GenAI赋能古籍数字化的可信度评估与可信治理策略研究*

付雅明　孙炫宇
（上海大学文化遗产与信息管理学院，上海 200444）

摘要：生成式人工智能（Generative Artificial Intelligence，GenAI）在赋能公共图书馆古籍数字化的同时，也带来了数据失真、模型“幻觉”与权责模糊等可信风险，故亟须加强系统性可信治理。本研究首先分析GenAI赋能古籍数字化过程中引发的可信风险，并构建适配古籍特性的涵盖基础层、内容层、知识层与服务层的可信度评估框架；然后，据此构建包含13项评估指标的GenAI赋能古籍数字化可信度评估指标体系，实现对生成内容的结构化、可操作的可信度评估；最后，提出建立分层响应机制、加强资源分级管理与权责协同的制度保障、构建可信治理长效机制三个方面的可信治理策略，形成“可信风险分析—可信度评估—可信治理”的完整闭环，旨在为GenAI在古籍数字化过程中的合规化、高质量的应用提供理论与实践参考。

关键词：GenAI；古籍数字化；可信风险；可信度评估；可信治理
中图分类号：G250　DOI：10.3772/j.issn.1673–2286.2026.05.008
引文格式：付雅明，孙炫宇. GenAI赋能古籍数字化的可信度评估与可信治理策略研究[J]. 数字图书馆论坛，2026，22（5）：83–90.

古籍数字化是图书馆传承中华优秀传统文化、拓展文化服务智慧化场景的核心举措。在国家文化数字化战略的驱动下，古籍数字化工作不仅是实现古籍资源永久保存与全民共享的技术与数据基石，更是智慧图书馆建设的关键环节。生成式人工智能（Generative Artificial Intelligence，GenAI）赋能古籍数字化是指在GenAI技术辅助下将纸质古籍转化为数字文本与知识资源，并延伸至智能化内容生成、知识挖掘与终端服务应用的广义数字化过程，涵盖了从文本处理到知识服务的全流程。古籍数字化内容的可信度是体现其价值与生命力的根本保障。《关于推进新时代古籍工作的意见》明确提出要“积极开展古籍文本结构化、知识体系化、利用智能化的研究和实践，加速推动古籍整理利用转型升级”[1]。《关于推进实施国家文化数字化战略的意见》要求“在数据采集加工、交易分发、传输存储及数据治理等环节，制定文化数据安全标准，强化中华文化数据库数据入库标准，构建完善的文化数据安全监管体系”[2]。以上政策分别从智能化转型方向与文化数据可信治理两个维度为古籍数字化全流程的质量管控与核查评估确立了政策导向。

由于GenAI存在“幻觉”问题[3]、领域知识局限性问题[4]，以及缺乏自我反思能力与深度语境甄别能力[5]，其生成的内容易出现事实性错误、机械性偏差乃至文化误读，因此，GenAI赋能古籍数字化后生成内容是否可信成为亟待解决的重要问题。尽管GenAI在我国图书馆领域的应用已初具规模，但针对GenAI赋能

收稿日期：2026–03–27
*本研究得到教育部人文社会科学研究一般项目“基于生成式人工智能的数字文化遗产叙事模式与可信治理研究”（编号：25YJCZH049）资助。

古籍数字化的可信风险研究仍显匮乏。基于此，本研究构建GenAI赋能古籍数字化可信度评估框架和可信度评估指标体系，并提出可信治理策略。

1 相关研究

基于大语言模型（Large Language Model，LLM）的GenAI推向市场以来，学者们针对LLM、人工智能生成内容（Artificial Intelligence Generated Content，AIGC）的可信问题展开研究。钱明辉等[6]从LLM可信的概念及内涵、影响因素、面临的挑战以及实现LLM可信的策略四个方面对LLM可信进行全面分析，其中包括了从技术、社会伦理、用户角度对AIGC可信的内涵界定：技术角度强调AIGC本身的可信性；社会伦理视角强调AIGC对所有服务对象一视同仁，不因服务群体的差异性而出现结果偏差；用户视角强调用户对AIGC的信任程度。

学界亦逐步开展面向AIGC的可信风险评估方法与管控机制构建的研究。杨艳妮等[7]构建包含系统可信度、媒介可信度、内容可信度三个维度的AIGC可信度评估体系，研究发现内容可信度对AIGC可信度的影响高于系统可信度和媒介可信度，并据此提出增强AIGC可信度的优化策略。宋筱璇等[8]借鉴信息安全管理ISO 27000标准，将信息资源管理各个环节的AIGC视为信息资产，从资产、威胁、脆弱性三个关键要素出发进行风险识别和评估，通过构建风险价值矩阵计算风险值，评估信息资源管理AIGC风险，进而提出通过减少风险事件发生时造成的损失对风险进行管控的策略。刘冰等[9]揭示人工智能（Artificial Intelligence，AI）生成信息的信息生态污染风险，并对风险影响因素及作用机理进行系统研究，指出AI信息标识率是实现高效能风险治理的决定性因素，继而提出全面提升AI生成信息标识率和构建全社会共同参与的多元协同治理体系的风险管控对策。

目前，学术界尚未出现专门针对GenAI在古籍数字化中应用的可信度评估研究，但在GenAI应用于古籍保护的风险识别与治理方面已有初步探索。胡婷等[10]指出GenAI深度介入古籍修复、数字化处理及智能传播全流程，在重构古籍活化路径的同时也面临古籍历史真实性消解、人才断层与经费匮乏、古籍传播权力失衡、制度滞后与权责模糊等一系列风险，并提出了包括构建古籍数据治理体系、健全人才培养与资金保障机制、建立公平开放的古籍资源传播生态、完善古籍保护应用相关法律体系在内的风险治理策略。

现有研究已在AIGC通用可信度评估框架与风险管控体系方面取得初步成果，但多立足于通用信息资源管理或知识服务场景，尚未充分回应古籍文本所特有的历史性、语境封闭性、语言异质性及文化敏感性等深层挑战。古籍保护领域的现有研究多聚焦于宏观层面的风险类型归纳与原则性治理建议，尚未深入到具体技术环节分层级、可操作的可信度评估。因此，本研究旨在梳理GenAI赋能古籍数字化过程中引发的可信风险，在此基础上构建适配古籍特性的可信度风险评估框架和评估指标体系，并提出可信治理策略，为古籍智能处理的可信保障提供理论与实证依据。

2 可信风险分析

（1）古籍文本特性带来的语义失真风险。从古籍文献的文本特性角度分析，古籍文本以繁体字为主，同时夹杂大量异体字、通假字、古简体、民间俗体字，且其篇章结构与表达逻辑与现代汉语存在显著差异，文本释义高度依赖历史语境与文化背景。借助GenAI开展古籍文本的自动转写、句读或翻译时，其输出结果易因转换错误或无法准确把握语义深层结构而造成语义失真，进而影响内容可信度。GenAI主要通过学习大量文本数据生成新的文本，如果源数据中存在错误、矛盾、过时或误导性的信息，或存在社会、文化、历史或性别等方面的偏见，加之缺乏对生成内容真实性的实时验证机制，GenAI可能会输出看似有说服力实则不准确或完全虚构的信息[11]。

（2）GenAI技术局限引发多重可信风险。从GenAI自身的技术局限性角度分析，古籍数字化各个环节在应用GenAI时均潜藏多重可信风险。在基础文本处理环节，由于GenAI存在长文本解析能力不足、上下文语境理解偏差的问题，容易在古籍光学字符识别（Optical Character Recognition，OCR）、自动句读、文本规范化处理过程中产生识别偏差与断句失误，显著降低基础文本加工的准确率。在内容理解与生成环节，GenAI依赖概率预测而非事实推理，易产生“幻觉”或语义曲解，虚构古籍原文未记载的实体、事件或关系。在知识挖掘与构建环节，GenAI在实体识别与关系抽取时若缺乏领域知识约束，可能导致知识图谱逻辑错乱，在面对多实体、多关系的复杂查询时，常因无

法全面覆盖原文关键知识单元而遗漏核心要素，或凭空构造不存在于原始文献的信息，造成问答内容与史实不符。例如，使用GPT4进行古籍限定域关系抽取时，生成内容存在理解错误和在古籍原文中无法找到相应完整描述的问题[12]。在服务交互环节，智能问答系统若无法有效锚定原始文献依据，则难以保障回答内容的可溯源性与语境适配性。

（3）标准缺失与法律滞后导致的管控真空风险。行业标准缺失与法律制度滞后使得GenAI赋能古籍数字化的可信风险缺乏有效管控及约束。一方面，当前尚缺乏针对AI生成古籍内容的统一质量标准和责任认定机制；另一方面，著作权归属模糊、独创性认定缺位等问题亦带来法律与伦理风险，GenAI在跨平台调用多源古籍资源时，可能引发多重权利主张冲突，而现行法律对GenAI生成古籍内容的著作权授权规则尚无明确界定[13]。

3　可信度评估框架

前文从三个维度揭示了GenAI赋能古籍数字化过程中的可信风险，这些风险在文本识别、内容生成、知识挖掘与服务交互等环节呈现出不同的表现形式与影响程度。因此，有必要针对各个环节的风险特征，构建一个分层级的可信度评估框架，使评估指标精准对应不同环节的风险靶点，为后续的量化测评与差异化治理提供结构化分析工具。

3.1　评估目标

评估框架设计的第一步是明确评估目标。GenAI赋能古籍数字化的可信度评估需要充分贴合古籍数字化应用场景的独特性，因此本研究设定以下5项评估目标。

（1）语义忠实性。古籍文本高度依赖语境、训诂与历史背景，在古籍数字化实践中，这一诉求不应仅停留在字符转换层面，而须准确传递原始语义。该目标针对异体字、通假字识别错误以及模型“幻觉”等问题，强调AIGC内容输出必须忠实于原文意涵，避免因字形混淆或上下文误判造成知识扭曲。

（2）知识完整性。古籍承载的是体系化的传统知识网络，包括人名、地名、典章、制度、引书关系等多维要素。GenAI赋能古籍数字化过程中，不应因算法偏好、训练数据偏差或推理能力局限而遗漏关键实体或断裂逻辑链条，尤其在处理残缺文本时，修复内容应有据可依，而非凭空补全。

（3）形态保真性。古籍的版式、行款、批注、插图等物理特征是构成文献的重要组成部分。GenAI在结构化处理中若忽略这些非文本信息，或者将现代排版逻辑强加于古籍，将导致数字副本丧失版本学价值。因此，可信成果须在保留原始文献形态特征的前提下实现知识转化。

（4）规范适配性。当前古籍数字化兼具学术研究与公共文化服务双重目标，不同应用场景对内容深度、表达方式与交互形式提出差异化要求，既要符合相关的行业标准与规范，又要能适配智能问答、知识图谱、教育普及等多元场景的功能需求，避免“一刀切”式生成导致实用性不足或学术失范。

（5）风险可控性。鉴于AIGC在著作权归属、数据合规等方面存在显著不确定性，可信度评估还需内嵌风险预警功能，防范模型在跨源调用古籍资源时引发的多重权利冲突，规避因训练数据污染或文化盲区导致的伦理争议，确保生成内容在法律上可追溯、在责任上可界定，从而守住古籍数字化的文化安全与制度合规底线。

3.2　框架设计

本研究参考管涛等[14]借鉴生命周期理论构建可信数据空间逻辑框架的思路，以及陈鑫等[15]为实现AIGC可信性的全生命周期管理所提出的融合技术保障、过程管控与协同主体的三维协同框架，为实现GenAI赋能古籍数字化可信度评估的具体目标，自下而上设计包含基础层、内容层、知识层、服务层四级架构的可信度评估框架。

（1）基础层是可信度评估框架的底层支撑，是对GenAI赋能古籍数字化文本处理技术的评估。该层主要考查古籍文本识别与切分的准确度，古籍文本普遍存在异体字、无标点等典型特征，文本识别与切分是GenAI开展古籍数字化的首要环节，若该环节存在技术偏差，后续所有环节都会出现连锁性错误。

（2）内容层是连接基础层与知识层的关键枢纽，是对GenAI赋能古籍数字化文本内容的评估。在基础层后面设置该层级，是因为即使基础层的内容可信，仍可能因GenAI自身技术局限导致在内容生成、残缺内容

修复环节出现生成无文献依据的虚假内容或语义偏差问题。

（3）知识层是对古籍知识挖掘成果的深层逻辑校验，其核心定位为保障知识体系的学术严谨性与逻辑完整性。在内容层后设置该层级，是因为古籍数字化不仅是文本的转写，更是知识的抽取与重组，而GenAI在知识关联挖掘与构建时易出现逻辑混乱、关系错配等问题，因此需评估知识图谱中实体、关系的历史逻辑与学术合理性，避免知识传播偏差。

（4）服务层是对终端应用效果的校验，其核心定位为检验前三层评估结果的实际服务价值，核心作用是确保GenAI赋能古籍数字化的成果适配多元服务场景。将其置于框架最顶层的原因在于，古籍数字化的最终目标是向用户提供经过数字化加工处理的古籍资源以实现资源共享与高效利用，GenAI赋能古籍数字化后，一般以智能化问答的方式提供服务，由于基础层、内容层、知识层可能出现的问题会导致服务层输出的答案出现缺乏溯源依据或语境不符等问题。

4　可信度评估指标体系

本研究以上述评估框架为基础，构建一套分层递进、可量化测评的可信度评估指标体系（见表1）。可信度评估指标体系以四层评估框架为纲，以具体、可测的指标为目，既关注技术实现的底层可靠性，也强调学术内容的高层价值，以期形成从数据可信到知识可信再到服务可信的闭环。四个维度之间既有层级侧重，又存在交叉支撑。由于GenAI赋能古籍数字化最终体现为可供用户使用的终端服务系统（以下简称“系统”），本研究构建的可信度评估指标体系是对该系统及其生成古籍数字化内容的多角度、立体化的可信度评判标准。

表1　可信度评估指标体系

搭配评估框架层级	评估指标	评估方法
基础层	异体字识别准确率	以专家标注的高难度异体字测试集进行批量测试，并结合人工复核统计误识率与漏识率
	源文献保真度	专家逐条对比
	古汉语分词精度	人工抽样比对、专家盲评等
	处理过程可溯性	通过检查元数据完整性、工作流文档规范性及第三方复现成功率来综合判定
内容层	文本结构合规性	由领域专家对生成文本结构进行核对
	文献征引准确性	构建引文网络分析其连通性与支持强度，或由专家评估其论证链条的严密性
	文本语义一致性	将GenAI生成文本与目标古籍的权威整理本（如中华书局点校本）进行比对、语义相似度计算，辅以专家判断进行综合评估
知识层	实体定义规范性	借助实体消歧技术统一实体含义，同时依托权威工具书、数据库进行交叉验证，统计事实性错误的数量与严重程度
	关系抽取准确性	构建专家标注的关系三元组标准集合，将GenAI输出与之进行匹配，计算准确率、召回率与F1值，并对新增关系进行专家逻辑一致性核验
	本体模型适配性	将GenAI构建的本体结构与古籍原生分类体系及现代通用本体进行比对，由专家进行评估
服务层	任务目标达成度	根据不同应用场景分别设定对应的量化指标（如校勘建议采纳率、用户理解测试得分、资源管理粒度、通俗化转化效果、知识元调取速度、多维链接覆盖率等），通过专家评分与用户调研相结合的方式进行综合衡量
	语境适配性	由专家对GenAI回答的问题类型匹配度、出处标注完整度、表述规范性等进行分项打分
	人机协同友好性	使用用户调研法，并使用任务完成时间、修正成本等指标衡量

4.1　基础层可信度评估指标

基础层可信度评估关注生成内容所依赖的原始材料、处理流程及技术底座的可靠性，具体包括4项评估指标。①异体字识别准确率：评估系统对古籍中异形字、异构字的精准识别能力，是文本转写的基础指标。古籍异体字数量庞大、字形演变复杂，同一个汉字在不同历史时期、不同刻本中可能存在数十种变体写法，如“體”字在敦煌写本中常写作“躰”，“禮”字亦存在多种简省写法。系统若缺乏对异体字的系统训练与上下文感知能力，极易将异体字误识别为其他常见字或直接输出乱码，从而导致后续文本切分、语义理解乃至知识抽取的连锁性错误。②源文献保真度：评估系统生成内容对原始古籍载体（如刻本、写本、拓片）信息的忠实还原

程度。该指标强调系统对古籍特有要素（如异体字、避讳字、行款格式、墨色差异、物理破损痕迹）的保留与标注能力[16]。在残缺内容修复方面，可参照国家标准《古籍修复技术规范与质量要求》（GB/T 21712-2008），从修复流程、技术要求等方面规范GenAI背景下残缺古籍的修复工作，为数字化修复提供基础[17]。具体而言，不得擅自增删原文，不得以现代通行字形替代古籍原字，不得忽略原本中的墨迹污损或虫蛀残缺痕迹。对于残缺部分的补遗，应严格遵循“修旧如旧”原则，补入内容须有同版本或其他权威版本的可供比照的依据，并在输出中明确标记补遗范围与来源。③古汉语分词精度：评估系统对无标点古汉语文本进行语义切分的合理性，该指标直接决定后续内容理解与知识挖掘效果。古汉语以单音节词为主，一词多义、兼类词现象极为普遍，且缺乏现代汉语的分词标记（如空格、标点），使得自动分词面临巨大挑战。例如，“王命尹氏及王子虎”中“王子虎”既可被正确切分为专有人名“王子虎”，也可能被误切分为“王/子虎”，这类由词边界不明引发的切分歧义正是古汉语分词精度评估的重点。④处理过程可溯性：评估古籍数字化从原始图像到最终生成文本的全过程是否完整、透明、可复现。例如，处理过程中应保留详细操作日志，包括图像预处理参数、模型版本、训练数据来源、人工干预节点等关键信息的记录。

4.2　内容层可信度评估指标

内容层可信度评估聚焦生成文本自身的语言形式、结构组织与表达合规性，确保其符合古籍整理的基本学术规范，具体包括3项评估指标。①文本结构合规性：评估系统生成内容是否遵循特定古籍类型或版本（如宋刻本、明抄本）的典型结构特征。具体而言，卷次划分方面，考查系统是否按照原典的卷、册、篇、章层级进行正确分割，避免出现跨卷合并或分卷错乱；篇目标题方面，考查系统是否完整保留原典的篇题、章题、子目等标题信息，且标题层级与正文对应关系是否正确；注疏体例方面，考查系统是否准确区分经文、传文、注文、疏文等不同层次的文本，并按原典惯例排列。②文献征引准确性：评估系统生成或修复的古籍文本是否有可靠的文献支撑，避免出现无中生有的编造内容。该指标针对生成内容中的引文、典故或观点，考查其是否提供了准确的出处，并能在相关传世文献中找到依据。这不仅要求引文无误，更强调引文与上下文的逻辑契合度。③文本语义一致性：评估系统生成文本在词汇选择、句式构造、语气、语体等方面是否与目标古籍的时代、作者及文体特征保持一致；评估其在残缺内容修复、文本补遗等任务中对古籍原典体例与原始内涵的忠实还原程度，不因模型“幻觉”而产生偏离原意的语义漂移。例如，验证智能体生成的古籍文本片段是否符合原典体例，修复的内容是否有文献支撑[18]。

4.3　知识层可信度评估指标

知识层可信度评估旨在检验生成内容所承载的历史信息、文献关联与学术推论是否真实、合理、有据，具体包括3项评估指标。①实体定义规范性：评估系统对古籍中人物、地名、官职、事件等实体进行标准化定义的准确程度，避免实体名称不一致、指代模糊等问题。在这一评估环节，可借助实体消歧技术解决人名、地名、书名等方面存在的歧义问题。例如，基于实体语境和时间知识构建消歧规则，将其应用于《春秋经传引得》的人名实体歧义处理，确保实体指向一致性[19]。②关系抽取准确性：评估系统从古籍文本中提取实体间语义关系的正确性，确保人物归属、事件时序、地理沿革等关联符合古籍记载逻辑。③本体模型适配性：评估系统构建的知识图谱在本体结构设计上是否契合古籍文献的叙事特征与分类体系，避免现代知识框架对古籍原始逻辑的强行套用。

4.4　服务层可信度评估指标

服务层可信度评估从用户视角出发，评估生成内容在具体古籍研究或公众传播场景中的实用性、易用性与启发性，具体包括4项评估指标。①任务目标达成度：评估系统生成内容在具体应用场景中是否达到预设目标。在辅助校勘任务中，考查系统是否有效识别并修正底本讹误，可通过校勘建议采纳率进行量化；在知识普及场景中，考查系统是否以通俗但不失严谨的方式解释难点，如对古籍内容的介绍是否结合古籍知识背景，能否通过标注年代、历史人物、年号、称谓等知识节点提供云讲解服务，当读者指向相关章节时能否自动弹出注释以辅助准确理解文本内涵[20]；在学术研究场景中，可参照上海图书馆“中文古籍联合目录及循证平台”、北京大学“经籍指掌——中国历代典籍目录分析系统”等案例，考查系统是否以细粒度的资源管理能力和专

业研究工具适配学者对内容深度与分析功能的核心需求；在公共传播场景中，考查系统是否通过对古籍原典内容进行通俗化转化与轻量化呈现（如在线阅读、虚拟展览、音视频解读等形式）来降低大众理解古籍文化的门槛，能否满足大众对古籍文化的认知、传播与体验需求[21]；在智能问答场景中，考查系统是否将古籍知识拆解为标准化、碎片化的知识元，能否实现知识内容的快速调取与精准应答以满足即时性、针对性的古籍知识查询需求；在知识图谱构建场景中，考查系统是否以结构化、关联化的形式呈现古籍数字资源，能否通过人物、事件、典籍、时空等核心要素的多维链接为知识图谱的搭建提供可关联、可整合、可运算的基础数据[22]。②语境适配性：评估系统对不同类型问题的回应是否契合问题本身的语义类型与古籍知识语境。具体而言，对于事实查询类问题，系统给出的回答应追求精确性与可溯源性，回答须关联至原始古籍文本的具体出处，避免无源回答或虚构引用；对于考证溯源类问题，系统给出的回答应体现严谨性与论证逻辑，回答须展示推理链条与文献依据；对于阐释分析类问题，系统给出的回答应有据有理，表述风格应符合古籍学术话语体系而非现代口语化表达。③人机协同友好性：评估系统生成结果是否便于学者介入、修正与再创造。具体考查内容包括：输出格式是否支持主流学术软件，以便用户直接导入个人研究环境进行二次加工；不确定性内容是否被明确标注（如置信度区间、多候选方案、分歧版本提示），以帮助用户判断信息的可靠程度并自主选择；交互界面是否直观等。

5　可信治理策略

在GenAI赋能古籍数字化过程中，单纯依赖技术优化或流程规范难以应对日益复杂的可信性风险，本研究构建的四层可信度评估指标体系能成为GenAI赋能古籍数字化的可信风险识别工具，但识别本身并不等同于治理。本研究面向不同层级、多元执行主体提出可信治理策略，形成“可信风险分析—可信度评估—可信治理”闭环：由技术开发方负责系统层面的技术响应与机制部署，由古籍专家、图书馆员、数字人文学者等构成的学术共同体承担专业审核与学术裁断职责，行业主管部门与机构管理层提供制度保障与资源协调。由此推动评估结果切实转化为差异化的治理行动，实现对古籍数字化可信度的系统性保障。

5.1　建立分层响应机制

可信治理的首要前提是将可信度评估结果转化为具体行动指令，针对不同层级暴露的薄弱环节，应实施差异化的干预措施。若源文献保真度或处理过程可溯性评估得分偏低，表明存在原始数据失真或流程黑箱风险。对此，古籍数字化项目实施单位的管理层应强制启动双盲校验机制，由两位独立专家对GenAI输出的文本与原始图像进行比对，并要求系统自动生成完整的处理日志（含模型版本、模型参数、人工干预节点），作为成果发布的必要附件。同时，学术共同体与行业主管部门（如国家古籍保护中心、省级文旅厅）推动建立古籍数字资源元数据强制标准，确保所有公开成果具备可追溯、可复现的技术基础。

当内容层与知识层评估显示文献征引准确性或文本语义一致性出现显著偏差时，往往指向模型“幻觉”问题或知识库缺失问题。对此，技术开发方应在系统中部署动态知识增强机制，在生成过程中实时调用权威古籍知识库，通过检索增强生成技术为模型提供相关文献依据，从而将生成内容限定在可信史料范围之内；对涉及重大史实判断或争议性文本修复的任务，古籍数字化项目实施单位的管理层应强制引入专家会审流程，并要求系统输出多套候选方案及置信度说明，供专家研判决策提供参考。

若服务层评估显示任务目标达成度或人机协同友好性评估不佳，则需技术开发方优化人机交互设计。例如，在校勘辅助场景中，系统应支持用户一键采纳或修正AI建议，并自动记录修正行为用于模型迭代。

5.2　加强资源分级管理与权责协同的制度保障

技术响应需辅以制度支撑。在分级数据治理与著作权合规方面，古籍收藏机构（如图书馆、博物馆、档案馆）的管理层应依据古籍的珍稀性、敏感性及著作权状态，实施动态分级访问控制。对孤本、善本等高价值资源，古籍收藏机构的管理层采用申请、审批、限时访问、操作留痕的闭环管理机制，防止核心数据泄露。同时，技术开发方可利用自然语言处理技术自动提取作者、刊刻年代、出版者等元数据，结合现行著作权法规，建立古籍著作权状态智能甄别系统，辅助判定古籍数字资源的著作权归属与开放权限，为合规共享提供技术支持。

在跨学科协同治理机制方面，学术共同体与行业主管部门应组建由古籍校勘专家、数字人文学者、AI工程师与图书馆员构成的专项工作组。专项工作组在项目初期即制定“AI辅助古籍处理操作规程”，明确各环节的学术审定标准与技术交付要求。其中，古籍校勘专家主导文本准确性、文献完整性的最终裁定，尤其在异体字辨识、语义还原等专业争议上拥有否决权；数字人文学者负责评估知识组织的学术逻辑与语义关联的合理性，确保知识图谱、实体关系等数字化成果符合人文学科的研究范式与叙事逻辑；图书馆员负责对接用户需求与资源管理实践，从服务场景出发提出可用性要求与用户体验标准，并协调古籍原件的调取与数字化流程管理；AI工程师则负责保障系统透明性与可干预性。

在公众参与和素养提升方面，可信生态的构建离不开用户端的理性参与。一方面，文化和旅游部及各级公共图书馆行政主管部门应会同国家古籍保护中心、省级古籍保护中心将用户的信息素养教育融入大语言模型与古籍智能信息处理的实践[23]中，通过培训、指南等形式提升用户对GenAI能力边界的认知，增强用户信息甄别与批判性使用能力；另一方面，古籍收藏机构可组织搭建众包式古籍标注平台，引导社会力量参与校对、关键词标注等工作，既扩充训练数据，又促进知识传播。在此过程中，为保证标注质量，所有工作成果均须经专家审核后方可被采纳。

5.3　构建可信治理长效机制

可信治理不应是静态的合规检查，而应依托一个动态演进的多方协同生态持续运转，具体可从三个层面推进长效机制的构建。在评估反馈层面，建议建立常态化的古籍AI生成内容可信度评估报告制度。可由行业主管部门或学术共同体牵头，定期发布涵盖不同GenAI工具赋能古籍数字化的可信度测评结果，公开典型错误案例与改进路径。这不仅能为学者选择模型提供参考，也能促使技术厂商优化模型在古籍场景下的表现。评估报告的编制宜采用开放式标准，邀请高校、图书馆、出版社等多方参与评估，确保评估结果的公信力与专业性。

在评估预警层面，行业主管部门应牵头设立由专家组成的、跨机构的古籍AI伦理委员会。伦理委员会应由古籍文献学、AI、信息资源管理、科技伦理等领域的专家学者共同组成，负责审议GenAI在古籍数字化领域的新技术应用方案，对潜在的学术风险进行前置研判，并出具风险提示或使用建议。古籍AI伦理委员会的介入宜早不宜迟，应在技术部署前而非问题爆发后进行干预，从而将风险控制在萌芽阶段。

在持续迭代层面，学术共同体与行业主管部门应着力推动高质量古籍语料与评估基准的开源共享。当前制约GenAI古籍数字化可信度的深层瓶颈之一，在于缺乏经专家校勘、版本标注清晰的公开训练语料与标准化测试集。本研究建议由国家级古籍收藏与研究机构牵头，联合高校与科技企业，分批开放经过脱敏与且不存在著作权争议的古籍语料资源，并同步建设覆盖异体字识别、分词、实体抽取、知识关联等任务的权威评估基准。此外，还应通过社区共建共治的模式，吸引更多学者参与语料补充、基准更新与模型评测，形成语料、评估、优化的循环反馈。

6　结语

古籍数字资源是中华优秀传统文化传承传播的核心载体，是极具战略意义的文化数字化资产。作为古籍活态传承的关键路径，古籍数字化不仅为珍贵典籍的永久保存与广泛传播提供了有效支撑，更成为拓展公共文化服务边界的重要引擎。针对GenAI背景下古籍数字化成果可信度不足、风险防控薄弱的现实问题，本研究首先系统梳理GenAI应用于古籍数字化中各环节所引发的风险，进而构建GenAI赋能古籍数字化可信度评估框架与评估指标体系，最终提出可信治理策略，为保障GenAI赋能下的古籍数字化工作既能高效推进，又能坚守可信、可靠、可用的质量底线，提供理论支撑与实践指南。

本研究的局限在于提出的评估指标主要基于理论推演与文献分析，尚未通过大规模实证测试加以验证。各层级指标的具体阈值设定、权重分配以及跨层级的综合评分方法，仍有待通过实际数据集与专家评审相结合的实验研究加以校准。在研究对象方面，本研究以通用型LLM为研究对象，未充分覆盖特定模型类型在古籍数字化过程中可信度表现上的差异性，后续研究将针对特定模型类型开展比较分析。

参考文献

[1]　关于推进新时代古籍工作的意见[EB/OL]. [2026-03-27]. https://

www.gov.cn/zhengce/2022-04/11/content_5684555.htm.

[2] 关于推进实施国家文化数字化战略的意见[EB/OL]. [2026-03-27]. https://www.gov.cn/xinwen/2022-05/22/content_5691759.htm.

[3] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[4] HUANG L, YU W, MA W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[5] SHAO A. New sources of inaccuracy? A conceptual framework for studying AI hallucinations[J]. Harvard Kennedy School Misinformation Review, 2025, 6(4): 1-13.

[6] 钱明辉, 李胡蓉, 杨建梁. 大语言模型可信: 内涵、影响、挑战与对策[J]. 图书情报工作, 2024, 68(20): 69-86.

[7] 杨艳妮, 武雷龙, 向安玲, 等. 形式信任与递进归谬: 人工智能生成内容可信度评估[J]. 新媒体与社会, 2025(2): 24-143, 464-465.

[8] 宋筱璇, 刘浏, 黄水清. 信息资源管理AI生成内容风险管控的基本思路与框架[J]. 图书情报工作, 2025, 69(20): 30-38.

[9] 刘冰, 侯志江. 生成式AI诱发信息生态污染的风险机理及管控要素仿真研究[J]. 情报理论与实践, 2026, 49(2): 92-102.

[10] 胡婷, 杨昊, 王为. 生成式人工智能在古籍保护中的创新应用、风险样态与治理策略[J]. 图书馆工作与研究, 2026(1): 87-93.

[11] 陈雨, 杨璐颖, 冯锐. 从内容生产到秩序重塑: GenAI出版的内容生产风险与规制研究[J]. 出版广角, 2024(22): 61-67.

[12] 刘畅, 张琪, 王东波, 等. 基于大语言模型技术的古籍限定域关系抽取及应用研究[J]. 情报学报, 2025, 44(2): 200-219.

[13] 张焕培. DeepSeek驱动图书馆智慧化转型: 理论逻辑、安全风险与优化路径[J]. 情报杂志, 2025, 44(8): 190-197.

[14] 管涛, 许茂恒. 可信数据空间的构建逻辑与路径: 基于全生命周期理论[J]. 图书与情报, 2025(5): 123-132.

[15] 陈鑫, 王浩, 刘倩, 等. 情报思维下AIGC可信性的全生命周期保障框架与机制研究[J]. 情报理论与实践, 2026, 49(8): 95-103, 124.

[16] 李敏瑜, 李永卉, 卢章平. 古方志历史名人资源知识图谱构建与知识发现研究[J]. 图书馆杂志, 2026, 45(1): 61-73.

[17] 李雅, 李雅辉, 周毅. 古籍数字化标准建设的框架、内容与策略研究[J]. 图书情报工作, 2026, 70(11): 110-124.

[18] 张强, 关书朋, 许芳红. 智能体技术驱动下先秦元典性文献知识图谱构建研究[J]. 南通大学学报(社会科学版), 2026, 42(1): 142-154.

[19] 刘洋, 王东波. 古籍智能信息处理研究现状[J]. 图书情报工作, 2024, 68(23): 120-138.

[20] 傅宝珍. 知识服务背景下古籍VR阅读推广研究[J]. 图书馆工作与研究, 2022(1): 108-115.

[21] 雷珏莹, 王晓光. 可供性理论下古籍数字平台的价值重构和提升策略[J]. 出版发行研究, 2023, (11): 19-24.

[22] 肖超, 高兆强. 出版社古籍数字出版商业模式研究: 基于商业模式画布理论[J]. 科技与出版, 2021(11): 74-84.

[23] 张海, 赵雪, 王东波. 大语言模型下古籍智能信息处理: 构成要素、框架体系与实践路径研究[J]. 信息资源管理学报, 2024, 14(5): 36-44.

作者简介

付雅明, 女, 博士, 助理研究员, 研究方向: 数字人文与知识组织, E-mail: ymingfu@126.com。
孙炫宇, 女, 硕士研究生, 研究方向: 数字人文与知识组织。

Research on Trustworthiness Assessment and Credibility Governance Strategy of Ancient Book Digitization Empowered by GenAI

FU YaMing SUN XuanYu
(School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, P. R. China)

Abstract: While Generative Artificial Intelligence (GenAI) empowers the digitalization of ancient texts in public libraries, it also introduces trustworthiness risks such as data distortion, model hallucination, and ambiguity of rights and responsibilities, which urgently require systematic governance. This study first analyzes the trustworthiness risks arising from the application of GenAI in the digitalization of ancient texts, and then constructs a four-layer trustworthiness assessment framework tailored to the characteristics of ancient texts, covering the foundation layer, content layer, knowledge layer, and service layer. Based on this framework, a trustworthiness assessment system comprising 13 evaluation metrics is designed to enable structured and operable judgment of generated content. Finally, three trust governance strategies are proposed: establishing a tiered response mechanism, strengthening institutional safeguards for graded resource management and collaborative responsibility allocation, and building a long-term trust governance mechanism, thereby forming a complete closed loop of trust risk analysis—trustworthiness assessment—trust governance. The aim is to provide theoretical and practical references for the compliant and high-quality application of GenAI in the digitalization of ancient texts.

Keywords: GenAI; Digitization of Ancient Books; Trustworthiness Risks; Credibility Assessment; Credibility Governance