基于“人在回路”的生成式人工智能在图书编目中的应用
——以Kimi和DeepSeek技术适配性对比为例

吴诺曼　赖伟　胡琳
（四川大学图书馆，成都 610065）

摘要：生成式人工智能为图书馆编目业务提供了新的技术路径，但其专业应用面临技术适配性挑战。本研究基于“人在回路”（HITL）机制设计了“基线测试-模型对比-定向优化”三阶段实验框架，系统评估了Kimi与DeepSeek两款国产大模型在西文图书MARC21编目中的表现。实验结果表明：在无专业干预条件下，两款模型原生能力存在显著差异（Kimi F1=7.41%，DeepSeek F1=51.30%）；经统一提示词引导后，DeepSeek（F1=83.00%）综合表现优异，显著优于Kimi（F1=63.50%）；进一步实施精细化提示工程后，DeepSeek（F1=95.16%）性能实现跃升。通过人类编目员的动态校验与反馈，生成式人工智能可突破初始技术限制，实现从通用对话到专业编目任务的适配。研究验证了“人在回路”机制在模型选型与性能优化中的有效性，提出了建立提示词知识库、实施字段分级管控等实践建议，为图书馆构建人机协同的智能编目体系提供了参考方案。

关键词：生成式人工智能；人在回路；智能化编目；Kimi；DeepSeek
中图分类号：G254.36　DOI：10.3772/j.issn.1673-2286.2025.10.003
引文格式：吴诺曼，赖伟，胡琳. 基于“人在回路”的生成式人工智能在图书编目中的应用：以Kimi和DeepSeek技术适配性对比为例[J]. 数字图书馆论坛，2025，21（10）：23-29.

在数字化与智能化浪潮的推动下，图书馆业务正经历着深刻的变革。图书编目作为信息资源组织与服务的基石，其效率与质量直接关系到馆藏资源的可发现性与可利用性。然而，传统人工编目流程烦琐、耗时费力，在面对海量且持续增长的文献资源时，其效率瓶颈日益凸显。近年来，生成式人工智能技术的突破性发展，为编目业务的自动化、智能化升级提供了新的可能。图书馆应克服技术焦虑，理性评估人工智能的“辅助”而非“替代”角色[1]，利用生成式人工智能在文本识别、自然语言处理等方面的强大能力[2]提升编目效率，优化资源服务。

在此背景下，四川大学图书馆于2024年启动了“AI+专业馆员”培养计划，支持馆员在工作中找到人工智能助力业务发展的生长点。该馆文科专款采购的西文图书多为2022—2024年出版，其原始编目比例高（2024年约为15%），要求编目员逐字段录入、查询与核对，导致工作量繁重、新书上架周期变长。为此，该馆成立了“AI+编目自动化”小组，以RDA西文编目实践为起点，引入“人在回路”（HITL）协同机制，探索生成式人工智能在编目业务中的技术适配路径。其目标不仅在于提升图书的原始编目效率，更深远的价值在于将编目员从重复性劳动中解放出来，使其能够转向知识组织、深度标引等更具创造力的任务，从而提升图书馆业务的整体价值。

收稿日期：2025-09-26

1 研究现状

1.1 既有智能化编目方案的多元路径与应用壁垒

图书编目业务的智能化转型，是图书馆应对资源激增与提升服务效能的必然选择。早期的探索实践，即智能化编目，主·要依托传统的自然语言处理与计算机视觉等技术，形成了多元的技术路径，并在特定场景下取得了显著成效。在主题标引与分类领域，芬兰国家图书馆开发的Annif开源工具包集成了多种文本分类算法，可实现主题词与分类号的自动分配[3]；德国国家图书馆采用文本分类算法实现资源的自动分类，生成杜威十进制分类号和主题词规范[4]；上海交通大学图书馆通过支持向量机与概率算法构建了外文电子书刊自动分类专家系统[5]。在集成化流程方面，比利时皇家图书馆构建了从扫描、信息提取到数据检索的完整自动化编目流程[6]；与之类似，广东省立中山图书馆的“采编图灵”则采用基于神经网络的专用人工智能模型，通过扫描图书关键页面自动识别并提取ISBN、题名等核心元数据，实现了采编流程的高度自动化[7]。然而，上述方案通常依赖专用算法、高质量训练数据与可观的算力支持，其技术门槛与实施成本均较高，客观上形成了普通图书馆难以跨越的应用壁垒。

1.2 生成式人工智能的技术突破与范式革新

生成式人工智能，特别是大语言模型技术的突破，为解决上述高门槛问题带来了范式变革的契机。其核心优势在于，通过提示工程（Prompt Engineering）即可实现非专业开发者的低代码应用，为人机协同提供了更友好的交互范式[8]。这一特性使其成为图书馆界探索的新方向：圣地亚哥州立大学图书馆利用ChatGPT为政府文档自动生成摘要与非控制主题词[9]；俄克拉荷马州立大学图书馆将其用于学位论文关键词与摘要的自动创建[10]；卡普兰诺大学图书馆则基于ChatGPT，开发了名为CatalogerGPT的工具，有效提升了MARC记录的生成效率[11]。这些实践表明，生成式人工智能为编目业务智能化提供了一条低代码、轻量化的新路径。

1.3 HITL协同范式的共识与深化方向

生成式人工智能在带来便捷的同时，输出结果的准确性与专业性仍存有不确定性。正因如此，HITL作为一种关键的人机协同范式，其重要性日益凸显。HITL是一种人机协同的设计理念，强调人在系统闭环中的主动参与，通过人的感知、判断和操作来弥补机器的局限，已在人工智能、控制工程、教育技术、虚拟仿真等多个领域获得广泛应用[12]。它不仅提升了系统的性能和可靠性，也增强了可解释性和用户体验。美国国会图书馆在其“探索计算描述”项目[13]中构建了HITL的编目模式，系统地将人工智能生成的元数据交由人工复核；比利时皇家图书馆[6]与卡普兰诺大学图书馆[11]的实践同样表明，编目员的验证与反馈是流程中不可或缺的环节。然而，现有研究对人工参与和完全无人参与两种模式下的编目效能差异，尚缺乏贯穿全程的系统性量化对比，难以精确评估HITL在不同阶段的实际价值。构建一个编目员主动参与、深度互动，并在提示优化、过程指导与结果校验全流程中发挥能动性的协同机制，仍是当前研究有待深入探索的领域。

2 研究方法

2.1 实验方案

研究旨在系统探究HITL模式下，生成式人工智能辅助西文图书编目的可行性与效能。为考察大模型的初始编目能力和人机协同模式下的工作效果，设计了包含基线能力测评、模型对比实验、定向优化验证的递进实验框架（见图1）。

2.2 实验对象

2.2.1 编目对象与测试集构建

实验以西文图书的原始编目为核心场景，旨在解决业务效率瓶颈最突出环节的问题。为在保证评估信度的同时，兼顾实验的可操作性，构建了一个包含50条书目记录的测试集。其构建遵循目的性抽样原则，以确保样本在关键维度上具有代表性，从而较为全面地评估模型的泛化能力。学科覆盖方面，约1/2的样本来源于

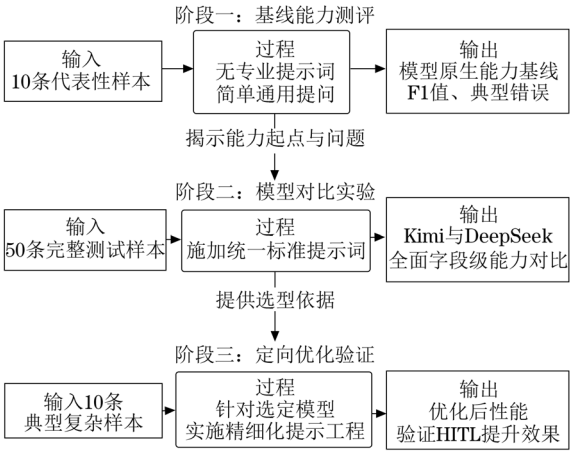

图1　总体实验方案流程

四川大学图书馆文科专款采购的待编新书，同时从既有馆藏中选取了占总数约1/2的理工、医学西文图书，用以测试模型的跨学科认知与泛化能力。出版时间跨度方面，样本主要为2022—2024年的出版物，并且有意识地纳入了占总数约1/10的出版于1970年之前的图书，用以检验模型对陈旧版式、历史拼写规则及普遍缺失现代标识符（如ISBN）等情况的处理能力。文献类型与编目复杂度方面，样本覆盖了10种典型、复杂的编目场景（见表1），包括专著、会议录、双语出版、多卷书、丛书、影印本、改编教材等，确保测试能有效反映模型在题名解析、责任者认定、版本甄别、层级关系处理等方面的能力。

表1　测试记录中10种典型西文图书样例

序号	分类特征	书名	测试重点
1	旧版专著	Tensor calculus	老旧版权页识别
2	新版专著	A provincial history of the Ottoman Empire: Cyprus and the Eastern Mediterranean in the nineteenth century	副题名解析、多作者处理
3	会议录	Prospects in international investment law and policy: World Trade Forum	会议名称标准化提取
4	双语出版	Robinson Crusoe 鲁滨逊漂流记	并列题名与翻译字段处理
5	多卷书	Encyclopaedia of computational law（Vol.1–4）	分卷题名提取
6	团体报告	Annual report on cardiovascular health and diseases in China（2022）	机构名称规范控制
7	改编影印教材	Textbook of otolaryngology, head and neck surgery 3rd ed.	版本项与多责任者信息提取
8	多版本图书	How to be an academic superhero: establishing and sustaining a successful career in the social sciences, arts and humanities	版本差异识别（如平装/精装/电子版以及不同版次）
9	稀有版本影印	Dialektik der natur	影印说明识别
10	丛书分册	From Twitter to Capitol Hill: far-right authoritarian populist discourses, social media and critical pedagogy	丛书项识别与层级关系处理

2.2.2　人工智能模型与编目员的选定

选取Kimi与DeepSeek两款国产大模型，技术选型主要基于以下3个因素。①业务适配性：预实验表明，国产模型对涉及双语出版、影印版等本土化场景的语义理解优于部分国际模型。②基础设施与合规性：国际大模型的网络访问稳定性难以满足常态化业务需求，且存在数据合规风险。③技术特性互补：Kimi以长文本解析见长，DeepSeek在结构化数据生成方面表现优异，二者对比可揭示不同技术路线的适配特征。

HITL中“人”的角色由具备CALIS三级编目员资质的馆员担当，其职责贯穿全流程，包括设计提示模板、执行字段级校验、标注错误模式、优化提示策略。

2.3　实验流程

第一阶段：基线能力测评（无人工干预，简单通用提问）。确立模型在无专业干预下的初始编目能力基线。对基线测试子集中的10条记录，仅向模型提出通用问题“请为这本书生成MARC21格式的编目数据”，并上传图书题名页、封面等图像。不提供任何专业提示或范例。收集模型的原始输出，用于后续对比分析。

第二阶段：模型对比实验（统一提示词）。在相同条件下对比两款模型的技术适配性。对50条测试集样本，使用统一的基础提示词模板，该模板包含指令：“这是一本图书的题名页、版权页和封面照片，稍后我将上传MARC21格式的范例，你按格式为该图书编目。”MARC21格式范例的核心是为模型提供明确的任务框架、专业规则与输出格式约束，形式如下。

020 L $$a9781032463414$$qhardcover

020 L $$z9781000884456$$qPDF ebook

020 L $$a9781000884432$$qelectronic book

24510 L $$aSocial equity in the public administration

classroom/$$cedited by Michaela E. Abbott，Bruce D. McDonald III，William Hatcher.

264 1 L $$aLondon；$$aNew York：$$bRoutledge，$$c2023.

336 L $$atext$$2rdacontent

337 L $$aunmediated$$2rdamedia

338 L $$avolume$$2rdacarrier

6500 L $$aSocial justice.$$0Social equity.

6500 L $$aPublic administration$$xStudy and teaching.

7001 L $$aAbbott，Michaela E.，$$eeditor.

7001 L $$aMcDonald，Bruce D.，III，$$eeditor.

7001 L $$aHatcher，William，$$dCollege teacher，$$eeditor.

按照表2规则进行逐字段校验与标注，计算两款模型的精确率、召回率和整体F1值。

表2　人工智能编目测试判定规则

类别	定义	编目场景示例
真阳性（TP）	字段应存在且被正确生成（内容与格式均符合规则）	个人名称正确录入100字段，名称规范通过LCNAF验证
假阳性（FP）	字段被生成但不应存在或内容/格式错误	①无ISBN却生成020字段；②245$b副题名解析错误
假阴性（FN）	字段应存在但未被生成	应著录700字段的合著者信息缺失

第三阶段：定向优化验证（精细化提示工程）。验证HITL机制对模型性能的持续提升能力。基于第二阶段结果，选择表现更优的模型进行优化，其措施为：提供特定文献类型（如会议录、多卷书）的编目范例；针对高频错误添加精准规则（如“汉语人名须添加中文子字段”“禁止把出版者著录到710”）。使用10条典型复杂样本评估优化效果。

2.4　数据收集与评估

（1）数据收集类型。量化数据：记录各模型在各字段的真阳性、假阳性、假阴性数量。质性数据：详细记录典型的错误模式及其发生语境。效率数据：记录人工校对耗时，与传统编目对比。

（2）效能评估指标。采用F1值作为核心评估指标。F1值是机器学习领域广泛采用的模型性能评估指标，通过综合精确率与召回率，为分类任务提供平衡性评价。F1值是精确率（P，正确生成字段数与总生成字段数之比）与召回率（R，正确生成字段数与应生成字段数之比）的调和平均数，避免因字段生成难度不均而产生的评估偏差。精确率、召回率及F1值计算公式分别如式（1）~式（3）所示。

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}} \tag{1}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}} \tag{2}$$

$$F_1=2\times\frac{P\times R}{P+R} \tag{3}$$

式中：N_{TP}、N_{FP}、N_{FN}为真阳性、假阳性、假阴性字段数量。

（3）测试字段选择。实验测评12个核心字段（见表3），排除了信息获取不完整的字段（如008、300、504、650），以及高度确定性字段（如336、337、338可达到100%准确，将其排除能更准确地甄别模型在复杂场景下的真实能力差异）。

表3　人工智能编目测试字段

字段描述	字段	字段描述	字段
ISBN	020	版本说明	250
个人名称	100	出版发行项	264
团体名称	110	丛编说明	490
会议名称	111	原版附注	534
题名与责任说明	245	个人名称附加项	700
变异题名	246	团体名称附加项	710

3　实验结果与讨论

3.1　阶段一：基线能力测评——模型原生编目能力评估

为确立生成式人工智能在图书编目任务中的原生能力基准，在无任何专业干预的条件下对Kimi和DeepSeek进行基线能力测试，结果如表4所示。可见，在未经过专业提示词引导的情况下，编目效能存在显著差异，且均未达到业务可用的基本要求。

Kimi模型表现出极低的综合编目能力（F1=7.41%）。其输出呈现出严重的结构化缺失，具体表现

表4　两款模型编目能力的基线测试结果

模型	F1值/%	精确率/%	召回率/%	关键发现
Kimi	7.41	4.23	30.00	输出基本不可用，字段混淆，格式错误严重
DeepSeek	51.30	35.10	95.20	能识别大部分应著录字段，但精确率极低

为：混用中西文编目字段，而非标准的MARC21格式；字段标签使用混乱，如将题名或出版者信息错误标记为1××字段；对基本编目概念的理解存在根本性偏差，几乎无法产出符合要求的编目记录。

相比之下，DeepSeek模型展现出相对较好的字段识别能力（召回率95.20%），能够识别出大部分应当著录的字段信息。然而，该模型存在严重的精确率问题（精确率35.10%），主要表现为：生成大量冗余字段，如在已经正确著录264字段的情况下，又频繁将出版者信息著录到710字段；产生内容性错误，如错误解析并列题名、分卷题名、团体责任者名称等。

基线测试结果表明，生成式人工智能在缺乏专业引导的情况下，无法直接应用于正式的编目工作环境。

3.2　阶段二：模型对比实验——标准化提示下的效能测试

在确立基线能力后，进入核心的模型对比阶段。通过对50条测试样本施加统一的标准化提示词，编目员对Kimi和DeepSeek进行了全面的编目效能评估。两款模型的整体性能对比如表5所示，显示出DeepSeek在综合效能上的显著优势。

从整体性能来看，DeepSeek在F1值、精确率和召回率3个指标上均显著优于Kimi。特别是其98.40%的召回率，意味着该模型能够近乎完整地捕捉所有应著录字段信息，字段遗漏风险极低。与此同时，72.00%的精确率表明其输出结果具有较高的直接可用性，能够有效减少人工校对工作量。

表5　模型整体性能对比

模型	F1值/%	精确率/%	召回率/%	性能评价
Kimi	63.50	56.20	73.20	基本可用，复杂字段处理能力弱
DeepSeek	83.00	72.00	98.40	综合表现优异，具备主力工具潜力

为了深入解析两款模型的技术特性差异，将两款模型在12个核心字段上的详细性能进行对比，结果如表6所示。

阶段二的对比实验表明，DeepSeek在综合性能、专业场景覆盖度和输出稳定性方面具有明显优势，而Kimi则因能力不均衡，仅适用于简单编目任务。该发现为图书馆工具选型提供了明确依据——DeepSeek是更可靠的编目辅助工具，并且为下一阶段针对该优势模型的定向优化奠定了基础。

3.3　阶段三：定向优化验证——基于DeepSeek的精细化提示工程

基于阶段二的对比结果，选定DeepSeek开展第三阶段的定向优化验证。通过对10条包含复杂编目场景的样本进行精细化提示工程实验，进一步验证HITL机制对模型性能的持续提升能力。DeepSeek典型错误与针对性优化策略如表7所示。

表6　Kimi与DeepSeek编目字段效能对比

字段描述	字段	F1值/%		差距分析
		Kimi	DeepSeek	
ISBN	020	81.97	100.00	DeepSeek完美识别，Kimi存在漏录与识别错误
个人名称	100	80.00	94.70	DeepSeek在名称规范控制上更精准
团体名称	110	0	88.90	Kimi基本无法处理，DeepSeek表现稳健
会议名称	111	33.30	88.90	Kimi能力严重缺失
题名与责任说明	245	63.01	94.70	DeepSeek在子字段拆分与著录上远胜Kimi
变异题名	246	0	85.70	Kimi完全缺失该能力
版本说明	250	0	90.90	Kimi漏录严重，DeepSeek能准确识别
出版发行项	264	59.15	75.00	两者均须优化格式，但DeepSeek基础更好
丛编说明	490	66.67	93.30	DeepSeek对丛书关系的理解力更优
原版附注	534	0	0	两者效果均不理想，但DeepSeek未遗漏字段
个人名称附加项	700	59.46	90.50	DeepSeek在复杂责任方式著录上更可靠
团体名称附加项	710	0	20.60	两者效果均不理想，但DeepSeek具备基础能力

表7　DeepSeek典型错误与针对性优化策略

错误类型	典型字段	优化前问题描述	HITL优化策略	预期改进目标
规范控制不足	100、700	汉语人名缺少中文名称形式	增加指令“汉语人名须添加$9子字段，提供中文名称”	提高个人名称字段的规范一致性
字段概念混淆	710	将出版者误录为团体责任者	明确约束“禁止将出版者著录到710字段”	降低710字段的误用率
复杂类型处理	111、245、534	会议录、多卷书、影印书等著录不规范	提供特定文献类型的标准编目范例	提升对复杂文献类型的处理能力
格式细节偏差	264	并列出版地标点使用不规范	预设格式模板，明确标点使用规则	提升出版信息的格式规范性

经过一轮精细化提示工程优化后，DeepSeek的编目效能实现了显著提升。DeepSeek优化前后性能对比如表8所示，优化后的DeepSeek模型保持高召回率（98.33%），说明字段捕捉能力基本稳定；精确率得到大幅度提升，从72.00%提升至92.18%，增长超过20个百分点，说明错误字段大幅减少；F1值从83.00%提升至95.16%，说明综合编码能力得到提升。

表8　DeepSeek优化前后性能对比

类别	F1值/%	精确率/%	召回率/%
优化前	83.00	72.00	98.40
优化后	95.16	92.18	98.33

本阶段的优化效果验证了HITL机制在模型性能优化中的核心价值。通过精细化的提示工程，成功地将人类编目员的专业知识转化为模型的可执行规则，实现了从“可用”到“好用”的提升。精确率的显著提升意味着人工校对工作量的大幅降低，使编目员能够将精力更多地集中在更高层次的语义判断与规范控制上。

在时间方面，编目员反馈表明，对DeepSeek生成的记录进行校对和修改，每条记录耗时5~10分钟；相比之下，从零开始进行原始编目，需要30~40分钟，实验带来超过70%的时间节省，体现了HITL模式在提升效率方面的巨大价值。

4　结语

研究通过引入HITL协同机制，通过3个递进阶段的实验，系统探索了生成式人工智能在RDA西文图书编目中的技术适配路径。实验结果表明：在无专业干预条件下，两款模型原生能力存在显著差异（Kimi F1=7.41%，DeepSeek F1=51.30%）；经统一提示词引导后，DeepSeek综合表现优异（F1=83.00%），显著优于Kimi（F1=63.50%）；进一步实施精细化提示工程后，DeepSeek性能实现跃升（F1=95.16%）。研究结果表明，图书馆可通过系统性工作有效提升人工智能工具的实用价值。基于此，建议各机构在具体实践中建立本地化的提示词与范例知识库，将优化策略转化为可复用的数字资产。同时，可根据字段级效能数据实施分级管控：对高F1值字段（如020、100）采用抽样校对，对复杂字段（如111、534）保持全量审核，从而实现编目效率与质量的动态平衡。需要明确的是，生成式人工智能的定位是辅助工具而非专业替代，其在主题标引、规范控制等领域仍需人工把关。这种分工协作模式既发挥人工智能的处理效率优势，又保留专业馆员的质量控制角色，为图书馆智能编目转型提供可行路径。

研究的局限性主要体现在：实验样本限于西文图书，未来须拓展至中文图书与多模态资源；通用大模型的“会话隔离”机制制约了知识的长期积累；此外，主题标引字段（650）与分类字段未被纳入测评范围，因其特殊性，研究采取了“可接受”“过于宽泛”“过于狭窄”“错误”四级主观评价，无法通过F1值进行量化统计。尽管如此，实验发现人工智能仅依据题名页等有限信息生成的主题词仍具重要参考价值，能够有效缩小候选范围，为编目员节省大量筛选时间，体现了人机协同在提升主题标引效率方面的潜力。

基于研究成果，下一步将着力开发编目专用智能体，重点突破两大技术瓶颈：通过持久化记忆模块固化优化策略，解决通用模型“会话隔离”问题；整合知识库与专业工具，提升在主题标引等复杂任务上的表现。这种轻量化方案相比传统定制开发更具推广价值。

研究验证了HITL机制的实际价值：通过建立“评估-选型-优化”的递进工作流程，不仅推动了生成式人工智能在编目任务中从不可用、可用到好用的渐进发展，也为探索编目员角色转型（从基础数据加工转向提示词设计、质量管控与系统优化等更高层次工作）提供了实践依据。

参考文献

[1]　吴建中，程焕文，周德明，等. 理念革新　技术重塑：人工智能

时代图书馆的专业发展与创新实践专家笔谈[J]. 图书馆杂志，2025，44（8）：4-19.

[2] 刘炜，林海青. 人工智能与波普尔“世界4”：智慧时代图书馆的认知重构[J]. 图书馆论坛，2025，45（9）：72-82.

[3] Annif. Annif: automated subject indexing and classification [EB/OL]. [2025-07-20]. https://annif.org/.

[4] MÖDDEN E. Artificial intelligence，machine learning and bibliographic control：DDC short numbers—towards machine-based classifying[J]. JLIS.it，2022，13（1）：256-264.

[5] 罗鹏程，王继民，聂磊. 基于生成式大语言模型的文献资源自动分类研究[J]. 情报理论与实践，2024，47（12）：174-182.

[6] Royal Library of Belgium. One automatic cataloging flow：tests and first results[EB/OL]. [2025-07-20]. https://repository.ifla.0rg/handle/123456789/2686.

[7] 肖燕. 图书采分编的自动化技术运用：以广东省立中山图书馆“采编图灵”为例[J]. 图书馆论坛，2024，44（4）：20-28.

[8] 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.

[9] San Diego State University. Enhancing cataloging of electronic government documents with programming and OpenAI[EB/OL]. [2025-07-20]. https://osf.io/39bws.

[10] Oklahoma State University. Teaching student workers to use generative AI to create metadata[EB/OL]. [2025-07-20]. https://www.xcdsystem.com/tla/program/pAJExb8/index.cfmpgid=806&sid=381828-abid=113425.

[11] GREENLY G. CatalogerGPT[EB/OL]. [2025-07-20]. https://glog.greenly.wixsite.com/catalogergpt.

[12] 李晓欢，郑钧柏，康嘉文，等. 基于人在回路的纵向联邦学习模型可解释性研究[J]. 智能科学与技术学报，2024，6（1）：64-75.

[13] POTTER A，SACCUCCI C. Could artificial intelligence help catalog thousands of digital library books? An interview with Abigail Potter and Caroline Saccucci[EB/OL]. [2025-07-20]. https://blogs.loc.gov/thesignal/2024/11/could-artificial-intelligence-help-catalog-thousands-of-digital-library-books-an-interview-with-abigail-potter-and-caroline-saccucci/.

作者简介

吴诺曼，女，硕士，馆员，研究方向：文献编目、知识组织、阅读推广，E-mail：1138462@qq.com。
赖伟，男，硕士，馆员，研究方向：数字图书馆、数据分析。
胡琳，男，硕士，研究馆员，研究方向：信息素养、人工智能素养。

Application of Generative Artificial Intelligence Based on “Human-in-the-Loop” in Bibliographic Cataloging: Comparative Study on Technical Adaptability of Kimi and DeepSeek

WU NuoMan　LAI Wei　HU Lin
(Sichuan University Library, Chengdu 610065, P. R. China)

Abstract: Generative artificial intelligence offers new technological pathways for library cataloging, yet its professional application faces technical adaptability challenges. Based on the “human-in-the-loop” (HITL) mechanism, this study designs a three-phase experimental framework—baseline testing, model comparison, and targeted optimization—to systematically evaluate the performance of two domestic large language models, Kimi and DeepSeek, in MARC21 cataloging of Western-language books. Experimental results indicate that without professional intervention, the native capabilities of the two models differ significantly (Kimi F1=7.41%, DeepSeek F1=51.30%). Under unified prompt guidance, DeepSeek demonstrates superior overall performance (F1=83.00%), significantly outperforming Kimi (F1=63.50%). After implementing refined prompt engineering, DeepSeek’s performance achieves a notable improvement (F1=95.16%). Through dynamic calibration and feedback from human catalogers, generative artificial intelligence can overcome initial technical limitations and adapt from general dialogue to professional cataloging tasks. This study validates the effectiveness of the HITL mechanism in model selection and performance optimization, and proposes practical recommendations, such as establishing a prompt knowledge base and implementing hierarchical field management, providing a reference solution for building a human-machine collaborative intelligent cataloging system in libraries.

Keywords: Generative Artificial Intelligence; Human-in-the-Loop; Intelligent Cataloging; Kimi; DeepSeek

（责任编辑：常春）