arXiv:2601.11425v1 [cs.CV] 16 Jan 2026

PUBMED-OCR：PMC 开放获取 OCR 标注

Hunter Heidenreich、Yosheb Getachew、Olivia Dinica、Ben Elliott Roots.aiai-ml@roots.ai

🤗 **数据集：**huggingface.co/datasets/rootsautomation/pubmed-ocr

摘要

PubMed-OCR 是一个以光学字符识别为核心的科学文章语料库，源自 PubMed Central 开放获取 PDF。每张页面图像均使用 Google Cloud Vision 进行标注，并以紧凑的 JSON 模式发布，其中包含单词、行和段落级别的边界框。该语料库涵盖 20.95 万篇文章（150 万页；13 亿个单词），支持布局感知建模、基于坐标的问答，以及对依赖 OCR 的处理流程进行评估。我们分析了语料库的特征（例如期刊覆盖范围和检测到的版面特征），并讨论其局限性，包括依赖单一 OCR 引擎以及使用启发式方法进行行重建。我们发布这些数据和模式，以促进后续研究，并欢迎他人扩展。

1 引言

PDF 文件和扫描文档广泛存在于商业、政府、教育、研究和医疗保健领域。要实现能够切实减少重复性工作的人工智能副驾驶，系统必须可靠地理解现实世界中的文档 [1]。

开放数据是实现这一目标的核心。通过 democratizing access and standardizing formats，开放语料库能够促成更好的模型和算法 [2], 支持可复现性 [3], 扩大参与范围，并产生更强的基准测试。这一动态在广泛领域的大语言模型训练集 [4, 5, 6, 7, 8] 以及大规模多模态资源 [9, 10, 11]中都有体现。在文档处理领域，开放数据集一再推动进步，从 *IIT-CDIP* [12, 13]及其衍生数据集 [14,15, 16, 17, 18, 19] ，到针对复杂科学文章的源自 PubMed 的资源 [20, 21, 22, 23])。

一种相互促进的模式清晰可见：开放数据带来更强的模型，而更强的模型又推动更多开放数据的创建。尽管闭源系统通常比开放模型得到更大规模的扩展，但它们的输出可以被发布，以加速社区进步，*OCR-IDL* [24]便是一个例子。开放模型也可以组合成集成，以提高最先进的系统的数据质量。例如，DeepSeek-OCR [25] 通过利用 PP-DocLayout [26], 、MinerU [27], 、GOT-OCR2.0 [28],和 PaddleOCR [29]扩展了监督信号。

这些影响不止于文档处理。光学字符识别的进步有助于改进大语言模型训练语料库 [30],，因为光学字符识别提供了将光信号转换为离散文本的转换层。若转换不忠实，文本信息对于纯文本模型而言实际上仍不可见。

本文。 我们介绍了*PubMed-OCR*，它基于以往 PubMed 资源所使用的同一开放获取数据库构建。不同于将从数字 PDF 中提取的文本或区域与 JATS XML 对齐的方法（这一过程容易受到解析器噪声、启发式方法依赖以及扫描文档中文本遗漏的影响），我们直接使用商业光学字符识别系统（Google Vision OCR）标注页面图像，以生成词级、行级和段落级监督信号。我们提供语料库统计数据和定性示例，并发布该资源，以支持模型开发、基准测试集整理及相关研究。

2 相关工作

2.1 基于 PMCOA 的布局与表格数据集

PubMed Central 开放获取（PMCOA）长期以来一直是文档理解研究的基础资源，因为它同时提供 PDF 文件和机器可读的期刊文章标记套件（JATS）XML。大多数以往的数据集都利用这种配对关系，通过启发式或半监督匹配，将从 PDF 文件中提取的文本和区域与 XML 对齐。

GROTOAP [20] 和*GROTOAP2* [21] 为 PMCOA 提供了早期的大规模真实标注。*GROTOAP2*提供分层 XML 标注——将页面分解为区域、行、单词和字符——并采用双点边界框和 22 种区域标签（例如标题、摘要、正文、参考文献）。

PubLayNet [22] 通过将 PMCOA XML 与 PDFMiner [31] 输出对齐，扩大了布局监督规模，生成了覆盖 ∼360K 页的 0.5M 条区域标注。区域被映射到五个规范类别，并按期刊划分数据集；验证集和测试集的筛选标准更严格，其中包括抽样规则，以限制任何单一期刊占比过高。

作为布局分割的补充，*PubTables-1M* [23]旨在实现表格理解：对 575K 页和 948K 个表格进行了标注，涵盖表格、行、列和单元格层级；提供 PDF 和图像坐标中的边界框，以及用于下游解析的词级边界框。

相比之下，我们的语料库原生支持 OCR：我们完全绕过 PMCOA XML，直接使用高质量的 OCR 引擎从页面图像中生成单词级、行级和段落级监督信号，从而避免继承自 PDF 解析器的对齐错误，并支持对非数字化页面进行 OCR（即文档中包含扫描件且没有文本覆盖层的页面）。

2.2 通用领域的 OCR 和布局资源

IIT-CDIP [12, 13] 汇集了由 UCSF IDL 托管的 ∼7M 份烟草诉讼文件（TIFF 扫描件 +文本），其中包含大量真实世界噪声（手写内容、污渍、扫描伪影）。重要的是，这些文本已经线性化，缺少单词或行的边界框。后续工作在 IIT-CDIP 的子集上叠加了 OCR 和结构信息。例如，DESSURT [32]发布了 Tesseract [33] 的输出（单词/行），以及通过 PubLayNet/PrimaNet [34]得到的块级/段落级区域。基于这一来源整理的广泛使用的基准测试包括*RVL-CDIP*（文档分 [35], *FUNSD* （表单理解） [15], *Tobacco-3482*/*Tobacco-800* （分类、页面流分 [16, 12, 17, 18], 以及 *DocVQA* [19]。

OCR-IDL [24] 延续了这一脉络，使用商业引擎（Amazon Textract）为 UCSF 行业文献库提供大规模 OCR 标注。该版本涵盖 >26M 页（从一个拥有超过 7000 万份文档的文献库中抽样选取），可用于评估依赖商业级 OCR 的系统，而无需捆绑专有模型。

TabMe++ [36] 使用 Azure OCR 重新处理了*TabMe* 页面流分割基准 [37] ，替换了质量较差的 Tesseract 输出。尽管 TabMe 属于*IIT-CDIP/IDL* 体系，*TabMe++* 仍展示了更高质量的 OCR 对下游分割与分类的影响。

一些资源以文档布局为目标，但不提供 OCR。*DocBank* [38] 将 LaTEX 与 PDF 对齐，生成包含词元级标签和 PDF 派生词框的 50 万页数据集。*DocLayNet* [39]贡献了 8.1 万个人工标注页面，这些页面取自真实场景，涵盖 11 种区域类别；更早的数据集包括*Marmot* [40]和*PRImA* 布局基准 [41]。同属基于解析器的数据集，*PDFA*（*PDF Association*数据集） [42] 是从 SafeDocs CC-MAIN-2021-31 爬取数据中选出的大规模子集， [43], 以 Web 规模（数百万份文档）提供数字 PDF 词语/行、推断的阅读顺序和布局元数据。与优先采用 OCR 的语料库不同，PDFA 从 PDF 对象中提取文本，因此基本不包含扫描页面或纯图像页面——这是任何基于解析器的方法都会带来的副作用。

通用领域语料库展示了大规模定位式光学字符识别和布局监督对下游任务的实用价值；然而，科学文章带来了独特的挑战（密集的数学内容、细粒度的参考文献以及大量的表格/图示使用）。*PubMed-OCR* 通过在 PMCOA 页面上采用以光学字符识别为先的监督信号来弥补这一空白。

2.3 文本识别与光学字符识别的比较

我们将纯文本识别——即按照阅读顺序将页面序列化为单一序列——与有依据的/结构化光学字符识别区分开来；后者会生成带有边界框的单词、行和段落。前者适用于将文档摄入文本语料库，以用于大语言模型预训练 [30, 44], 而有依据的输出则保留空间溯源信息，从而支持布局感知建模和可验证归因。

数据集	领域	引擎	规模 Docs	页	粒度（边界框） 文本块	行	单词	字符
IIT-CDIP	UCSF IDL	–	6.5M	35.5M	✗	✗	✗	✗
OCR-IDL	UCSF IDL	Tesseract	–	825.1K	✓	✓	✓	✓
TabMe++	UCSF IDL	Amazon Textract†	4.6M	26M	✗	✓	✓	✗
PDFA	CC	Azure OCR†	44.8K	122.5K	✗	✓	✓	✗
GROTOAP2	PMCOA	PDF解析（无光学字符识别）	2.2M	18M	✗	✓	✓	✗
PubTables-1M	PMCOA	PDF解析（无光学字符识别）	13.2K	119.3K	✓	✓	✓	✓
***PubMed-OCR*（我们的方法）**	PMCOA	PDF解析（无光学字符识别）	–	575.3K	✗	✗	✓	✗
		Google Vision OCR†	**209.5K**	**1.5M**	✓	✓	✓	✗

表1：按规模和标注粒度对文本资源的比较。商业引擎标有 †.

布局感知模型会显式使用文本与版面来改进文档理解，既可以借助图像编码器，也可以将布局标记与文本一同编码。通常，布局信息以光学字符识别生成的边界框形式被摄入。代表性方法包括 LayoutLM [45, 46] 和 LiLT [47], , 以及更新的以大语言模型为中心的方法，例如 LayoutLLM（布局指令微调） [48], 、DocLLM（仅使用布局标记的大语言模型） [49], 和 LayTextLLM（将边界框标记与文本交错排列） [50],；这些方法无需庞大的视觉主干网络便展现出优异的效果。

有依据的输出还支持有依据的响应生成：答案会与细粒度证据（引文，以及在可用时页面上的坐标）一同生成。近期关于归属式／有依据的生成的研究通过学习以跨度级粒度附加引文 [51] 并评估引文质量 [52], , 提高了可验证性；与此同时，有依据的推理也取得了进展，将文本与边界框坐标交错处理 [53]。

由于有依据的光学字符识别在需要时可以确定性地线性化，因此它是更详尽但也更灵活的标注形式。因此，我们的语料库采用有依据的设置，并在词、行和段落层面提供监督信号。

3 PubMed-OCR 数据集

3.1 数据收集

我们通过官方 FTP/OAI 端点下载了 PMCOA PDF 文件，并将再分发范围限定为许可证允许共享衍生制品的文章。在 ∼2M 个 PDF 文件中， ∼60% 符合这一条件（∼1.2M）。我们从中均匀随机抽取 209.5k 篇文档，并使用 Google Vision API（2024 年 12 月 19 日版本）为每一页添加标注，费用为每 1000 页 1.50 美元。总费用为 ∼2.3k 美元（若对全部内容进行光学字符识别，费用约为其 5 倍，即 12k 美元）。我们仅纳入其 PMCOA 许可证允许再分发的文章。对于每篇文档，我们都会提供 OCR JSON；在许可允许的情况下，也会提供原始 PDF。元数据 CSV 记录了许可证（例如 CC BY、CC BY-SA、CC BY-NC、CC BY-NC-SA）、PMCID/PMID 的直接链接，以及允许的用途（例如，是否允许商业用途 =true/false）。OCR 标注采用与源文章相同的条款授权。

3.2 光学字符识别处理与规范化

我们将每个 PDF 页面渲染为分辨率为 150 每英寸点数的图像，并对图像字节运行 Google Cloud Vision 的文档_文本_检测。将页面图像传入 API 前，不进行手动倾斜校正。对于返回的完整_文本_标注，我们遍历各页的 → 文本块、 → 段落和 → 单词，提取每个单词的文本及其四顶点多边形。随后将顶点规范化为由{top-left, bottom-right}定义的轴对齐边界框。段落文本由各单词拼接而成；段落边界框则是包含所有单词顶点的轴对齐矩形。

行重建。 Google Vision API 仅返回单词和段落的边界框。我们使用粗略的启发式方法，将垂直对齐的单词聚类，从而推导出行：

1. 对于每个单词 w，设 $y_{\min}(w)$ 和 $y_{\max}(w)$ 分别为其各顶点的最小值和最大值， $x_{\min}(w)$、$x_{\max}(w)$ 为最小值/最大值 x。2. 维护以 ($\bar{y}_{\min}$， $\bar{y}_{\max}$) 为代表的行组。仅当 $|y_{\min}(w) - \bar{y}_{\min}| \leq 5$且 $|y_{\max}(w) - \bar{y}_{\max}| \leq 5$ 像素时，单词才会加入现有行组；否则创建新组。

3. 为避免跨列或跨段落合并，我们会根据原始 Google Vision OCR 返回的段落索引，将包含来自不同段落的单词的行组拆分，使每一行都只属于一个段落。4. 在每个行组内，按 $x_{\min}(w)$（从左到右）对单词排序，并将其连接起来构成行文本。行边界框为 $\left[\min_w x_{\min}(w),\ \min_w y_{\min}(w),\ \max_w x_{\max}(w),\ \max_w y_{\max}(w)\right]$。

标准化输出。 对于每页，我们会生成两个标准文件：一个 JSON 和原始 PDF。每个页面 JSON 都包含 text.words、text.lines 和 text.paragraphs，其中每个项目的多边形都会转换为轴对齐的边界框 $[X_1, Y_1, X_3, Y_3]$（左上角、右下角）。我们还会包含用于生成光学字符识别的基本图像元数据（路径、宽度、高度、每英寸点数），以确保结果可复现。

3.3 数据统计

	PubMed-OCR 数量	每份文档	每页	*OCR-IDL* 数量	每份文档	每页
文档	209.5K	-	-	4.6M	-	-
页	1.5M	7.4	-	26M	6	-
段落	61.0M	291.3	39.5	-	-	-
行	164.3M	784.9	106.3	46M	101.3	17.5
单词	1.3B	6229.6	844.0	166M	360.8	62.5

表2：*PubMed-OCR* 语料库统计数据（左）与*OCR-IDL* 报告的统计数据（右）。我们按已发布的数据报告*OCR-IDL*统计数据，但需要指出，文档/页数及其每份文档/每页统计数据表明，单词和行的数量比稿件所述多一个数量级。

与先前语料库的比较。 表1展示了*PubMed-OCR* 在广泛使用的文档资源中的定位。*IIT-CDIP*绝对规模最大，但其原生形式完全缺少边界框；Tesseract 处理等叠加方案仅为其中一个包含82.5万页的子集添加边界框 [32]。*OCR-IDL* 和*TabMe++* 展现了商业OCR在UCSF IDL领域大规模应用的价值，但未提供段落级或字符级边界框。解析器派生的PMCOA数据集（*GROTOAP2*、*PubTables-1M*、*PDFA*）从数字 PDF 中而非页面图像中提取文本和区域；对于非数字文档，这一过程可能导致召回率降低。相比之下，*PubMed-OCR* 以OCR优先的方式处理PMCOA，并提供段落级、行级和词级边界框，弥补了解析器派生的PMCOA资源与其他领域中OCR优先语料库之间的空缺。

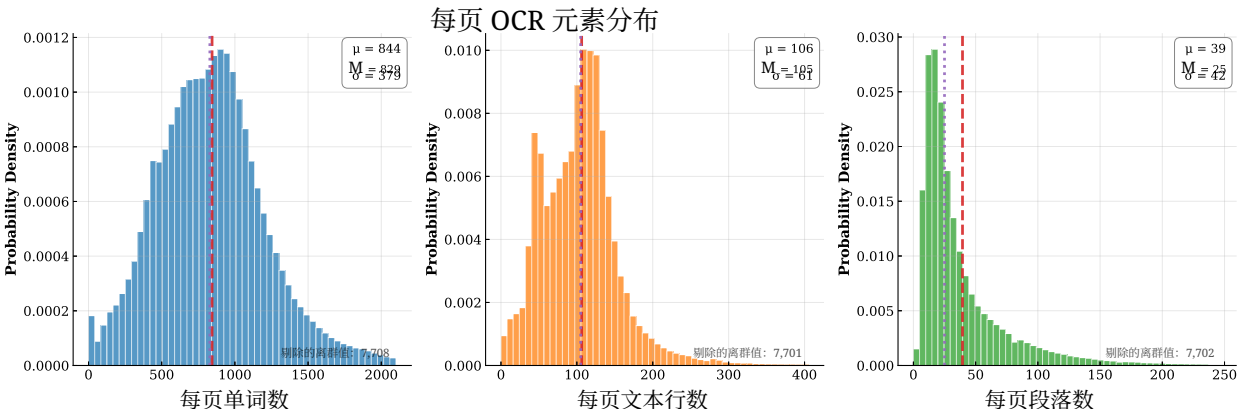


图1：每页单词数（左）、行数（中）和段落数（右）的分布。μ 表示均值，M 表示中位数，σ 表示标准差。每种分布均截断于第99.5百分位数或以下，以展示核心概率质量，而非长尾。

语料库概览（本研究）. 如表2所示，该数据集包含 209.5K 份文档和 1.5M 页（均值为每份文档 7.4 页）。平均而言，每页包含 39.5 个段落、106.3 行和 844 个单词，相当于每份文档包含 291.3 个段落、784.9 行和 6,229.6 个单词。将这些统计数据与报告的统计数据进行比较，

OCR-IDL [24], 我们发现，尽管*PubMed-OCR* 的文档和页数更少，但其行标注数量几乎是前者的 4 倍，词标注数量则是 10 倍。图1和图2显示，每页和每份文档的数量分布均呈右偏长尾，反映了短篇通讯与长篇文章并存的情况。这种规模与有依据的细粒度（带框的段落/行/单词）相结合，旨在支持布局感知建模、基于页面坐标的文档问答，以及对异质文章长度的稳健评估。

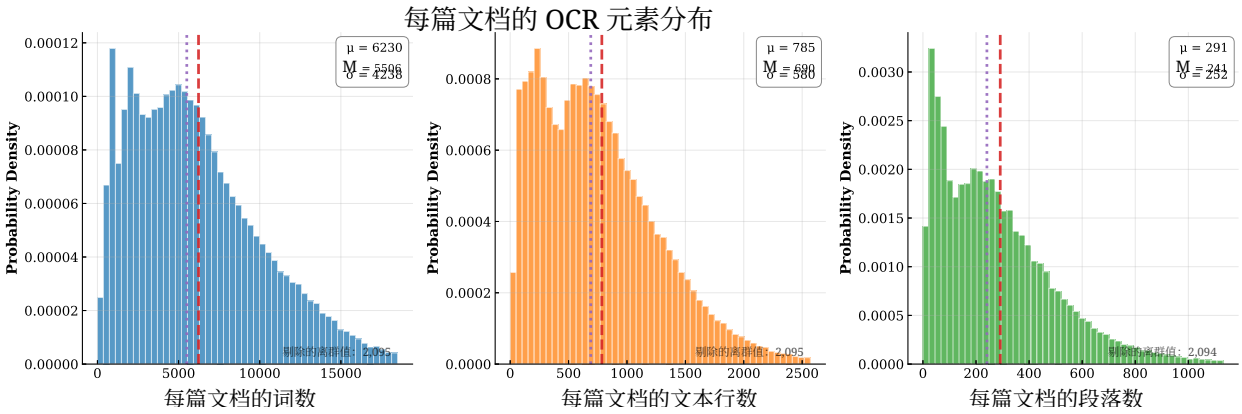


图 2：每篇文档的单词数（左）、行数（中）和段落数（右）的分布。 μ 表示均值， M 表示中位数， σ 表示标准差。每种分布均截断于第 99 百分位数或以下，以便直观呈现核心概率质量，而非长尾部分。

期刊分布。 PMCOA 的构成集中于少数高发期刊。排名前三的期刊——《细胞生物学杂志》（9.7%）、《实验医学杂志》（9.4%）和 《核酸研究》（3.9%）——约占文档总数的 23%。尽管分布有所偏斜，我们的数据集中仍涵盖了 2,478 种期刊。其中仅出现一篇文档的期刊共有 637 种，占 2,478 种期刊的约 25.7%，占文档总数的 0.3%。图 3 展示了按文档数量排名前 20 的期刊。

3.4 定性分析

为更好地了解数据集的定性特征，我们随机均匀抽取 40,000 页，并将其输入预训练的版面检测模块进行处理。为此，我们使用 PP-DocLayout_plus-L，它会将页面区域标记为 20 个不同类别。较有意思的特征包括公式（出现在 ∼25% 的页面中）、图像（出现在 ∼22% 的页面中），以及图表和表格（分别出现在 ∼16% 和 ∼18% 的页面中）。表 3 展示了这 20 类版面特征的分布。

作为定性示例，我们在图 4 中展示了两个页面。第一个页面来自一份较早的文档，右上角盖有印章。页面中既有密集的文本特征，也有嵌入文本中的公式。第二张图像展示了表格数据、算法定义，以及其他常见的文本和标题特征。我们在附录 A 中提供了一些额外样例。

4 结论

我们介绍了 *PubMed-OCR*，这是一个从 PubMed Central 开放获取子集衍生而来的 OCR 优先语料库，可直接从页面图像中提取段落级、行级和词级边界框，并以精简、标准化的 JSON 格式呈现。该资源绕过了脆弱的 PDF/XML 对齐流程，补充了基于解析器的 PMCOA 衍生数据集，并支持布局感知建模、基于依据的问答，以及面向科学文献的可归因生成。我们还提供语料库级别的统计数据和分布信息，以刻画不同期刊和文章长度的规模与多样性，使该数据集成为训练和评估的实用基础。

尽管实用，该语料库仍存在局限。目前它依赖单一的 OCR 引擎，并且根据词级边界框通过启发式方法重建行标注，这可能会在阅读顺序和分组方面引入偏差。其中不包括字符级边界框，也不包含数学表达式或图表结构的显式表示；其覆盖范围则受到 PMCOA 的许可证和期刊分布影响。在报告结果和设计实验划分时，应考虑这些限制。

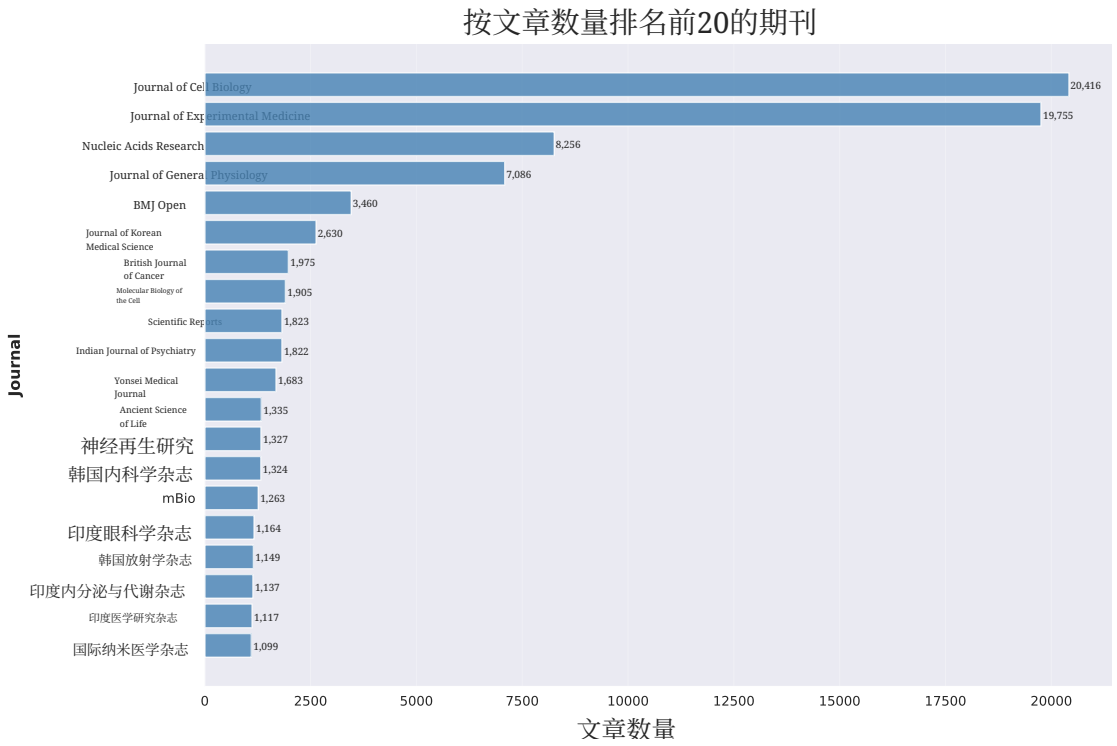


图 3：排名前 20 的期刊 收录于*PubMed-OCR*中的期刊。排名前 3 的期刊占所有收录文档的 ∼23% .

总而言之，*PubMed-OCR* 提供了一个可复现、开放访问的数据集，适用于需要在科学文章中准确关联文本与布局的研究。我们发布这一资源，旨在支持稳健评估，并促进不依赖专有流程的公平比较；同时也欢迎社区对其进行审查、扩展并在此基础上开展工作。

参考文献

[1] Weishi Wang、Hengchang Hu、Zhijie Zhang、Zhaochen Li、Hongxin Shao 和 Daniel Dahlmeier。大语言模型时代的文档智能：一项综述。*arXiv* 预印本 *arXiv:2510.13366*，2025年。[2] Sayash Kapoor、Rishi Bommasani、Kevin Klyman、Shayne Longpre、Ashwin Ramaswami、Peter Cihon、Aspen K Hopkins、Kevin Bankston、Stella Biderman、Miranda Bogen 等。开放基础模型的社会影响。*CoRR*，2024年。[3] Joelle Pineau、Philippe Vincent-Lamarre、Koustuv Sinha、Vincent Larivière、Alina Beygelzimer、Florence d'Alché Buc、Emily Fox 和 Hugo Larochelle。提升机器学习研究的可复现性（NeurIPS 2019 可复现性计划报告）。机器学习研究杂志，22(164):1–20，2021年。[4] Leo Gao、Stella Biderman、Sid Black、Laurence Golding、Travis Hoppe、Charles Foster、Jason Phang、Horace He、Anish Thite、Noa Nabeshima 等。The Pile：一个用于语言建模的多样化文本数据集，规模达 800 GB。*arXiv* 预印本 *arXiv:2101.00027*，2020年。[5] Hugo Laurençon、Lucile Saulnier、Thomas Wang、Christopher Akiki、Albert Villanova del Moral、Teven Le Scao、Leandro Von Werra、Chenghao Mou、Eduardo González Ponferrada、Huu Nguyen 等。BigScience ROOTS 语料库：一个 1.6 TB 的复合多语言数据集。神经信息处理系统进展，35:31809– 31826，2022年。[6] Denis Kocetkov、Raymond Li、Loubna Ben Allal、Jia Li、Chenghao Mou、Carlos Muñoz Ferrandis、Yacine Jernite、Margaret Mitchell、Sean Hughes、Thomas Wolf 等。The Stack：3 TB 采用宽松许可证的源代码。*arXiv* 预印本 *arXiv:2211.15533*，2022年。[7] Luca Soldaini、Rodney Kinney、Akshita Bhagia、Dustin Schwenk、David Atkinson、Russell Authur、Ben Bogin、Khyathi Chandu、Jennifer Dumas、Yanai Elazar 等。Dolma：一个包含三万亿词元的开放语料库，用于语言

版面特征	检测到的数量	包含该特征的页面百分比
数字	36,961	92.09
text	223,264	87.43
页眉	44,349	79.66
页脚	36,709	64.96
段落标题	58,312	63.65
图标题	28,101	43.99
公式	69,480	24.83
图像	13,426	22.54
参考文献内容	154,944	19.87
参考文献	13,319	19.80
表格	9,277	18.34
图表	15,421	16.59
文档标题	5,286	12.81
脚注	5,529	10.80
摘要	4,379	10.32
旁注	1,466	2.65
公式编号	1,238	1.18
算法	28	0.06
内容	9	0.02
seal	6	0.01

表 3：从 *PubMed-OCR* 的随机 4 万页样本中检测到的版面特征。检测到的数量表示在整个数据集中发现的各类版面特征的数量，而包含该特征的页面百分比表示样本中至少出现一次某种版面特征的页面所占百分比。版面特征是使用 PP-DocLayout_plus-L 检测的；该模型预测图像、表格、图表和公式的出现率较高。请注意，这些结果取决于模型，不应视为金标准标签。

模型预训练研究。在计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），页 15725–15788，2024 年。

[8] Maurice Weber、Dan Fu、Quentin Anthony、Yonatan Oren、Shane Adams、Anton Alexandrov、Xiaozhong Lyu、Huu Nguyen、Xiaozhe Yao、Virginia Adams 等。Redpajama：用于训练大语言模型的开放数据集。神经信息处理系统进展，37：116462–116492，2024 年。

[9] Christoph Schuhmann、Robert Kaczmarczyk、Aran Komatsuzaki、Aarush Katta、Richard Vencu、Romain Beaumont、Jenia Jitsev、Theo Coombes 和 Clayton Mullis。Laion-400m：由 CLIP筛选的 4 亿图文对开放数据集。载于*NeurIPS* 数据中心人工智能研讨会，数字 FZJ-2022-00923。于利希超级计算中心，2021 年。

[10] Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman 等。《Laion-5b：用于训练下一代图文模型的开放式大规模数据集》。《神经信息处理系统进展》，35：25278–25294，2022年。

[11] Jiaxi Gu、Xiaojun Meng、Guansong Lu、Lu Hou、Niu Minzhe、Xiaodan Liang、Lewei Yao、Runhui Huang、Wei Zhang、Xin Jiang 等。《悟空：一亿规模的中文跨模态预训练基准》。《神经信息处理系统进展》，35：26418–26431，2022年。

[12] David Lewis、Gady Agam、Shlomo Argamon、Ophir Frieder、David Grossman 和 Jefferson Heard。《构建用于复杂文档信息处理的测试集合》。载于第*29*届 *ACM SIGIR* 信息检索研究与发展国际年会论文集，第665–666页，2006年。

[13] Ian Soboroff。复杂文档信息处理（cdip）数据集。https://doi.org/10.18434/ mds2-2531，2022年。访问日期：2025年10月28日。

[14] Adam W Harley、Alex Ufkes 和 Konstantinos G Derpanis。用于文档图像分类和检索的深度卷积网络评估。载于*2015*年 第*13*届国际文档分析与识别会议（*ICDAR*），第991–995页。IEEE，2015年。

[15] Guillaume Jaume、Hazim Kemal Ekenel 和 Jean-Philippe Thiran。FUNSD：用于理解含噪扫描文档中表单的数据集。载于*2019*年国际文档分析与识别会议研讨会（*ICDARW*），第 2 卷，第 1–6 页。IEEE，2019年。

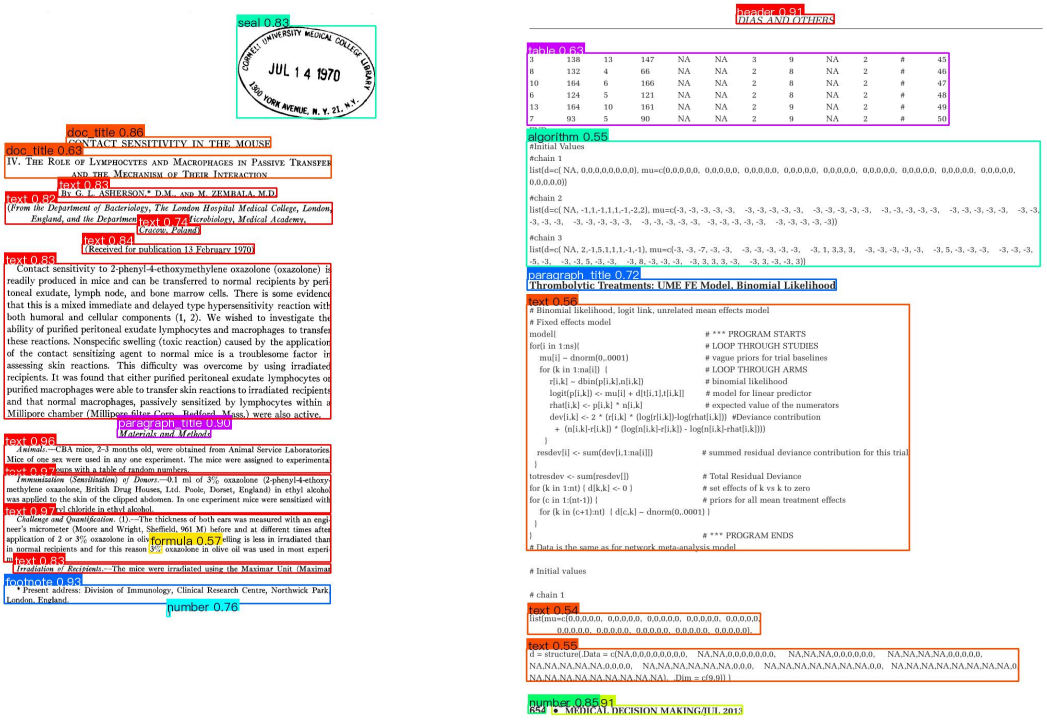


图 4：展示了 *PubMed-OCR* 中两个示例页面，并叠加了 PP-DocLayout 预测的版面检测类别。左侧页面有一枚印章，旁边是高密度文本（文本中嵌有公式）。右侧页面包含许多表格输出、代码片段和其他文本。

[16] Guangyu Zhu 和 David Doermann。自动文档标志检测。载于第九届国际文档分析与识别会议（*ICDAR 2007*），第 2 卷，第 864–868 页。IEEE，2007年。[17] G. Agam、S. Argamon、O. Frieder、D. Grossman 和 D. Lewis。复杂文档图像处理（CDIP）测试集项目。技术报告，伊利诺伊理工学院，2006年。[18] 加州大学旧金山分校。烟草历史文档库（LTDL），2007年。[19] Minesh Mathew、Dimosthenis Karatzas 和 CV Jawahar。DocVQA：用于文档图像视觉问答的数据集。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第 2200–2209 页，2021年。[20] Dominika Tkaczyk、Artur Czeczko、Krzysztof Rusek、Lukasz Bolikowski 和 Roman Bogacewicz。GROTOAP：开放获取出版物的真实标注。载于第 *12* 届 *ACM/IEEE-CS*数字图书馆联合会议论文集，第 381–382 页，2012年。[21] Dominika Tkaczyk 和 Pawel Szostek。GROTOAP2——创建大型科学文章真实标注数据集的方法。*D-Lib Magazine*，20(11/12)，2014年。[22] Xu Zhong、Jianbin Tang 和 Antonio Jimeno Yepes。PubLayNet：迄今最大的文档版面分析数据集。载于*2019*年国际文档分析与识别会议（*ICDAR*），第 1015–1022 页。IEEE，2019年。[23] Brandon Smock、Rohith Pesala 和 Robin Abraham。PubTables-1M：面向非结构化文档的综合表格提取。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4634–4642 页，2022年。[24] Ali Furkan Biten、Ruben Tito、Lluis Gomez、Ernest Valveny 和 Dimosthenis Karatzas。OCR-IDL：行业文档库数据集的 OCR 标注。载于欧洲计算机视觉会议，第 241–252 页。施普林格，2022年。[25] Haoran Wei、Yaofeng Sun 和 Yukun Li。DeepSeek-OCR：上下文光学压缩。*arXiv* 预印本 *arXiv:2510.18234*，2025年。

[26] Ting Sun、Cheng Cui、Yuning Du 和 Yi Liu。PP-DocLayout：一种用于加速大规模数据构建的统一文档版面检测模型。*arXiv*预印本 *arXiv:2503.17213*，2025年。[27] Bin Wang、Chao Xu、Xiaomeng Zhao、Linke Ouyang、Fan Wu、Zhiyuan Zhao、Rui Xu、Kaiwen Liu、Yuan Qu、Fukai Shang 等。MinerU：一种用于精确提取文档内容的开源解决方案。*CoRR*，2024年。[28] Haoran Wei、Chenglong Liu、Jinyue Chen、Jia Wang、Lingyu Kong、Yanming Xu、Zheng Ge、Liang Zhao、Jianjian Sun、Yuang Peng 等。通用OCR理论：通过统一的端到端模型迈向OCR-2.0。*arXiv*预印本 *arXiv:2409.01704*，2024年。[29] Cheng Cui、Ting Sun、Manhui Lin、Tingquan Gao、Yubo Zhang、Jiaxuan Liu、Xueqing Wang、Zelun Zhang、Changda Zhou、Hongen Liu、Yue Zhang、Wenyu Lv、Kui Huang、Yichao Zhang、Jing Zhang、Jun Zhang、Yi Liu、Dianhai Yu 和 Yanjun Ma。PaddleOCR 3.0技术报告，2025年。[30] Hynek Kydlíˇcek、Guilherme Penedo 和 Leandro von Werra。FinePDFs。https://huggingface.co/datasets/ HuggingFaceFW/finepdfs，2025年。[31] pdfminer.six开发者。pdfminer.six：一个用于从PDF文档中提取信息的Python库。https: // github.com/pdfminer/pdfminer.six，2025年。版本：20250506。MIT许可证。[32] Brian Davis、Bryan Morse、Brian Price、Chris Tensmeyer、Curtis Wigington 和 Vlad Morariu。使用DESSURT进行端到端文档识别与理解。载于欧洲计算机视觉会议，第280–296页。施普林格，2022年。[33] Anthony Kay。Tesseract：一款开源光学字符识别引擎。*Linux Journal*，2007(159):2，2007年。[34] Brian Davis。IIT-CDIP数据集的Tesseract OCR。https://doi.org/10.5281/zenodo.6540454，2022年5月。[35] Adam W Harley、Alex Ufkes 和 Konstantinos G Derpanis。深度卷积网络在文档图像分类与检索中的评估。载于国际文档分析与识别会议 *(ICDAR)*，2015年。[36] Hunter Heidenreich、Ratish Dalvi、Rohith Mukku、Nikhil Verma 和 Neven Piˇculjan。页面流分割的大语言模型，2024年8月。[37] Thisanaporn Mungmeeprued、Yuxin Ma、Nisarg Mehta 和 Aldo Lipani。为这份文档文件夹添加标签：商业文档的页面流分割。载于第*22*届*ACM*文档工程研讨会论文集，第1–10页，2022年。[38] Minghao Li、Yiheng Xu、Lei Cui、Shaohan Huang、Furu Wei、Zhoujun Li 和 Ming Zhou。DocBank：文档版面分析基准数据集。载于第*28*届计算语言学国际会议论文集，第949–960页，2020年。[39] Birgit Pfitzmann、Christoph Auer、Michele Dolfi、Ahmed S Nassar 和 Peter Staar。DocLayNet：一个由大量人工标注构成的文档版面分割数据集。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，第3743–3751页，2022年。[40] Jing Fang、Xin Tao、Zhi Tang、Ruiheng Qiu 和 Ying Liu。用于表格检测评估的数据集、真实标注与性能指标。载于第*10*届*IAPR*国际文档分析系统研讨会论文集 *(DAS)*，第445–449页，澳大利亚昆士兰州黄金海岸，2012年3月。IEEE计算机学会。[41] Apostolos Antonacopoulos、David Bridson、Christos Papadopoulos 和 Stefan Pletschacher。用于评估文档版面分析性能的真实数据集。载于*2009*年第*10*届国际文档分析与识别会议，第296–300页。IEEE，2009年。[42] Pablo Montalvo 和 Ross Wightman。PDF Association数据集（PDFA）：英语WebDataset分片。https: //huggingface.co/datasets/pixparse/pdfa-eng-wds，2024年。Hugging Face数据集卡片。访问日期：2025-10-31。[43] Safedocs（cc-main-2021-31-pdf-untruncated）。https://digitalcorpora.org/corpora/file-corpora/ cc-main-2021-31-pdf-untruncated/，2023年。Digital Corpora。访问日期：2025-10-31。[44] Jake Poznanski、Aman Rangapur、Jon Borchardt、Jason Dunkelberger、Regan Huff、Daniel Lin、Christopher Wilhelm、Kyle Lo 和 Luca Soldaini。olmOCR：借助视觉语言模型解锁PDF中的数万亿词元。[45] Yiheng Xu、Minghao Li、Lei Cui、Shaohan Huang、Furu Wei 和 Ming Zhou。LayoutLM：面向文档图像理解的文本与版面预训练。载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第1192–1200页，2020年。[46] Yupan Huang、Tengchao Lv、Lei Cui、Yutong Lu 和 Furu Wei。LayoutLMv3：通过统一的文本与图像掩码进行文档人工智能预训练。载于第*30*届*ACM*国际多媒体会议论文集，第4083–4091页，2022年。

[47] Jiapeng Wang、Lianwen Jin 和 Kai Ding。LiLT：一种简单而有效的语言无关布局变换器，用于结构化文档理解。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第7747–7757页，2022年。[48] Chuwei Luo、Yufan Shen、Zhaoqing Zhu、Qi Zheng、Zhi Yu 和 Cong Yao。LayoutLLM：利用大语言模型进行布局指令微调，以实现文档理解。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15630–15640页，2024年。[49] Dongsheng Wang、Natraj Raman、Mathieu Sibue、Zhiqiang Ma、Petr Babkin、Simerjot Kaur、Yulong Pei、Armineh Nourbakhsh 和 Xiaomo Liu。DocLLM：一种用于多模态文档理解的布局感知生成式语言模型。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第8529–8548页，2024年。[50] Jinghui Lu、Haiyang Yu、Yanjie Wang、Yongjie Ye、Jingqun Tang、Ziwei Yang、Binghong Wu、Qi Liu、Hao Feng、Han Wang 等。一个边界框抵得上一个词元：在用于文档理解的大语言模型中交错使用版面与文本。*arXiv*预印本 *arXiv:2407.01976*，2024年。[51] Lei Huang、Xiaocheng Feng、Weitao Ma、Yuxuan Gu、Weihong Zhong、Xiachong Feng、Weijiang Yu、Weihua Peng、Duyu Tang、Dandan Tu 等。为带来源归属的大语言模型学习细粒度、有依据的引文。*arXiv*预印本 *arXiv:2408.04568*，2024年。[52] Yumo Xu、Peng Qi、Jifan Chen、Kunlun Liu、Rujun Han、Lan Liu、Bonan Min、Vittorio Castelli、Arshit Gupta 和 Zhiguo Wang。Citeeval：以原则为导向的来源归属引文评估。*arXiv*预印本 *arXiv:2506.01829*，2025年。[53] Yue Fan、Xuehai He、Diji Yang、Kaizhi Zheng、Ching-Chen Kuo、Yuting Zheng、Sravana Jyothi Narayanaraju、Xinze Guan 和 Xin Eric Wang。Grit：教多模态大语言模型借助图像进行思考。*arXiv*预印本 *arXiv:2505.15879*，2025年。

更多示例

图 5–8 展示了几个来自 PP-DocLayout 的额外样本及其布局检测标注。

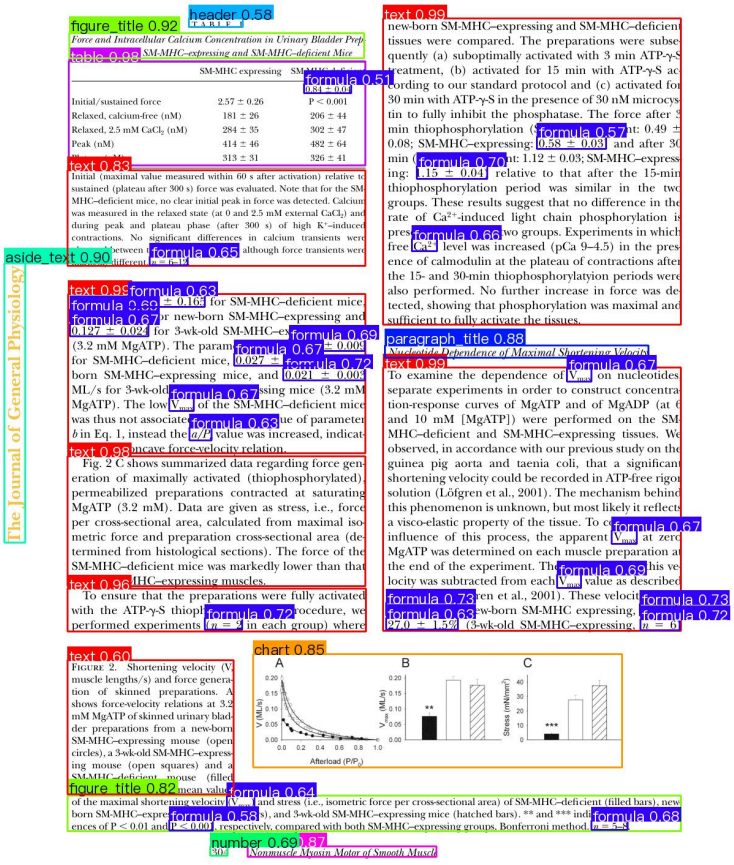


图5：一页来自*PubMed-OCR*的样例，展示了各种特征：旁注、图表、说明文字和公式。

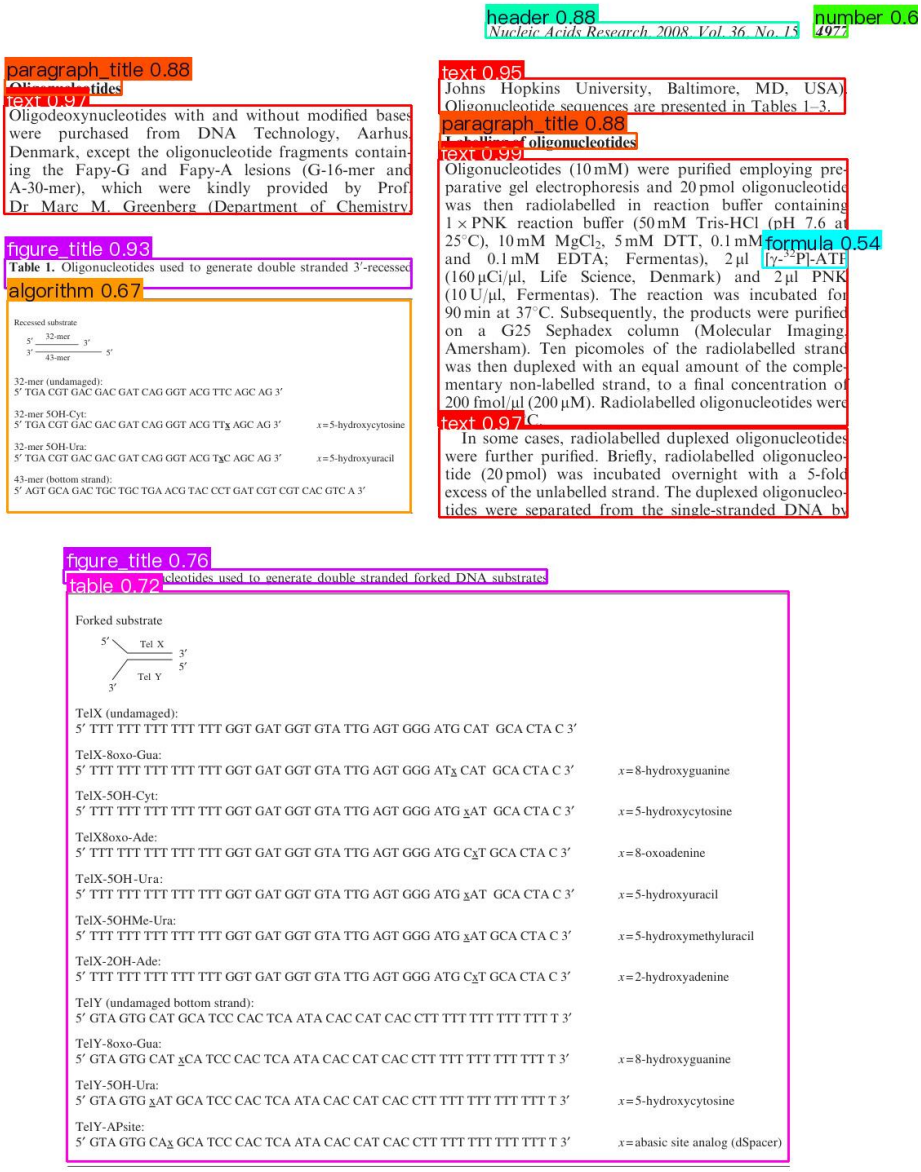


图6：一个包含复杂科学表格的示例。第二个表格部分被误识别为算法。引入带有版面标注的 *PubMed-OCR* 可能是一种有价值的扩充，有助于改进当前一代版面模型。

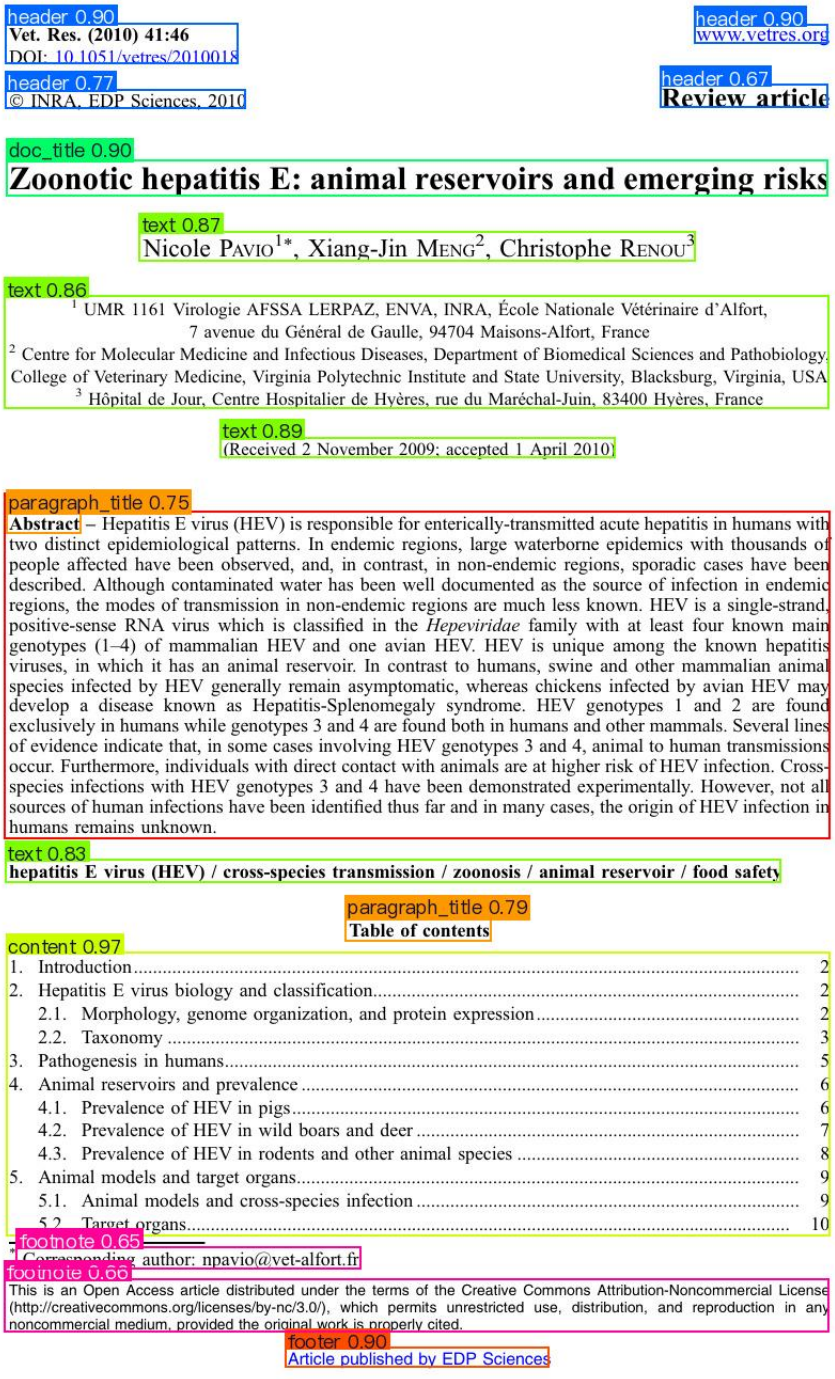


图7：一页包含密集文本和目录的示例页面，以及对段落和文档标题的结构化检测。

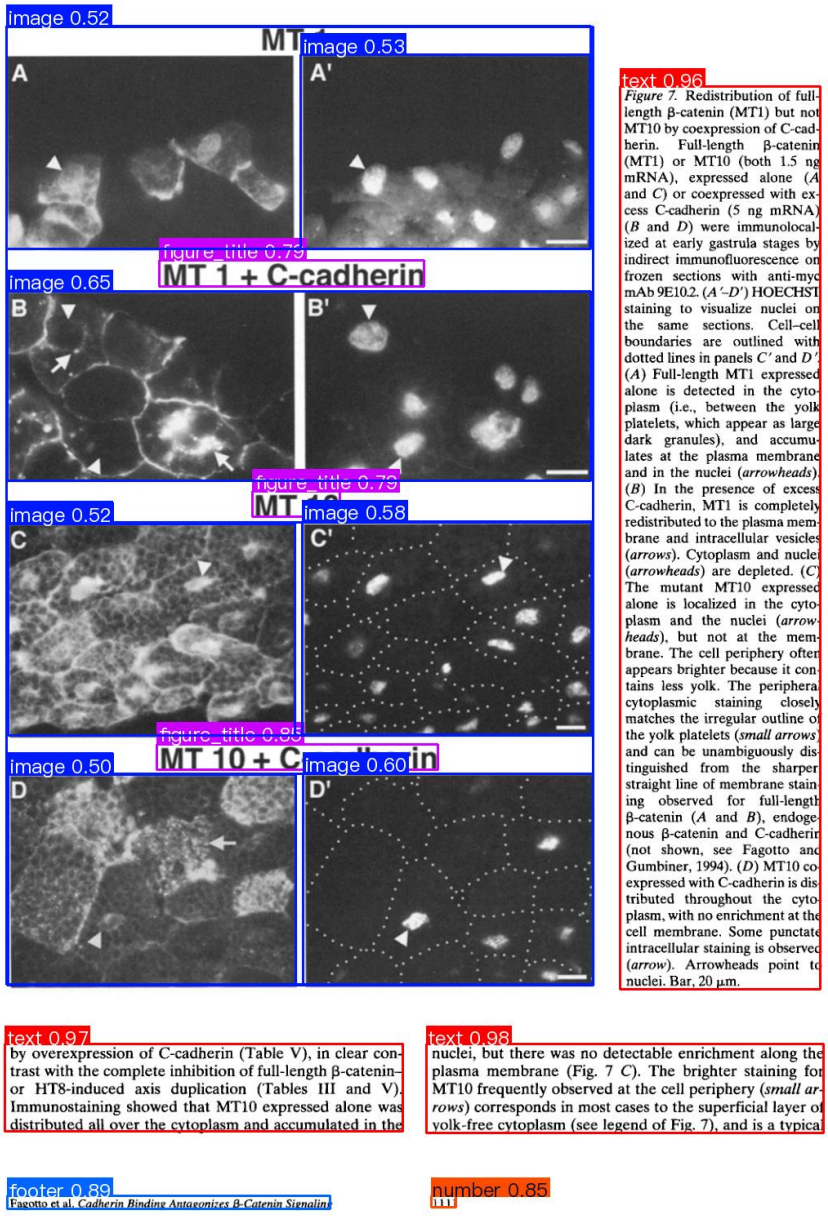


图 8：一个示例　一个图像密集的页面，配有结构化说明文字，介绍细节及预期解读。