检索增强生成（RAG）驱动的知识服务：原理、范式及评估*

王　亮
北京印刷学院经济管理学院，102600，北京

摘　要　文章从通用人工智能在专业知识服务领域的局限性入手，分析其在语料来源广度与深度、知识迁移泛化与精确性、参数规模与知识覆盖矛盾以及知识时效性与动态性等方面的不足；提出以检索增强生成（RAG）技术为核心的解决方案，通过结合大语言模型的语义理解与生成能力以及专业知识库的权威性与精确性，将检索与生成功能有机结合，平衡知识服务的形式多样性与内容精准性。在此基础上，文章提出基于RAG技术的知识服务系统构建范式、实施路径和效果评估方法，为出版业实施RAG知识服务提供建议。

关键词　知识服务；检索增强生成；RAG；人工智能；评估
DOI:10.16510/j.cnki.kjycb.20250409.002

国务院印发的《新一代人工智能发展规划》明确提出要推动人工智能技术在知识服务领域的深度应用，建设知识服务技术体系，“重点突破知识加工、深度搜索和可视交互核心技术”，形成“多学科和多数据类型的跨媒体知识图谱”[1]。人工智能领域知名专家、斯坦福大学教授尼尔斯·约翰·尼尔森也将人工智能定义为“人工智能是关于知识的学科——怎样表示知识以及怎样获得知识并使用知识的学科”[2]。随着计算能力的提升和海量数据的积累，人工智能技术取得了突破性进展，大语言模型（Large Language Models，LLMs）在文本生成、翻译、摘要、问答等任务中表现出极高的语言理解与生成能力，为知识服务领域的创新发展奠定了基础，但其在专业知识服务领域的表现却存在不足。一个典型的例子是，大语言模型虽然能够提供一些看似合理的解答，但由于缺乏对权威文献的深度理解，往往无法达到实践所需的精确性和可靠性。造成这种问题的核心原因在于：大语言模型的训练语料以通用文本为主，而非专业领域的高质量、权威知识。如果想解决这个问题，可以使用专业语料对大模型进行重新训练或微调，但技术和算力成本都很高，这对于以出版传媒企业为代表的知识服务提供者而言都是有较大难度和风险的。检索增强生成（Retrieval Augmented Generation，RAG）由通用大语言模型负责语义理解和逻辑生成，由向量化的知识数据库负责专业知识获取，将相关检索结果与用户提问共同输入大语言模型[3]，二者协同作用，以极小的技术和算力需求获取准确的知识服务结果。

1　通用人工智能在专业知识服务领域的局限性分析

要深入探讨通用人工智能在专业知识服务领域的不足，我们需要从技术原理和工作流程出发进行分析。以GPT为例，其核心技术是基于自注意力（Transformer）架构的深度学习模型，通过预训练-微调（Pretraining-Finetuning）的方式进行构建。[4] 预训练阶段，模型通过海量数据学习词语、句子乃至上下文之间的统计关系；微调阶段，模型可以通过少量标注数据进行特定任务的能力强化。这种训练范式虽赋予了大语言模型强大的语义理解和语言生成能力，但也带来了以下问题。

* 基金项目：国家社会科学基金项目“基于联盟区块链的自媒体侵权监管和版权引导机制研究”（21BXW037）的阶段性成果。

1.1 语料来源的广度与深度问题

大语言模型的核心优势在于其预训练阶段所依赖的海量语料。这些语料大多来源于开放网络，数据覆盖从日常用语到科普知识的广泛主题，为模型提供了多样化的语料来源。但这种语料来源更注重逻辑，在专业性和权威性上存在明显不足，导致模型在专业知识服务领域中表现出“逻辑严谨、精确欠佳；广度有余、深度不足”等问题。相比之下，专业领域的核心知识资源——如图书、期刊、报纸中经过严格审校的优质内容——通常并未在开放网络中免费提供。这些权威文献往往以受限访问的形式存在，通常需要专业组织认证、付费订阅或通过特定渠道获取。因此，模型在预训练阶段难以接触到这些高质量的专业语料，导致其在涉及专业问题时的生成质量难以达到人类专家的水平。另外，开放网络语料的内容形式以非结构化文本为主，缺乏系统性和深度。

1.2 知识迁移的泛化与精确性问题

大语言模型主要依赖上下文的统计关系进行语言生成，而非直接“理解”或“内化”知识。这种基于概率的生成机制使其在处理日常语言任务时表现出较强的泛化能力，也就是我们常说的“正确的废话”。在泛化能力方面，大语言模型能够通过其庞大的参数规模，捕捉训练语料中的语言模式和统计关系，从而生成看似合理的回答。例如，在医学领域，药物相互作用问题是一项高度复杂的专业任务。药物之间的相互作用不仅依赖于其化学和药理学特性，还可能受到剂量、患者个体差异以及其他药物的干扰等多种因素的影响。大语言模型可能会根据其在训练语料中学到的语言模式生成看似合理的回答，但这些回答可能并不符合药理学知识或实际临床实践的要求。在内容生成逻辑方面，大语言模型的逻辑处理和推理能力仍不尽如人意。例如，在法律领域，涉及法律条文解释、案件判例分析或合同审阅等任务时，模型需要精准理解法律术语的含义、条文之间的逻辑关系以及上下文中的细微差异。但是，由于模型仅通过统计关系生成语言输出，而非真正理解这些内容，其生成的答案可能缺乏严谨性。

1.3 参数规模与知识覆盖的矛盾问题

随着大语言模型的发展，其参数规模不断增加。例如，最新的GPT-o1可能拥有3 000亿参数[5]。这种参数规模的扩大虽然可以让大模型在语言模式捕捉和生成能力方面显著提升，但却不能直接提高知识覆盖面，尤其是当训练语料中缺乏某些领域的高质量数据时，这些领域的知识在模型中的表现可能严重不足。首先就是“知识盲区”问题，训练语料的质量和多样性直接决定了模型的知识覆盖范围。尽管大语言模型能够通过大规模参数对语料中的语言模式进行复杂的记忆和泛化，但它无法生成超过训练数据范围的知识。这种依赖语料的特点使得模型在未被充分覆盖的领域中表现出明显的“知识盲区”。其次是效率和成本问题，大语言模型需要强大的算力和存储做支撑，其成本之高显而易见。现在的人工智能开发者为了提高其通用性，会将算力和存储资源偏重于通用语料处理。这种资源分配方式决定了高级别大模型无法针对某些特定领域进行深入学习。最后是认知问题，通用人工智能与人类认知能力的差异也进一步凸显了其局限性。人类专家在某一领域内的知识积累往往是基于长期的学习和实践，而这种积累的核心在于对知识的深刻理解和灵活应用。相比之下，大语言模型的大规模参数更多地用于泛化预测，而非形成对知识的内在理解，这就导致了大语言模型在问题理解和答案生成两方面都存在缺陷。

1.4 知识的时效性与动态性问题

专业领域的知识具有高度的动态性。目前的生成式人工智能技术范式决定了大语言模型的训练过程是静态的。一旦训练完成便固定下来，所谓的“学习能力”则更多是在语意理解和生成逻辑上。这种静态的构建逻辑与专业领域知识的动态性之间的矛盾，极大地限制了大语言模型在专业知识服务领域的实用性。虽然可以通过微调

或补充训练的方式对模型进行更新，但这种方法存在诸多困难。首先就是成本问题，微调需要重新收集、清洗和标注大量新数据，这不仅成本高昂，而且训练周期很长。其次是微调后的模型可能会导致“灾难性遗忘”（Catastrophic Forgetting）问题[6]，即在学习新知识时丢失部分旧知识。最后是时效性问题，实时更新的技术难点也使大语言模型难以适应快速变化的专业环境，大语言模型由于无法直接访问实时数据，往往无法为用户提供与最新研究一致的建议或答案。

上述问题让基于大语言模型的通用人工智能在面向专业知识服务领域时难以满足用户的高精度需求。学术界和产业界不断探索通过引入外部知识库或专业内容资源来弥补通用模型的不足的可能性，其中RAG是目前最优方案之一。

2 RAG：一种知识服务新路径

针对如何将专业知识内容融入大语言模型，RAG技术提供了一种新路径。它结合大语言模型的理解、生成能力和专业知识库的权威性、精确性优势，通过将检索与生成相结合，实现了知识服务的形式多样性与内容精准性之间的平衡。

2.1 RAG的原理及技术框架

RAG的核心思想是将大语言模型与外部知识库相结合，在生成内容之前先通过检索模块从外部知识库中获取相关信息，再利用语言模型对检索到的信息进行加工和生成。如图1所示，RAG主要分为向量化、检索和生成三个阶段。

向量化（Embedding）：向量化是RAG流程的第一步，也是其技术基础。在这一阶段，知识所有者需要将知识库中的数据以及用户输入的查询关键词转换为可比较的高维向量表示。这些工作可以由嵌入模型（Embedding Model）自动完成。从技术角度来说，向量化的目地是捕捉文本的语义信息，无论用户查询和知识库在语言表达上如何不同，只要它们的语义相近，其向量表示就会具有较高的相似度。向量化的质量直接影响后续

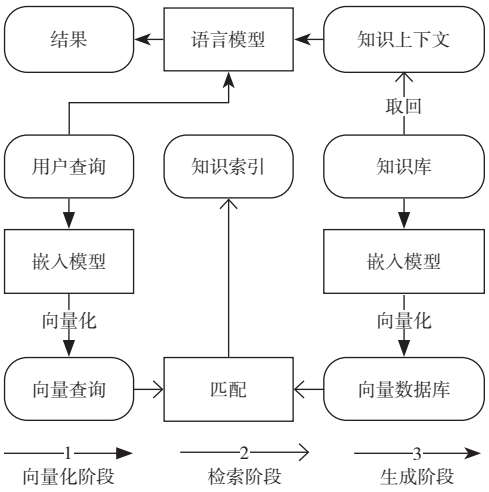

图1 RAG的工作原理示意图

检索阶段的准确性，因此选择合适的嵌入模型至关重要。

检索（Retrieval）：向量化是RAG流程的第二步，其任务是从向量化后的知识库中找到与用户查询最相关的内容。在这一阶段，用户输入的查询关键词将会被采用相同的嵌入模型进行向量化，并利用相似度搜索算法与向量知识库中的数据相匹配，然后将匹配的结果与用户查询组合成复合提示词。检索阶段的效率和准确性通常依赖于向量检索技术，其主要目的是将用户查询与知识库中的相关内容进行有效匹配，从而为生成阶段提供可靠的知识来源。

生成（Generation）：生成是RAG流程的最后一步，负责根据第二步生成的复合提示词和大语言模型的固有能力生成最终的输出文本。这一阶段的核心在于如何有效融合检索到的知识和用户输入，因此大语言模型的理解和生成能力决定了本阶段的准确性和效率，也决定了输出内容的流畅性和逻辑性。

2.2 RAG应用于知识服务的关键问题探讨

RAG的技术特点决定了它能够解决前文提到的通用人工智能在知识服务中的局限性。

2.2.1 专业知识覆盖

知识服务是一种“知识封装”，它利用不同的媒介技术系统对知识内容进行产品化组织并为用户提供知识服务。[7]通过RAG模块，知识服务

系统可以实时从知识库中获取专业信息，让系统的知识覆盖范围不局限于大语言模型。在多领域知识融合上，RAG可以根据用户需求，跨知识领域获取资源并进行整合生成——这在交叉学科领域是尤为重要的。对于专业性极强的知识服务场景，RAG也可以通过指定检索范围，获得更为精准的知识覆盖范围。

在知识覆盖方面，RAG也存在局限性。在知识深度方面，虽然RAG可以通过检索扩展知识覆盖范围，但对于需要深度推理或专业判断的复杂问题，其生成模块可能缺乏足够的能力进行准确的解答——特别是当检索到的信息存在模糊或冲突时；在检索和生成之间的协调方面，通过知识库检索到的专业知识通常是片段化的，在整合并生成连贯的文本的过程可能导致信息失真或遗漏重要细节，影响知识覆盖的完整性；在知识冲突风险方面，当面对多领域或跨学科问题时，RAG可能检索到不一致甚至矛盾的知识，这对生成模块的理解和推理能力提出了更高的要求。

2.2.2 知识动态更新

RAG的检索和生成模块相对独立，这极大提升了其动态捕获最新的知识和信息的能力。相较于传统模型依赖静态语料库的方式，RAG系统的动态获取能力可以确保生成的内容的时效性，这在舆情分析、突发事件处理等知识服务应用场景中尤为重要。

但RAG系统的动态更新能力很大程度上取决于外部知识库的质量和更新速度。在处理新闻观点或新兴科学概念这类知识服务过程中，因为它们本身可能存在矛盾或争议，会导致检索过程返回冗余甚至相互矛盾的信息，从而使结果的出现偏差或遗漏。高频动态更新也会带来成本和效率方面的问题，为了实现知识动态更新，RAG需要频繁地访问外部资源，这可能带来高计算成本和延迟，这可能会让RAG系统在性能上难以满足高并发场景的需求。

2.2.3 精准性和可验证性

在可追溯性方面，RAG系统可以提供可追溯的知识来源并精准定位知识在知识库中的具体位置，这种特性使生成结果具有较高的可验证性。在可信度方面，由于RAG系统并非完全依赖模型内部的参数和推断，因此能有效降低“幻觉”现象。在知识透明化方面，RAG系统支持证据链的透明化，有助于用户直接验证答案的来源，特别是在高精度要求的知识服务领域。

但在信息偏差和冲突信息处理方面，RAG也存在一些问题。RAG需要对从知识库中获取的内容进行整合，在这一过程中难免会对信息源头的精准定位产生一定的负面影响。当遇到外部数据之间存在矛盾或不一致时，RAG可能难以正确判断哪些信息更为可信。这种情况下，生成的答案可能会模糊化处理或直接包含冲突内容，从而降低精确性和可验证性。

2.2.4 多模态信息处理

多模态是指不限于文字，还包含图像、音频、视频等为主要表现形式的资源[8]，RAG的向量化技术允许其集成多模态检索模块，以处理音视频、图等非文本数据。随着以DeepSeek为代表的国产大模型的崛起，这一特质将被进一步拓展。例如DeepSeek提供V3（用于生成）、R1（用于推理）、VL（用于多模态信息处理）等多个模型，为多模态知识服务奠定了深厚的基础，这在需要结合文本与非文本信息的场景（如医学影像解读、视频教程生成或音频数据分析）中作用尤为明显。

但多模态数据处理相对复杂，通常具有高度异质性。这种复杂性可能导致技术实现难度增加，特别是在需要高精度和低延迟的知识服务场景中。另外，RAG虽然可以检索多模态数据，但多模态内容生成的能力仍然不足。多模态数据处理也对数据标注和标准化提出了较高要求，对算力资源的需求相较文本处理有大幅增长，这些问题都可能会对系统研发成本和性能造成压力。

3 面向知识服务的RAG系统建设范式

以出版企业为代表的知识服务提供者应该如何构建RAG工作范式呢？专业知识库建设、嵌

入模型选择、大语言模型选择是其中最为重要的三点。其中“专业知识库建设”决定了RAG知识服务系统的数据可用性和易用性；“嵌入模型选择”决定了用户需求与专业知识库的精准匹配度；“大语言模型”则决定了复合关键词的理解能力和最终内容输出的逻辑性，如图2所示。

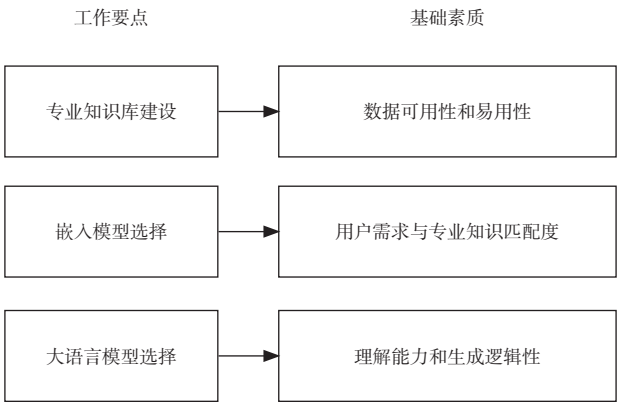

图2　RAG知识服务系统工作要点和基础素质对应关系

3.1　专业知识库建设：数字化与规范化的双轮驱动

出版企业在构建RAG知识服务体系时，数据建设是基础和核心工作，主要体现在数字化建设与规范化建设两个方面。

专业知识库的数字化建设是基础。出版传媒企业应该对现有专业知识内容进行系统性数字化加工，包括科学分类、高精度数字化、高质量碎片化等。特别是对于专业领域的内容，应确保数字化过程中保留高精度的结构化信息，如公式、图表、代码块等，以便后续检索和生成环节的处理。

专业知识库的规范化建设是核心。为确保知识服务系统的可扩展性与语义一致性，出版媒企业需要对知识数据进行标准化处理。包括采用统一的元数据标准或行业特定标准来进行结构化标引、通过构建领域本体来定义知识之间的关系、通过知识图谱技术实现语义关联等。规范化建设有助于提升知识服务的准确性，同时为后续嵌入模型和大模型的应用奠定语义基础。

3.2　嵌入模型选择：面向知识服务的向量化策略

知识服务的核心能力之一是精准检索。在基于RAG的知识服务中，嵌入模型负责将文本数据转化为高维向量表示，以支持高效的语义检索。当前主流的嵌入模型包括以下几类，出版传媒企业可根据自身需求选择适配的方案。

以BGE为代表的开源模型：BGE，即BAAI General Embedding，它由BAAI团队开发，主要用于文本向量化。BGE系列模型的主要特点包括多语言支持（包括专门的中文和英文模型）和多版本支持（包括针对不同需求规模的版本）等，可以满足不同应用场景的需求[9]。另外，BGE系里模型是开源的（MIT许可），这意味着其在RAG知识服务系统开发中“零成本”。

以阿里云的Text-Embedding为代表的国产商业模型：国内主流云服务和人工智能提供商均提供商业化嵌入模型，例如阿里云的Text-Embedding系列模型。其特点是可将文本、图像、音频、视频等数据类型表示为数学空间中的向量，通过计算向量之间的距离或夹角判定数据的相似度，从而作用于分类、检索、推荐等任务[10]。由于这些模型是商业模型，一般具有功能强大、灵活易用的特点，但需要支付一定的费用。

以OpenAI的Text-Embedding-Ada-002为代表的国外商业模型：Ada-002是目前广泛使用的通用嵌入模型，支持语义检索、分类和聚类等任务。其优势在于部署便捷、性能稳定，并且能够以较低的成本生成高质量的向量表示[11]。其与国产商业模型在原理和使用上并无本质区别，主要区别在于模型本身在功能和性能上具有不同特点。

在实际应用中，出版传媒企业可根据数据特点和应用场景选择合适的嵌入模型。对于领域内容较为通用的知识服务，建议使用通用嵌入式模型；而对于技术术语丰富、专业性极强的领域，可以考虑在开源模型的基础上微调优化，以满足个性化和定制化需求。也可以考虑将通用模型与开源调优模型相结合，通过集成方式进一步提升

向量化效果。

3.3 大语言模型选择：面向结果生成的模型匹配

大语言模型在基于RAG的知识服务中主要负责对检索到的知识进行增强生成，以提供专业、精准且高可读性的内容输出。知识服务提供者在选择大语言模型时，可参考嵌入模型的分类和选择依据。但无论如何，业务需求、数据隐私要求和预算限制是选择大语言模型的三个基本依据。根据我国出版传媒企业的普遍特点，给出以下建议。

如果所提供的知识服务涉及多语言知识处理、多语言结果生成等国际化需求，如学术出版物的全球化发行或跨语言技术资料构建，则可以考虑以GPT-o1为代表的最新大模型。这类模型在多语言生成内容的流畅性、逻辑性和准确性方面具有一定优势，能够满足高标准的知识服务需求。但同时也需要考虑成本、数据安全性及技术部署合规性和难度等问题。

对于专注于中文知识服务或面向国内市场的知识服务者，以DeepSeek、通义千问等为代表的国产大语言模型是更适配的选择。这类模型对中文语言环境和本地化需求的优化，使其在生成中文专业内容、处理复杂中文术语及本地化语境相关任务中表现出色。此外，这些模型在上线时就已经通过了国内相关法律法规要求，这对于以中文知识服务为主的出版传媒企业而言更具竞争力。

对于在知识服务中涉及高度敏感度数据、高隐私、高定制化需求的应用场景，则开源大模型是最佳选择。开源大模型的本地部署能力不仅确保数据隐私，还支持企业根据特定领域的需求对模型进行微调以提升性能，另外，开源模型的低成本特性非常适合预算有限但追求高定制化的企业。

4 RAG驱动的知识服务效果评估

人工智能系统的效果评估是验证其有效性和可靠性的重要环节。常见的评估方法包括回归模型评估和分类模型评估等。回归模型评估方法主要用于预测连续型变量，其目标是最小化预测值与真实值之间的误差。常用于根据连续数据值预测模型本身的性能。分类模型评估方法主要用于处理离散型标签的预测任务，其目标是尽可能准确地预测样本所属的类别。常用于人工智能系统能力的定性分析[12]。

本文所研究的评估目的，主要是验证前文所述的面向知识服务的RAG系统建设范式在采用不同模型时的效果差异，从而评估其性能。因此，评估实验更适合采用分类模型进行定性评估，目的是将评估视角定位在系统整体性能而非大语言模型本身的性能。

4.1 分类模型评估指标

分类模型应用于人工智能系统性能评估时，通常使用以下指标。

准确率（Accuracy）：分类正确的样本数占总样本数的比例，是最直观和常用的分类模型评估指标之一。见式1。

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad （式1）$$

精确率（Precision）：在被分类为正类的样本中，真正为正类的样本所占的比例。它关注的是预测为正类的样本中有多少是真正的正类。见式2。

$$Precision = \frac{TP}{TP+FP} \quad （式2）$$

召回率（Recall）：指实际为正类的样本中，被正确分类为正类的样本所占的比例。它关注的是正类样本是否被充分地检测出来。见式3。

$$Recall = \frac{TP}{TP+FN} \quad （式3）$$

F1Score：精确率和召回率的调和平均数，它综合考虑了精确率和召回率，试图在两者之间找到一个平衡。见式4。

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad （式4）$$

4.2 数据集及实验步骤

为了尽量真实地模拟出版行业知识服务应

用场景，同时避免版权问题，我们选择了三种文档，分别模拟专业出版物、教育出版物和大众出版物。本实验选择网络上可公开获取的非结构化数据《三星SM-W9025移动通信设备用户手册》模拟专业出版物，由于其为PDF格式，含有大量文字图表且版式复杂，可以模拟相对苛刻的知识服务数据条件；选择《公民科学素质科普知识问答500题》模拟教育出版物，它的内容表现形式为选择题和判断题，错误答案会为结果选择提出较高的要求，着重考察大语言模型的理解和推理能力；选择《百位中国著名人物介绍》模拟大众出版物，因其专业知识和大模型内置知识存在叠加，着重考察知识甄别和内容生成组织能力。为了使读者能够还原和验证实验过程，也能结合自身选用的模型做定制化评估，本实验选择主流国产嵌入模型和大语言模型，同时将实验程序的数据源、实验样例、实验结果和源代码共享至码云（https://gitee.com/mjst/ragkstest），供读者通过Git等方式下载使用。具体实验步骤如下。

第一步，采用阿里云提供的DashScope Text-Embedding-v2嵌入模型进行实验数据和用户查询的向量化，并设置相似度阈值为0.2，用于保证一定的相关度。

第二步，应用阿里云“百炼”大模型应用平台，采用DeepSeek-V3、通义千问-Max-Latest和通义千问-Turbo-Latest三种大语言模型分别进行语义理解和内容生成，温度系数均设置为0.85，用于调控生成结果的多样性。并将知识数据来源参数分别设置为“大模型+知识库”和“仅知识库”，分别采集数据。

第三步，分别从三种出版物模拟数据源中用人工形式摘编出50个问题，正例和负例为一组，共计25组，表示其类别和答案。这些问题经过一定改编，以验证逻辑分析和处理能力。

第四步，用三种大语言模型分别运行样例数据，计算准确率、精确率，召回率和F1Score，从而评估不同大语言模型在评估实验中的表现。

4.3 实验结果

为了避免大语言模型生成结果多样性带来的偏差，我们将上述测试数据的50个问题为一组，每组运行10次，结果如表1所示。

通过实验结果可以初步分析：从应用场景角度，模拟专业出版物效果最佳，且数据来源权重对结果影响不大；模拟大众出版物效果最差，尤其是在大模型和知识库作为共同知识来源的环境下，这说明源数据的规范性和数据来源权重设置的重要性。从大模型选择角度，DeepSeek-V3在除了以大模型和知识库作为共同数据来源的模拟教育出版物场景外均表现优秀。这可能是由于源数据结构和模型并非面向推理导致的，相信偏重推理的DeepSeek-R1模型会有更佳表现，但遗憾的是由于公共服务不稳定，未能完成基于DeepSeek-R1的评估；通义千问-Max-Latest各方面表现均衡；通义千问-Turbo-Latest则速度最快。

需要强调的是，本实验所使用的大语言模型的定位和特点不同，有的偏重性能，有的偏重速度。因此，本实验主要展示评估方法和过程，用于知识服务提供者在自身应用场景下综合考量性能、成本和速度，从而得出模型选型的最优解，而非优劣。

5 出版业实施RAG知识服务的策略与建议

出版业作为优质知识资源的创造者和传播者，是知识服务应用的重要领域。根据前文的实验结果，结合出版业的实际情况，提出实施RAG知识服务的策略与建议如下。

5.1 基础设施构建

建议出版企业积极构建支持RAG知识服务的大模型本地化部署基础设施。出版企业拥有大量高价值的专业内容和版权资产，其安全性和私密性至关重要。本地化部署可以有效避免数据泄漏和外部依赖，同时提升模型的安全性和运行效率。在本文的实验过程中，也遇到了公共服务不稳定的情况，构建本地部署的基础设施也可以有效解决使用第三方服务制约的问题。出版企业实

表1　实验结果集合

	大语言模型 LLMs	准确率 Accuracy	精确率 Precision	召回率 Recall	F1Score
模拟专业出版物（大模型+知识库）	通义千问-Max-Latest	0.96	1	0.9 259	0.9 615
	通义千问-Turbo-Latest	0.906	0.924	0.8 919	0.9 077
	DeepSeek-V3	0.918	1	0.8 591	0.9 242
模拟教育出版物（大模型+知识库）	通义千问-Max-Latest	0.918	0.88	0.9 524	0.9 148
	通义千问-Turbo-Latest	0.87	0.88	0.8 627	0.8 713
	DeepSeek-V3	0.84	0.808	0.8 632	0.8 347
模拟大众出版物（大模型+知识库）	通义千问-Max-Latest	0.894	0.96	0.8 481	0.9 006
	通义千问-Turbo-Latest	0.912	1	0.8 503	0.9 191
	DeepSeek-V3	0.938	1	0.8 897	0.9 416
模拟专业出版物（仅知识库）	通义千问-Max-Latest	0.96	1	0.9 259	0.9 615
	通义千问-Turbo-Latest	0.892	1	0.8 224	0.9 025
	DeepSeek-V3	0.918	1	0.8 591	0.9 242
模拟教育出版物（仅知识库）	通义千问-Max-Latest	0.92	0.88	0.9 565	0.9 167
	通义千问-Turbo-Latest	0.862	0.852	0.8 694	0.8 606
	DeepSeek-V3	0.89	0.82	0.9 535	0.8 817
模拟大众出版物（仅知识库）	通义千问-Max-Latest	0.914	0.96	0.8 791	0.9 178
	通义千问-Turbo-Latest	0.916	1	0.8 562	0.9 225
	DeepSeek-V3	0.96	1	0.9 259	0.9 615

施RAG知识服务的主要基础设施应该至少包括算力和分布式存储系统，以支持系统构建和运维，但应特别注重性能价格比，避免浪费。为了降低部署成本，出版企业也可以探索与云服务商合作，构建混合云基础设施，在本地运行核心数据和模型，同时利用云平台的弹性计算资源处理高并发任务。出版企业还需要制定严格的数据接入与访问控制策略，确保在多部门协作中数据流动的安全性和合规性。

5.2　工作流程优化

出版企业的知识生成流程通常包括数据收集、编辑、排版、校对和发布等多个环节，其中许多步骤涉及重复性的数据加工工作。为了更好地为RAG知识服务系统提供优质知识库，应该有效优化这些工作流程，降低重复劳动以提升效率。例如在编辑流程中，是否可以考虑进行必要的关键数据标引？在校对环节是否可以不仅考虑文字本身的问题，也考虑针对知识服务的数据规范性检查和内容一致性验证？在排版环节是否可以兼顾输出形式，让信息不仅服务于纸质出版物，更可以方便快捷地用于知识库？是否可以考虑进行跨部门协作，优化数据流转和审批流程，通过减少重复性劳动并提升知识生成的自动化程度？

5.3　数据准备

数据是RAG知识服务的核心资源，其质量直接决定了系统的性能和用户体验。为确保数据的高效利用和准确性，出版企业需要重点关注数据标准化建设、数据标引和数据权重设计。数据标准化是基础工作，建议针对RAG知识服务制定统

一的数据格式和标准——尤其是针对图像、音视频等多模态数据——以便各类内容能够被RAG系统高效读取和处理。在上一章的实验中也可以发现，在模拟专业出版物的实验中，采用“大模型+知识库”和“仅知识库”作为数据来源的实验数据差别不大；而在模拟教育和大众出版物的实验中却有一定差别。这种现象产生的原因可能是大语言模型本身的学习语料包含更多通识知识而非专业知识。不同类型的出版企业在实施RAG知识服务的过程中，可以设置不同的数据权重，以实现来自知识库的专业性和来自大模型的通识性之间的平衡。

5.4 版权保护

版权保护是出版业实施RAG知识服务时必须考虑的问题。出版内容是高价值的版权资产，如果处理不当，可能导致侵权风险或版权收益的流失。出版企业需要构建完善的版权管理与保护机制，确保知识服务的合规性和版权资产的安全性。可以通过加密技术限制内容的访问权限；可以通过区块链技术实现记录和溯源；可以通过本地化部署尽量降低数据外泄风险等。

6 结语

我国在推动人工智能与行业深度融合方面出台了多项政策，强调“新质生产力”在经济社会发展中的重要作用。在这些政策背景下，知识服务领域的专业内容如何与通用人工智能结合成为一个重要的研究方向。以出版企业为例，作为专业知识的主要生产与传播机构，其积累了大量高质量的专业内容资源。这些内容不仅具有高度的权威性和可靠性，还涵盖了许多大语言模型训练语料中未能覆盖的细分领域。RAG可将知识服务领域的专业内容融入大语言模型，不仅能够拓宽应用形式、提升其在专业场景中的表现，还能推动知识生产和传播的数字化转型，为经济社会发展提供新动能。基于海量专业优质数据的学习与推理，也能够从跨领域知识中发现潜在关联，助力用户探索新知识。这无疑将为知识获取、生成与共享开辟全新路径。

参考文献

[1] 国务院关于印发新一代人工智能发展规划的通知[EB/OL]. [2025-02-08]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[2] 贾同兴. 人工智能与情报检索[M]. 北京：北京图书馆出版社，1997. 15-103.

[3] 文森，钱力，胡懋地，等. 基于大语言模型的问答技术研究进展综述[J]. 数据分析与知识发现，2024，8（6）：16-29.

[4] 张新新，丁靖佳. 生成式智能出版的技术原理与流程革新[J]. 图书情报知识，2023，40（5）：68-76.

[5] Medec: a benchmark for medical error detection and correction in clinical notes[EB/OL].[2025-02-08]. https://arxiv.org/abs/2412.19260.

[6] 朱飞，张煦尧，刘成林. 类别增量学习研究进展和性能评价[J]. 自动化学报，2023，49（3）：635-660.

[7] 易龙. 从数字出版到智能出版：知识封装方式的演进[J]. 出版科学，2023，31（1）：81-90.

[8] 崔浩男. 多模态档案知识服务平台的基本特征与价值取向：基于国内外20个案例的分析[J]. 档案学通讯，2024（1）：70-78.

[9] 新一代通用向量模型BGE-M3：一站式支持多语言、长文本和多种检索方式[EB/OL]. [2025-02-08]. https://hub.baai.ac.cn/view/34816.

[10] 阿里云计算有限公司. Embedding模型[EB/OL]. [2025-02-12]. https://help.aliyun.com/zh/model-studio/user-guide/embedding.

[11] OpenAI. 嵌入指南（Embeddings Guide）[EB/OL]. [2025-02-12]. https://www.openaidoc.com.cn/docs/guides/embeddings.

[12] CSDN. 机器学习：回归模型和分类模型的评估方法介绍[EB/OL]. [2025-02-12]. https://blog.csdn.net/rubyw/article/details/142828639.

Retrieval-Augmented Generation (RAG)-Driven Knowledge Service: Principles, Paradigms, and Evaluation

WANG Liang
School of Economics and Management, Beijing Institute of Graphic Communication, 102600, Beijing, China

Abstract This paper examines the limitations of artificial general intelligence (AGI) in professional knowledge service domains, particularly its inability to reconcile the breadth of general-purpose corpora with the depth required for specialized expertise, a challenge exacerbated by the static nature of training data and the inherent trade-offs between generalization and precision. These limitations stem from the AGI's dependence on open-source, nonspecialized training data, which exclude high-value, peer-reviewed resources, and its static architecture, which struggles to adapt to the dynamic evolution of domain knowledge. To address these challenges, this study proposes Retrieval-Augmented Generation (RAG), a hybrid framework that integrates the semantic comprehension and generative fluency of LLMs with the authority and precision of structured knowledge bases. RAG operates through three interconnected phases: vectorization, where domain-specific texts and user queries are transformed into high-dimensional embeddings to capture semantic nuances; retrieval, which employs similarity search algorithms to extract contextually relevant knowledge snippets from vectorized databases, ensuring alignment with professional standards; and generation, where LLMs synthesize retrieved content with user inputs to produce outputs that balance readability with factual accuracy. The implementation of a RAG requires meticulous attention to knowledge base construction, digitization and standardization of domain content through metadata tagging, ontology development, and knowledge graph integration to ensure semantic consistency. Model selection further influences performance: open-source options such as BGE offer flexibility for niche domains but may lack scalability, whereas commercial solutions such as Aliyun's text embedding provide robust multilingual support at higher costs. LLM selection must align with application needs: models such as DeepSeek-V3 excel in Chinese-language contexts because of localized optimization, whereas GPT-4 proves advantageous for multilingual tasks despite privacy concerns. Experimental validation via simulated datasets—professional technical manuals, educational quizzes, and popular biographies—demonstrated RAG's efficacy. In professional scenarios, the RAG algorithm achieves excellent accuracy by leveraging structured knowledge bases. However, in educational and popular contexts, accuracy has decreased slightly, but it is still acceptable. For the publishing industry, a RAG offers transformative potential but demands strategic adaptations. Infrastructure localization is paramount for safeguarding proprietary content; hybrid cloud architectures can balance cost efficiency with data security, whereas blockchain integration ensures immutable copyright tracking. Workflow optimization should automate metadata tagging during editorial processes and integrate consistency checks into proofreading stages, reducing manual labor. Data standardization must address multimodal challenges—e.g., aligning image annotations with textual descriptions—to support emerging applications such as interactive textbooks. Copyright protection requires granular access controls and encryption, particularly for subscription-based services. Despite these advancements, the RAG algorithm faces unresolved challenges: multimodal data integration remains computationally intensive, real-time updates strain system latency, and conflicting knowledge sources necessitate advanced conflict-resolution frameworks. Future research should explore adaptive retrieval algorithms, federated learning for decentralized knowledge bases, and hybrid human-AI validation mechanisms to increase reliability. By bridging AGI's generative capabilities with domain expertise, the RAG not only elevates the precision and adaptability of knowledge services but also catalyzes innovation in digital publishing, enabling industries to harness their authoritative content as dynamic, interactive assets in an increasingly data-driven world.

Keywords knowledge service; retrieval-augmented generation; RAG; artificial intelligence; evaluation

（责任编辑：郭田珍）