生成式人工智能视角下的人机协同档案整理路径研究

——以清华大学图书馆藏民国馆史文档为例

董琳* 许澜洋

摘要 清华大学图书馆收藏有民国时期馆史文件档案超万页，这些文档常见字迹模糊缺损、手写体笔迹辨识困难、多语种等情况，长期以来其整理工作面临低效能、低产出的问题，已经数字化的文献也多以关键词和图片方式呈现，缺少全文信息。近年来生成式人工智能的发展为突破馆史文档的整理难题提供了新的机遇。文章基于清华大学图书馆藏民国馆史文档整理工作，以其中的复杂文档为例，聚焦字符识别环节，通过与简单人机交互模式进行对照研究，阐述了人机协同模式下，人与AI的分工协作关系，以及人机合作的原理、工作重点与流程。研究表明，人机协同模式下所得字符识别结果集较纯人工或简单人机交互模式，具有明显的效率优势和应用价值。

关键词 人工智能 人机协同 字符识别 民国 文件 档案

分类号 G254 G272

DOI 10.16603/j.issn1002－1027.2026.04.011

引用本文格式 董琳，许澜洋.生成式人工智能视角下的人机协同档案整理路径研究——以清华大学图书馆藏民国馆史文档为例[J].大学图书馆学报，2026，44(4)：92－102.

1 引言

2025年8月，国务院发布《关于深入实施“人工智能＋”行动的意见》[1]，明确提出行动目标为“推动人工智能与经济社会各行业各领域广泛深度融合，重塑人类生产生活范式，促进生产力革命性跃迁和生产关系深层次变革，加快形成人机协同、跨界融合、共创分享的智能经济和智能社会新形态”。

图书馆作为人类文明的存藏地之一，长期致力于知识、信息等图文资料的存藏、整理和服务工作，致力于为用户文化生活提供高质量的文献资源。在国家推进人工智能网络布局的背景下，AI技术正在深刻改变图书馆的运作模式和服务方式，推动其从传统书库到智慧图书馆的转型升级。随着生成式人工智能(GenAI)的出现，AI以其高效的文本处理能力，在图书馆采购、编目、数字化加工、智能咨询等业务环节受到关注和应用。2024年，美国蒙大拿州立大学(Montana State University)团队统计发现：元数据提取、自然语言处理、图像识别位列AI在图书情报与档案领域具体应用方向的前三位，而光学字符(OCR)识别的应用频次位居第八位[2]，印证了AI在文本处理方面的价值。

就文献数字化而言，图书馆在近现代发展历程中积累了大量的经验，其在识别规范文字(如印刷体文字)、制定编目规则等方面具有不可替代的专业优势。但同时，图书馆、档案馆等文献收藏单位还保存大量复杂文件档案(以下简称复杂文档)，因年代久远、字迹模糊，加之手写体、多语种等多重复杂因素叠加，长期面临整理困难。笔者作为清华大学图书馆民国馆史文档整理工作的亲历者，基于对人工成本和时间效率等因素的考量，试图从人机协同的角度，探讨馆员与AI在复杂文档字符识别工作中发挥的作用和工作机制。

2 复杂文档字符识别的现状

2.1 复杂文档字符识别的人工困境

复杂文档多见于图书馆、档案馆等历史文献收

* 通讯作者：董琳，ORCID：0000-0002-8063-8973，邮箱：donglin@tsinghua.edu.cn。

藏机构。收藏的目的在于利用，在提供使用之前，收藏机构通常需要对文档进行识读等整理加工。目前，各收藏单位的识读方法和收效各异，传统 OCR 技术识别复杂文档（尤其是手写体复杂文档）准确率低。国内外除个别具有研发实力的存藏机构尝试使用自主研发的 AI 模型进行手写体文字识别外，绝大多数文档收藏单位和个人仍普遍依赖人工识读，间或辅助通用的文字识别或查询软件，这一现象在古典文献的整理校对工作中尤为常见。然而，人工识别效率低，准确率因人而异且稳定性较低，导致复杂文档的数字化和整理工作进展缓慢，大量的复杂文档只能以图片辅助关键词著录的整理方式呈现给读者，缺乏对内容的全面、深入揭示。

本文以清华大学图书馆藏民国馆史文档为例，该批文档较为完整地保存了清华大学图书馆自建馆以来至 1936 年，在内部工作、参与图书馆界活动中所产生的各界往来信笺等，对于再现中国近代新式图书馆建设历史面貌具有重要价值。前期虽已对中文文档进行了关键词人工著录，但因准确率偏低等问题并未全部向读者开放；英文文档则因多重识别困难长期处于搁置状态，无法有效开发和利用。

具体而言，文档字符识别困难主要集中在以下五个方面。其一，字体因素。如手写体字体多样和字符变形。该批文档通讯对象广泛，涵盖清华大学教职员、学生，国内外各大高校等教育机构、政府，以及出版商、印书局等组织或个人，留下了大量笔法、形态、颜色各异的手写字符及字符的各种变形体。其二，语种因素。文档以中文、英文为主，混杂了德文、法文等多种语言。对于识别人员的语言能力要求极高，一份多语种文档通常需要多人合作才能实现完整阅读。其三，版式因素。繁体中文采用自右向左竖排版，行文习惯与其他语种不同；手写文档中还常存在多板块书写、多人笔迹叠加等情况。其四，纸张因素。如纸张破损导致内容缺失，纸张或墨迹褪色等带来的字迹模糊。其五，保存状态因素。相关内容文档离散存放，难以实现同类型文档的批量处理或语义联系。笔者整理过程中发现：上述多种因素常叠加出现，带来更加复杂的识别困难，这些多重困难因素叠加的文档即本文所称复杂文档。清华大学图书馆馆藏民国馆史文档绝大多数属于复杂文档。

目前尚未发现测定人工识别手写体文字速度的直观数据，可借助人工校对工作效率的相关数据为例来说明。根据实验测定，比较理想的眼动速度为平均每个字符停留 0.9 秒（包括异同比较、是非判断和他校时间在内）。根据中国出版工作者协会校对委员会推算，每个工作日实际校对时间为 6 小时，合计 21600 秒，校对日定额以 25000 字左右为宜[3]。然而，校对速度又因识读人员的知识储备、文档的文字类型和是否需要誊写产生差异。本研究以实际工作中的手写英文、夹杂简繁体汉字为例进行测算，人工识读速度为 250 字/小时（识读人员具备较为熟练的古文字和手写体辨别知识储备）。按每个工作日 6 小时计算，日识读量约为 1500 字。清华大学图书馆馆史文档合计近 1.5 万页，按 100 字/页估算，总字数约 150 万字，一位熟练识读人员完成人工识读誊写约需 1000 个工作日。实际工作中，工作人员常借助各类 OCR 识别软件和知识点查询软件以提高效率，破除识别障碍，或将全文著录转为关键词著录以降低工作量。

2.2 国内外复杂文档文字识别的 AI 探索

文字识别效率和准确率一直是复杂文档人工识别或传统 OCR 技术识别的瓶颈。GenAI 在计算机视觉与自然语言方面的处理能力，为实现文本的高效解析提供了丰富的可扩展空间。针对手写文本的各种字符变型，国内外已研究开发了多种基于深度学习模型的 OCR 技术平台。

1998 年，美国新泽西州雷德班克（Red Bank）AT&T 语音与图像处理服务研究实验室认为卷积神经网络技术在手写字符识别方面优于其他技术。该网络已投入商业应用，每天可处理数百万张商业和个人支票的识别[4]。其后，瑞典国家档案馆（The Swedish National Archives）针对手写和敏感数据文档的转录需求，开发了 Swedish Lion Libre 转录模型，对馆藏 1600 年至 1900 年手写瑞典语文本完成大规模转录[5]。荷兰国家档案馆（The National Archives of the Netherlands）利用 Transkribus 手写文本识别技术构建自定义人工智能模型，对 17 世纪和 18 世纪荷兰东印度公司档案和 19 世纪公证人档案等超 300 万页扫描文件进行识别处理，并在此基础上利用人工智能实现对姓名、地点等关键信息的标注[6]。梵蒂冈秘密档案馆（Vatican Secret Archives，VSA）研发的 In Codice Ratio 项目，通过“数字化图像预处理、训练集采集、字符识别、字符转录”

四阶段框架,将卷积神经网络与统计语言模型结合,推进了对中世纪手稿的 AI 深度处理[7]。

国内研究团队也针对手写体文字识别进行了攻坚。武汉大学刘明忠、贾永红团队通过 Inception 结构构建卷积神经层网络层,进行手写图像多特征捕捉,并通过循环神经网络和 CTC 算法进行特征预测,提升手写汉字长本文识别精度[8]。黑龙江省档案馆孙凯明运用 Pytorch、SQL Server 和 OpenCV 开发面向满文档案图像的手写体满文智能识别软件,突破传统字符分割方法对特殊语言的适配瓶颈,并作为核心技术运用到馆藏清代黑龙江将军衙门档案数字化工作中[9]。

2.3 人工知识系统与大语言模型的辩证认识

现有研究已初步展现 GenAI 技术在图书情报及档案处理领域的广泛潜力。如中国教育出版研究中心公众号介绍了 Recaps、Bookworm、AI 问书等国内外多款 AI 辅助阅读产品[10]。张谙宁等调查发现欧洲、美洲的多个国家及中国已将 AI 生成元数据应用于图书馆资源建设[11]。蒲云强研究了美国国际战略研究中心(Center for Strategic and International Studies,CSIS)在智库建设中的 AI 应用,应用范围涵盖研究报告分析、图像和音频生成等多个领域[12]。广东省档案局科技项目"'大模型+档案'应用场景及基于多模态深度学习档案智能体研究"指出,"AI 大模型凭借强大的计算能力与智能算法为档案编研注入创新动能,数据处理效率提升、知识挖掘深度拓展、成果生成模式创新重塑了档案编研的业务流程与服务效能,展现出技术赋能带来的显著优势"[13]。

既然 AI 已经很强大,人工知识系统是否还有利用价值?实践发现,AI 在档案文献识别、整理中存在"技术、伦理、人文"三个层面的应用断层,弥合这些断层需要人机协作。在技术层面,现有研究多聚焦算法性能、针对特定小样本的模型训练和精度提升,而该类技术端研发者往往较少考虑真实馆藏中存在的大量多语种手稿等情况,使得最需要 AI 介入的复杂场景往往缺乏开源的技术支持。在伦理层面,AI 生成的文档整理结果可能出现语义误判、错误联想、数据安全等伦理问题。在人文层面,AI 核心的学习、关联、生成能力,已超越传统文本识别、信息整理等工具的技术限度,但目前国内外 AI 可调用的可信历史数据有限,限制了 AI 自动深度挖掘历史文档背后的学术脉络、社会关系等历史文化细节的能力。因此,目前 AI 仍无法完全替代人工实现历史文档的自动化整理,制定科学合理的人机协同机制,是有效应对复杂文档整理场景的有效路径。

3 字符识别方法和数据

3.1 路线设计

为了探究 AI 在支持图书馆复杂文档业务场景下字符识别的效用问题,本文以清华大学图书馆藏民国馆史文档整理业务中的字符识别工作为任务场景,围绕图像分区解析、关键信息提取、文献深度学习、笔迹特征学习四大类任务,设计了若干人机策略组合,具体如表 1 所示。四大类任务采用不同的策略,旨在展示 AI 在简单人机交互模式或人机协同模式下应对多样化需求时的识别性能。

任务执行选取了 5 份具有代表性的民国时期复杂文档,以展示识别过程及结果,文档详细情况见表 2。这 5 份复杂文档涉及版式、笔迹、语种、清晰度等多重复杂识别难点,如图 1 所示。研究邀请多名具备较强文字识别能力的工作人员对 5 份文档进行通读识别和内容研究,并将人工结果确定为标准文档。通过将 AI 在简单人机交互模式、人机协同模式下生成的任务结果与标准文档进行对比,以评估了解 AI 在两种模式下的执行效果,进而明确人与 AI 的分工与合作方式。

鉴于此批历史文档以中、英文为主的多语言混合特性,在兼顾分析信度和效度的基础上,本研究选择了以 ChatGPT(GPT-5 免费版)为主的具备图像识读和处理能力的国外开源大模型,以及以"豆包"AI 为主的中文语料强化的国产模型进行探索。

3.2 结果及讨论

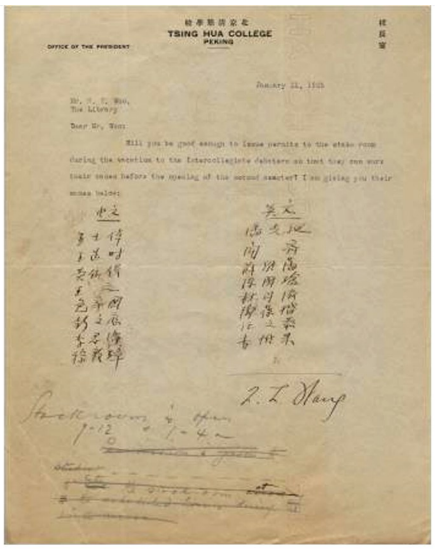
简单人机交互和人机协同模式存在本质区别。简单人机交互模式下,AI 是执行人工预设简单指令的工具;而人机协同模式下,AI 是提供各种可能性的合作者,人工基于自身知识储备指导 AI 进行工作,双方通过多轮策略迭代,优化工作结果,必要时由人工为 AI 提供可靠的学习材料。

表 1　复杂文档文字识别任务

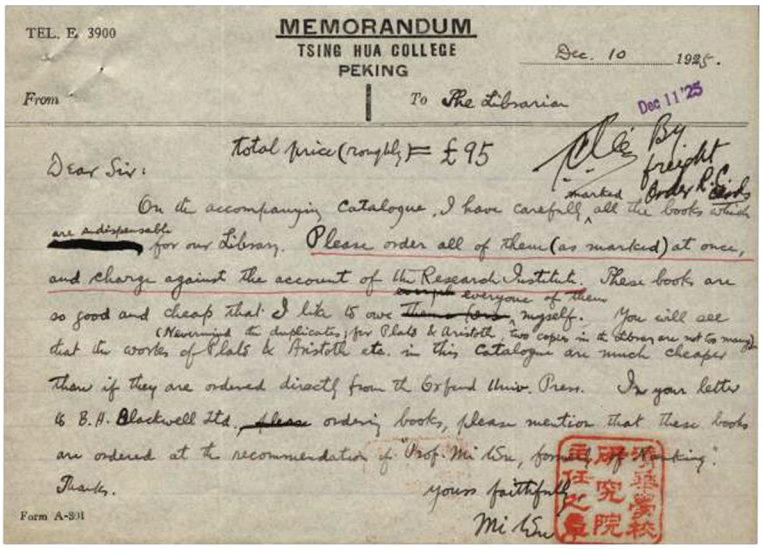
任务类型	任务描述	任务执行步骤(简单人机交互模式)	任务执行步骤(人机协同模式)
图像分区解析	对单页或多页复杂文档进行初步结构化处理，识别不同信息区域	1. 将需要识别的文档进行单页扫描或拍照 2. 将待识别的文档图片上传 AI 3. 让 AI 自动进行图像分区	1. 将需要识别的文档进行单页扫描或拍照； 2. 将待识别的文档图片上传 AI 3. 人工观察文档特点，结合文档特点，输入分区提示词 4. 让 AI 按提示词进行图像分区 5. 让 AI 提供分区可信度及困难提示 6. 人机对步骤 3、4、5 的迭代优化 7. 让 AI 输出分区结果和分区关键语义总结等
关键信息提取	按分区进行文本提取和初步翻译，对于未能识别的信息进行重难点划定	让 AI 对可以标准化识别处理的分区进行识别、翻译	1. 让 AI 对可以标准化识别处理的分区进行识别或翻译 2. 人机对步骤 1 的迭代优化 3. 人工给 AI 提供人名、地名、机构名称等专有名词等易错信息的提示 4. 让 AI 按提示进行错误调整或标注 5. 让 AI 根据提示对未能识别的信息进行重难点划定并记录
文献深度学习	依托可信知识，对 AI 进行针对性训练，以提升其针对特定历史语境下未能识别内容的识别精度和语义理解深度	AI 在独立模式下无法进行此项任务	1. 人工选择可信权威知识或文献 2. 让 AI 学习步骤 1 获取的内容并更新知识库 3. 人工提示 AI 重新进行特定重难点区域内容的识别或翻译 4. 人机对步骤 1、2、3 的迭代优化
笔记特征学习	提取特定人物或来源的独特笔记(以手写为主)的图像特征，实现对同一批文档中模糊、潦草或变体笔记的推理识别	AI 在独立模式下无法进行此项任务	1. 人工确定笔记学习样本 2. 人工引导 AI 锲定学习笔迹的独特特征并存入个人知识库 3. 人工引导推理更多文本内容 4. 人机对步骤 1、2、3 进行迭代优化

表 2　5 份复杂文档详情

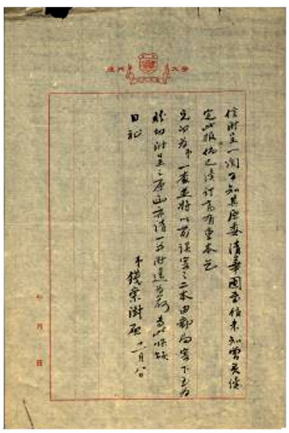
编号	1 号文档	2 号文档	3 号文档	4 号文档	5 号文档
主题	清华学校校长室写给图书馆职员吴汉章的信	清华学校研究院写给图书馆主任的信	钱崇澍写给戴志骞的信	钱崇澍写给戴志骞的信	钱崇澍写给戴志骞的信
书写年	1925	1925	未标明	未标明	1926
语种	中文简体 中文繁体 英文	中文繁体 英文	中文	中文简体 中文繁体 英文	中文简体 中文繁体 英文
版式	横排分列	横排	竖排	竖排	横排
字体	打字机字体 手写体	手写体 刻印	手写体	手写体	手写体



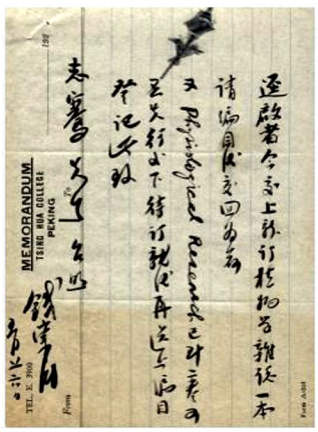
1号文档



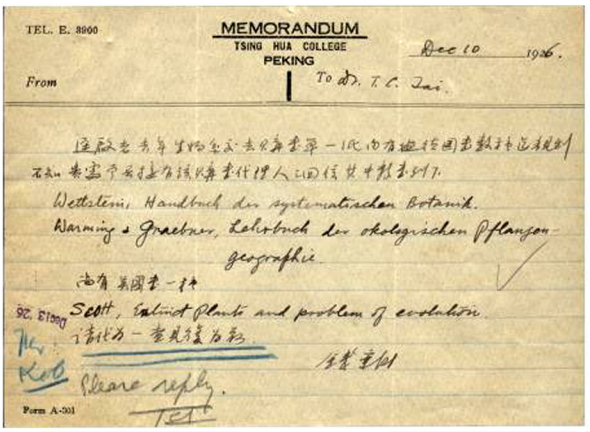
2号文档



3号文档



4号文档



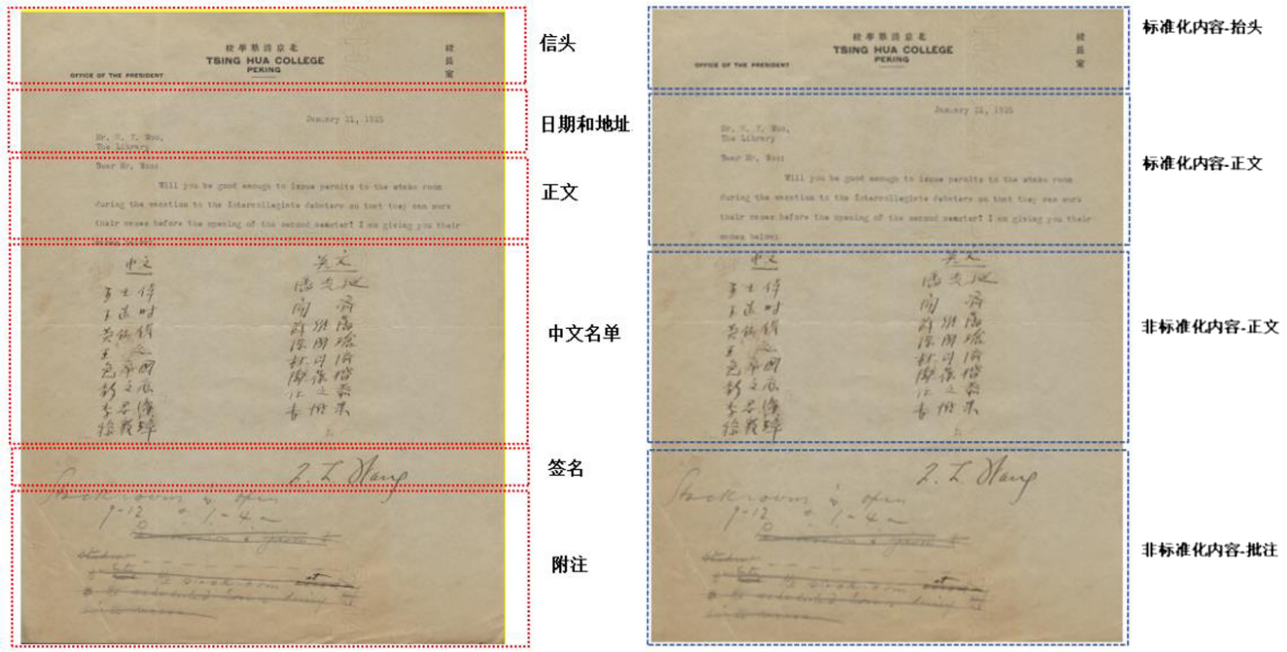
5号文档

图 1　5 份复杂文档扫描图

(1)图像分区解析

分区机制有助于任务按区域执行。简单人机交互模式下，人工向 ChatGPT 发出简单的“图像分区”任务指令，要求将 1 号文档进行分区，得到结果如图 2(a)所示。该模式策略下，人工虽然并未给出分区特点和分区标准的提示，GPT 仍能自动将字体识别为“手写体”“印刷体”“打字机字体”三类，并按照字符位置给出了较为准确的分区结果。此外，GPT 在未收到“解析分区内容”指令的情况下，自行判断给出了简单的分区内容摘录和总结。人机交互模式下，分区结果在考虑字体、位置的基础上，还提供了“标准化”和“非标准化”的聚类，如图 2(b)所示。

对上述简单人机交互和人机协同模式下的“图像分区解析”结果的主要异同分析见表 3。在人机协同模式下，分区结果将简单人机交互模式下的信件日期、称谓和正文部分进行了合并，由于这两部分同属于书信正文，且语种和字体相同(均为英文打字机字体)，合并之后可以有效避免内容识别的分散和碎片化。1 号文档中出现了四种字体，如果将“印刷体”和“打字机字体”等规范字体定义为标准化文字，与之对应的“手写字体”与“繁体/异体中文”等定义为非标准化文字，显然非标准化文字是此类文档整理工作中的核心难点。在人机协同模式下，图 2(b)分区结果依据识别困难程度的不同，给出了“标准化内容”和“非标准化内容”提示。



（a）简单人机交互模式　　（b）人机协同模式

图 2　ChatGPT 对 1 号文档进行分区解析的结果

表 3　两种模式下 1 号文档的“图像分区解析”结果异同

合作方式	策略	分区情况	分区依据	分区语义总结	困难提示	优化方法提示
简单人机交互	任务指令	分为信头、日期和地址、正文、中文名单、签名、附注 6 个区	AI 自动以字体和位置为主要依据	有。多为简单摘录文中词句	无	无
人机协同	任务指令＋提示词＋限定词＋多轮迭代＋选择性增加交互任务	分为标准化内容—抬头、标准化内容—正文、非标准化内容—正文、非标准化内容—批注 4 个区	以人工给出的提示词、限定词为依据	有。较为准确的内容描述	有	有

人机协同模式下，组织任务提示词是简单人机交互模式下不涉及、且需要人工思考决策的一项重要任务。人工首先观察文档和 AI 给出的分区结果，进而确定分区的依据，并以此为基础拟订与 AI 沟通的分区提示词。人工可围绕“字体”“位置”“颜色”“排版”等组织完整的提示词。在 AI 给出分区结果后，如需进一步优化，人工还可根据文档的具体特征为提示词增加限定词，如使用“左上、右上、左下、右下”等方位限定词描述位置特征，并根据需要不断迭代优化。图 2(b)中 1 号文档分区结果中出现的“标准化”和“非标准化”分区，就是 AI 在人工分区提示下，给出的合理化结果。

对于内容丰富的单页文本，人机协同可增加对应区域语义总结、可信度评估、困难提示、数据优化等解析任务。人机协同给出的“标准化”和“非标准化”分区结果，既是分区结果，也是困难提示。多款开源大模型在未经针对性深度学习的前提下，对出现的专有名称，尤其是手写体名称识读均存在较大偏差。以 1 号文档出现的手写体人名为例，在人工引导下，多款 AI 产品均能结合信件上下文判定该部分为特定群体之姓名，并给出群体特征提示，如：“……此为校际辩论队队员姓名……”。而对于识读困难的信息，主流开源大模型还能够提供推进整理的基本建议提示，如：“……部分姓名因手写辨识难度，需结合历史背景确认……”。

显然，人机协同模式所获得的图像分区解析结

果更加准确、针对性更强。通过锚定“非标准化内容”和语义总结、困难提示,该模式可以为后续人工判断工作的重点与难点提供有效依据。

(2)关键信息提取

本阶段的任务基于“(1)图像分区解析”的结果集,针对2号文档不同分区分别进行文本提取和必要的初步翻译。该文档是一封民国初期的手写英文信函,格式为书信体,落款处盖有朱文机构名章,具体提取策略如下。①简单人机交互:读取正文区域的全部英文原文。②人机协同:读取正文区域的全部英文原文+准确率测算+误读原因判断+图片内容背景提示+错误提示+重新定位并二次读取文字。

通过策略①的交互,研究者发现ChatGPT能够快速输出图片文字,但文中存在大量基于文意的联想性描述。经统计错词率达40%,漏词率为4%。

依据策略②,由人工进一步分析判断高误读率的原因,除了英文书写连笔造成的少量误读以外,主要误读和语义联想来自于涂改和补充文字造成的笔迹密集区域。例如信件书写者将“You will see that woks of Plato&Aristotle etc.”等内容密集书写于一行之中,ChatGPT将其联想误读为“If you will take these books...”。通过人工引导ChatGPT对该区域进行重新定位与二次读取,顺利消除了这部分误读。

值得注意的是,历史文档中常涉及人名、地名、机构名等专有名词,也是AI判断的难点。大模型往往过度参考训练语料,加之手写体潦草难以辨认,AI易将其替换为常见姓名或历史名人,导致误读或误译,影响对文档的理解。如策略①将2号历史文档中出现的“Mi Wu(吴宓)”判断为“Mi Chih”。人工基于自身知识体系,结合策略①中得到的结果集,对文档所涉专有名词的特征和词义形成准确判断后,通过关键错误内容介绍等提示性语句对AI进行引导(如策略②中输入提示词“这是民国教授吴宓写给图书馆订购柏拉图和亚里士多德图书的信”),从而使误读的“Mi Wu(吴宓)”得以纠正。此外,人工也可要求AI在此阶段划定或保留关键信息(如输入提示词“输出/翻译时保留所有词首字母大写的人名、机构名等专有名词”),从而减少部分误读。

最后,策略②成功使错词率和漏词率下降到9%和1%。残存的错误主要源于笔迹潦草和清晰度受限,但已不影响整体内容语义的判断。借助人工自身知识储备,并通过多次迭代,错误率可基本消除。随后,人工可将已确认的经验点和尚未明确的内容标记为“重难点”记录并反馈至AI,由AI记录至个人知识库,留待后续批量学习使用。

(3)文献深度学习

本阶段旨在针对“(1)图像分区解析”和“(2)关键信息提取”两个阶段划定的“非标准化”内容以及其他重难点内容等,依托权威历史文献和人工专业知识素养,辅助AI进行针对性训练,以提升AI在特定档案卷宗和历史语境下的识别精度和语义理解深度,从而突破通用模型在此类场景中的局限性。

在1号文档“非标准化”正文中出现的人员名单为中文简繁体夹杂的手写字体,是目前主流OCR工具软件和人工智能大模型识别的常见难点。本研究使用“豆包”AI,采用两种人机策略进行识别。策略①简单人机交互:由机器读取“非标准化”正文中人员名单,读取“标准化”正文英文原文,进行语义提取。策略②人机协同:由机器读取“非标准化”正文中人员名单,读取“标准化”正文英文原文,进行语义提取,人工查找学习资料,输入学习资料,引导AI学习,机器重新读取人员名单。

由表4可见,策略①得到的人员名单错误字以斜体字标示,错误率是34%(以人工识读结果为标准对照统计)。目前主流人工智能大模型的训练数据虽然已经包含诸多清华历史名人,但对清华学生等普通个体的信息覆盖明显不足,如未经针对性历史文献深度学习,易出现误认和联想的情况。

表4 两种人机模式下1号文档手写体人员名单识别对比

人工原文	王士倬	王造时	黄仕俊	王之	包华國	彭文應	李忍濤	徐敦璋
简单人机交互	□王傅	□选时	黄*供侯*	王之	*色辛*国	彭文应	李*君伟*	徐*毅*璋
人机协同	王士倬	王造时	黄仕俊	王之	包华国	彭文应	李忍涛	徐敦璋
人工原文	潘光迥	闻□齐	薛维藩	陈国瑢	林同濟	陶葆楷	任之恭	李惟果
简单人机交互	潘*克*迥	闻□齐	薛*张蘅*	陈国瑢	林丹*侑*	陶葆楷	任之恭	*专忧*果
人机协同	潘光迥	闻亦齐	薛维藩	陈国玱	林同济	陶葆楷	任之恭	李惟果

策略②中，人工对AI输出的结果集进行深度校验，不仅保证了质量，也深化了对AI能力边界和优化方向的了解。根据AI反馈的识别困难提示，经过人工搜集历史文献，选择了清晰刊印的《清华同学录》（民国二十六年即1937年4月刊行）及对应年份的《清华年报》《清华周刊》等官方出版物，由AI进行针对性学习后重新识别。经此过程，生成结果的精度显著提高，达到接近精准识别的水平，进一步验证了该类针对性文献辅助方法的有效性。

值得注意的是，针对“潘光迥”“闻亦齐”二人，经针对性文献训练的AI能够基于人物关系和在校时间等背景信息，将此二人与同时期更为人知的“潘光旦”“闻一多”二人有效区分，并对此前人工识别的输出结果“闻□齐”进行修正（历史文档中只留下了“闻”“齐”两个字，中间以空格分隔），充分展示了AI作为人工协作者的推理与纠错能力。

深度学习阶段是将AI从简单工具转变为协作者的关键。在此阶段，人工挖掘文献内容的能力在AI的帮助下得到了提升，而AI也在人工引导下突破通用模型的知识局限，展现了强大的深度学习能力与在复杂文档处理场景下的应用潜力。

（4）笔迹特征学习与批量精校整理

在复杂文档整理场景中，特定人物或来源的独特笔迹（尤其是手写体）的识别是另一核心难点。笔迹特征学习指利用AI图像识别和知识库管理能力，将清晰的笔迹样本转化为可存储、可复用的“图像特征知识”，实现对同一批档案中模糊、潦草或变体笔迹的推理识别，显著提升识别效率与精度。

以3、4、5号历史文档为例，这三封清华生物学教授钱崇澍写给清华大学图书馆主任戴志骞的信件，呈现了早期图书馆的书籍推荐与跨校交流工作细节。信件语言混合了中文与英文，版式包含了横版与竖版，且手写笔迹清晰程度不一，传统OCR技术识读困难。

从3号历史文档入手，原文经人工识读并校对，录入结果如下：“信附呈一阅可知其原委 清華圖書馆未知曾买续定此报 倘已续订當有重本 乞兄即为弟一查 并将以前误寄之二本由郵局寄下 至为盼切 附呈之原函并请一并附還为荷 专此顺颂日祉 弟钱崇澍启 十一月八日”。

“豆包”AI在未经笔迹学习时，生成结果如下：“信附呈一阅，乃知其原委。清华园电仍未到，至为系念。报稿已读，订否，希裁夺见示。兄需书一套，兹特以前漾寄二本，由邮局寄下，乞照收，附呈之原书并请一并附还为荷 专此即颂日祉。弟钱崇澍，十二月十六”。

简单人机交互模式下，AI的生成结果为基于少数文字延展出的语义串联，与信件原文相差甚远，若抛开原文，其语句符合中文语言习惯，在处理中极易形成“幻觉”造成误读，进而影响对历史事件的理解，值得警惕。

鉴于三封信中都有钱崇澍的亲笔签名，研究以“钱崇澍”手写签名作为笔迹学习的突破口，人工向AI输入“钱崇澍”手写签名并提示其对笔迹书写进行识别和学习，AI思考如下：“他的字体有着流畅的笔势，笔画间连带自然，签名‘钱崇澍’书写时，线条灵动且有独特的行笔节奏……虽……字迹潦草，但能感受到那种随性又不失韵味的书写风格，已记住这种带有个人书法特质、笔画连贯洒脱的样式。”

经由签名为突破口的多封信件字迹学习后，AI生成结果如下：“信附呈一阅可知其原委，清华图书馆未知曾买续定此报，***似已续订务望查本乞见即为弟一查***，并将以前误寄之二本由邮局寄下，至为盼切，附呈之原函并请一并附还为荷，专此顺颂日祉。弟钱崇澍启，十一月八日”。

将AI学习后的结果与人工识别的结果对比，将识别错误用粗斜体标出。可见经过笔迹学习，AI能够建立对特定书写者字迹的图像化记忆，进而实现对模糊笔迹和竖版手写繁体中文的推理与识别突破。在此过程中，识别和生成内容的精度显著提升。尽管行文个别细节处仍有误读，但对整体语义影响较小。识读结果对于初步整理阶段的语义判断具有参考价值，若进一步结合人工校读，将实现更高的精度。

本阶段中，人工通过精准筛选特征样本，引导AI构建可计算的特征模型，并严格评估其效用，进而完成文档整理和知识生产。面对差异化的任务，人工所具备的深厚AI素养是该阶段任务顺利执行的关键保障。同时，AI展现出强大的图像特征学习、抽象建模与跨场景迁移推理能力，将原本依赖人工个体经验的笔迹识别任务，转化为基于特征模型的可计算、可共享、可迭代的过程。这一转变不仅显著提升了当前档案整理的精度与效率，更将个人笔

迹等细节特征转化为可复用的共享知识模型，为后续开展大规模、可持续的复杂档案整理工作，以及基于档案的深度历史研究，奠定了人机协同、知识共建共享的坚实基础。

4 研究结论

4.1 生成式人工智能视角下的人机协同档案整理路径

笔者在对复杂历史文档整理工作经验进行梳理和总结的基础上，提出了生成式人工智能视角下的人机协同档案整理路径，如图3所示。通过建立“AI预处理—人工校验—模型迭代”的闭环工作思路，系统整合图书馆员的专业能力与人工智能的技术优势，旨在有效提升复杂历史文档整理工作的效率和质量。图3展示了图书馆开展此项工作的核心实施路径及各环节在“人”“机”两个维度上的相互衔接。该流程框架弥合了主流人工智能大模型技术与图书馆文档整理实务工作场景的断层，推动AI从工具属性向协同主体的角色转变，为高校图书馆在人工智能时代开展历史文档智能化整理工作提供了一种可操作的方法论参考。

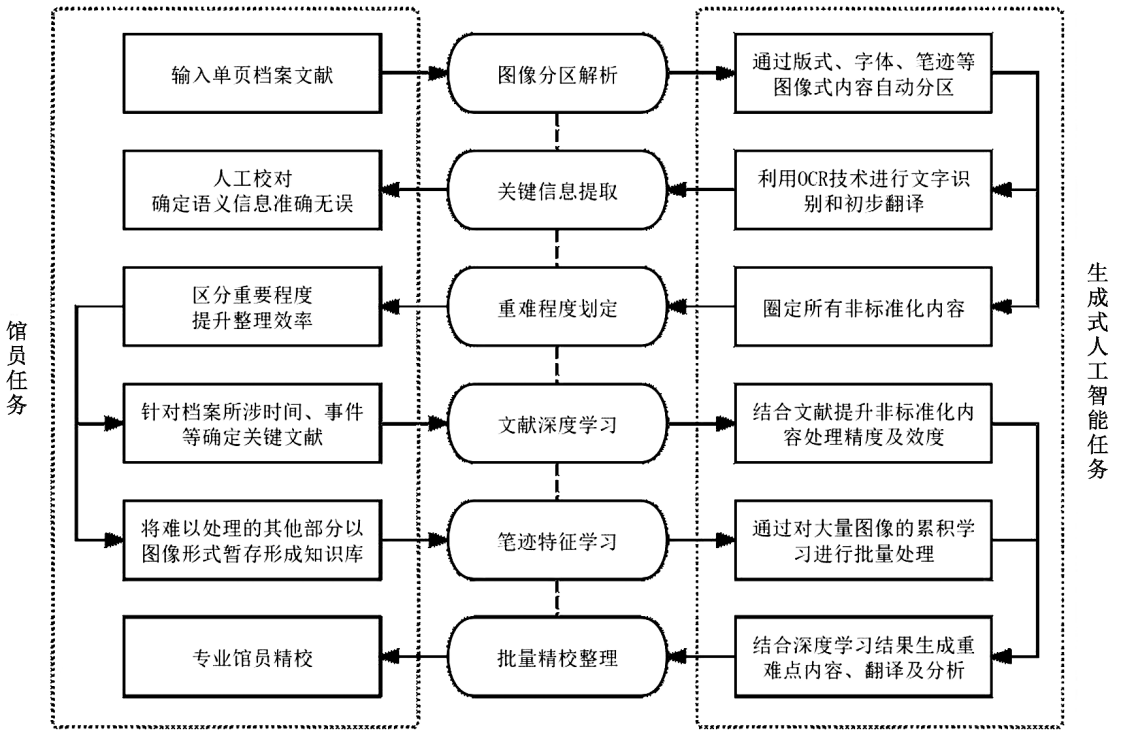


图3 生成式人工智能视角下的人机协同档案整理路径

4.2 重点总结及反思

作为图书馆的基础业务，文献整理智能化将进一步推动图书馆智慧化发展。本文聚焦复杂历史文档整理工作中文字识别的四个核心环节，结合前期清华大学图书馆对2000余套包含中西文民国文档的智能化整理探索，证实了利用GenAI整理复杂历史文档是一种可行的操作方法。

构建人机协同的动态工作流，其核心价值体现在文档整理工作效能的提升。若暂不考虑输出质量，仅就人工与人机协同所投入的时间成本进行大致对比：据本文2.1章节中的人工识别经验估算，人工处理2000余套文档需要1名熟练人工投入1000个工作日（每天6小时，合计6000小时）；而在人机协同模式下，同等规模文档投入时间折合为1人108个工作日（该人工每天工作4小时，合计432小时）。实际工作中，效率必须建立在准确率的基础上。通过上述工作机制，结合后续文档精校（对前期人机协同整理后的文档内容进行校对）人员的反馈，前期人机协同整理文档的正确率最高可达90%及以上，其中英文手写文档的整体识别准确率普遍高于中文手写体文档。

通过纯人工处理和人机协同模式进行文档整理

都需要人的智力投入，但两者对于参与人员有不同的素质要求。纯人工识读复杂历史文档，要求参与人员具备较丰富的人文学科知识背景、语言能力和手写体文字识读经验。人机协同模式则更注重工作人员的AI素养，涉及撰写AI提示词、人工知识投喂、人工判断AI幻觉等重要工作步骤。这些工作步骤形成了人机协同的关键特性：在各环节中，人与AI始终处于持续的互动状态，不断进行知识的积累与共享。在这种互动中，AI逐步形成学习记忆，继而在后续大规模整理中降低人工介入。总体来看，额外的人机互动步骤属于“前期投入，长期收益”的优化模式。

此外，人机协同工作模式为高校图书馆传统业务的智慧化转型提供了三重启示，同时回归并解答了本文2.3节中提出的人工智能在“技术—伦理—人文”三个层面的应用断层的问题。

（1）技术层面，主流AI大模型应用亟须突破单向的工具化逻辑，转向馆员、人工智能、馆藏资源三者的深度协同发展。突破口在于培养并提升馆员的AI素养，建立并深化馆员对AI从理论到实践方法的全面认知，在实务工作中实现人机分工合作边界的动态调整，以有效应对各种复杂智能化场景。GenAI的潜力将在与人类专业知识的深度互动中被充分激发。

（2）伦理层面，随着主流AI大模型应用于图书馆批量处理工作，现有研究中阐述的AI语义幻觉、联想认知偏见、数据安全等伦理问题将在图书馆实务工作中显现。对此，人工判断纠错和正确的数据语料投喂仍是目前最科学的合作策略。馆员需在各阶段关注、优化AI的伦理意识，图书馆应建立有效的防护和精校机制。

（3）人文层面，AI参与图书馆工作的核心意义在于唤醒过往被忽视或难以处理的馆藏文献的多元文化价值，辅助馆员合作生产更多高质量的文化、学术产品，以激活文献价值、重构学术脉络、延续历史记忆。基于此，图书馆应大力推动图书馆员从文献保管者向知识生产者转型，将AI协同工作成果转化为特色知识库，并开放为教研基础设施，使馆藏文献尤其是特藏文献真正成为承载中华文明变迁的活态资源。

本研究构建的人机协同框架，以构建历史特色知识库为技术突破方向，以伦理校验机制规范技术应用，以挖掘人文价值引领技术目标，最终实现“技术－伦理－人文”三个层面的人机深度协同。本研究基于中、西文历史文档，聚焦于主流大模型在文档内容识别方面的应用，以期为图书馆历史文档整理工作提供参考。就整理工作而言，本研究所涉及的工作环节和资料种类还有待继续丰富和全面，相关手写笔迹、历史信息知识库及配套智能体的开发还需要继续完善，这些将是后续研究的重点探索方向。

参考文献

1 国务院. 国务院关于深入实施“人工智能＋”行动的意见[EB/OL]. [2025－09－17]. https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html.

2 Mannheimer S, Bond N, Young S W H, et al. Responsible AI practice in libraries and archives: a review of the literature[J]. Information Technology & Libraries. 2024, 43(3): 1－29.

3 周奇. 现代校对基本理论与方法[M]. 北京：中国书籍出版社，2020：115.

4 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE. 1998, 86(11): 2278－2324.

5 瑞典：推出转录手写文档的AI模型和开源软件[J]. 陕西档案，2025(1)：61.

6 荷兰：国家档案馆使用手写文本识别技术自动转录超300万页文件[J]. 陕西档案，2023(1)：13.

7 陈蝶. 人工智能技术在档案工作中的应用实践与成功经验——以In Codice Ratio项目为例[J]. 档案天地，2024(2)：40－44，47.

8 刘明忠，贾永红. 基于Inception结构的手写汉字档案文本识别方法[J]. 武汉大学学报(信息科学版)，2022，47(4)：632－638.

9 孙凯明，孙磊，王刚，等. 面向满文档案图像的手写体满文智能识别软件设计与实现[J]. 自动化技术与应用，2024，43(1)：91－94.

10 E出版. 亚马逊Kindle推出阅读新功能：AI辅助阅读会成为出版的下一个风口吗？[EB/OL]. [2025－04－14]. https://mp.weixin.qq.com/s/sOWo9z9P9V3BH1Dqyyb_tg.

11 张谙宁，叶兰，周文琦，等. AI生成元数据赋能图书馆资源建设的实践与启示——基于国内外案例调查[J]. 大学图书馆学报，2025，43(4)：90－104.

12 蒲云强，唐川，敖丽娟. 人工智能赋能智库建设[J]. 智库理论与实践，2025，10(2)：93－105.

13 崔忠菲，罗人芳，邹如雁. AI大模型下档案智慧编研的知识再生产路径[EB/OL]. [2025－10－17]. https://mp.weixin.qq.com/s/IcifkQAV99NvXHnpc1cHDQ.

作者贡献说明：

董琳：提出论文选题，设计研究思路，论文撰写与修改

许澜洋：研究内容讨论，历史文档分析，文献素材筛选，参与论文撰写与修改

作者单位：董琳，清华大学图书馆，北京，100084

许澜洋，清华大学社会学系，北京，100084

收稿日期：2025年9月22日

修回日期：2025年11月13日

（责任编辑：王菲）

Research on Human-in-the-Loop Pathways for Archival Organization from a Generative AI Perspective: A Case Study of Republican-Era Institutional Archives in Tsinghua University Library

DONG Lin XU Lanyang

Abstract: Tsinghua University Library preserves more than ten thousand pages of archival documents of its institutional history during the Republican era of China. These materials are highly complex, characterized by blurred and deteriorated scripts, difficult-to-decipher handwritten text, and multilingual content. For decades, archival organization has been constrained by low efficiency and limited output. Even after digitization, most materials have been accessible only as images with keyword-based metadata, while full-text content remains unavailable. Recent advances in generative artificial intelligence (GenAI) offer new possibilities for addressing these long-standing challenges in processing complex historical archives. To tackle the core issues in archival digitization including the limited accuracy of traditional optical character recognition (OCR) technologies and the inefficiency of purely manual transcription, this study examines a human-in-the-loop workflow grounded in real library practice. Drawing upon the completion of nearly 2,000 sets of Republican-era archival documents at Tsinghua University Library, the research aims to design and validate an innovative workflow that dynamically integrates human expertise with automated recognition capabilities through iterative optimization, thereby achieving both high efficiency and high accuracy in character recognition. Methodologically, the study first constructs a test corpus comprising multiple types of complex historical archival materials. It then proposes a closed-loop workflow that includes image region segmentation and analysis, key information extraction, human-in-the-loop verification, and feedback-driven learning. The core innovation lies in the development of accuracy-validated human-AI strategy combinations, which enable a comparative evaluation of large AI models under three operational modes: full human-AI collaboration, simple human-AI interaction, and traditional manual processing. Experimental results demonstrate that the human-AI collaborative model significantly outperforms simple human-AI interaction in character recognition accuracy, achieving accuracy rates of 90% or higher. Simultaneously, the required level of human intervention is reduced by approximately 93%, substantially surpassing traditional manual transcription in efficiency. The findings indicate that when guided by professional judgment and domain knowledge, GenAI systems can function as effective collaborative agents in archival processing. The proposed mechanism facilitates intelligent task allocation and closed-loop feedback, effectively integrating machine processing speed with human precision. It provides a scalable and cost-effective solution for large-scale digitization initiatives involving complex archival materials. The study further suggests that the effectiveness of this collaborative mechanism depends critically on participants' AI literacy across three interconnected dimensions: technology, ethics, and the humanities. Future research will explore diversified collaborative strategies and the development of supporting knowledge bases and intelligent agents to enhance system adaptability and sustainability.

Keywords: Artificial Intelligence; Human-in-the-Loop; Character Recognition; Republican Era; Documents: Archives