

企业 RAG 知识库建设与使用白皮书

知识治理 检索架构 引用质量与持续评估

面向企业知识负责人 技术负责人及应用建设团队

讨论资料如何转化为可信 可用且可持续维护的知识服务

广州数娱信息科技有限公司

2026 年 10 月 29 日 版本 1.0

执行摘要

企业 RAG 知识库的建设目标，是在用户权限和业务语境内找到可靠资料，生成有依据、可核验且适合实际工作的回答。把文档转为向量并接入大模型，只完成了部分工程。资料是否可信、版本是否有效、权限是否同步、回答是否真的被引用支持，往往决定系统能否长期使用。

本白皮书建议将知识库作为持续运行的知识服务建设。项目从业务问题和资料责任人开始，依次建立来源登记、结构化解析、权限与版本控制、混合检索、重排和证据输出。评测同时检查检索、回答、引用、拒答和权限，分别分析每一层的失败原因。

企业应先用关键词检索和简单 RAG 建立可复现基线，再决定是否引入图谱、复杂查询规划或模型微调。艺术文博、多语种内容、项目资料和企业制度可以共用技术底座，但需要不同的元数据、知识关系、审核方式与质量要求。

研究方法 with 适用范围

本文以检索与生成领域论文、公开技术实验和官方工程文档为研究资料，检索截至 2026 年 10 月 29 日。公开研究的任务、语料和模型版本各不相同，不能直接横向排列为产品能力榜单。本文架构、评测配额与验收建议属于实施方法，不是行业统一标准。

本文未对企业客户实施问卷调查或统一性能实测。场景、数据分布和成本均在相应位置标注为设计示例或假设算例。知识问答产生的建议不能替代有资质人员在鉴定、合同、财务等场景中的专业判断。

阅读导航

01 RAG 的价值与边界

03 知识治理与入库

05 权限与知识生命周期

07 评测与验收

09 运维与采购清单

02 公开研究与技术选择

04 检索与回答架构

06 艺术文博及企业使用场景

08 成本与实施路径

参考资料

01 RAG 的价值与边界

把外部证据带入回答

检索增强生成即 Retrieval Augmented Generation，通常简称 RAG。系统先从外部知识中找出相关资料，再将资料作为上下文提供给生成模型。原始 RAG 研究将参数化模型与外部检索记忆结合，用于知识密集型任务。[R1] 企业工程中的文档接入、混合搜索和权限过滤是这一思想的应用扩展，不等于原论文系统的直接复制。

RAG 可以帮助使用较新资料、展示来源及限制回答范围，但不保证回答真实。检索阶段可能漏掉有效资料；资料本身可能错误；生成阶段可能过度推断或引用不匹配。因此“接入知识库”不能作为“消除幻觉”的同义表达。

根据知识特征选择方法

需求特征	优先考虑	关键条件
少量稳定资料	直接提供上下文	权限和输入长度可控
查找原文或精确编号	关键词与结构化检索	文档结构和字段完整
多文档知识问答	RAG	有可信语料与可评测问题
实时库存 价格或业务状态	数据库或业务 API	数据口径与权限受控
跨实体关系及全库主题分析	图谱或图 RAG 试点	关系价值高于维护成本
固定格式与领域表达	提示模板 必要时微调	有质量稳定的训练样本

RAG、微调和实时工具查询可以组合。微调适合改善行为、格式或特定任务表现，不能可靠替代持续变化的事实库；数据库查询适合精确数值与状态，不能把复杂权限和计算完全交给自由文本生成。

02 公开研究与技术选择

检索能力必须在目标语料上验证

BEIR 在 18 个不同检索数据集上比较多种检索架构，发现 BM25 是稳健基线，部分重排和晚交互方法表现较强但计算成本更高。[R2] 这为企业保留关键词基线提供依据，并不能证明某个组合在所有中文资料上最佳。编号、专名、简称和相近版本往往需要不同检索策略。

Anthropic 在其 Contextual Retrieval 实验中报告，结合上下文增强向量与 BM25 后，前 20 个片段的检索失败率从 5.7% 降至 2.9%，相对下降约 49%。[R4] 该指标描述特定实验的检索表现，并非回答准确率、业务正确率或企业通用改善幅度。它支持测试“补充片段语境”的方法，但收益仍需在目标资料上验证。

更长上下文与更复杂架构均有条件

Lost in the Middle 研究发现，相关证据在上下文中的位置会影响其所测试模型的表现。[R3] 该研究不能证明所有当前长上下文模型具有相同退化程度；它提示企业测试不同信息位置、噪声比例和上下文长度，避免仅依据窗口容量采购。

GraphRAG 研究面向全库主题等全局总结问题，通过实体关系和社区摘要组织知识。[R5] 对“某件展品的材质是什么”这类局部事实，先测试常规检索通常更容易建立基线。只有多跳关系、主题综合或跨文献关联确有需求时，才值得承担实体消歧、抽取质量与更新维护成本。

Self RAG 研究通过专门训练与反思机制按需检索和判断生成内容。[R9] 企业可以借鉴检索必要性判断和证据检查的思想，但增加一段自检提示并不等于复现该训练方法。

03 知识治理与入库

资料目录先于向量索引

企业需要为每类资料确定来源、维护人、使用对象、有效期和授权条件。正式制度、历史版本、客户交付物、个人工作笔记应区别管理。多个来源冲突时，应制定优先级或展示分歧；模型不能自行把最新上传时间等同于业务上最新有效版本。

元数据字段	示例或用途	缺失后的典型问题
document_id 与 version	稳定文档编号及版本	更新后无法判断来源
source 与 owner	来源系统及责任人	错误内容无人维护
valid_from 与 valid_to	生效及失效时间	历史规则被当成现行规则
access_scope	企业 部门 项目或角色范围	跨项目泄露资料
entity_id 与 aliases	藏品编号 产品型号及别名	同名异物或简称混淆
citation_anchor	页码 章节 时间码或图号	引用无法定位
review_status 与 rights	审定状态及素材权利	内部资料误用于对外发布

解析质量决定能否找到证据

扫描件需要 OCR 并记录置信度；低质量页应进入人工修订队列。表格要保留表头、单位、合并单元格含义和来源页码；图注不能与图片对象脱离；音视频转写需要保留时间码和说话人信息。解析错误通常无法通过更换大模型补救。

切分以语义和版面结构为依据。制度条款保留适用范围与例外；产品参数保留型号、数值和单位；研究文章保留作者、年代与限定词。较长条目可用父子片段连接，检索小片段后补充必要上文。片段大小与重叠比例应通过评测选择，不能把固定 token 数写成普适规则。

知识可信度与权利需要独立记录

来源权威、内容最新和具有使用权，是不同属性。已获授权的资料可能过期；可信资料可能仅供内部查阅；公开可浏览的图片也不必然允许商业再利用。知识系统应让“能检索到”“能生成内部草稿”“能对外发布”形成不同权限条件。

04 检索与回答架构



图1 本文建议的入库链路及查询链路 原始出处与权限贯穿两条链路

查询先经过身份与业务范围约束

系统从已认证会话取得身份，确定企业、部门、项目和资料范围，之后进行查询理解。用户查询可以被改写为检索表达，但原始意图和限制条件需要保留。“截至去年年底”的时间范围不能被改写丢失，馆藏编号也不能被语义扩写替换。

召回层并行使用关键词、向量及必要的结构化条件。合并结果后去重，并按来源、有效版本与相关性处理。重排模型可以重新排序候选证据，但不能授予访问权限。生成模型只接收允许访问且与当前任务有关的片段。

回答应携带可定位的证据

回答可以采用“结论、证据、适用条件、未确认事项”的结构。系统必须区分原文事实、跨资料综合和推断。引用应定位到具体版本、页面、章节或时间点；一个真实存在的链接并不能证明它支持当前论断。

例如提问“这个项目是否可以继续沿用上一版报价”，系统先核对项目、有效期、产品版本和授权范围。若仅找到历史报价，宜说明“已找到历史记录，但未找到当前有效确认”，而不是根据相似文本推断仍可使用。

检索失败需要可解释处理

没有足够证据时，系统应拒答、询问或转交责任人。拒答原因可区分“未找到资料”“资料冲突”“超出知识范围”和“当前身份无权访问”，但对无权用户不暴露受限文档标题、存在性及内容摘要。需要在用户体验和防止信息侧漏之间确定具体规则。

上线初期可比较四种配置：关键词检索、向量检索、混合检索、混合检索加重排。固定语料快照、问题集和预算，每次只改变一个主要变量，并记录检索质量、回答质量、时延与费用。用失败类别决定优化方向，避免同时调整全部参数后无法归因。

05 权限与知识生命周期

权限不能只靠回答提示词

权限应在检索服务和业务服务中执行，覆盖召回、重排、生成上下文、缓存和引用回读。Microsoft 的工程文档说明了基于身份元数据进行文档级过滤的模式。[R8] 索引中的权限字段只是执行依据的一部分，认证身份和每次查询的过滤仍需可信后端保证。

共享缓存要把用户或授权范围、知识版本和策略版本纳入键值，避免另一名用户读取前一个人的回答。权限撤销后，需要同步索引、缓存、会话与下游副本；根据资料敏感度确定撤销生效时限，并实际测试撤权后的访问。

更新需要可重建与可回退

入库记录应保留原文哈希、解析器版本、切分策略、embedding 模型和索引版本。更换向量模型通常需要重建相关索引；新旧向量不宜未经验证混用。建立影子索引，使用冻结评测集比对后再切换，保留必要回退能力。

文档更新时，需要处理旧片段、引用、摘要和问答缓存。删除原文后只删除文件、却保留其向量或生成摘要，会形成残留。是否保留审计证据应按企业制度另行处理；知识检索可见性和审计留存不应混为一谈。

日志同样属于数据资产

日志中可能存在用户问题、客户名称、原文片段和内部决策。应确定采集范围、脱敏方式、查看权限与留存周期。若采用外部模型或重排服务，要列明哪些数据离开原环境、如何处理以及由谁承担维护责任。私有化部署也需要接口权限、访问控制和运行审计，不能单凭部署地点判断安全。

06 艺术文博及企业使用场景

艺术文博知识应用设计示例

为器物、作者、时代、材质、技法、纹饰、收藏机构、文献和图像建立可追溯标识。历史名称、异体字和多语种译名进入别名表；同名实体通过馆藏编号和来源消歧。已确认归属、学术争议和策展解释应保留不同知识状态。

用户询问某器物纹饰含义时，系统先确认对象，再分别检索馆藏说明与研究资料；证据不足时说明解释范围。图像相似检索可用于寻找候选对象，但不能直接生成真伪鉴定结论。面向公众的讲解需要引用审定资料，内部研究界面可以展示不同观点及来源。

多语种场景要同时评估检索和生成。可以采用跨语种向量、查询翻译或双语词表，但须测试专业名词、纪年、音译和文化概念。中文正确率不能直接代表其他语种表现；每种重点语言应有独立样本及人工审核人员。

内部生产力应用设计示例

知识库可作为项目助手、方案工具和内容工作台的公共能力。用户输入项目背景后，系统查找历史交付物、可复用模板和决策依据，生成带来源的草稿。界面应允许查看原文、纠错和申请知识更新，并明确草稿与正式交付物的区别。

这类系统需要检验“是否找到正确资料”以及“是否减少实际工作负担”。重复问题、无人维护资料 and 频繁人工补充说明，可能反映源头治理问题。不能把所有问题归因于模型理解能力。

实时事实使用业务接口

当用户询问库存、项目进度或实时价格时，回答应读取相应业务系统并标记查询时间。静态文档可提供规则解释，数据库提供当前状态。接口失败时显示失败或最后更新时间，不应使用过期文档伪装实时结果。

使用者需要给出对象与核验要求

有效提问通常包含对象、时间、任务和所需输出。相比“帮我介绍这个项目”，“请基于项目 P017 截至本月生效的审定资料，整理面向公众的 200 字介绍，并列出每项关键事实的出处”更容易得到可核验结果。用户不能通过提示词扩大自己的资料权限。

使用目的	建议提问方式	使用后的核验
查找事实	指明对象 编号 时间与所需字段	回读相应版本和单位
比较资料	指定比较对象 维度及有效日期	检查是否遗漏例外与冲突
生成草稿	说明受众 用途 长度和引用要求	审核事实 权利及对外口径
多语种讲解	指明语言 术语表与知识范围	由具备领域能力的人员复核

发现回答错误时，反馈应保留问题、答案、引用和错误位置。知识维护人判断是资料有误还是系统误读；用户不宜只在单次对话中纠正后就认为全库已经更新。对话反馈、正式知识修订和重新入库需要明确衔接。

07 评测与验收

从问题分布建立测试集

建议以 200 个问题作为启动规划示例：单文档事实 60 个、跨文档综合 40 个、表格与编号 30 个、版本或时间 20 个、权限与撤权 20 个、无答案或冲突 20 个、多语种及专业名词 10 个。若多语种是主要业务，需显著增加各语种样本，不能沿用该示例配额。

标注记录至少包括问题、用户角色、应答或拒答、有效资料版本、支持证据及关键得分项。多文档题的标准答案可以有多种表达，但必须具备必要事实。开发集与冻结测试集分开，模型生成问题只能补充覆盖，不能替代业务人员提供的真实问题。

指标	本文建议口径	用途
Recall at k	前 k 个结果覆盖的已标注相关证据数除以相关证据总数 再按问题平均	检查证据漏召回
证据命中率	前 k 个结果至少包含一条必要证据的问题占比	不能替代多证据召回率
回答正确率	关键事实全部满足评分要求的问题占比	衡量业务正确性
引用支持率	被引用证据实际支持的论断数除以有引用论断数	检查引而不证
引用完整率	有有效引用的应引用事实数除以全部应引用事实数	检查关键事实无来源
正确拒答率	应拒答且正确拒答的问题数除以应拒答问题数	检查无依据回答
过度拒答率	可回答却拒答的问题数除以可回答问题数	防止以拒答换高分

权限泄露和错误外发应单独计数，不能合并进文本平均分。时延报告 P50、P95 与超时率；查询成本按完整回答计入召回、重排和生成。根据任务重要性设置验收阈值，本文不提供跨行业统一合格线。

自动评分需要人工校准

RAGAS 与 ARES 提供了拆分检索与生成质量的评估思路。[R6][R7] 自动评分便于迭代，但裁判模型可能偏好表达流畅的回答，也可能误判领域事实。抽样人工复核、保存原始证据并分析评分分歧，才能把自动分数用于工程决策。

用消融实验定位瓶颈

先固定生成模型比较检索配置；再固定证据比较模型；最后测试完整系统。若给定标准证据仍答错，优化检索不能直接解决；若标准答案所在文档未解析成功，提高 top k 也无效；若问题本身超出资料范围，正确动作是澄清或拒答。

示例验收记录应包含配置版本、语料快照、问题编号、召回片段、答案、引用映射、评分、耗时和费用。对关键错误保存可重现最小样例，作为后续版本的回归任务。

08 成本与实施路径

资料规模和查询量共同决定成本

建设成本包括资料清理、解析、元数据、身份集成和界面开发；运行成本包括向量生成、索引服务、重排、模型调用和监控；持续成本还包括知识维护、人工评测与错误处理。资料量大但查询稀少的场景，可能主要承担存储与治理成本。

单次回答成本 = 检索与重排费用 + 模型输入输出费用 + 按使用量分摊的基础服务费用。每个有效回答成本还要把失败重试和无效回答计入分子。供应商应列出计费假设，避免只展示一轮理想问答的 token 成本。

以下为假设算例：每月 10,000 次查询，80%形成可用回答，每个可用回答平均减少 4 分钟检索整理时间，人工综合成本 90 元每小时，潜在工时价值为 48,000 元。若每月技术运行与知识维护共 18,000 元，初始建设投入摊销前净值为 30,000 元。

平均实际节省时间	有效回答数	对应工时价值	扣除月度成本后的净值
2 分钟	8,000 次	24,000 元	6,000 元
4 分钟	8,000 次	48,000 元	30,000 元
6 分钟	8,000 次	72,000 元	54,000 元

假设查询对应真实工作需求，核验与返工已计入节省时间，而且质量不下降。若员工原本不会完成该任务，不能把全部回答时间折算为节省；若员工没有实际使用，理论问答量也不能计入收益。

按知识范围逐步实施

第一阶段选一个部门 and 一类稳定资料，确定维护人、基线问题及权限；第二阶段完成解析、引用和简单检索，暴露资料问题；第三阶段按评测引入混合检索或重排；第四阶段与内部工具结合并运行灰度；第五阶段再扩展图谱、多语种或复杂流程。

每阶段保留可迁移资料与索引配置。扩展范围前检查维护能力：增加资料 and 用户往往意味着增加权限组合、版本冲突与人工审核，而不只是增加向量数量。

09 运维与采购清单

知识服务需要持续负责人

建议建立业务知识负责人、系统负责人和使用部门之间的处理机制。用户反馈“答案错了”后，先确认来源、版本和任务，再决定修订原文、标注争议、调整检索或修改生成策略。错误处理记录可形成质量改进队列。

按周观察失败类型和高频问题，按版本运行回归评测，按资料更新要求执行同步检查。关键知识可设置发布前审核和定期复审。无效资料、过期规则与来源失效链接应定期处理，防止语料规模增长掩盖有效知识下降。

交付项目	应核验的内容
资料与授权清单	来源 权利 维护人 有效期 可见范围
入库管线	OCR 结构解析 切分 元数据 错误队列
查询与引用	过滤 混合召回 重排 引用回读 拒答
评测与验收	数据集 口径 基线 版本对比 失败分析
生命周期	更新 撤权 删除 重建 回退与缓存失效
数据退出	原文 元数据 配置 日志与评测资料的导出

采购演示宜加入同名对象、历史版本、撤销权限和无答案问题，观察系统如何解释边界。只上传数份整理良好的资料并提问已知答案，无法验证企业长期使用所需的治理能力。

数娱科技的知识库服务可与企业生产力工具、工作流和多模态应用组合。项目以明确知识范围和可验收任务为起点，EVIPX 作为可选产品平台能力之一。平台复用与定制范围、部署方式及维护责任在项目方案中分别说明。

参考资料

[R1] Lewis P 等 Retrieval Augmented Generation for Knowledge Intensive NLP Tasks 2020

<https://arxiv.org/abs/2005.11401>

[R2] Thakur N 等 BEIR A Heterogenous Benchmark for Zero shot Evaluation of Information Retrieval Models

2021 <https://arxiv.org/abs/2104.08663>

[R3] Liu N F 等 Lost in the Middle How Language Models Use Long Contexts 2023

<https://arxiv.org/abs/2307.03172>

[R4] Anthropic Introducing Contextual Retrieval 2024 厂商实验 <https://www.anthropic.com/research/contextual-retrieval>

[R5] Edge D 等 From Local to Global A Graph RAG Approach to Query Focused Summarization 2024

<https://arxiv.org/abs/2404.16130>

[R6] Es S 等 RAGAS Automated Evaluation of Retrieval Augmented Generation 2023

<https://arxiv.org/abs/2309.15217>

[R7] Saad Falcon J 等 ARES An Automated Evaluation Framework for Retrieval Augmented Generation Systems

2023 <https://arxiv.org/abs/2311.09476>

[R8] Microsoft Security filters for trimming results in Azure AI Search 官方工程文档 <https://learn.microsoft.com/en-us/azure/search/search-security-trimming-for-azure-search>

[R9] Asai A 等 Self RAG Learning to Retrieve Generate and Critique through Self Reflection 2023

<https://arxiv.org/abs/2310.11511>

以上网页访问日期均为 2026 年 10 月 29 日。论文和厂商实验的语料、版本及指标定义不同，不能直接用于产品排名或合同效果承诺。

关于数娱科技

广州数娱信息科技有限公司围绕企业 AI 应用与文旅数字化开展业务。企业 AI 服务包括知识应用、生产力工具、工作流及多模态应用等方向，具体范围按项目确定。EVIPX 是 AI 板块中的产品平台能力。

官网 <https://shuyukeji.cn/> 产品平台 <https://evipx.com/>

业务联系 market@shuyukeji.cn 020 38080881