OCR 评估方法与指标综述及历史文献的不可见性

Fitsum Sileshi Beyene
fsb5110@psu.edu 宾夕法尼亚州立大学，大学公园，宾夕法尼亚州，美国

Christopher L. Dancy
cdancy@psu.edu 宾夕法尼亚州立大学，大学公园，宾夕法尼亚州，美国

arXiv:2603.25761v1 [cs.CV] 26 Mar 2026

摘要

光学字符识别（OCR）和文档理解系统日益依赖大型视觉与视觉语言模型，但评估仍然侧重于现代、西方和机构类文献。这种侧重掩盖了系统在历史档案和边缘化档案中的表现；在这些档案中，版面布局、字体排印和材料劣化都会影响解读。本研究考察了OCR和文档理解系统的评估方式，特别关注黑人历史报纸。我们采用PRISMA框架，回顾了2006年至2025年间发表的OCR和文档理解论文及基准数据集。我们探讨了这些研究如何报告视觉转换器和多模态OCR系统的训练数据、基准设计和评估指标。审查过程中我们发现，黑人报纸及其他由社区制作的历史文献很少出现在已报告的训练数据或评估基准中。大多数评估侧重于现代版面布局中的字符准确率和任务成功率，很少捕捉历史报纸中常见的结构性失效，包括栏位坍塌、排印错误和幻觉文本。为阐明这些发现，我们结合以往的实证研究以及重要黑人报刊馆藏的档案统计数据，说明评估差距如何导致结构性不可见和表征伤害。我们提出，这些差距源于组织层面（中观）和制度层面（宏观）的行为与结构，并受到基准激励机制和数据治理决策的影响。

CCS概念

• **计算方法** → 人工智能；机器学习；• **信息系统** → 信息检索；• **应用计算** → 艺术与人文。

关键词

光学字符识别评估、文档理解、历史文献、基准、版面分析、计算机视觉

允许免费制作本作品的全部或部分内容的数字版或纸质副本，供个人或课堂使用，前提是这些副本不得用于营利或商业用途，也不得为此分发，且副本首页须附有本声明和完整引文。对于本作品中由作者以外的他人拥有版权的部分，必须尊重其版权。注明出处后可进行摘要引用。若要以其他方式复制、再版、发布到服务器或分发至列表，须事先获得特定许可和/或支付费用。请向 permissions@acm.org 申请许可。*FAccT '26*，蒙特利尔，加© 2026 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM https://doi.org/XXXXXXX.XXXXXXX

1 引言

评估技术的设计不仅仅是衡量机器学习系统；它还塑造了这类系统被构建来识别的内容。事实上，对于何为可接受的认识表征，与标准 和主流人类体裁之间的联系，往往隐含在人工智能与机器学习系统工程流程中 [53]。在光学字符识别（OCR）和文档理解领域，标准、指标以及（生成和管理的）数据集共同决定了哪些文档可读，以及哪些历史能够被计算 [37]。我们认为，当前的 OCR 评估系统始终无法识别具有文化相关性的失效模式，实际上使整类历史文本无法被旨在阅读它们的系统访问。

过去十年间，光学字符识别已从更偏重字符级的识别，发展到全面整合的文档理解。视觉语言模型（VLM）和基于 Transformer 的架构有望从文档图像中提取的不仅是文本，还包括语义结构、空间关系和上下文含义 [21, 28, 54]。然而，评估技术并未跟上这一发展。用于评估这些系统的主流指标，如字符错误率（CER）和词错误率（WER），将文档保真度简化为纯粹的字符串编辑距离——把文档理解视为严格按序进行的转录活动——模糊了误读字符与误读栏列之间、转录错误与版面布局丢失之间的界限 [37]。此外，OCRBench、OmniDocBench 和 DocVQA 等顶尖基准 [32, 33, 39] 严重依赖科学论文、企业表格和当代数字 PDF，导致历史档案和社区档案长期以来未得到充分代表 [32, 39]。尽管有人可能会主张，鉴于某些视觉语言模型开发过程中隐含的价值观存在问题，摆脱对可读性的追求具有优势，但批判性地审视人工智能工程生命周期的这一环节，确实为创建具有文化相关性的数字档案所需的新型批判性、解放性系统提供了契机。

在这篇综述中，我们有意聚焦于黑人数字档案，而非试图广泛涵盖数字档案。除了与这些档案反映出一系列不同社会物理限制的方式相关的重要社会技术考量之外，我们所关注的早期黑人报刊数字档案对于当前的社会政治时刻也至关重要。让这些档案更易辨读（以批判、负责的方式）且更易获取，能使更广泛的受众在当下加以利用（例如，参见最高法院大法官凯坦吉·布朗·杰克逊在近期一份异议意见中使用有色人种大会项目 [11] [22]，这是这些文献通过数字化变得易于辨读所蕴含力量的重要例证）。这些文献包括《自由报》（1827年）、《北极星报》（1847年）和《芝加哥

Defender》（1905年）等历史报纸。它们不仅是有趣的历史背景记录，也是美国政治、社会和文化史的重要档案。正如Benjamin Talton所指出的："如果没有黑人报刊，就无法讲述美国黑人的故事，因为黑人报刊报道了主流白人报纸未曾报道的事情，并真正尊重黑人作为人的尊严 [8]。"这些报纸恰恰呈现了当前评估系统所忽略的挑战，例如缩微胶片保存造成的扫描件劣化、现代训练语料库中缺失的维多利亚时代字体和哥特字体，以及版面布局复杂的多栏页面——其空间编排承载着修辞和政治意义 [3, 17, 25]。

其后果是，评估系统本身存在系统性盲点——被视为最先进的模型虽以 PubLayNet [56], DocBank [29], 和 DocLayNet [41]等数据集训练，但用于黑人历史报纸时却表现不一致，部分原因在于所用评估系统并不要求模型在这些材料上取得成功。这些材料具有特定的社会文化语境，影响其物理形态、数字化过程及相关元数据的生成和记录方式。即使标准准确率指标可能显示字符错误率（CER） [37],令人满意，这些模型往往仍会悄然抹去编辑们用来强化修辞、促进社会凝聚力并彰显政治身份的多栏结构。这种错误不仅是技术问题，也是认识论问题。

在本文中，我们为文档理解领域的公平性、问责性和透明度研究作出三项贡献：

- 对光学字符识别和文档理解评估系统进行系统分析，重点考察其训练数据、基准和指标是如何呈现与实施的。
- 本文指出，当前评估系统往往无法检测历史报纸中常见的结构和版面布局错误，尽管其报告显示字符级准确率很高；本文也为这一论点提供了支持。
- 本文以历史黑人报刊资料为案例研究，展示这些评估盲点如何导致具有结构意义的错误持续未能被常用评估技术报告。

或许，与人工智能伦理相关的讨论相比，这个问题更为简单。我们认为，应重新审视光学字符识别系统评估中的关键迭代步骤；不过，关于人工智能伦理以及更高层面的人工智能工程流程，请参见 [5, 53] 中的相关讨论。挑战本身并非技术性的，而在于资源分配、数据集多样性，以及对文化意义的认知——这也可以理解为一项价值观方面的挑战 [4]。通过展示评估选择如何反映人们对哪些文档重要的假设，我们希望为机器学习研究人员提供一套术语，用于批判性审视自己的基准，同时也提供一个可推广至黑人报纸以外领域的范例，例如社区支持的原住民档案等其他重要且具有文化相关性的数字档案，以及标准评估实践悄然失效的其他领域。

2 相关工作

本研究承接了计算机视觉（用于光学字符识别）和数字人文等领域的大量前期研究。我们梳理了四个相互关联的研究领域中的既有工作：光学字符识别系统与文档理解架构的演进、评估指标和基准的开发、历史文献数字化研究，以及关于机器学习中的偏见与表征的批判性研究。

2.1 从字符识别到文档理解

过去十年间，光学字符识别研究的发展方向显示出一种根本性转变：从单个字符识别转向全面的文档理解。传统的光学字符识别系统（例如 Tesseract）将文本提取视为一个由二值化、分割和字符分类构成的流程，并将版面分析作为预处理的中间步骤 [46]。

在这些重要系统之后，LayoutLM 模型系列通过在同一 Transformer 架构中对文本、版面布局和视觉方面进行建模，代表了架构上的转变 [54]。LayoutLMv2 纳入了融合空间表征与文本表征的跨模态预训练目标 [55], ；而 LayoutLMv3 则结合文本掩码和图像掩码，在不依赖预先提取的光学字符识别特征的情况下学习文档结构 [21]。这些模型表明，将版面布局视为具有语义重要性，而不只是位置信息，能够提升文档理解任务的结果。

近期，大多数研究都致力于绕过传统光学字符识别流程的端到端系统。例如，TrOCR 表明，在合成数据上预训练的 Transformer 编码器-解码器模型无需进行其他特定于任务的架构调整，也能取得良好的手写文本识别效果 [28]。随后，Donut [23]和 Nougat [6], 将这种文档解析技术拓展至直接从文档图像生成结构化输出的视觉Transformer。这些模型中，GOT-OCR 2.0 将场景文本、文档和数学符号等多种 OCR模态整合到单一架构中，但在历史材料上的效果仍明显不足 [52]。

显然，视觉语言模型的发展代表了当前光学字符识别系统研究的前沿。Qwen2.5-VL 和 PaddleOCR-VL 等系统采用大规模多模态预训练，将光学字符识别作为更广泛的视觉推理流程的一部分来完成 [1, 13]。此外，olmOCR 项目利用互联网档案馆的内容提供训练数据，其中包含一些较早的文本，从而专注于大规模 PDF 解读 [42]。然而，正如下一节所示，向文档理解的转变并未伴随着评估这些系统的方式发生相应变化。

2.2 评估指标：编辑距离

长期以来，光学字符识别系统的评估一直基于字符错误率（CER）和词错误率（WER），二者均源自预测字符串与基准真值字符串之间的 Levenshtein 编辑距离。这些指标计算简单且易于理解；字符错误率为 5% 意味着每二十个字符中就有一个需要更正。字符错误率/词错误率几乎被光学字符识别研究普遍采用，许多系统只报告这些指标 [37, 41]。这些限制之所以重要，是因为编辑距离算法将所有错误一视同仁，无论是词内发生替换，还是换行符打乱了段落结构。它们

没有惩罚版面错误的机制；如果某个模型成功转录了每个字符，却打乱了栏顺序，那么它仍能获得完美的字符错误率，尽管结果很差。

尽管改进后的指标开始着手应对历史数字档案中一些与社会和文化背景相关的差异，但评估各种 OCR 系统时仍可能面临基准真值缺失的问题。考虑到许多此类档案难以实现数字化，缺少用于有监督评估的相关数据和元数据也就不足为奇了。Beyene 和 Dancy [3] 提出了一种无监督评估框架，在缺少数字化档案的真实标注转录文本时，可用于开展某些评估。作者结合语义连贯性分数、区域熵差异和文本冗余分数，对一份缺少配套基准真值的历史数字档案中的各类 OCR 系统进行性能评估。虽然这只是完善评估实践、使其更能应对与文化相关的数字历史档案现实情况所需的一部分，但考虑到这些数字档案在个人、组织和机构层面的实际情况（包括黑人历史数字档案在转录方面通常资源较少），这项研究确实提供了一种有用的工具。

尽管取得了这些进展，指标多元主义仍属例外，而非常态。据我们的调查，报告光学字符识别结果的系统中，只有少数除了字符错误率/词错误率之外还包含结构指标，而专门关注历史文本版面保留的系统更是寥寥无几。

2.3 基准测试与预训练数据

用于评估文档理解系统的基准，也设定了哪些文本具有价值的预期。IIT-CDIP 是许多当前系统的主要预训练语料库，包含约 700 万份文档，共约 4200 万页扫描页（图像），这些材料均通过烟草诉讼证据披露收集而来 [27]。IIT-CDIP 涵盖多种文档类别（信函、表格、备忘录和报告），规模庞大且类型多样，但它属于一种小众机构格式（即 20 世纪中后期的美国商业文件），这会影响模型所采纳的视觉和语言模式，并使之成为“标准”。现代基准扩大了覆盖范围，但仍存在一些缺口。OCRBench v2 包含来自 31 种场景的 11,500 张照片，涵盖大量科学记号和场景文本，但未包含历史文档类别 [32]。OmniDocBench [39]提供了 1,355 份带标注的PDF，并针对模糊、水印等退化因素进行基于属性的评估；然而，它显然面向“现代时期的文档”，而非历史文档。近期发布且仍在扩充的 olmOCR-Bench 包含 7,010 项单元测试，着重评估版面保真度，也纳入了一些数字化印刷材料 [42], ，因此在主要基准中最接近历史文档覆盖范围，但仍缺少对边缘化社群档案中具有社群或文化相关性的内容进行有针对性的表征。

一些基准关注的是文档问答，而非光学字符识别准确率本身。DocVQA [33], DUDE [51], 和 DUE[7] 评估模型从文档中提取特定信息的能力，其中光学字符识别作为上游中间环节。尽管这些标准评估的是实际效用，但如果模型根据不完整或劣化的文本生成了正确答案，它们可能会掩盖光学字符识别的缺陷。

近期，KITAB-Bench 专门针对阿拉伯语文献遗产而创建，例如手稿和历史行政文献，并通过人在回路评估加以改进 [18]。Newspapers Navigator 数据集从 Chronicling America 的历史报纸档案中提取视觉内容，重点关注图像分类 [24]。American Stories 研究数据集大规模提供了 1920 年以前美国报纸的结构化文章文本，表明历史报纸处理是可行的 [15]。

鉴于黑人报纸的档案价值，它们未被纳入评估基准这一点值得注意。Chronicling America 目前收录了从 1777年到 1963 [30] 年的 344 种非裔美国人报纸出版物。此前，这些馆藏的光学字符识别错误率平均为 18% [9], ；美国国会图书馆于 2025年启动了一种光学字符识别重新处理方法，以减少早期数字化批次中的噪声[31]。霍华德大学的黑人新闻档案拥有 2,000 多种报刊名和 100,000期报纸，存放在 2,847 卷微缩胶卷上 [20]；然而，截至 2025 年末，这些档案中仍有相当大一部分正在作为一个多年项目的一部分进行数字化，该项目旨在将馆藏的 60% [35]上线 [50]。因此，所用的基准系统反映了多个层面（个人、组织和机构）的选择，其中包括影响哪些历史能够通过计算方式变得可读的机构和资金决策。

2.4 历史文献数字化

多年来，研究历史文献数字化的学者一直意识到，主流 OCR 评估所忽略的种种挑战。Holley 的 [19] 开创性研究展示了众包纠正澳大利亚报纸 OCR 错误的方式，揭示了标准方法在处理历史字体排印时的不足，并涉及大规模人工监督。近期对历史文献数据集的调查指出了多种导致文献劣化的因素，例如斑点霉变、透印、照明不均和物理损坏；这些因素使档案材料有别于现代记录 [36]。

尽管存在这些潜在障碍，专用系统已证明，只要经过适当训练，历史文献的光学字符识别在技术上是可行的。OCR4all通过在领域专用基准真值上进行微调，将19世纪哥特字体的错误率降低了84% [43]。1800年至1941年间，80%的德国文献使用哥特体字体，而现代系统若未经专门训练，处理这类材料时表现仍然很差（字符错误率为50%至70%）；这表明训练数据构成直接决定了哪些历史时期能够被识读 [47]。OCR后校正已成为一种互补方法。Nguyen等人 [37] 开展了一项调查，记录了从基于规则的模式匹配到神经序列到序列模型等各种解决方案，其中基于T5的方法在历史材料上的错误率降低幅度达到28%至40% [44]。Beshirov等人 [2] 还展示了针对保加利亚历史记录的语言特定后校正，说明语言资源会限制哪些历史传统能够从校正流程中受益。这一系列研究表明，处理历史文献具备相应的技术能力。这种能力与系统典型性能之间持续存在的差距，源于

训练数据集和评估基准中内嵌的优先级，而非技术限制。

2.5 Bias，表征与文化抹除

在数字人文领域，研究者探讨了算法系统如何与档案空缺相互作用。Gallon 的 [16] “黑人数字人文”倡议强调，必须考量黑人文化资产生成、保存和开放获取所处的具体条件。黑人报刊尤其面临经济困境和政治迫害，这影响了文献记录实践以及档案存续状况，而通常采用的光学字符识别系统并非为应对这些情况而设计。近期相关研究已开始将这些批判性视角与技术实践联系起来。Smith 等人 [45] 考察历史报纸中的文本复用，展示了光学字符识别质量如何影响哪些文本能够通过计算方式追踪，哪些文本则无法被算法分析所见。Casey [10] 倡导采用“半球性”视角进行报纸数字化，着重关注跨国黑人报刊网络，而非将各出版物孤立看待。Taurino 和 Smith [48] 对“作为档案科学的机器学习”的定义指出，训练数据整理本身就是一种档案实践，可能影响计算系统从过去中学习到的内容。事实上，一些当前的机器学习系统（尤其是生成式系统）被视为历史知识系统的权威形式，延续了可能存在问题的信息集中式节点，既强化了有问题的社会结构（另见 [14, 38]），也会以概率方式生成错误的信息序列，因此必须审慎看待这种作为具有重要意义的档案实践 。

在机器学习与档案相结合的背景下，理解这种档案工程实践尤为重要，而所谓的“过度历史化”问题更凸显了这一点：视觉语言模型会生成符合历史背景却并不准确的字符，这将这些担忧与当前模型的行为联系起来。一项关于 GPT-4o Vision 处理 18 至 19 世纪俄罗斯手稿的研究发现，59% 的错误涉及插入不合时代的字符，这表明模型学会的是生成“历史感”，而非准确转录历史文献 [26]。这种失败表明，生成式人工智能会造成数字档案失真，并可能由此生成看似可信却并不真实的历史记录。

我们以这些基础为起点，通过系统综述将评估指标与表征结果联系起来。我们不仅指出黑人报纸在基准测试中代表性不足，还揭示了这种代表性不足是如何由以下因素造成的：忽视版面布局的指标选择、排除历史字体排印的训练数据集，以及在评估过程中完全排除社区档案的机构结构。这种排除既妨碍人们在批判性使用潜在 OCR 工具的同时深入了解数字化档案，也增加了遇到先前提及的过度历史化等问题的可能性 [26]。

3 系统性综述方法

我们依据 PRISMA 2020 原则 [40] 对光学字符识别评估方法进行了系统调查和综述；图1展示了本次调查的工作流程。我们的分析涵盖 2006年–2025年，涉及文档理解领域的 Transformer 革命直至当前的视觉语言模型，并采用 PRISMA 为以评估为重点的机器学习研究提供透明的纳入、排除和报告标准。本研究围绕三个研究问题展开：

(1) 最先进的光学字符识别系统使用哪些数据进行训练？评估研究如何报告这些数据？(2) 将最先进水平的光学字符识别系统应用于黑人历史字体排印和版面布局时，会出现哪些失效模式？(3) 需要哪些评估、数据和资源转变，才能确保文化结构得到保留且准确无误？

3.1 检索与筛选

我们检索了六个数据库，共获得 80 篇论文：ACM 数字图书馆（n=15）、IEEE Xplore（n=12）、arXiv（n=18）、ACL 文集（n=10）、Springer（n=8）以及其他来源（Google 学术、Semantic Scholar 等）（n=17）。我们将使用的检索式分为四类：主要评估（例如，“光学字符识别评估指标 文档理解”“文档人工智能基准评估”）、历史与偏见意识（例如，“历史文献 OCR 评估”“报纸数字化 OCR 评估”）、特定模型（例如，“视觉 Transformer 光学字符识别评估”）以及负对照（验证）；检索式完整列表见附录 A。删除 2 条重复项后，剩余 78 条唯一记录进入筛选。

筛选后，我们应用纳入标准，保留符合以下条件的论文：（1）将使用定量指标展示或评估光学字符识别系统作为主要贡献；（2）提出或分析评估基准；或（3）概述光学字符识别评估流程。

3.2 数据提取

对于每个系统，我们提取了模型架构、训练数据来源、数据规模与透明度、语言覆盖范围以及历史文档覆盖情况。表1展示了对13个主要系统的分析，并指出各系统在报告透明度方面存在差异：许多系统仅将训练数据描述为“未公开”或“网络规模”。对于每个基准，我们记录了其支持的语言、文档类型、历史文档覆盖范围、退化测试，以及是否明确纳入代表性不足的档案或社区档案。表3概述了主要基准，并指出基准范围与历史档案需求之间的差距。

4 系统性综述结果

4.1 当前 OCR 评估衡量什么，又遗漏了什么

表1列出了13种 OCR 系统的训练来源，其中有六种模型依赖 IIT-CDIP 数据库，该数据库收录了4200万份20世纪中叶的烟草诉讼文档。其余模型使用合成数据、科学类 PDF（PubMed Central）或组成各异、覆盖网络规模的语料库。olmOCR 2 使用一种名为 olmOCR-mix-1025 的专用数据组合，其中包括来自互联网档案馆和 S2PDF 的广泛数据，以及一组20,000份历史文档和手写文档。该模型经过优化，能够处理档案材料中常见的复杂、劣化扫描件，例如

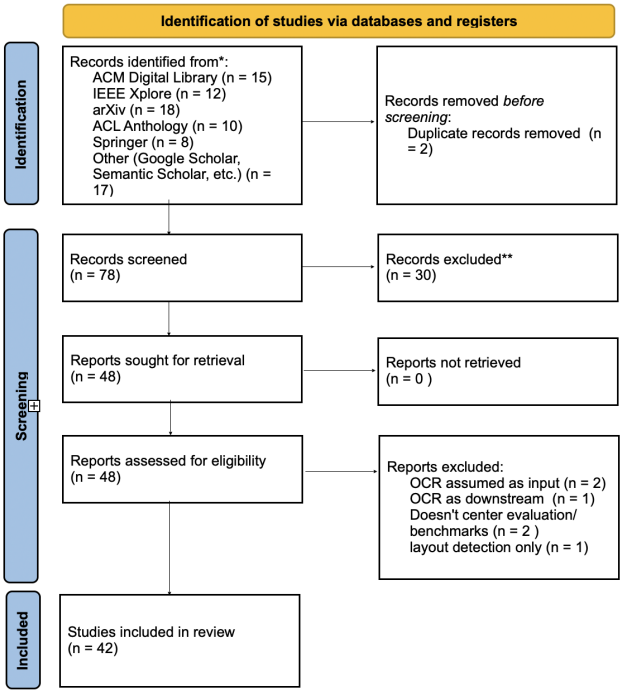


图1：用于系统性回顾文档理解中 OCR 评估方法与指标（2006–2025年）的 PRISMA 2020 流程图。

美国国会图书馆的19世纪书信，并采用可验证奖励强化学习（RLVR）。报告中列出的训练数据包括科学类出版物（360K+）、企业与法律文档（4200万页）以及现代数字 PDF。从这些数据来源的分布可以清楚看出，历史黑人报纸、社区出版物、哥特体和黑体字体，以及非西方文档版式，都是缺失或未报告的模型训练数据示例。

表2显示，评估覆盖范围并不均衡：只有三个系统报告了历史扫描件的定量结果。例如，olmOCR 2报告了旧版数学文档扫描件82.3%的准确率，但在一般历史扫描件上的准确率降至47.7%。相比之下，GOT-OCR 2.0在同一olmOCR-Bench上的准确率仅为22.1%，而TrOCR即使经过微调，字符错误率仍为5.7%。大多数有关历史字体排印的条目都标注为“未评估”或“需要微调”，而现代能力则有更详尽的记录。

表3总结了七个基准，并按语言覆盖范围、文档类型、是否纳入历史文献以及退化测试进行分类，凸显了这些基准对历史文献和社区制作文档的关注有限。例如，CC-OCR（ICCV 2025）支持28种以上语言，侧重现代应用；而OmniDocBench（CVPR 2025）虽评估模糊、旋转和水印，却没有档案研究重点。KITAB-Bench独树一帜，明确纳入历史阿拉伯语文档，并采用人在回路审核。

表1：主要光学字符识别系统的训练数据来源与历史文档覆盖范围（2021年–2025年）

模型	Year	训练来源	规模	参数	文字体系	历史封面 age
olmOCR 2	2025	网页 PDF 和互联网档案	~260 千页	~7B	EN+	大量 (IA 扫描件；未筛选)
PaddleOCR-VL	2025	合成与网络语料-pora	未公开	~0.9B	109+	有限（无历史-cal eval）
Qwen2.5-VL	2025	网络多模态 cor-pora	网络规模	70亿/72B	多种	未报告
GOT-OCR 2.0	2024	合成 + 现代 web	~5百万配对数据	~580百万	EN	无（合成文档 cus）
Surya	2024	扫描文档图像	未公开	未公开。	90+	有限（无基准测试-标记）
DocFormer v2	2024	文档图像（IIT-CDIP）	~11M 张图像	200–500M	EN	未报告
Nougat	2023	PubMed PDFs 集中式	~1M 文档	~350M	EN	None （学术 PDF文件）
UDoc	2021	IIT-CDIP 区域	未公开	未公开。	EN	None
LayoutLMv3	2022	IIT-CDIP	~11M 文档	133/368M	EN	部分 (诉讼文档)
DiT	2022	IIT-CDIP	~42M 张图像	33–675M	EN	部分（结构 fo-cus）
Donut	2022	IIT-CDIP +Synth-DoG	~13M 张图像	~200M	4 语言	无（合成）
TrOCR	2021	合成 + IAM	未公开	~558M	EN	仅微调
Tesseract v5	2021	合成字体	各不相同	N/A	100+	需要自定义训练要

表2：关键文档分析系统（2025年）的比较评估

系统	架构	历史扫描件	印刷体-phy	复杂版面布局	手写文字
生成式视觉语言模型（端到端文本生成）					
olmOCR 2	视觉语言模型（70亿）	**82.3%**（数学） 47.7%（通用）	支持	83.7%（多种 col)	支持（20k 页）
Qwen2.5-VL	视觉语言模型（70亿+）	65.5%（基准测试）	高（生成能力）	**强**	非凡 (手写测试子集)
GOT-OCR 2.0	编码器-Dec.	22.1%（噪声下失败）	Poor	Poor （re-已报告）	未评估
Nougat	Transformer	差（产生幻觉）	N/A	学术 Only	Poor
Donut	Transformer	差	Poor	Poor	Weak
TrOCR（L）	Transformer	5–7% 字符错误率（微调）	微调型-可	N/A （行-级）	**2.89%** **CER** (IAM)
判别式 / 混合式（检测与识别）					
Surya	混合式（分割）	81%（历史泰米尔语）†	中等	中等	未评估
PaddleOCR v5	两阶段	中等（现代）	中等	良好（表格）	高（82–89% 准确率）
Tesseract v5	LSTM	高词错误率（性能下降）	需要微调	Poor	Poor
视觉文档理解（需要外部光学字符识别）					
DocFormer v2	多模态	依赖输入 *OCR*	依赖输入光学字符识别	中等	N/A
LayoutLMv3	多模态	依赖输入 *OCR*	依赖输入入 *OCR*	**SOTA**	N/A
UDoc / DiT	图像变换器	依赖输入 *OCR*	依赖输入 *OCR*	强	N/A

†“依赖输入 *OCR*”表示对版面布局进行分类、但自身不生成文本的模型。

4.2 黑人报刊作为案例研究

为了进一步检验 OCR 系统开发在文化相关数字化历史档案方面的不足，我们将综述中提到的三个模型进行比较，让它们识读《The Weekly Advocate》（1837）的一页内容。这是最早的非裔美国人报纸之一，

表3：主要光学字符识别和文档理解基准的评估覆盖范围

基准	语言	文档类型	标注部分-范式	历史／档案覆盖范围	代表性不足档案
CC-OCR	多种 (28+)	网页 PDF，场景，文档	纯文本	无（现代网络爬取）	无特定侧重
OmniDoc	英语，中文	学术，幻灯片，书籍	检测与文本	现代时期（原生数字化）	没有针对性侧重
OCRBench	英语、中文	31 场景（混合）	视觉问答 / 键值	有限（孤立 IAM/HKR）	未明确关注
DUDE	主要为英语	多页商业	问答对	当代（扫描件）	无特定关注重点
DUE	英语	表格、图表	问答／分类	None	无特定关注重点
DocVQA	英语	行业表单（烟草）	问答对（ANLS）	历史诉讼（20世纪）	单领域
KITAB*	阿拉伯语	手稿，编年史	转录	高（主要 Focus）	阿拉伯遗产

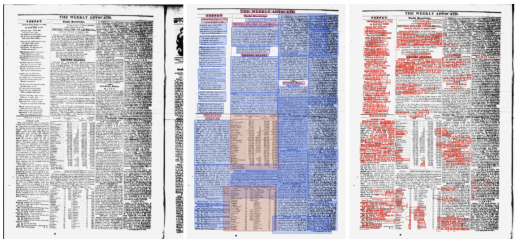
*指 *OpenITI/KITAB* 项目中针对古典阿拉伯语文本的评估子集。

该报刊于纽约市。在此，我们通过这一案例研究超越整体基准测试分数，观察评估方面的不足会如何导致这些 OCR 系统在自动转录和解读档案时出现具体问题。我们测试了表2所列的三个系统：Tesseract v5（一种传统混合式流水线）、Surya（一种以版面为中心的现代系统）和 olmOCR 2（一种最先进的视觉语言模型）。我们在测试中观察到每个系统都呈现出不同的失效模式，详见表4。

表4：*The Weekly Advocate*（1837）中的失效模式

错误机制	模型	技术故障	档案后果
几何折叠（线性偏差）	Tesseract v5	误识别竖排栏 分栏线为文本；默认采用从左到右的“Z 字形阅读模式”阅读顺序	不同内容的语义合并 内容（社论诗歌 con-与市政报告融合）；颗粒度搜索变得无效
解码不稳定性	Surya	高对比度的 19 世纪字体触发 导致词元坍缩；生成字符序列的重复序列	引入“垃圾词元”
幻觉—纠正性 tion	olmOCR 2（最先进的视觉语言模型）	用视觉证据覆盖了高概率标记（“知识渗漏？”）	实体被替换成了虚构文本；

图2：原始版面与 Surya 版面分析和 Tesseract 的对比



我们的分析表明，已报告的最先进水平（SOTA）光学字符识别（OCR）系统的性能，与这些系统用于历史档案文献时的实用性之间存在显著差距；这些文献体现了社会、文化和物质现实，例如历史黑人报纸所呈现的现实。我们指出，当前的评估实践在很大程度上缺乏捕捉历史文献准确性所需的“指标多元主义”。尽管已有具备版面布局感知能力的模型和特定任务的版面评估，但它们通常是在网络规模或企业级资料集（如 IIT-CDIP ）上训练的，随后再与基准进行比较；这些基准优先考虑现代版面布局和视觉上明显的退化，而非历史结构和阅读顺序。我们的案例研究表明，尽管模型报告的字符识别得分令人满意，却常常生成幻觉内容，或无法保留多栏黑人报纸的阅读顺序。当前评估方法无法识别表征伤害，因此模型即便忽视了历史黑人印刷文化所塑造的社会和物质限制，仍可能被认定为“成功”。

5 讨论

5.1 评估差距

目前可用的评估流程存在算法性隐形问题：基准测试并未报告历史上代表性不足的文献所特有的失效模式。尽管 Chronicling America 明确聚焦于美国的数字档案，但与 DocVQA 等基准相比，它仍未得到充分利用。我们的案例研究通过评估黑人数字档案中的一份报纸《The Weekly Advocate》，印证了这一观察。该报展现出多样的版面布局和字体排印，也凸显了光学字符识别模型出现的失效如何被忽视。因此，若不有意将此类文献纳入评估系统和训练流程，这些系统就会继续把黑人文化遗产视为例外，而非核心领域。

5.2 合成数据与真实数据集

合成数据增强与历史性劣化之间存在重大差异。光学字符识别模型通常能够很好地处理高斯噪声或人工斑点噪声等合成扰动。相比之下，早期黑人报刊具有与材料特性相关的伪影，包括缩微胶片上的光照渐变、物理翘曲以及不可逆的二值化损失。正如我们的案例研究所示，那些看似能够抵御合成噪声的模型，在面对这类结构性劣化时往往会彻底失效，造成一种误导性的可靠感，而这种可靠性并不适用于真实世界中的档案数字化。

5.3 技术准确性与文化和历史保护

我们观察到，标准的OCR评估标准可能会造成一种难以化解的矛盾：究竟是优化文本的“准确性”或“整洁性”，还是保留历史文献作为文化档案。这些历史文献的版面布局通过复杂的多栏版式体现其编排逻辑，最大限度地利用稀缺的印刷空间，并且能够传达额外的含义 [25]。最先进水平的系统随后

表5：部分黑人报刊文献的结构、时间与档案特征。

出版物／系列	年份	版面布局拓扑	报头／页眉	非文本元素	字体排印复杂度	扫描退化
The Weekly Advocate 与《有色美国人报》的《**有色美国人报**》（过渡期性报纸）	1837—1841年	4 栏，标准网格 经常被……打断 广告和跨栏 标题	高复杂度、丝带 横幅，带有多种 嵌入文字（例如，“Established for ...”和粗体衬线 带副文本的标题	高方差、厚重的水平线- 水平分栏线、木刻版画 美国国会大厦的雕刻 和图像象形字 见于广告中	中等至高，现代 衬线字体（迪多风格）、紧密行距， 古体连字	透印
Mirror of Liberty（季刊杂志）	1838–1841	2栏，较宽的一栏 宽度、杂志宽高比 比	复杂度低、简单 字体排印式标题、大纲 块状字体	极简，主要为文本- 粗重、稀疏分隔线	中等、字号较大， 清晰的字体排印层级	低对比度、字符 缺失
《**北极星报**》（废奴主义者）	1847–1851	7 栏，极其密集 版面布 局	中等，具有印刷术风格 报头常被 扫描伪影	结构性强调，细线 以及断续的垂直分隔线	极高，极小字号 字号、排版紧凑， 斜体用于强调	模糊、斑点噪声、倾斜
The Impartial Citizen	1861–1864	5栏，紧密的网格 页边距略宽	高，哥特体／黑体 报头与……不同 罗马体正文	标准竖排栏 分栏 线	高密度文本块 对行分割构成挑战	来自……的光照渐变 缩微胶片
The Weekly Anglo-African	1859–1861	7 栏，页眉居中 结构	中等，详细的报头- 带有嵌入文字的页眉	嵌入式艺术作品、格言 嵌入在雕刻中	混用展示字体 页眉与正文的对比 text	线条比其他内容更简洁 样本

强制采用线性阅读顺序，或试图减少页边批注等“噪声”，以提高字符级准确率，却抹去了文献的语义结构。因此，未来的基准必须超越词错误率（WER），纳入具备版面布局感知能力且注重保真度的评估标准，从而维护档案记录的完整性。

6 局限性与未来工作

我们的研究采用了一个范围较小且具有特定性的案例研究，旨在凸显结构上的多样性，而非统计准确性。我们着重进行定性失败分析，而不是使用真实标注转录文本；这也反映了历史档案的限制，使得难以直接通过字符级准确率指标进行比较。我们还聚焦于美国的十九世纪黑人报纸及其特有的劣化情况和排版模式，这些发现未必适用于其他历史或语言背景。最后，我们的评估优先考虑档案层面的影响，而非特定任务的表现，因此布局与保真度感知指标的开发仍有待未来研究。

此外，如何以最佳方式鼓励人们接触这些重要历史档案，仍是一个关键问题。尽管理解这些档案的内容及其含义（参见[22]）十分重要，也确实有益，但面对档案可能遭到的不当使用，究竟应让这些档案开放到何种程度，仍是一个值得探讨的问题，其中包括过度历史化等潜在问题[26]。在探索各种黑人解放主义馆藏[12], 所面临的这些关键问题时，我们希望谨记意义如何跨越时间传递，同时关注意义曾如何被刻意遮蔽[34], 以及如何在这项工作中延续这些传统（及其局限）。

7 结论

要为早期黑人报刊打造有效的光学字符识别系统，面临的重大挑战不仅在于技术能力，还涉及资源分配、数据集多样性以及对文化意义的认识。尽管现代系统能够解析复杂的历史版面布局，但它们很少针对黑人报纸多样的排版与文档条件进行评估，这会导致编辑逻辑遭到抹除，并引发对社群优先事项的质疑。本文主张，评估实践应将黑人数字档案视为关键测试数据集的一部分，而非边缘案例，以开展具有批判性且对文化负责的光学字符识别评估和基准系统。尽管纳入这些档案存在局限（而纳入行为本身也可能具有掠夺性[14, 49]），但这一改变或许能帮助我们更深入地理解这些历史档案，并探索不同的未来。

8 生成式人工智能使用声明

尽管生成式人工智能系统是本文研究主题的一部分（针对 OCR 系统），但据作者所知，他们在撰写本文时并未以任何方式使用生成式人工智能。

参考文献

[1] Shuai Bai、Kai Chen、Xiaohui Liu、Jianfeng Wang、Wenqi Ge、Shiqi Song 和 Jingdong Lin。2025年。Qwen2.5-VL 技术报告。*arXiv*（2025年）。https://arxiv.org/abs/ 2502 13923 。[2] Aleksandar Beshirov、Siyana Hadzhieva、Ivan Koychev 和 Milena Dobreva。2022年。DuoSearch：一种用于保加利亚历史文献的新型搜索引擎。载于第*44*届欧洲信息检索会议（*ECIR 2022*）论文集。Springer，沙姆，瑞士，265 269 10 1007 978 3 030 99739 7_31 – 。doi: . / - - - - [3] Fitsum Sileshi Beyene 和 Christopher L. Dancy。2025。面向黑人数字档案、采用无监督评估的布局感知光学字符识别。载于*IEEE*技术与社会国际研讨会（*ISTAS 2025*）论文集。IEEE，1–5。doi:10.1109/ISTAS65609.2025.11269615[4] Abeba Birhane、Pratyusha Kalluri、Dallas Card、William Agnew、Ravit Dotan 和 Michelle Bao。2022年。机器学习研究中编码的价值观。载于*2022*年*ACM*公平性、问责制与透明度会议论文集（大韩民国首尔）（*FAccT '22*）。计算机协会，纽约，美国，173 184 10 1145 3531146 3533083 – 。doi: . / . Abeba Birhane、[5] Elayne Ruane、Thomas Laurent、Matthew S. Brown、Johnathan Flowers、Anthony Ventresque 和 Christopher L. Dancy。2022年。人工智能伦理中被遗忘的边缘。载于*2022*年*ACM*公平性、问责制与透明度会议论文集（大韩民国首尔）（*FAccT '22*）。计算机协会，纽约，美国，948–958。doi:10.1145/3531146.3533157[6] Lukas Blecher 和 Graham Neubig。2023年。Nougat：面向学术文档的神经光学理解。*arXiv*（2023年）。https://arxiv.org/abs/2308.13418[7] Łukasz Borchmann、Michał Pietruszka、Tomasz Stanisławek、Dawid Jurkiewicz、Michał Turski、Karolina Szyndler 和 Filip Graliński。2021年。DUE：端到端文档理解基准。载于*NeurIPS*数据集与基准赛道论文集。神经信息处理系统基金会。https://datasetsbenchmarksproceedings.neurips.cc/paper/2021/file/069059b7ef840f0c74a814ec9237b6ec-Paper.pdf[8] Curtis M. Bunn。2024年。霍华德大学正在将数千份黑人报纸数字化。https://www.nbcnews.com/news/nbcblk/howard-university-digitize-archive-thousands-black-newspapers-rcna17185 检索日期：2026年1月14日。[9] J ørgen Burchardt。2023年。光学字符识别生成的档案中的搜索可靠吗？数字报纸档案分析。经济史年鉴 / 经济史年鉴 64, 1 2023 31 54 10 1515 2023 0003 ( ), – 。doi: . / [10] jbwg- - John Casey。2022。。“我们需要一家报社——一家属于我们自己的报社”：废奴运动之后的黑人报刊。内战史 68, 2 2022 117 130 10 1353 2022 0010 ( ), – 。doi: . / [11] cwh. . Colored Conventions Project。[ ]无日期。。关于 Colored Conventions Project。https://omeka.coloredconventions.org/about 检索日期：2026年1月14日。[12] Payton Croskey、Fabian Offert、Jennifer Jacobs 和 Kai M. Thaler。2025年。解放性馆藏与合乎伦理的人工智能：从黑人社区档案和数据集中重新构想人工智能开发。载于*2025*年*ACM*公平性、问责制与透明度会议论文集（*FAccT '25*）。计算机协会，纽约，美国，900–913。doi:10.1145/3715275.3732058[13] Cheng Cui、Tianxiang Sun、Shiyao Liang、Tao Gao、Zhe Zhang、Jiajun Liu 和 Yu Ma。2025年。PaddleOCR-VL：借助0.9B超紧凑视觉语言模型提升多语言文档解析能力。*arXiv*（2025年）。https://arxiv.org/abs/2510. 14528[14] C. L. Dancy 和 P. K. Saucier。2022年。人工智能与黑人身份：迈向超越偏见与表征。*IEEE*《技术与社会汇刊》3，1（2022年），31–40。doi:10.1109/TTS.2021.3125998[15] Melissa Dell、Jacob Carlson、Tom Bryan、Emily Silcock、Abhishek Arora、Zejiang Shen、Luca D'Amico-Wong、Quan Le、Pablo Querubin 和 Leander Heldring。2023年。美国故事：美国历史报纸的大规模结构化文本数据集。载于*NeurIPS*数据集与基准赛道论文集。神经信息处理系统基金会。https://proceedings.neurips.cc/paper_files/paper/2023/file/ffeb860479ccae44d84c0de32acd693d-Paper-Datasets_和_Benchmarks.pdf[16] Kim Gallon。2016年。论证黑人数字人文的必要性。https://dhdebates.gc.cuny.edu/read/untitled/section/fa10e2e1-0c3d-4519-a958-d823aac989eb 载于 *数字人文论争*。[17] Kim Gallon。2021年。黑人报刊。https://oxfordre.com/americanhistory/view/10.1093/acrefore/9780199329175.001.0001/acrefore-9780199329175-e-851 检索日期：2026年1月14日。[18] Ahmed Heakl、Muhammad Abdullah Sohail、Mukul Ranjan、Rania Elbadry、Ghazi Shazan Ahmad、Mohamed El-Geish、Omar Maher、Zhiqiang Shen、Fahad Shahbaz Khan 和 Salman Khan。2025年。KITAB-Bench：用于阿拉伯语光学字符识别和文档理解的综合性多领域基准。载于计算语言学协会：*ACL 2025*论文集。计算语言学协会，22006 22024 10 18653 1 2025 1135 – 。doi: . /v / .findings-acl。[19] Rose Holley。2009。。效果能有多好？分析并提升大规模历史报纸数字化项目中的光学字符识别准确率。*D Lib Mag.* 15（2009 2568435 ）。[20] https://api.semanticscholar.org/CorpusID: 霍华德大学穆兰-斯平加恩研究中心。2025年。黑人新闻档案。https://msrc.howard.edu/black-press-archive。馆藏包含超过2,000个报纸标题、100,000期和2,847卷微缩胶卷。[21] Yupan Huang、Tengchao Lv、Lei Cui、Yutong Lu 和 Furu Wei。2022年。Lay- outLMv3：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（*MM '22*）。ACM，纽约，美国纽约州，4083–4091。doi:10.1145/3503161.3548112[22] Jackson, Ketanji Brown。2024年。案件编号 *23–1275*中的反对意见。美国最高法院。https://www.supremecourt.gov/opinions/24pdf/23- 1275_e2pg.pdf 反对意见。[23] Geewook Kim、Teakgyu Hong、Moonbin Yim、JeongYeon Nam、Jinyoung Park、Jinyeong Yim、Wonseok Hwang、Sangdoo Yun、Dongyoon Han 和 Seunghyun Park。2022年。无 OCR 文档理解 Transformer。载于计算机视觉——*ECCV 2022*：第*17*届欧洲会议，特拉维夫，以色列，*2022*年*10*月*23*日至*27*日，论文集，第*XXVIII*部分（特拉维夫，以色列）。施普林格出版社，柏林，海德堡，498–517。doi:10.1007/978-3-031-19815-1 29_[24] Benjamin Charles Germain Lee、Jaime Mears、Eileen Jakeway、Meghan Ferriter、Chris Adams、Nathan Yarasavage、Deborah Thomas、Kate Zwaard 和 Daniel S. Weld。2020年。报纸导航数据集：从 Chronicling America 的 1600 万个历史报纸页面中提取新闻标题和视觉内容。载于第*29*届*ACM*国际信息与知识管理会议论文集（*CIKM '20*）。ACM，3055–3062。doi:10.1145/3340531.3412767[25] B. C. G. Lee、J. Ortiz Baco、S. H. Salter 和 J. Casey。2021年。探索版面编排：针对多族裔期刊视觉版式的解释性机器学习方法。载于*CHR 2021*：计算人文研究会议，。[26] Maria Levchenko。2024年。借助多模态大语言模型构建历史语料库：18世纪俄罗斯书籍中的认知缺口与误读。载于*ACH* 文集。计算机与人文协会，美国。https: //doi.org/10.63744/SKoZVUHQbtE7/[27] David Lewis、Gady Agam、Shlomo Argamon、Ophir Frieder、David Grossman 和 Jonathan Heard。2006年。构建复杂文档信息处理测试集。载于第*29*届*ACM*国际信息检索研究与开发年会论文集（美国华盛顿州西雅图）（*SIGIR '06*）。ACM，纽约，美国纽约州，665–666。doi:10.1145/1148170.1148307[28] Minghao Li、Tengchao Lv、Jingye Chen、Lei Cui、Yijuan Lu、Dinei Florencio、Cha Zhang、Zhoujun Li 和 Furu Wei。2023年。TrOCR：基于预训练模型的 Transformer 光学字符识别。载于第三十七届 *AAAI* 人工智能会议论文集（*AAAI'23*）。AAAI 出版社，文章编号1469，共9页。doi:10.1609/aaai.v37i11.26538[29] Minghao Li、Yiheng Xu、Lei Cui、Shaohan Huang、Furu Wei、Zhoujun Li 和 Ming Zhou。2020年。DocBank：文档版面分析基准数据集。载于第*28*届国际计算语言学会议论文集（*COLING 2020*）。国际计算语言学委员会，西班牙巴塞罗那（线上），949 960 10 18653 1 2020 82 – 。doi: . /v / . [30] coling-main. 2025美国国会图书馆。Chronicling America：非裔美国人报纸。https://www.loc.gov/collections/chronicling-america/titles/。按主题族裔筛选：非裔美国人；出版日期为1777年至1963年。[31] 美国国会图书馆。2025年。通过 OCR 再处理改进 Chronicling America 的搜索。https://blogs.loc.gov/headlinesandheroes/2025/04/ocr- reprocessing/。[32] Yuhang Liu、Zhe Li、Minghao Huang 等。2024年。OCRBench：大型多模态模型中 OCR 的隐藏奥秘。中国科学：信息科学 67, 220102 2024 10 1007 11432 024 4235 6 文章（）。doi: . /s - - - [33] Minesh Mathew、Dimosthenis Karatzas 和 C. V. Jawahar。2021年。DocVQA：用于文档图像视觉问答的数据集。载于*IEEE* 冬季计算机视觉应用会议论文集（*WACV 2021*）。IEEE，2199–2208。doi:10.1109/WACV48630.2021.00225[34] R. Morrison。2022年。《丰盈的解体：黑人计算思想的未来史》。*DHQ*：数字人文季刊16卷3期（2022年）。[35] NBC 新闻经 BET 工作人员报道。2022年。霍华德大学获得拨款，用于归档黑人报纸。https://www.bet.com/article/u49dcv/howard-university-receives-grant-to-archive-black-newspapers。报道霍华德大学获得200万美元拨款，用于将黑人新闻档案中的60%数字化，内容依据 NBC 新闻报道。[36] Clemens Neudecker、Konstantin Baierer、Mike Gerber、Christian Clausner、Apostolos Antonacopoulos 和 Stefan Pletschacher。2021年。OCR 评估工具与指标综述。载于第*6*届历史文档成像与处理国际研讨会论文集（瑞士洛桑）（*HIP '21*）。ACM，纽约，美国纽约州，13–18。doi:10.1145/3476887.3476888[37] Thi Tuyet Hai Nguyen、Adam Jatowt、Mickael Coustaty 和 Antoine Doucet。2021年。OCR 后处理方法综述。*ACM* 计算调查54卷6期，文章 124 2021 37 10 1145 3453476（7月），页数。doi: . / S. U. Noble。[38] 2018。《压迫算法：搜索引擎如何强化种族主义》。纽约大学出版社，纽约，美国纽约州。[39] Linke Ouyang、Yuan Qu、Hongbin Zhou、Jiawei Zhu、Rui Zhang、Qunshu Lin、Bin Wang、Zhiyuan Zhao、Man Jiang、Xiaomeng Zhao、Jin Shi、Fan Wu、Pei Chu，

Minghao Liu、Zhenxiang Li、Chao Xu、Bo Zhang、Botian Shi、Zhongying Tu 和 Conghui He。2025年。OmniDocBench：通过全面标注对多样化 PDF 文档解析进行基准测试。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集（*CVPR 2025*）。IEEE，24838–24848。doi:10.1109/CVPR52734.2025.02313

[40] Matthew J. Page、Joanne E. McKenzie、Patrick M. Bossuyt、Isabelle Boutron、Tammy C. Hoffmann、Cynthia D. Mulrow、Larissa Shamseer、Jennifer M. Tetzlaff、Elie A. Akl、Sue E. Brennan、Roger Chou、Julie Glanville、Jeremy M. Grimshaw、Asbjørn Hróbjartsson、Manoj M. Lalu、Tianjing Li、Elizabeth W. Loder、Evan Mayo- Wilson、Steve McDonald、Luke A. McGuinness、Lesley A. Stewart、James Thomas、Andrea C. Tricco、Vivian A. Welch、Penny Whiting 和 David Moher。2021年。《PRISMA 2020 声明：系统综述报告指南更新》。*BMJ* 372，文章 n71（2021年）。doi:10.1136/bmj.n71

[41] Birgit Pfitzmann、Christoph Auer、Michele Dolfi、Ahmed S.Nassar 和 Peter Staar。2022年。DocLayNet：用于文档版面分割的大型人工标注数据集。载于第 *28* 届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集（*KDD '22*）。计算机协会，纽约，纽约州，美国，3743–3751。doi:10.1145/3534678.3539043

[42] Adam Pozanski 等。2025。olmOCR：利用联网档案馆数据扩展 PDF 理解能力。*arXiv*（2025年）。https://arxiv.org/abs/2502.18443

[43] Christian Reul、Dennis Christ、Alexander Hartelt、Nico Balbach、Maximilian Wehner、Christoph Wick、Christine Grundig、Andreas Büttner 和 Frank Puppe。2019年。OCR4all——一款为历史印刷品提供（半）自动 OCR 工作流的开源工具。*Applied Sciences* 9 卷 22 期，文章编号 4853（2019年）。doi: 10 3390 9224853 。/app Shruti Rijhwani、Antonios Anastasopoulos 和 Graham Neubig。[44] 2020年。濒危语言文本的 OCR 后校正。载于*2020*年自然语言处理实证方法会议论文集。计算语言学协会，线上，, 5931 5942 10 18653 1 2020 478 – 。doi: 。/v / 。emnlp-main。David A. Smith 、 Ryan Cordell 和 Elizabeth Maddock Dillon。[45] 。传染性文本：十九世纪报纸中的文本复用建模。载于*2013* 年 *IEEE* 国际大数据会议论文集。IEEE，86–94。doi:10.1109/BigData. 2013 6691675 。R. Smith。 。Tesseract OCR 引擎概述。载于第九届国际文档分析与识别会议（*ICDAR 2007*），第 2 卷。629–633。doi:10.1109/ICDAR.2007.4376991[46] 2007 Uwe Springmann、Christian Reul、Stefanie Dipper 和 Johannes Baiter。2018年。用于训练 OCR 引擎识别德语哥特体和早期现代拉丁语历史文献的基准真值。*arXiv*（ 2018 1809 05501 ）。[47] https://arxiv.org/abs/ 。Giulia Taurino 和 David A. Smith。[48] 2022 。作为档案科学的机器学习：人工智能、文化数据和档案修复背后的叙事。载于*NeurIPS* 人工智能文化研讨会。https://ai-cultures.github.io/papers/machine_learning_as_an_archiva.pdf[49] K.-Y. Taylor。2019年。逐利的种族歧视：银行和房地产行业如何削弱黑人住房所有权。北卡罗来纳大学出版社，教堂山，北卡罗来纳州。[50] The Crowley Company。2024年。保存声音：将霍华德大学黑人报纸馆藏数字化。https://thecrowleycompany.com/preserving-voices-digitizing-howard-universitys-black-newspaper-collection/。介绍黑人新闻档案的多年数字化工作。[51] Jordy Van Landeghem、Rafał Powalski、Rubèn Tito、Dawid Jurkiewicz、Matthew Blaschko、Łukasz Borchmann、Mickaël Coustaty、Sien Moens、Michał Pietruszka、Bertrand Ackaert、Tomasz Stanisławek、Paweł Józiak 和 Ernest Valveny。2023年。文档理解数据集与评估（DUDE）。载于*IEEE/CVF* 国际计算机视觉会议论文集。IEEE，19471–19483。doi:10.1109/ICCV51070.2023.01789[52] Hao Wei、Chao Liu、Jingye Chen、Jianfeng Wang、Lingpeng Kong、Yiheng Xu 和 Xiaohui Zhang。2024年。通用 OCR 理论：通过统一端到端模型迈向 OCR-2.0。*arXiv*（ 2024 2409 01704 ）。https://arxiv.org/abs/ 。Deja Workman 和 Christopher L. Dancy。[53] 2023 。识别人工智能开发流程中可能存在的人的切入点：人工智能开发中的人类与反黑人主义。载于*2023* 年计算机支持的协同工作与社交计算会议伴随出版物（*CSCW '23 Companion*）。计算机协会，纽约，纽约州，美国，348–353。doi:10.1145/3584931.3606981[54] Yiheng Xu、Minghao Li、Lei Cui、Shaohan Huang、Furu Wei 和 Ming Zhou。2020年。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* 国际知识发现与数据挖掘会议论文集（*KDD '20*）。计算机协会，纽约，纽约州，美国，1192–1200。doi:10.1145/3394486.3403172[55] Yang Xu、Yiheng Xu、Tengchao Lv、Lei Cui、Furu Wei、Guoxin Wang、Yijuan Lu、Dinei Florencio、Cha Zhang、Wanxiang Che、Min Zhang 和 Lidong Zhou。2021年。LayoutLMv2：面向视觉丰富文档理解的多模态预训练。载于计算语言学协会第 *59* 届年会暨第 *11* 届自然语言处理国际联合会议论文集（*ACL-IJCNLP 2021*）。计算语言学协会。doi:10.18653/v1/2021.acl-long.201

[56] Xingjiao Zhong、Jianbin Tang 和 Antonio Jimeno Yepes . 2019. PubLayNet：文档版面分析领域迄今最大的数据集。载于论文集第 *15* 届文档分析与识别国际会议（*ICDAR 2019*）。IEEE。https://conferences.computer.org/icdar/2019/pdfs/ICDAR2019-5vPIU32iQjjaLtHlc8g8pO/7doDqZ3WIuLF2RHKwraE4k/1rPcoTBWZJldKF8sPWivxO.pdf

A 附录 A：PRISMA 2020 系统综述所用的检索式

表6：PRISMA 2020 光学字符识别评估文献系统综述中使用的检索式。

查询类别		检索式
主要查询	评估	“光学字符识别评估指标 文档理解” “文档人工智能基准评估-估” “OCR 基准数据集同行评审” “文档版面分析基准” “视觉语言模型光学字符识别评估”
历史与偏差感知查询		“历史文档光学字符识别评估” “光学字符识别偏差 历史文献” “报纸数字化 光学字符识别评估度” “版面保留 光学字符识别评估度” “文化遗产文献分析 OCR”
模型类别查询	具体	“视觉 Transformer 光学字符识别评估” “多模态文档理解基准” “端到端文档理解 ing 评估”
排除检查查询	/ 合理性-	“合成 文档光学字符识别基准测试-标记” “手写文本识别评估数据集”