine	Projects Manager

42 / JOURNAL OF MANAGEMENT IN ENGINEERING / NOVEMBER/DECEMBER 1999
J. Manage. Eng. 1999.15:41-46.

JSON HTML

```

page : 1,
"element_type": "paragraphs",
"text": "TABLE 1. Contractors' Background Information",
"continued": "false",
"styles": {
  "font_size": 6,
  "margin_top": 12,
  "margin_bottom": 12,
  "margin_top": 13,
  "margin_bottom": 7
}
}
{
"index": 26,
"page": 1,
"element_type": "tables",
"continued": "false",
"cells": {
  "0_0": {
    "text": "Contractor number"
  },
  "0_1": {
    "text": "Years in business"
  },
  "0_2": {
    "text": "Number of employees"
  },
  "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.

1 Original Page:

Management Discussion and Analysis

Revenue by Segment

	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1989	1988	1987	1986	1985	1984	1983	1982	1981	1980	1979	1978	1977	1976	1975	1974	1973	1972	1971	1970	1969	1968	1967	1966	1965	1964	1963	1962	1961	1960	1959	1958	1957	1956	1955	1954	1953	1952	1951	1950	1949	1948	1947	1946	1945	1944	1943	1942	1941	1940	1939	1938	1937	1936	1935	1934	1933	1932	1931	1930	1929	1928	1927	1926	1925	1924	1923	1922	1921	1920	1919	1918	1917	1916	1915	1914	1913	1912	1911	1910	1909	1908	1907	1906	1905	1904	1903	1902	1901	1900	1899	1898	1897	1896	1895	1894	1893	1892	1891	1890	1889	1888	1887	1886	1885	1884	1883	1882	1881	1880	1879	1878	1877	1876	1875	1874	1873	1872	1871	1870	1869	1868	1867	1866	1865	1864	1863	1862	1861	1860	1859	1858	1857	1856	1855	1854	1853	1852	1851	1850	1849	1848	1847	1846	1845	1844	1843	1842	1841	1840	1839	1838	1837	1836	1835	1834	1833	1832	1831	1830	1829	1828	1827	1826	1825	1824	1823	1822	1821	1820	1819	1818	1817	1816	1815	1814	1813	1812	1811	1810	1809	1808	1807	1806	1805	1804	1803	1802	1801	1800	1799	1798	1797	1796	1795	1794	1793	1792	1791	1790	1789	1788	1787	1786	1785	1784	1783	1782	1781	1780	1779	1778	1777	1776	1775	1774	1773	1772	1771	1770	1769	1768	1767	1766	1765	1764	1763	1762	1761	1760	1759	1758	1757	1756	1755	1754	1753	1752	1751	1750	1749	1748	1747	1746	1745	1744	1743	1742	1741	1740	1739	1738	1737	1736	1735	1734	1733	1732	1731	1730	1729	1728	1727	1726	1725	1724	1723	1722	1721	1720	1719	1718	1717	1716	1715	1714	1713	1712	1711	1710	1709	1708	1707	1706	1705	1704	1703	1702	1701	1700	1699	1698	1697	1696	1695	1694	1693	1692	1691	1690	1689	1688	1687	1686	1685	1684	1683	1682	1681	1680	1679	1678	1677	1676	1675	1674	1673	1672	1671	1670	1669	1668	1667	1666	1665	1664	1663	1662	1661	1660	1659	1658	1657	1656	1655	1654	1653	1652	1651	1650	1649	1648	1647	1646	1645	1644	1643	1642	1641	1640	1639	1638	1637	1636	1635	1634	1633	1632	1631	1630	1629	1628	1627	1626	1625	1624	1623	1622	1621	1620	1619	1618	1617	1616	1615	1614	1613	1612	1611	1610	1609	1608	1607	1606	1605	1604	1603	1602	1601	1600	1599	1598	1597	1596	1595	1594	1593	1592	1591	1590	1589	1588	1587	1586	1585	1584	1583	1582	1581	1580	1579	1578	1577	1576	1575	1574	1573	1572	1571	1570	1569	1568	1567	1566	1565	1564	1563	1562	1561	1560	1559	1558	1557	1556	1555	1554	1553	1552	1551	1550	1549	1548	1547	1546	1545	1544	1543	1542	1541	1540	1539	1538	1537	1536	1535	1534	1533	1532	1531	1530	1529	1528	1527	1526	1525	1524	1523	1522	1521	1520	1519	1518	1517	1516	1515	1514	1513	1512	1511	1510	1509	1508	1507	1506	1505	1504	1503	1502	1501	1500	1499	1498	1497	1496	1495	1494	1493	1492	1491	1490	1489	1488	1487	1486	1485	1484	1483	1482	1481	1480	1479	1478	1477	1476	1475	1474	1473	1472	1471	1470	1469	1468	1467	1466	1465	1464	1463	1462	1461	1460	1459	1458	1457	1456	1455	1454	1453	1452	1451	1450	1449	1448	1447	1446	1445	1444	1443	1442	1441	1440	1439	1438	1437	1436	1435	1434	1433	1432	1431	1430	1429	1428	1427	1426	1425	1424	1423	1422	1421	1420	1419	1418	1417	1416	1415	1414	1413	1412	1411	1410	1409	1408	1407	1406	1405	1404	1403	1402	1401	1400	1399	1398	1397	1396	1395	1394	1393	1392	1391	1390	1389	1388	1387	1386	1385	1384	1383	1382	1381	1380	1379	1378	1377	1376	1375	1374	1373	1372	1371	1370	1369	1368	1367	1366	1365	1364	1363	1362	1361	1360	1359	1358	1357	1356	1355	1354	1353	1352	1351	1350	1349	1348	1347	1346	1345	1344	1343	1342	1341	1340	1339	1338	1337	1336	1335	1334	1333	1332	1331	1330	1329	1328	1327	1326	1325	1324	1323	1322	1321	1320	1319	1318	1317	1316	1315	1314	1313	1312	1311	1310	1309	1308	1307	1306	1305	1304	1303	1302	1301	1300	1299	1298	1297	1296	1295	1294	1293	1292	1291	1290	1289	1288	1287	1286	1285	1284	1283	1282	1281	1280	1279	1278	1277	1276	1275	1274	1273	1272	1271	1270	1269	1268	1267	1266	1265	1264	1263	1262	1261	1260	1259	1258	1257	1256	1255	1254	1253	1252	1251	1250	1249	1248	1247	1246	1245	1244	1243	1242	1241	1240	1239	1238	1237	1236	1235	1234	1233	1232	1231	1230	1229	1228	1227	1226	1225	1224	1223	1222	1221	1220	1219	1218	1217	1216	1215	1214	1213	1212	1211	1210	1209	1208	1207	1206	1205	1204	1203	1202	1201	1200	1199	1198	1197	1196	1195	1194	1193	1192	1191	1190	1189	1188	1187	1186	1185	1184	1183	1182	1181	1180	1179	1178	1177	1176	1175	1174	1173	1172	1171	1170	1169	1168	1167	1166	1165	1164	1163	1162	1161	1160	1159	1158	1157	1156	1155	1154	1153	1152	1151	1150	1149	1148	1147	1146	1145	1144	1143	1142	1141	1140	1139	1138	1137	1136	1135	1134	1133	1132	1131	1130	1129	1128	1127	1126	1125	1124	1123	1122	1121	1120	1119	1118	1117	1116	1115	1114	1113	1112	1111	1110	1109	1108	1107	1106	1105	1104	1103	1102	1101	1100	1099	1098	1097	1096	1095	1094	1093	1092	1091	1090	1089	1088	1087	1086	1085	1084	1083	1082	1081	1080	1079	1078	1077	1076	1075	1074	1073	1072	1071	1070	1069	1068	1067	1066	1065	1064	1063	1062	1061	1060	1059	1058	1057	1056	1055	1054	1053	1052	1051	1050	1049	1048	1047	1046	1045	1044	1043	1042	1041	1040	1039	1038	1037	1036	1035	1034	1033	1032	1031	1030	1029	1028	1027	1026	1025	1024	1023	1022	1021	1020	1019	1018	1017	1016	1015	1014	1013	1012	1011	1010	1009	1008	1007	1006	1005	1004	1003	1002	1001	1000	999	998	997	996	995	994	993	992	991	990	989	988	987	986	985	984	983	982	981	980	979	978	977	976	975	974	973	972	971	970	969	968	967	966	965	964	963	962	961	960	959	958	957	956	955	954	953	952	951	950	949	948	947	946	945	944	943	942	941	940	939	938	937	936	935	934	933	932	931	930	929	928	927	926	925	924	923	922	921	920	919	918	917	916	915	914	913	912	911	910	909	908	907	906	905	904	903	902	901	900	899	898	897	896	895	894	893	892	891	890	889	888	887	886	885	884	883	882	881	880	879	878	877	876	875	874	873	872	871	870	869	868	867	866	865	864	863	862	861	860	859	858	857	856	855	854	853	852	851	850	849	848	847	846	845	844	843	842	841	840	839	838	837	836	835	834	833	832	831	830	829	828	827	826	825	824	823	822	821	820	819	818	817	816	815	814	813	812	811	810	809	808	807	806	805	804	803	802	801	800	799	798	797	796	795	794	793	792	791	790	789	788	787	786	785	784	783	782	781	780	779	778	777	776	775	774	773	772	771	770	769	768	767	766	765	764	763	762	761	760	759	758	757	756	755	754	753	752	751	750	749	748	747	746	745	744	743	742	741	740	739	738	737	736	735	734	733	732	731	730	729	728	727	726	725	724	723	722	721	720	719	718	717	716	715	714	713	712	711	710	709	708	707	706	705	704	703	702	701	700	699	698	697	696	695	694	693	692	691	690	689	688	687	686	685	684	683	682	681	680	679	678	677	676	675	674	673	672	671	670	669	668	667	666	665	664	663	662	661	660	659	658	657	656	655	654	653	652	651	650	649	648	647	646	645	644	643	642	641	640	639	638	637	636	635	634	633	632	631	630	629	628	627	626	625	624	623	622	621	620	619	618	617	616	61
--	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	----

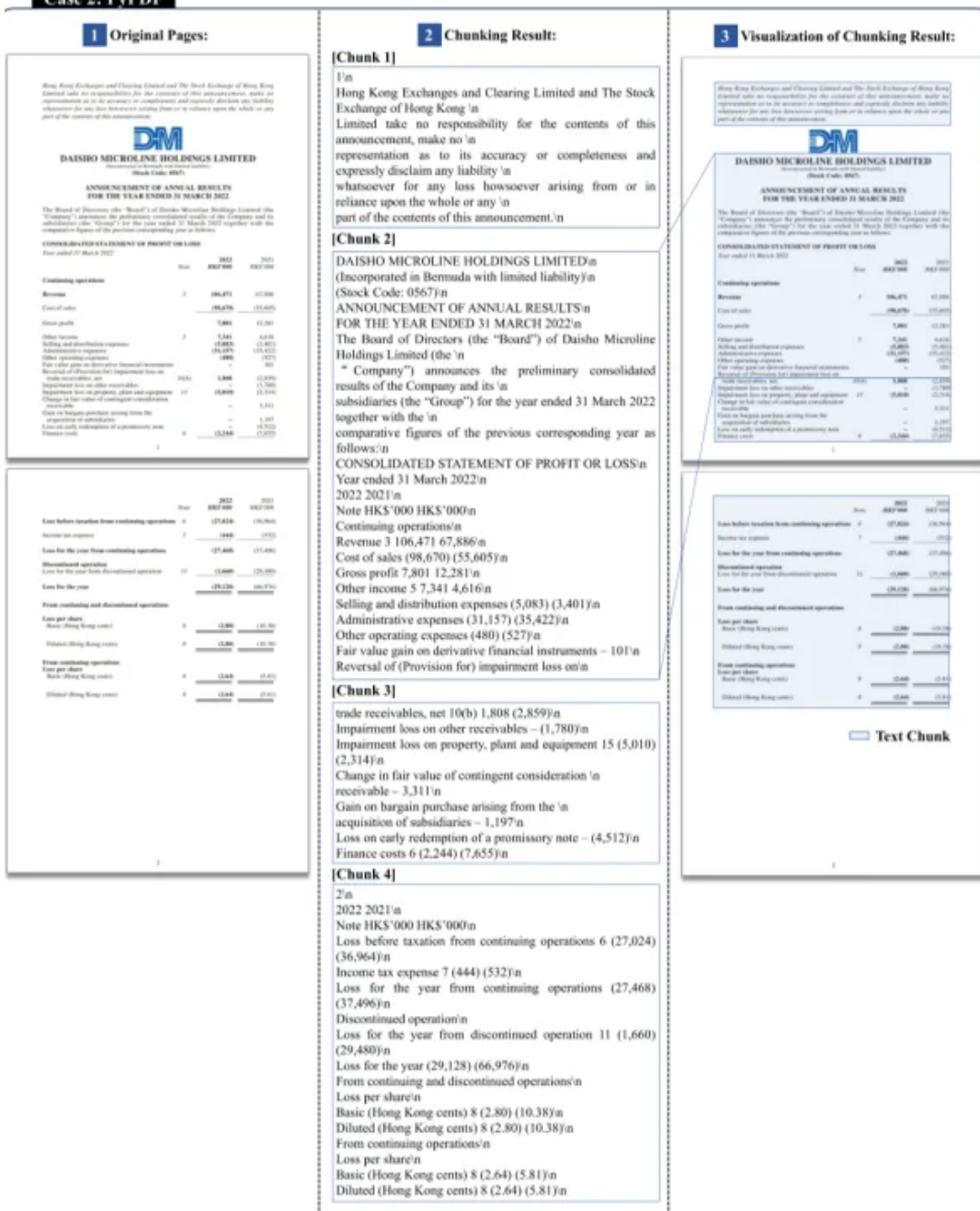


Figure 15. Parsing and chunking results of PyPDF on Case 2 (original document: [4]). Zoom in to see the details.

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	<ol style="list-style-type: none"> 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. 	<ol style="list-style-type: none"> 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。