

基于Qwen大模型构建RAG智能问答智能体的应用实践

徐前进 郭肖和 刘易斯 查德飞 史鸿晖

中国电信股份有限公司安徽分公司, 中国·安徽 合肥 230031

摘要: 海量异构知识资产的高效挖掘与利用成为电信行业亟待破解的核心问题。本文依托千问大模型, 融合检索增强生成 (RAG) 技术构建电信知识智能问答智能体, 系统阐述其整体架构设计、核心技术实现路径及业务赋能实际成效。通过知识精细化预处理、Milvus 向量数据库搭建、组合式文本智能切分及生成环节优化等关键举措, 有效解决了向量召回精度不足、多模态文档解析困难、问题与答案间语义鸿沟等行业共性难题。实际应用结果表明, 该智能问答系统的专业问答准确率达 90%, 显著提升了电信业务运营效率, 为电信行业知识问答的智能化升级提供了可落地的实践参考。

关键词: 检索增强生成; 大语言模型; 智能体; 智能问答

Application Practice of Building RAG Question-Answering Agent Based on Qwen Large Model

Xu Qianjin, Guo Xiaohe, Liu Yisi, Zha Defei, Shi Honghui

China Telecom Anhui Branch, China Anhui Hefei 230031

Abstract: The efficient mining and utilization of massive heterogeneous knowledge assets have become a core problem urgently to be solved in the telecommunications industry. Relying on the Qwen large model, this paper constructs an intelligent question-answering agent for telecommunications knowledge by integrating Retrieval-Augmented Generation (RAG) technology, and systematically elaborates its overall architecture design, core technology implementation paths and practical effects of business empowerment. Through key measures such as refined knowledge preprocessing, construction of Milvus vector database, combined intelligent text segmentation and optimization of generation links, common industry problems such as insufficient vector recall accuracy, difficulty in multimodal document parsing and semantic gap between questions and answers are effectively solved. Practical application results show that the professional question-answering accuracy of this intelligent question-answering system reaches 90%, which significantly improves the operation efficiency of telecommunications services and provides a practical reference for the intelligent upgrading of knowledge question answering in the telecommunications industry.

Keywords: Retrieval-augmented generation; Large language model; Agent; Intelligent question answering

0 引言

近年来, 大型语言模型 (Large Language Models, LLMs) 参数量呈爆发式增长, 其学习能力、泛化能力与上下文理解能力得到大幅提升, 无需针对特定任务进行定制化适配, 即可满足多行业领域的通用自然语言处理需求^[1-2]。

2025 年 8 月, 国务院印发《深入实施“人工智能+”行动的意见》, 明确提出到 2027 年实现人工智能与重点领域深度融合, 智能体应用普及率超 70% 的发展目标。2025 年行业大模型技术持续突破, DeepSeek、Kimi、豆包等国产 AI 工具百花齐放, 推动人工智能从技术成熟走向行业应用爆发期, 形成“技术普惠、产业升级、治理增强”三位一体的发展格局, 为企业智能化转型筑牢了政策与产业双

重支撑^[3-4]。在此背景下, 电信运营商在长期运营中积累了海量格式异构、存储分散且专业术语密集的核心数据资产, 如何借助大语言模型等关键技术, 将这些资产转化为以数据为驱动、业务变革为核心的智能化运营体系, 进而实现生产力跃迁与决策效能优化, 成为电信行业数字化转型过程中亟待解决的关键问题。

随着大模型以及检索增强生成 (Retrieval Enhanced Generative, RAG) 技术的快速发展, 依托该技术实现垂直领域知识的快速检索与精准查询, 已经是目前各行业智能化发展的趋势^[5]。本文以千问大语言模型为基础, 完成电信知识的向量化处理, 融合 RAG 技术构建“小易 AI”智能问答智能体, 详细阐述通信行业场景下 RAG 与大模

型融合的关键技术实现方案，并分析其业务赋能实际成效。

1 检索增强生成技术概述

大模型是基于深度学习技术的大规模自然语言处理模型，但仅能基于训练数据获取知识，而无法处理超出训练数据范围的特定领域问题。为了弥补这一不足，RAG 技术可以增强 LLM 的生成能力，有效减少 LLM 幻觉现象，提高模型在垂直领域中的回答精度^[6-8]。它主要由知识库、检索器、生成器三个部分组成。基本的流程如图 1 所示。

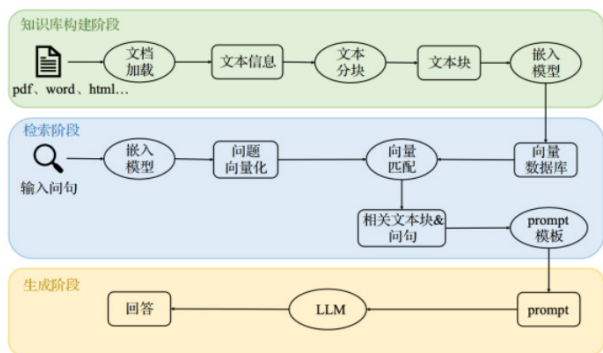


图1 RAG基本流程图

1.1 知识库构建

在向量知识库构建阶段，需要将不同格式的文档如 word、pdf、markdown 等通过文档解析器进行解析转化为纯文本格式。为了提升模型检索的性能，需要将纯文本格式的文档进行切片处理，形成一个个小片段，这些片段称为文本块，文本切片的方式多种多样，不同切片方式得到的文本块都不完全相同，同时影响最终的问答效果。文本分块后需要将文本块转换为高维度的语义表征信息，这一过程需要使用文本嵌入模型，它将上下文映射到高维度向量空间从而得到编码向量，编码向量的维度由嵌入模型定义。完成文本嵌入后需要构建索引，将原始文本块、文本块向量和一些元数据信息，如作者、文章名字、摘要等信息存储进向量数据库构建索引，完成向量知识库的构建。

1.2 检索阶段

在检索阶段，RAG 系统会对用户的提问进行处理，利用文本嵌入模型将其映射到高维潜在语义空间中，生成稠密的查询向量，之后系统将该查询向量索引化并输送至向量数据库进行相似度检索。该过程通常采用近似最近邻搜索（approximate nearest neighbor, ANN）从而在海量数据中实现高效检索，通过计算查询向量与库中预存文本块向量之间的向量距离来比较文本块与用户提问的相关性得到相似度得分。基于上述计算结果，系统将依据相似度得分进行降序排列，召回并返回得分最高的前 K 个文本块，从而为大模型提供精准的参考回答文本块。

1.3 生成阶段

将用户提问、检索到的向量化文本块及系统提示指令，基于大模型与向量化检索的协同机制进行结构化整合。首先，通过计算用户查询向量与知识库向量之间的余弦相似度，系统筛选出相似度最高的前 K 个文本块作为候选上下文。随后，将这些文本块、用户原始提问以及预设的指令约束（如“请严格依据检索内容生成答案”）按预定模板拼接为完整的模型输入。调用大模型时，通常设置较低的温度参数以增强回答的确定性与一致性。模型将依据指令引导，在向量相似度最高的文本块中定位相关信息，并生成准确、可溯源的答案。整个流程以向量相似度为关键反馈依据，确保生成内容与检索知识在语义空间中的紧密对齐。

2 架构与相关技术

2.1 技术架构

基于 Qwen 大模型及多模态大模型基模，通过关键词提取、重排序小模型核心能力，打造“分级分域、智能切萃、即传即问”多模态知识智能萃取，构建“小易 AI”智能问答智能体，整体架构如图 2 所示。

2.2 RAG 实现

2.2.1 向量库构建

构建向量数据库是 RAG 的核心环节之一，需要在检索性能和存储效率之间权衡，向量数据库筛选方面，目前市场上向量数据库有 Chroma、Milvus、Pinecone 等，具体特点如表 1。

表1 向量数据库的特点

产品	开源情况	查询性能	使用规模	核心定位
Chroma	Apache-2.0	毫秒级	千万级	轻量级/原型
Milvus	Apache-2.0	微秒-毫秒级	十亿级	企业级
Pinecone	闭源商业	毫秒级	亿级	全托管

小易 AI 智能问答智能体面向人员使用量约为 3 万人，未来将整合省内各类知识数据，从知识向量化与管理的维度出发，当前系统基于 2 万余条初期知识（未来可扩展至多业务域），通过 BGE-large-zh 模型将其编码为 1024 维稠密向量，并构建稀疏向量表示，形成“稠密 + 稀疏”的混合语义表征体系。在向量存储与检索架构上，选用 Milvus 向量数据库，主要基于以下向量化能力优势：

企业级稳定性与扩展性：支持企业级高可用部署，满足日均 3 万至未来全省量级用户的稳定服务要求，具备平滑扩展至十亿级向量的横向扩容能力，为后续多域知识融合提供底层支撑。

原生高效混合检索：依托 Milvus 原生支持的混合检索

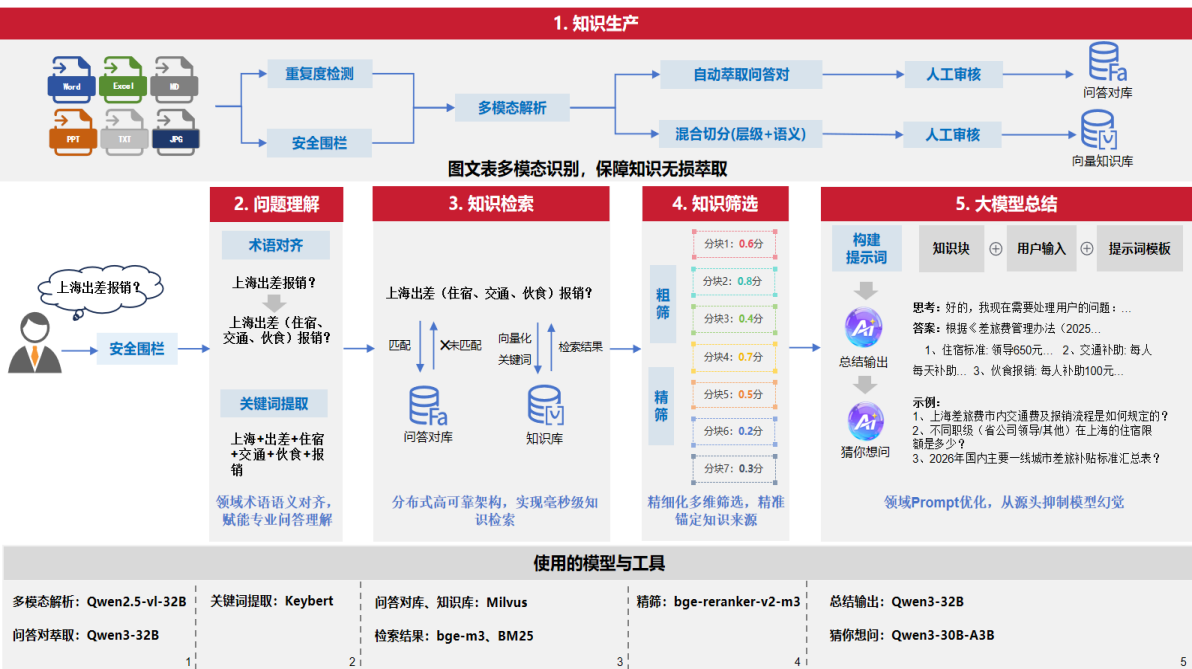


图2 整体架构图



图3 多模态文件解析流程图

机制，深度融合语义向量检索与关键词检索，通过对稠密向量与稀疏向量进行加权融合，兼顾语义相关性匹配与实体术语精确召回，显著提升多场景问答准确率。

多租户向量隔离：支持基于部门、区域、业务域等维度建立独立 Collection，实现向量数据在存储层面的逻辑隔离。该机制在保障多租户数据安全的同时，有效限定检索空间，降低跨域误召回，提升检索精准度与响应效率。

2.2.2 多模态知识萃取与预处理

在知识采编阶段，小易 AI 智能问答智能体接入 AI 安全围栏，经重复度与安全检测、多模态解析，文档自动转化为高质量的问答对和知识切片，最终通过人工审核校验入库，确保图文表格无损萃取。

针对 PDF、图片、Excel 等多模态知识，调用 Qwen2.5-vl-32B 视觉 - 语言多模态大模型，对各类型格式文档进行图像理解、文档分析等多模态解析及问答对生成，以确保知识无损萃取，如图 3 所示。

此外，对全部 2 万条数据执行了去重处理，基于字段完全匹配规则剔除重复条目，数据质量得到有效提升。经去重后，最终保留有效知识数据 1.9 万条。

2.2.3 知识解析和切分

知识解析与切分是 RAG 处理流程中的关键环节，其质量直接决定问答系统的最终效果。在多模态知识解析方面，针对图片内容构建了高效并行的解析机制：调用 Qwen 多模态大模型，实现对 PPT、PDF、活动海报、通知公告等富含信息图像的有效筛选与识别，确保仅对高质量、高信息量图像进行后续处理。在信息提取环节，多模态大模型可对图像关键文本进行结构化抽取与内容摘要，实现图文信息的深度融合表示。该设计在保障解析准确性与信息完整度的同时，通过高并发架构显著提升了解析效率，为后续向量化与检索提供高质量、低延迟的多模态文本输入。

在文本切分方面，针对 Markdown 文本提出了一种结合标题切分与语义切分的组合优化方案。如图 4 所示混合切片方案先基于 Markdown 的章节标题进行切分，并在每个文本块中保留其父级标题，以保证语义的连贯性。当某个文本块的长度超过预设的最大长度限制时，系统将对其进行语义切分；语义切分后的每个子文本块将保留父文本块的标题，并附加索引序列。标题切分能够快速生成文本

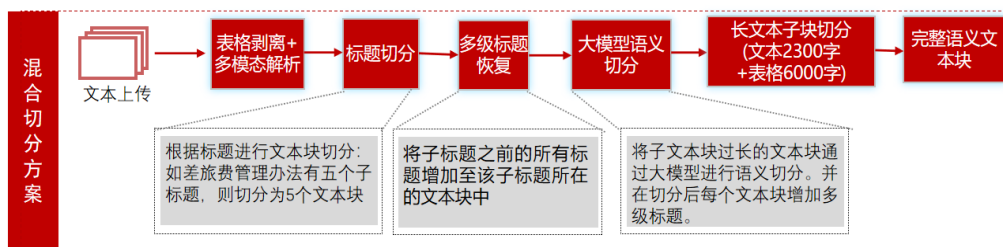


图4 混合切片方案

区域医疗-县域医共体信息平台介绍

建设内容

【图片总结】：该图片介绍了县域医共体信息平台的产品身份卡，旨在通过信息化手段有效落实分级诊疗政策，构建以县级公立医院为龙头、乡镇卫生院为枢纽、村卫生室为基础的医疗共同体。平台以管理共同体、利益共同体、责任共同体和服务共同体为核心，推动医疗服务模式从以治病为中心转向以健康为中心。平台基于湖仓一体架构的医疗大数据平台，整合自研及生态场景化应用与能力，形成行业综合解决方案及标准化产品，并通过集成/被集成两类模式推广。其核心功能包括数据实时对接、标准化医共体监管指标、统一标准和安全体系以及湖仓一体化大数据底座。重点信息：面向客户：区县卫健委、区县中心医院；核心能力：数据实时对接、标准化医共体监管指标、统一标准和安全体系、湖仓一体化大数据底座；建设价值：通过数据中心实现医共体内医疗业务数据的互通共享，通过统一门户实现基层与县级医院基础业务一体化，通过管理中心实现医共体综合管理标准化；平台架构：基于湖仓一体架构的医疗大数据平台为核心，涵盖全健康大数据平台、县域医共体及基层一体化平台、传染病监测预警与应急指挥平台、医疗影像云平台、医保卫士平台等；应用场景：数据治理、数据资源、数据服务、数据可视化分析、AI服务等；行业数据源：医院、卫健委/局、医保局、疾控中心、乡镇卫生院、社区卫生服务中心、急救中心、妇幼保健院；推广模式：集成/被集成两类模式；目标：助力完成中国医改，推动医疗服务模式转型，实现全民健康信息化。](http://134.95.222.252:20401/learn/api/getFile?filePath=/nfs/learnupload/floaBoard/202403/w8235209_20240321100847903.jpg&fileName=幻灯片9.jpg)

【图片总结】：该图片对比了县级全民健康平台与医共体信息平台在建设主体、服务对象、建设目标、业务场景和监管指标等方面的区别。县级全民健康平台由区县卫健局牵头建设，覆盖区域内所有人员，旨在实现公共卫生、医疗服务、药品管理、医疗保障等领域的综合互通；而医共体信息平台由区县中心医院或区县卫健局牵头建设，主要服务于县域医共体单位内的机构和患者，以落实分级诊疗政策为核心。重点信息：1. **建设主体**：县级全民健康平台：区县卫健局牵头；医共体信息平台：区县中心医院或区县卫健局牵头。2. **服务对象**：县级全民健康平台：覆盖辖区内所有人员（包括医疗机构、医疗资源和居民等）；医共体信息平台：仅服务于县域医共体单位内的机构和患者。3. **建设目标**：县级全民健康平台：涵盖公共卫生、医疗服务、药品管理、医疗保障等领域的综合性信息平台，实现各类医疗健康数据的互联互通和共享互通；医共体信息平台：以县级公立医院为龙头，乡镇卫生院为枢纽，村卫生室为基础，构建县乡一体化管理模式的医共体，核心在于管理共同体、利益共同体、责任共同体和服务共同体。4. **业务场景**：县级全民健康平台：实现区域内健康医疗数据汇集、共享及档案CDA、交互服务标准统一，逐步实现医疗健康数据互通、信息共享和检查检验结果互认；医共体信息平台：整合区域内医院与其他医疗服务机构，重新构建整体性的全新医疗组织架构，统筹医共体内所有医疗资源。5. **监管指标**：县级全民健康平台：对县域整体卫生指标情况进行统计分析和展示，包含医疗服务、公共卫生、医疗保障、卫生资源等信息的指标数据汇总和监管；医共体信息平台：监测指标体系由有序就医格局基本形成、县域医疗卫生服务能力提升、医疗卫生资源有效利用、医保基金使用效能提升四个方面构成。](http://134.95.222.252:20401/learn/api/getFile?filePath=/nfs/learnupload/floaBoard/202403/w8235209_20240321100847959.jpg&fileName=幻灯片10.jpg)

图5 多模态表格解析效果示例

块，但容易产生过长的文本块；而语义切分虽然能保证文本块的语义完整性，有利于提高 RAG 问答效果，但在处理长文本时速度较慢，会增加文本入库的时间成本。因此，将标题和语义两者结合能够弥补各自的不足，在保证文本块质量的同时，既解决了标题切分产生的长文本问题，缓解了语义切分在处理超长文本时的耗时问题。

2.2.4 答案生成与优化

用户输入提问后，为解决语义鸿沟问题，通过指代消解、省略补全、情感语气消歧三大能力，提升系统对用户意图的深度理解，同时对用户问题进行术语对齐。术语对齐是解决语义鸿沟的常规解决方案之一，其作用是将一些专有名词转变为通俗语句。对用户的输入进行语义增强，经过语义对齐后的用户提问，将提高相关文本块的召回率。

在文本总结方面，采用 COT 思维链增强的提示词工程，形成一套总结逻辑体系，可以根据召回文本块进行总结回答、推理回答、信息提取回答、图文输出、基础闲聊能力等，当用户的提问召回的知识无法回答时，会倾向于拒答，而非胡言乱语。

3 智能问答智能体应用成效

3.1 功能与能力成效

小易 AI 智能问答智能体引入 Qwen32B 大模型、多模态大模型与向量化技术进行架构升级，构建分级分域、智能萃取、实时流转的 RAG（检索增强生成）知识问答体系，以高质量知识为基础，保障内容的准确性与时效性。知识解析与切分中使用的大模型 Qwen3-32B、多模态解析使用 Qwen2.5-VL-32B，多模态表格解析效果如图 5。

图像的多模态解析将图像内容提取出来后放到 markdown 图像格式中的图像描述中。组合文本切分效果如图 6。

小易 AI 智能问答智能体问答效果图 7 图文混合回答、图 8 基础闲聊回答、图 9 知识缺失时的拒答机制触发。具备图文输出能力、知识提取回答、支持思考模式开关、基础闲聊和自我认知。其问答大模型使用的是 Qwen3-32B 模型。

为了分权分域和减少知识的召回范围采用知识分库的方式进行知隔离和筛选，减少知识基础数量回答效果，展现形式如图 10。



图6 基于标题-语义的组合切分结果可视化



图7 图文混合回答



图8 闲聊回答

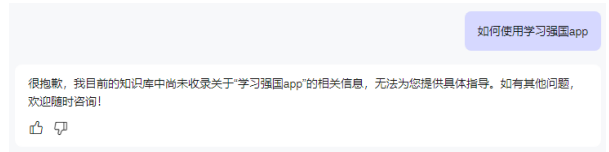


图9 知识缺失时的拒答机制触发



图10 知识库分库

3.2 业务应用成效

小易 AI 智能问答智能体覆盖企业规章制度、员工发展、部门职责、财务知识、差旅报销、党建规范、产品政策、装维知识、渠道知识等高频领域，不仅“找得全”，更能“想得对”。嵌入量子办公、OA 首页、易支撑等核心系统，平均响应时长缩短至 1.5 秒，应答准确率达 90%，覆盖超 3 万从业人员，支撑超 5000 次的日提问量，用户互动点赞率高达 97%。

4 结语

本文系统阐述了基于大模型构建检索增强生成 RAG 智能问答智能体的工程实践方法。依托 Qwen 大模型进行多模态知识萃取，通过 Markdown 格式重构与知识去重策略，从源头保障了知识质量与规范；在检索架构层面，采用 Milvus 向量数据库的混合检索与多租户分库机制，在实

现数据隔离的同时,显著提升了检索精准度与响应速度;在生成优化层面,引入术语对齐与思维链(CoT)增强的提示词工程,有效弥合了用户提问与知识内容间的语义鸿沟,并通过拒答机制从根本上规避模型幻觉,保障了回答的可靠性与可溯源性。

当前,该智能问答系统已在电信运营商多个核心业务场景成功落地并规模化应用,充分验证了本研究技术方案的可行性与实践价值。未来研究与优化工作将聚焦两大方向:其一,探索智能问答与 AI 智能体(AI Agent)的深度融合,推动系统从单纯的“知识问答”向“业务操作执行”的智能化跃迁;其二,深入研究多模态大模型与知识图谱的联合增强机制,进一步提升系统在复杂场景下的知识推理与问题解决能力。

本研究提出的混合检索架构、多模态知识解析方法及知识分域治理方案,为电信运营商及传统行业的知识管理智能化转型,提供了可复制、可复用的工程实践参考与技术路径借鉴。

参考文献:

[1] 郭晋阳,贺昌义,杨戈等. 大语言模型压缩综述[J]. 计算机科学与探索, 2026, 20(1): 1-20.

[2] 郭陆祥,王越余,李芊玥等. 大语言模型智能体操作系统研究综述[J]. 计算机科学, 2026, 53(1): 1-11.

[3] 秦小林,古徐,李弟诚等. 大语言模型综述与展望[J]. 计算机应用, 2025, 45(3): 685-696.

[4] 李瑞昌,宋学坤,朱红磊等. 生成式人工智能大模型技术探析[J]. 数字技术与应用, 2025(2): 16-18.

[5] 吴璇,付涛. 检索增强生成技术研究综述[J]. 计算机工程与应用, 2025, 61(20): 19-35.

[6] 裴炳森,李欣,胡凯茜等. 基于知识增强预训练模型的司法文本摘要生成[J]. 科学技术与工程, 2024, 24(20): 8587-8597.

[7] 张超然,梁素平,杨杰等. 基于大模型 RAG 架构的专利检索 Agent 研究[J]. 中国发明与专利, 2025, 22(2): 14-25.

[8] 李圣飞,卢昱杰,陈晓莹等. 基于检索增强生成的建筑工程管理知识问答模型实现[J]. 科学技术与工程, 2025, 25(25): 10840-10849.

作者简介:徐前进(1985.06-),男,安徽蒙城人,硕士研究生,研究方向:AI 应用、大数据分析挖掘。