内嵌RAG的科技文献发现系统测评方法研究*

陈阳　孟旭阳　陈熙
（中国科学技术信息研究所，北京 100038）

摘要：检索增强生成（Retrieval-Augmented Generation，RAG）技术已广泛应用于科技文献发现系统。但当前发现系统测评方法仍存在人工成本高、难以适配RAG架构特性等问题。因此，本研究提出一种内嵌RAG的科技文献发现系统测评方法，从检索质量、生成质量和溯源质量三个维度构建测评指标体系。首先利用自动化测评获得各指标的量化结果，再采用融合大模型智能评分与人工核对的混合评估方法进行评价，最后通过斯皮尔曼、肯德尔相关性检验验证指标量化结果与混合评估结果的一致性。本研究以国家科技图书文献中心（National Science and Technology Library，NSTL）发现系统为实证对象进行验证。研究结果表明，本研究提出的测评方法能有效识别系统在科技文献检索、生成和溯源方面的差异，且自动化测评结果与混合评估结果呈显著正相关，验证了本研究所构建指标体系的可靠性与有效性，为科技文献发现系统测评与迭代升级提供了实践支撑。

关键词：科技文献；检索增强生成；指标体系；自动化测评；NSTL
中图分类号：G251　DOI：10.3772/j.issn.1673−2286.2026.06.006
引文格式：陈阳，孟旭阳，陈熙. 内嵌RAG的科技文献发现系统测评方法研究[J]. 数字图书馆论坛，2026，22（6）：59−69.

科技文献发现系统从出版商、内容服务商、大学图书馆、公开网站等渠道提取各种元数据资源，利用抽取、映射、规范、融合等手段，将多源异构资源进行全面整合与深度揭示[1]。其发展历程可分为经典检索、语义检索与生成式人工智能检索三个阶段，每个阶段都以核心技术跃迁为驱动，在资源组织深度、检索匹配精度与知识服务能力上实现阶梯式升级[2]。

当前，以ChatGPT[3]、Deepseek[4]为代表的大语言模型（Large Language Model，LLM）正加速发现系统向生成式人工智能检索范式演进，但存在模型幻觉、知识过时及推理过程难以追溯等问题[5]。检索增强生成（Retrieval-Augmented Generation，RAG）技术依托外部知识库实时检索相关信息以增强LLM能力，有效弥补通用模型在垂直领域专业性不足的缺陷，为提升专业文献检索效率、优化知识问答质量、赋能内容智能生成提供了全新路径，成为数字图书馆推动服务智能化转型的核心方向。

然而，内嵌RAG技术的科技文献发现系统在应用成效测评时面临严峻挑战。一是嵌入RAG技术的系统具有“检索—生成”混合架构且依赖动态知识源，单独评估任一环节无法反映系统整体效能，且传统检索指标难以适配对检索质量进行细粒度评测的需求[6]。二是传统测评方式过度依赖人工评估，不仅评估效率低、成本高昂，且受限于评估者的主观偏见，难以实现大规模系统的快速迭代[7]。这些问题导致现有发现系统效能测评缺乏科学依据，制约了模型选型优化与系统迭

收稿日期：2025−05−20
*本研究得到国家社会科学基金青年项目“大模型赋能下的学术文献智慧服务新模式与应用研究”（编号：24CTQ008）资助。

代升级。

为此，本研究提出一种内嵌RAG的科技文献发现系统测评方法。从检索质量、生成质量和溯源质量三个维度构建测评指标体系，利用自动化测评获得各指标的量化结果；进一步融合大模型智能评分与人工核对的混合评估方法，并通过斯皮尔曼、肯德尔相关性检验验证指标量化结果与混合评估结果的一致性，从而实现对内嵌RAG的科技文献发现系统核心效能的精准与全面测评。

1 相关研究

1.1 RAG技术及应用

RAG技术由Lewis等[8]于2020年首次提出，其核心架构包含检索模块、融合模块与生成模块，通过动态检索外部知识库并将检索结果与输入提示融合，弥补了传统LLM在知识更新滞后与事实性错误方面的固有缺陷。RAG技术在提升检索召回率、重排序机制，以及生成器对长上下文的适应能力等方面表现出色[9-10]，还衍生出ModularRAG[11]、FlashRAG[12]、GraphRAG[13]、AgenticRAG[14]等多种范式。其中AgenticRAG引入了智能体的自主决策和编排机制，并融合数据库、模型微调、逻辑推理等多种技术，能适应各种复杂灵活的任务场景。

目前，RAG技术已广泛应用于问答系统、文本摘要、知识密集型任务等多种场景[15]，在多文档场景中表现尤为突出[16]，能显著提升生成内容的准确性和相关性。此外，胡伟群等[17]提出基于RAG的增量式生成方案，通过融合RAG与知识图谱的动态关联能力，以解决传统知识图谱生成中语义割裂与更新滞后的问题。尚渡新等[18]构建了一种融合RAG与LLM微调的学术短文本分类框架，进一步提升多标签分类性能。宋保仁等[19]深入剖析RAG在图书馆私有知识库构建中的应用，明确其在语义检索优化、多模态资源整合和用户导向服务提升方面的技术优势。张浩然等[20]提出动态细粒度检索增强生成（DF-RAG）方法，通过查询重写以及将外部文档纳入响应生成，实现动态细粒度的检索增强。

1.2 发现系统性能测评

随着技术的不断迭代，如何客观、准确地评估发现系统的性能成为核心问题。早期测评多沿用Recall@k、Precision@k、MRR、NDCG@k等通用型指标，或依赖BLEU[21]、ROUGE[22]等自然语言处理自动评估指标，难以精准衡量系统在知识挖掘深度、推理可靠性以及特定科研场景下的适配能力。欧荣[23]从检索范围、检索字段、检索方式、检索结果处理、个性化服务等角度对主流生物医学文献检索工具的性能进行对比测评。王莉等[24]采用Pooling技术对NSTL智能检索平台的扩展检索效果进行测评与分析。此外，宋琳琳等[25]从用户视角对国内20所大学图书馆的资源发现系统进行调查，发现用户更注重实用性、直观性和学术深度。

近年来，RAG技术在各类发现系统中逐步实现规模化落地，传统单一化、模块化的性能测评体系已难以适配新型技术架构的应用需求。在此背景下，面向RAG全流程测评的自动化评估框架应运而生，为系统量化测评提供了理论支撑。TruLens框架内置RAG三元组评估体系，有效解决了系统全流程量化评估的难题[26]。Ragas是面向RAG全链路测评的自动化框架，涵盖上下文精确度（Context Precision）、上下文召回率（Context Recall）、忠实度（Faithfulness）、答案相关性（Answer Relevancy）四大核心指标，可快速实现系统性能的批量化测评[27]。RAGEval框架则重点聚焦事实准确性问题，从完整性（Completeness）、幻觉度（Hallucination）和无关性（Irrelevance）三个维度搭建评估模型[28]。除此之外，CCRS[29]、DeepEval[30]等开源测评框架也逐步完善，可适配不同测评精度与服务场景的评估需求。然而，现阶段主流测评框架仍存在一定局限性，这类评估方法过度依赖提示词与模型评价器质量，可能存在一定的评估偏见。

综上，现有研究主要存在以下不足：一是内嵌RAG技术的发现系统包含检索和生成两个模块，但已有研究主要聚焦于检索性能或生成性能测评，缺乏覆盖系统全链路的评估方法，且针对专业领域的测评研究相对匮乏；二是现有发现系统自动化测评结果与人工判断一致性偏低，难以保障测评结论在真实场景中的实用价值。因此，构建兼顾效率与可信度的领域专用测评体系，仍是当前科技文献发现系统测评的重点研究方向。

2 研究设计

2.1 数据集选择及构建

目前，适用于科技文献问答场景的公开数据集主要包括Qasper数据集[31]、LongCite-45k数据集[32]等。其

中，Qasper数据集专为科技文献问答任务设计，包含5 049个问题，覆盖1 585篇自然语言处理论文，分为训练集（888篇论文）、验证集（281篇论文）、测试集（416篇论文），数据格式为“问题—支持上下文—答案”；LongCite-45k数据集由清华大学于2024年发布，包含44 600条细粒度引用问答数据。本研究以Qasper数据集为核心测评数据集，选取字段包括论文ID、问题（用户输入的问题）、检索上下文（模型根据问题从文献库中检索到的相关文档片段）、正确答案（人工标注的问题标准答案）。

2.2　测评指标体系构建

测评指标是衡量发现系统信息检索及生成内容质量的核心工具。高雅奇[33]设计包括准确性、响应时间和F1值在内的多维度评估指标，用于评估基于LLM和RAG的高校知识库智能问答系统的性能。王博楷等[34]利用Ragas自动化测评框架的忠实度、答案相关性、上下文相关性等指标，对基于RAG的供应链智能问答模型进行测评。Gao等[35]提出首个评估LLMs引文生成能力的自动评估基准ALCE（Automatic LLMs’Citation Evaluation），通过检索证据生成带引文的文本，构建流畅性、正确性和引文质量三维自动评估指标。Saad-Falcon等[36]提出自动化检索增强生成系统评估框架ARES（Automated RAG Evaluation System），从上下文相关性、答案忠实度和答案相关性三个指标进行准确评估。Ru等[37]提出RAGChecker评估框架，针对检索和生成模块设计一套诊断指标。

本研究在借鉴上述指标的基础上，结合RAG技术特点，从检索质量、生成质量和溯源质量三个方面构建测评指标体系（见表1）。指标体系构建主要遵循原则：①兼顾通用性与领域特异性，既保留检索生成系统的基础性能衡量维度，又强化对科技文献来源追溯能力的专属要求；②依托已有学术成果，基于现有自动化测评框架进行迁移与改进；③覆盖“检索—生成—溯源”全链路，确保指标体系实现对科技文献从资源获取到知识输出完整流程的质量把控。

2.3　测评方法

本研究采用“自动化测评+混合评估”的组合方式，对内嵌RAG的科技文献发现系统开展应用效果测评。在自动化测评环节，本研究主要通过调用发现系统内置的LLM接口，将测试数据集中的问题与对应上下文输入LLM，借助提示词引导LLM输出回答，再利用Ragas自动化测评框架及ROUGE指标对输出结果进行量化评估，完成对检索质量和生成质量的自动化测评。在此基础上，引入Factscore框架[38]、BERTScore指标[39]，围绕引用覆盖率与引用准确性两个维度完成溯源质量的量化测评。在混合评估环节，本研究采用大模型智能评估结合人工审核的方式，先由LLM对全量测评结果进行初步打分，再由科技文献领域专业人员对初评结果进行逐一复核，对评分偏差、模糊判定项进行校准与判定，保障评估结果的专业性与可信度。

表1　测评指标体系

评估维度	指标	说明
检索质量	上下文召回率	衡量检索到的上下文与人工标注答案的一致性程度
	上下文相关性	衡量检索到的上下文与用户问题的关联程度
	检索上下文与基准上下文一致性	衡量检索到的上下文与基准上下文的文本重合度与信息覆盖度
生成质量	答案相关性	衡量生成的答案与用户问题之间的相关程度
	忠实度	衡量生成的答案与检索上下文的事实一致性
	答案语义相似性	衡量生成的答案与人工标注答案的一致性，反映模型生成结果的准确性与完整性
溯源质量	引用覆盖率	衡量生成答案中的每一个关键事实陈述是否有明确支持，反映溯源信息的全面性
	引用准确性	衡量生成答案与引用片段的语义匹配度

2.3.1　提示词

本研究使用发现系统内置的DeepSeek-R1模型应用程序编程接口（Application Programming Interface，API）作为回答问题的LLM，经过反复试验，对比不同提示词下模型的输出结果，不断迭代优化问答指令任务。本研究采用的问答任务提示词如图1所示。

2.3.2　自动化测评

2.3.2.1　检索与生成质量测评

（1）ROUGE方法测评。一致性是衡量系统检索有

prompt="
你将收到一个问题和几段可能相关的上下文。请严格按照以下规则回答问题：
1.相关性优先：你的答案必须直接、精确地回答问题。不要添加任何与问题无关的背景信息、解释或额外内容。
2.基于上下文：答案的所有信息必须来自提供的上下文。
3.极度简洁：用最少的文字表达核心意思。删除所有修饰词、过渡句和重复信息。
4.格式要求：
如果问题是问“ 是什么 ”（如定义、名称、类别等），请用名词性短语或短句回答。
如果问题是问“ 如何 ”（如方法、过程等），请用动词开头的短语或步骤化的短句回答。
如果问题是问“ 多少 ”（如数量、百分比等），请直接给出数字和单位。
无法回答时：如果上下文不包含回答问题的信息，请仅输出“ 无法从上下文获取答案 ”。"

图1　问答任务提示词

效性的核心标尺，只有检索到的上下文与基准上下文高度一致，才能保证大模型基于检索内容生成的回答准确，避免出现“答非所问”。ROUGE[22]是基于文本重合度的自动化评价方法，本研究借鉴该指标计算发现系统检索到的上下文与基准上下文之间的相似度，进而测评检索到的上下文是否可靠。

ROUGE-N是基于n-gram匹配程度来评估相似性。其包含多个变体，广泛使用的有ROUGE-1、ROUGE-2、ROUGE-3，计算方法如式（1）所示。

$$\text{ROUGE-}N=\frac{\sum_{S_{\text{baseline context}}}\sum_{n\text{-gram}\in S_{\text{baseline context}}}\text{Count}_{\text{match}}\left(n\text{-gram}\right)}{\sum_{S_{\text{baseline context}}}\sum_{n\text{-gram}\in S_{\text{baseline context}}}\text{Count}\left(n\text{-gram}\right)} \quad (1)$$

其中，$S_{\text{baseline context}}$表示基准上下文集合，即从Qasper数据中获取的标准答案；n-gram为n元词（n=1，单词；n=2，二元词；n=3，三元词）；$\text{Count}_{\text{match}}\left(n\text{-gram}\right)$为大模型检索到的上下文与基准上下文中同时出现$n$-gram的个数；$\text{Count}\left(n\text{-gram}\right)$为基准上下文中出现的$n$-gram个数。

ROUGE-L是基于最长公共子序列（Longest Common Subsequence，LCS）来评估文本之间的相似度，通过计算召回率（R_{lcs}）、精确率（P_{lcs}），调和平均值（F_{lcs}）进行综合评价。三个指标的计算方法如式（2）~式（4）所示。

$$R_{\text{lcs}}=\frac{\text{LCS}\left(X,Y\right)}{m} \quad (2)$$

$$P_{\text{lcs}}=\frac{\text{LCS}\left(X,Y\right)}{n} \quad (3)$$

$$F_{\text{lcs}}=\frac{\left(1+a^2\right)R_{\text{lcs}}P_{\text{lcs}}}{R_{\text{lcs}}+a^2P_{\text{lcs}}} \quad (4)$$

其中，$\text{LCS}\left(X,Y\right)$为两类文本的最长公共子序列的长度；$m$为基准上下文的长度；$n$为检索到的上下文的长度；$P_{\text{lcs}}$为精确率；$R_{\text{lcs}}$为召回率；$F_{\text{lcs}}$为$P_{\text{lcs}}$和$R_{\text{lcs}}$的调和平均值；$a$为用于调节精确率和召回率权重的参数，默认为1。

（2）基于Ragas框架测评。Ragas是专为评估RAG技术应用的自动化测评框架，通过多维度指标分数全面评估系统性能。该框架具有两大核心优势：一是评估效率高，多数指标仅需问题与检索上下文即可完成评估，无需依赖标准答案；二是评估维度灵活可组合，既可计算综合评分，也可针对特定维度开展专项分析。本研究借鉴Ragas框架指标对发现系统的检索质量与生成质量开展测评。

①上下文召回率（Context Recall）。衡量检索到的上下文是否覆盖参考答案所需的所有关键信息，上下文召回率取值范围[0，1]，数值越接近1表明检索性能越好。上下文召回率的具体计算方法如式（5）所示。

$$\text{Context Recall}=\frac{\left|F\right|}{\left|B\right|} \quad (5)$$

其中，F为上下文覆盖的关键信息数量，B为参考答案中关键信息总数量。

②上下文相关性（Context Relevance）。衡量检索到的上下文与问题之间的相关性，上下文相关性取值范围[0，1]，数值越接近1表明相关性越好。上下文相关性的具体计算方法如式（6）所示。

$$\text{Context Relevance}=\frac{\left|S\right|}{\left|R\right|} \quad (6)$$

其中，S为上下文中与问题相关的片段数量，R为上下文中片段总数量。

③答案相关性（Answer Relevancy）。衡量生成的答案与问题之间的关联程度，答案相关性取值范围[0，1]，数值越接近1表明相关性越好。答案相关性的具体计算方法如式（7）所示。

$$\text{Answer Relevancy}=\frac{1}{n}\sum_{i=1}^{n}\text{sim}\left(q,q_i\right) \quad (7)$$

其中，n为生成的问题数量，q为原始问题，q_i为第i

个生成的问题，sim(·)为语义相似度计算函数。

④忠实度（Faithfulness）。衡量生成答案与给定上下文的事实一致性，忠实度取值范围[0，1]，数值越接近1表明一致性越好。忠实度的具体计算方法如式（8）所示。

$$\text{Faithfulness}=\frac{|V|}{|S|} \tag{8}$$

其中，V为上下文能够推出事实的数量，S为答案拆解出的事实总数量。

⑤答案语义相似性（Answer Semantic Similarity）。衡量生成答案与参考答案的之间的相似性，答案语义相似性取值范围[0，1]，数值越接近1表明生成的答案与标准答案之间的语义相似性越高。答案语义相似性的具体计算方法如式（9）所示。

$$\text{Answer Semantic Similarity}=\frac{1}{n}\sum_{i=1}^{n}\text{sim}\left(A,A_{\text{ref}}\right) \tag{9}$$

其中，A为生成答案，A_{ref}为参考答案，sim(·)为语义相似度计算函数。

2.3.2.2　溯源质量评估

本研究结合科技文献的严谨性、专业性需求，从引用覆盖率、引用准确性两个维度构建溯源质量评价指标。

（1）引用覆盖率（Citation Coverage）。引用覆盖率聚焦生成文本中最小可验证独立事实单元，通过引用支撑与否的量化比例，反映生成内容对检索知识的忠实度。引用覆盖率取值范围[0，1]，数值越接近1表示原子事实的引用支撑越充分。本研究以Factscore评估框架为基础，将模型生成的长文本拆解为若干最小可验证的原子事实单元，通过判断这些原子事实是否能被可靠知识源佐证，进而量化模型的事实可信程度。计算流程为：首先利用DeepSeek-R1模型对生成的长文本进行句内、句间拆分，确保每个单元仅包含1个可验证的独立事实单元；其次对每个原子事实进行“支持”或“不支持”的标签判定；最后计算引用覆盖率得分，引用覆盖率的计算方法如式（10）所示。

$$\text{Citation Coverage}=\frac{N_{\text{sup}}}{N_{\text{total}}} \tag{10}$$

其中，N_{sup}为被支持的原子事实数，N_{total}为总原子事实数。

（2）引用准确性（Citation Accuracy）。引用准确性衡量原子事实与引用片段的语义匹配度，引用准确性取值范围[0，1]，得分越接近1表示事实与引文的一致性越强。具体从以下三个步骤实现。

首先，进行语义相似度计算，对每个被支持的原子事实，提取其绑定的引用片段。

计算两者BERTScore的F1值作为其语义相似度得分，相关计算如下。

根据式（11）计算原子事实中单个词元与引用片段中单个词元之间的余弦相似度。

$$\cos\left(\boldsymbol{t}_i,\boldsymbol{v}_j\right)=\frac{\boldsymbol{t}_i\cdot\boldsymbol{v}_j}{\|\boldsymbol{t}_i\|_2\|\boldsymbol{v}_j\|_2} \tag{11}$$

其中，$\boldsymbol{t}_i$、$\boldsymbol{v}_j$分别为原子事实、引用片段中单个词元的语义向量，$\|\cdot\|_2$为L2范数（向量的模长）。

根据式（12）计算BERTScore精确率（P），即原子事实中每个词元在引用片段中找到最相似词元的平均相似度。

$$\text{BERTScore-}P=\frac{1}{|T|}\sum_{t_i\in T}\max_{v_j\in V}\cos\left(\boldsymbol{t}_i,\boldsymbol{v}_j\right) \tag{12}$$

其中，T为原子事实的词元集合，$|T|$为原子事实的词元总数量，V为引用片段的词元集合。

根据式（13）计算BERTScore召回率（R），即引用片段中每个词元在原子事实中找到最相似词元的平均相似度。

$$\text{BERTScore-}R=\frac{1}{|V|}\sum_{v_j\in V}\max_{t_i\in T}\cos(\boldsymbol{t}_i,\boldsymbol{v}_j) \tag{13}$$

其中，$|V|$为引用片段的词元总数量。

根据式（14）计算BERTScore-F1综合语义相似度得分，由精确率和召回率的调和平均得到，取值范围为[0，1]。

$$\text{BERTScore-Fl}=\frac{2\text{BERTScore-}P\cdot\text{BERTScore-}R}{\text{BERTScore-}P+\text{BERTScore-}R} \tag{14}$$

其次，进行阈值筛选。设定BERTScore-F1阈值为0.6，低于阈值的原子事实判定为“引用不准确”。

最后，根据式（15）计算引用准确性得分。

$$\text{Citation Accuracy}=\frac{\sum_{i=1}^{N_{\text{total}}}\text{I}\left(S_{\text{bert},i}\geqslant 0.6\right)}{N_{\text{total}}} \tag{15}$$

其中，N_{total}为被支持的原子事实总数，$S_{\text{bert},i}$为第i条原子事实对应的BERTScore-F1值，I(·)为指示函数，当括号内条件成立时取值为1，不成立时取值为0。

2.3.3 混合评估

为验证提出的测评指标体系的科学性与有效性，本研究采用LLM结合人工审核的混合评估方法。在模型选取上，Qwen3.5-Plus[40]以开源形式发布，可通过API服务使用，且支持训练和推理部署，可适配本研究混合评估的落地需求。因此，本研究选取该模型作为测评模型。先使用Qwen3.5-Plus按预设的测评维度（见表1）对各模型输出进行初步打分与排序，大模型评估提示词如图2所示。

完成初评后，再由科技文献领域的专业人员对自动评估结果进行抽样复核与校准，以修正偏差。最后以“机器初评+人工精审”的综合结果作为模型性能的评估依据。

> 假如你是科技文献领域的资深专家，请根据以下标准对发现系统的各项指标进行评估。参考内容如下：问题：XXX；正确答案：XXX；基准上下文：XXX；检索到的上下文：XXX；引用上下文：XXX；生成答案：XXX。评分标准见表2，各指标维度均设置阈值1-10，分数越高代表性能越好。请输出每条数据的每个指标分数，并给出解释。

图2 大模型评估提示词

表2 评分标准

指标	指标说明	评分
检索完整性	检索到的上下文与正确答案的一致程度	几乎未检索到有用信息：1~2分
		核心信息严重缺失：3~4分
		部分核心信息缺失，不影响基本理解：5~6分
		核心信息基本完整，仅次要细节缺失：7~8分
		所有关键信息完全覆盖：9~10分
检索相关性	检索到的上下文与问题的语义相关程度	检索结果完全偏离主题：1~2分
		大部分为噪声，少量相关：3~4分
		部分相关，但有大量噪声：5~6分
		大部分相关，少量噪声：7~8分
		所有片段高度相关，无冗余噪声：9~10分
检索上下文一致性	检索到的上下文与基准上下文在事实、表述、核心信息上的契合程度	检索片段与基准上下文无关：1~2分
		少量片段的核心信息一致：3~4分
		检索片段与基准上下文核心信息基本一致，但可能存在非核心信息偏差：5~6分
		检索片段与基准上下文核心信息、表述逻辑一致，仅细节表述差异：7~8分
		检索片段与基准上下文完全契合：9~10分
答案相关性	生成答案与问题的契合度，是否切题、不跑偏、不答非所问	答案与问题无关或提供无效信息：1~2分
		答案涉及问题主题，但未解决核心需求：3~4分
		答案方向正确，但存在信息冗余、模糊或部分错误：5~6分
		答案基本与问题契合，但细节不完整或需小幅优化：7~8分
		答案直接、完整解决用户问题，无冗余或错误信息：9~10分
生成忠实度	生成答案是否完全基于检索到的上下文，无编造、无幻觉	多处关键信息编造，答案不可信：1~2分
		少量关键信息编造，核心结论不可靠：3~4分
		少量关键信息编造，核心结论可靠：5~6分
		无关键信息编造，仅少量推断无直接依据：7~8分
		所有内容均有检索上下文支撑，无任何幻觉或冲突：9~10分
答案语义相似性	生成答案与正确答案的语义契合程度	核心结论与标准答案相反或无关：1~2分
		仅个别细节与标准答案一致：3~4分
		核心结论有重叠，但关键信息缺失或偏差：5~6分
		核心结论、关键信息与标准答案基本一致：7~8分
		核心结论、关键信息、逻辑与标准答案完全一致：9~10分

续表

指标	指标说明	评分
引用覆盖率	生成答案中的每一个关键事实陈述是否在引用上下文中有明确支持	生成答案中无引用上下文支持：1~2分
		生成答案中有少量引用上下文支持：3~4分
		生成答案基本有引用上下文支持：5~6分
		生成答案中多数有引用上下文支持：7~8分
		生成答案完全有引用上下文支持：9~10分
引用准确性	引用上下文与生成答案的语义匹配度	引用上下文未准确支撑生成答案：1~2分
		仅少量引用片段，大部分引用无法支撑答案内容：3~4分
		核心引用方向一致，但可能存在偏差：5~6分
		引用内容及上下文基本准确，能够有效支撑答案中的核心观点：7~8分
		引用内容、上下文完全准确，与答案内容一一对应、严谨支撑：9~10分

3　实证研究

3.1　实验操作与参数配置

国家科技图书文献中心（National Science and Technology Library，NSTL）是面向全国提供公益、普惠的科技文献知识服务机构，近年来积极开展科技文献智能化服务，将LLM和RAG嵌入传统学术文献发现过程，推动服务量级实现新的增长。因此，本研究选取NSTL作为实验平台，将Qasper测试论文全文作为语料库，通过NSTL发现系统的API获取实际检索的上下文并生成答案，进一步验证系统自动化测评效果。具体执行流程为：NSTL发现系统先对单篇论文执行文本切片处理，再基于测试问题检索匹配切片内容，最终输出检索上下文与回答结果。实验过程中严格将NSTL返回结果与Qasper数据集自带的证据（evidence）、标准答案（answer）逐条对齐，完成指标计算与效果测评。参数配置如表3所示。

表3　参数配置

框架	版本
LLM	DeepSeek-R1-671B
Embedding模型	Qwen3-Embedding
向量数据库	MongoDB
Python	3.8.10
LLM temperature	0.3
Embedding chunk	8192

3.2　测评数据集

为确保测评结果的代表性与可靠性，本研究从Qasper测试集中随机选取50篇论文，共124个“问题—上下文—答案”对作为测试对象，数据集示例（部分）如表4所示。

3.3　结果分析

3.3.1　自动化测评结果分析

3.3.1.1　检索质量评估结果

在检索质量评估环节，设置返回前N（N=3，5，10）条检索结果作为对比实验，从上下文召回率、上下文相关性、检索上下文一致性三个维度进行量化评估。检索质量测评结果如表5所示。

实验结果表明，随着返回结果数量N的增加，上下文召回率呈持续显著上升趋势，从N=3的0.53提升至N=10的0.78，增幅达47.17%。上下文相关性指标保持在0.68~0.70之间，未随N的增大出现明显波动。该结果表明，模型在提升召回覆盖度的同时，有效维持了返回结果的相关性质量，避免了为追求高召回而引入大量无关信息的问题。

在ROUGE指标方面，各子指标呈现一致变化规律：精确率随返回结果数量N增大持续下降，召回率持续上升，导致综合F值整体呈下降趋势。具体而言，当N=3时，所有ROUGE指标的F值均达到峰值，体现出最优的综合检索质量；随着N增加至5时，F值略有下降但仍保持在较高水平；当N达到10时，F值出现显著下滑，表明过多的返回结果会严重稀释检索精度。

综合来看，NSTL发现系统在N=3时实现了检索精准性与召回覆盖度的最佳平衡，是检索质量的最优阈

表4　数据集示例（部分）

问题	检索上下文	正确答案
What baselines did they compare with	The following baselines were used in our experiments. LDA BIBREF2. LDA is the classic topic model, and we used GibbsLDA++ for its implementation. Doc-NADE BIBREF24. Doc-NADE is a representative neural network based topic model. We used the open-source code provided by the authors. HTMM BIBREF9. HTMM models consider the sentence level Markov transitions. Similar to Doc-NADE, the implementation was provided by the authors. GMNTM BIBREF12. GMNTM considers models the order of words within a sentence by a feedforward neural network. We implemented GMNTM according the descriptions in their papers by our own	LDA; Doc-NADE; HTMM; GMNTM
How many tags are included in the ENE tag set	In the collection of the dataset articles, we targeted only Japanese Wikipedia articles, since our annotators were fluent Japanese speakers. The articles were selected from Japanese Wikipedia with the condition of being hyperlinked at least 100 times from other articles in Wikipedia. We also considered the Goodness scoring measures mentioned in BIBREF9 to remove some of the unuseful articles. The collected dataset contained 120 333 Japanese Wikipedia articles in different areas, covering 141 out of 200 ENE labels	141
How do they perform semi-supervised learning	Our semi-supervised text classification framewrok is inspired by works BIBREF18, BIBREF19. We assume the original classifier classify a sample into one of INLINEFORM0 possible classes. So we can do semi-supervised learning by simply adding samples from a generative network G to our dataset and labeling them to an extra class INLINEFORM1. And correspondingly the dimension of our classifier output increases from INLINEFORM2 to INLINEFORM3. The configuration of our generator network G is inspired by the architecture proposed in BIBREF16. And we modify the architecture to make it suitable to the text classification tasks. Table TABREF13 shows the configuration of each layer in the generator G. Lets assume the training batch size is INLINEFORM4 and the percentage of the generated samples among a batch training samples is INLINEFORM5. At each iteration of the training process, we first generate INLINEFORM6 samples from the generator G then we draw INLINEFORM7 samples from the real dataset. We then perform gradient descent on the AC-BLSTM and generative net G and finally update the parameters of both nets	On each step, a generative network is used to generate samples, then a classifier labels them to an extra class. A mix of generated data and real data is combined into a batch, then a gradient descent is performed on the batch, and the parameters are updated

表5　检索质量测评结果

返回值	上下文召回率	上下文相关性	检索上下文一致性											
			ROUGE-1			ROUGE-2			ROUGE-3			ROUGE-L		
			P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
N=3	0.53	0.68	29.11%	52.70%	34.77%	18.51%	37.68%	22.68%	15.77%	33.28%	19.59%	28.31%	51.45%	33.90%
N=5	0.68	0.69	23.69%	64.69%	32.69%	15.07%	50.14%	21.54%	12.79%	44.65%	18.47%	23.18%	63.60%	32.04%
N=10	0.78	0.70	16.29%	76.59%	25.69%	10.32%	63.19%	16.82%	8.78%	57.23%	14.43%	16.04%	75.63%	25.30%

值。在实际应用中，若场景对检索精度要求较高，应优先选择*N*=3；若需更高的召回覆盖度，可适度放宽至*N*=5；而*N*=10时虽能进一步提升召回率，但会牺牲检索精度，需根据具体业务需求进行权衡。

3.3.1.2　生成与溯源质量结果

在生成质量方面，NSTL发现系统的答案语义相似性得分为0.80，表明生成的答案能较好地贴合真实信息，体现出较强的事实一致性能力。答案相关性得分为0.64，处于中等水平，表明系统生成的内容与用户查询需求的匹配度一般。但忠实度得分只有0.59，表明系统生成的答案中有少量关键信息无原文支撑，存在捏造事实、幻觉等问题。在溯源质量方面，引用覆盖率得分为0.86，这意味着NSTL发现系统在生成答案过程中，能广泛调用相关参考来源进行支撑，溯源素材的覆盖范围较为全面，为答案的可追溯性提供了良好基础。但引用准确性为0.68，说明尽管NSTL发现系统能广泛调用参考来源，但引用内容与答案核心内容匹配度低，存在引用错误或偏差等问题，这极大削弱了答案的可信度与溯源价值。

3.3.2　混合评估结果分析

为验证自动化测评结果的可靠性，选取自动测评的124条“问题—上下文—答案”对进行混合评估，严格遵循表2评分标准，先调用Qwen3.5-Plus模型对各指标维度开展初步打分，再由科技文献领域专业人员对打分结果进行复核与校准。其中，自动化测评中检索质量

评估是对返回前3、5、10条检索结果进行测评，考虑到混合评估成本，在混合评估环节只选取返回效果较好的前3条检索结果作为测评样本。此外，由于混合评估得分量纲为1~10分，因此将自动化测评0~1分制折算为10分制，保持量纲的严谨性。混合评估与自动化测评结果对比如表6所示。

表6　混合评估与自动化测评结果对比

测评方法	检索质量			生成质量			溯源质量	
	检索完整性	检索相关性	检索上下文一致性	答案相关性	忠实度	答案语义相似性	引用覆盖率	引用准确性
混合评估	5.63	6.35	4.90	6.23	6.65	7.17	8.19	7.61
自动化测评	5.28	6.81	3.47	6.44	5.93	7.96	8.62	6.80

结果表明，自动化测评与混合评估结果总体接近，但部分指标存在差异。整体上混合评估在多数维度上表现更优，表明增加人工精审环节有效修正自动化测评的部分偏差，进一步提升评估结果的可靠性。在检索质量方面，混合评估整体得分偏低，平均分为5.63，其中检索相关性得分最高，说明NSTL发现系统在识别用户查询核心需求并匹配相关信息方面具备一定能力。但检索上下文一致性得分偏低，反映出检索结果可能存在信息冲突、逻辑断裂等问题。在生成质量方面，整体表现较为稳健，平均分为6.68，其中答案语义相似性得分最高，说明NSTL发现系统生成答案与参考答案的语义契合度较高，信息输出精准性良好。答案相关性得分为6.23，略低于同维度其他两项指标，反映出NSTL发现系统在精准匹配用户深层需求方面仍有提升空间，需进一步优化生成内容的针对性。在溯源质量方面的表现最为突出，平均得分为7.90，说明NSTL发现系统生成的大部分内容均可溯源，且引用信息与原始来源保持较高的一致性，溯源可信度较强。

为进一步验证自动化测评与混合评估的一致性，本研究利用SPSS工具，将124条自动化测评结果逐一与人工评估结果进行非参数相关性检验，每条样本对应的是8个指标自动化测评与混合评估的综合得分。非参数检验结果显示，肯德尔相关系数为0.56（$p<0.01$），表现为中等程度正相关；斯皮尔曼相关系数为0.74（$p<0.01$），表现为强正相关。因此，自动化测评结果与混合评估显著正相关，表明本研究提出的测评方法具有一定的可靠性与适用性，进一步验证了其在科技文献服务场景中的适配性优势。

4　结语

本研究提出一种内嵌RAG的科技文献发现系统测评方法，并以NSTL发现系统为实证对象开展全面评估。在指标构建方面，围绕RAG全链路服务模式，建立包含检索质量、生成质量和溯源质量三大核心维度的量化测评指标体系，实现了对科技文献发现系统的多维度测评。在测评方法方面，采用“自动化测评+混合评估”的组合方式，可批量、高效地完成系统各环节的性能评估，解决了传统人工评价效率低、一致性不足等问题。在实证研究方面，以内嵌RAG的NSTL发现系统为对象开展实证评估。实验结果表明，NSTL发现系统在生成质量和溯源质量方面表现突出，但其检索质量仍存在提升空间，核心原因可能是系统对关键文献内容的覆盖度不足，存在漏检问题，且检索算法对用户问题的语义理解精准度有限，导致部分检索结果存在冗余信息，或核心信息与问题关联度偏低等。此外，经对比验证，自动化测评结果与人工参与的混合评估呈显著正相关，说明二者具备较高的一致性，充分证明了本测评框架的科学性与有效性。由此可见，本研究提出的测评方法既能高效识别系统在检索、生成和溯源全环节的服务缺陷，又能对生成结果的溯源精准度进行严格校验，为科技文献发现系统的测评提供了高效、可复用的解决方案。

本研究仍存在一定局限性。一是本研究的数据集局限于现有公开测试集Qasper，其覆盖的学科领域与文献类型较为单一，后续应结合科技文献知识服务的实际需求，构建涵盖多学科、多类型文献的专用测评数据集，进一步提升测评体系的领域适配性与评估结果的泛化能力。二是本研究测评聚焦于单篇论文性能指标，后续可拓展研究边界，将本方法应用到多篇文献指标性能测评，进一步制定标准化综合评估框架，实现对科技文献发现系统应用成效的全面评估。

参考文献

[1]　曾建勋，丁遒劲．基于语义的国家科技信息发现服务体系研

究[J]. 中国图书馆学报，2017（4）：51–62.

[2] 吴利俊，辛继宾，王曼，等. 面向医学科技文献的智能检索与阅读分析系统构建：以复旦大学图书馆为例[J]. 大学图书情报学刊，2026，44（2）：39–44.

[3] Introducing ChatGPT[EB/OL]. [2026–05–14]. https：//openai.com/index/ChatGPT.

[4] Deepseek[EB/OL]. [2026–05–14]. https://www.deepseek.com.

[5] 袁乐，刘绍华，王禹，等. 大语言模型检索增强生成优化技术研究综述[J]. 计算机学报，2026，49（2）：383–422.

[6] YU H，GAN A，ZHANG K，et al. Evaluation of retrieval-augmented generation: a survey[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2405.07437.

[7] GAN A，YU H，ZHANG K，et al. Retrieval augmented generation evaluation in the era of Large Language Models: A comprehensive survey[J/OL]. arXiv preprint，2025. DOI:10.48550/arXiv.2504.14891.

[8] LEWIS P，PEREZ E，PIKTUS A，et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C]//Proceedings of the 34th Conference on Neural Information Processing Systems（NeurIPS 2020）. Red Hook：Curran Associates Inc.，2020：9459–9474.

[9] LI X，OUYANG J. How Does Knowledge Selection Help Retrieval Augmented Generation?[C]//The 2025 Conference on Empirical Methods in Natural Language Processing（EMNLP 2025）. Suzhou：Association for Computational Linguistics，2025：4104–4121.

[10] LIU Y，HU X，ZHANG S，et al. Fine-Grained guidance for retrievers: Leveraging LLMs' feedback in Retrieval-Augmented Generation[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2411.03957.

[11] GAO Y，XIONG Y，WANG M，et al. Modular RAG: Transforming RAG systems into LEGO-like reconfigurable frameworks[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2407.21059.

[12] JIN J，ZHU Y，YANG X，et al. FlashRAG: a modular toolkit for efficient Retrieval-Augmented Generation research[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2405.13576.

[13] EDGE D，TRINH H，CHENG N，et al. From local to global: a graph RAG approach to query-focused summarization[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2404.16130.

[14] SINGH A，EHTESHAM A，KUMAR S，et al. Agentic Retrieval-Augmented Generation: A survey on agentic RAG[J/OL]. arXiv preprint，2025. DOI:10.48550/arXiv.2501.09136.

[15] HUANG Y，HUANG J. A survey on retrieval-augmented text generation for Large Language Models[J]. ACM Computing Surveys，2026，58（12）：1–38.

[16] LYU Y，LI Z，NIU S，et al. CRUD-RAG: a comprehensive Chinese benchmark for Retrieval-Augmented Generation of Large Language Models[J]. ACM Transactions on Information Systems，2025，43（2）：1–32.

[17] 胡伟群，石钰莹，徐曦，等. 基于RAG技术的图书知识图谱生成及应用研究[J]. 图书馆界，2025（5）：39–44.

[18] 尚渡新，段宇锋，柏萍，等. 融合RAG和大语言模型微调的学术短文本学科分类研究[J]. 现代情报，2026，46（3）：18–29.

[19] 宋保仁，冷静. 检索增强生成技术在图书馆私有知识库构建中的应用与边界：以四川高校特色数据库为例[J]. 情报理论与实践，2026，49（4）：184–193.

[20] 张浩然，郝文宁，靳大尉，等. DF-RAG：基于查询重写和知识选择的检索增强生成方法[J]. 计算机科学，2025，52（11）：30–39.

[21] EVTIKHIEV M，BOGOMOLOV E，SOKOLOV Y，et al. Out of the BLEU: How should we assess quality of the code generation models?[J]. The Journal of Systems and Software，2023（203）：111741.

[22] GANESAN K. ROUGE 2.0: updated and improved measures for evaluation of summarization tasks[J/OL]. arXiv preprint，2018. DOI:10.48550/arXiv.1803.01937.

[23] 欧荣. PubMed，ISI-Medline，Google Scholar检索性能对比测评[J]. 医学信息学杂志，2009，30（12）：37–40.

[24] 王莉，梁冰. NSTL智能检索平台的扩展检索效果测评与分析[J]. 数字图书馆论坛，2015（11）：33–38.

[25] 宋琳琳，张明鑫，赖嘉儿，等. 用户对图书馆资源发现系统的评价：以聚合发现功能为切入点[J]. 图书馆论坛，2018，38（10）：85–92，65.

[26] TruLens: Evals and Tracing for AI Agents[EB/OL]. [2026–05–17]. https://www.trulens.org.

[27] ES S，JAMES J，ESPINOSA-ANKE L，et al. Ragas: automated evaluation of Retrieval Augmented Generation[J/OL]. arXiv preprint，2023. DOI:10.48550/arXiv.2309.15217.

[28] ZHU K，LUO Y，XU D，et al. RAGEval: scenario specific RAG evaluation dataset generation framework[J/OL]. arXiv preprint，2025. DOI:10.48550/arXiv.2408.01262.

[29] MUHAMED A. CCRS: A zero-shot LLM-as-a-judge framework for comprehensive RAG evaluation[J/OL]. arXiv preprint，2025. DOI:10.48550/arXiv.2506.20128.

[30] DeepEval: The LLM Evaluation Framework[EB/OL]. [2026−05−17]. https://deepeval.com.

[31] DASIGI P，LO K，BELTAGY I，et al. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Online：Association for Computational Linguistics，2021：4599−4610.

[32] ZHANG J，BAI Y，LV X，et al. Longcite: Enabling LLMS to generate fine-grained citations in long-context qa[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2409.02897.

[33] 高雅奇. 基于大语言模型和RAG技术的高校知识库智能问答系统构建与评价[J]. 电脑知识与技术，2024，20（29）：18−20，38.

[34] 王博楷，牛本，刘玲丽. 基于RAG的供应链智能问答模型[J]. 运筹与模糊学，2024，14（6）：637−644.

[35] GAO T，YEN H，YU J，et al. Enabling Large Language Models to Generate Text with Citations[C]//EMNLP 2023—2023 Conference on Empirical Methods in Natural Language Processing. Singapore：Association for Computational Linguistics，2023：6465−6488.

[36] SAAD-FALCON J，KHATTAB O，POTTS C，et al. ARES: An automated evaluation framework for Retrieval-Augmented Generation systems[J/OL]. arXiv preprint，2023. DOI:10.48550/arXiv.2311.09476.

[37] RU D，QIU L，Hu X，et al. RAGChecker: A Fine-grained framework for diagnosing Retrieval-Augmented Generation[J/OL]. arXiv preprint，2024. DOI:10.48550/arXiv.2408.08067.

[38] MIN S，KRISHNA K，LYU X，et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore：Association for Computational Linguistics，2023：12076−12100.

[39] ZHANG T，KISHORE V，WU F，et al. BERTScore: Evaluating text generation with BERT[J/OL]. arXiv preprint，2019. DOI:10.48550/arXiv.1904.09675.

[40] Qwen3.5-Plus[EB/OL]. [2026−06−09]. https://baike.baidu.com/item/Qwen3.5-Plus/67392693.

作者简介

陈阳，女，硕士，工程师，研究方向：发现系统建设，E-mail：cheny@istic.ac.cn。
孟旭阳，女，硕士，助理研究员，研究方向：自然语言处理。
陈熙，男，硕士，研究实习员，研究方向：智能检索。

Evaluation Methods for Scientific and Technical Literature Discovery System with Embedded RAG

CHEN Yang　MENG XuYang　CHEN Xi
(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract: Retrieval-Augmented Generation (RAG) technology has been widely applied in scientific and technical literature discovery systems. However, current evaluation methods for these systems still face issues such as high labor costs and difficulties in adapting to the RAG architecture. Therefore, this research proposes an evaluation approach for scientific and technical literature discovery systems with embedded RAG, constructing a metric system based on three dimensions: retrieval quality, generation quality, and traceability quality. First, quantitative results for each indicator are obtained through automated assessment. Next, a hybrid evaluation method combining intelligent scoring by a large language model with manual verification is employed. Finally, the consistency between the quantitative results of the indicators and the hybrid evaluation results is verified using Spearman's and Kendall's correlation tests. This research uses the discovery system of National Science and Technology library (NSTL) as its empirical subject. The results indicate that this evaluation method can effectively identify differences in the system's performance regarding scientific and technical literature retrieval, generation, and attribution. Furthermore, the indicator calculation results show a significant positive correlation with the hybrid evaluation results, validating the reliability and effectiveness of the indicator system developed in this study and providing practical support for the evaluation and iterative improvement of scientific and technical literature discovery systems.

Keywords: Scientific and Technical Literature; Retrieval-Augmented Generation; Indicator System; Automated Evaluation; National Science and Technology Library