DOI:10.16356/j.2097-6771.2026.01.022

面向航空标准的大语言模型迭代检索增强生成方法

秦晓瑞[1],何 柳[1],安 然[1],曾江辉[1],刘姝妍[1],王少枫[1],田 宇[2]
(1.中国航空综合技术研究所标准数据技术研究部,北京 100028;2.中国飞机强度研究所强度与结构完整性全国重点实验室,西安 710065)

摘要: 航空标准数据具有结构复杂、语义严谨和跨文档引用频繁等特点,为实现高效、精准的知识获取与问答应用带来挑战,本文提出一种面向航空标准的大语言模型(Large language models, LLMs)迭代检索增强生成(Retrieval-augmented generation, RAG)方法,设计了基于结构路径感知的标准向量知识库构建与检索机制,结合标准文档的章节结构与标题链条构建支持语境追溯的知识库,并提出基于关键词与语义融合的知识检索机制。在此基础上,设计LLM驱动的自动迭代检索与生成机制,使模型能够自主判断是否需要发起子问题拆解与深层意图识别,并结合多轮检索与动态调度策略,实现问题拆解、信息获取、自主判断与生成控制的一体化闭环,提升对多知识点聚合型、语义递进型等复杂标准问答任务的生成质量与覆盖深度。实验基于7 459份航空标准文档构建知识库,针对500条专家标注问答对,在4类涵盖不同参数规模、模型类型及中英文语言能力的主流开源大语言模型上开展对比实验。结果表明,对于中大型参数规模的大模型,此方法在回答准确性、覆盖度和表达质量等指标上均显著优于传统方法。在大模型DeepSeek-R1-70B上,双语评估替补(Bilingual evaluation understudy, BLEU)指标平均提升27.97%,模拟主观评分提升7.99%;在大模型Qwen-2.5-32B上,BLEU指标平均提升54.67%,模拟主观评分提升8.58%。本文所提方法不仅适用于航空标准场景,也可推广至适航规章、维修手册等其他航空结构化文档场景,以及法律、医疗等对回答效果、可信度与可溯源性要求极高的领域,为相关问答系统的构建提供通用的技术框架与实现路径。

关键词: 大语言模型;检索增强生成;航空标准;向量知识库;迭代机制
中图分类号: TP18 **文献标志码:** A **文章编号:** 1005-2615(2026)01-0235-14

Iterative Retrieval-Augmented Generation Method for Aviation Standards Based on Large Language Models

QIN Xiaorui[1], *HE Liu*[1], *AN Ran*[1], *ZENG Jianghui*[1], *LIU Shuyan*[1], *WANG Shaofeng*[1], *TIAN Yu*[2]
(1. Department of Standard Data Technology Research, China Aero-polytechnology Establishment, Beijing 100028, China; 2. National Key Laboratory of Strength and Structural Integrity, Aircraft Strength Research Institute of China, Xi'an 710065, China)

Abstract: Aviation standards are characterized by complex data structures, rigorous semantics, and frequent cross-document references, which pose significant challenges for achieving efficient and accurate knowledge acquisition and question-answering applications. In response, an iterative retrieval-augmented generation (RAG) method for aviation standards based on large language models (LLMs) is proposed. A structure-aware vectorized knowledge base construction and retrieval mechanism is designed, which leverages

基金项目: 中国飞机强度研究所强度与结构完整性全国重点实验室项目。
收稿日期: 2025-09-02;**修订日期:** 2025-11-03
通信作者: 曾江辉,男,研究员,E-mail:zengjh2014@sina.com。

引用格式: 秦晓瑞,何柳,安然,等. 面向航空标准的大语言模型迭代检索增强生成方法[J]. 南京航空航天大学学报(自然科学版),2026,58(1):235-248. QIN Xiaorui, HE Liu, AN Ran, et al. Iterative retrieval-augmented generation method for aviation standards based on large language models[J]. Journal of Nanjing University of Aeronautics & Astronautics(Natural Science Edition),2026, 58(1):235-248.

the hierarchical structure and title chains of standard documents to support context-traceable knowledge representation. A hybrid retrieval strategy combining keyword matching and semantic similarity is further introduced. On this basis, an LLM-driven automatic iterative retrieval and generation mechanism is developed, enabling the model to autonomously determine whether sub-question decomposition and deep intent recognition are required. By integrating multi-turn retrieval with dynamic scheduling, the proposed framework forms a closed-loop process of problem decomposition, information acquisition, self-assessment, and content generation, effectively improving generation quality and semantic coverage for complex queries such as multi-point aggregation and semantic progression. Experiments are conducted on a knowledge base constructed from 7 459 aviation standard documents and a benchmark set of 500 expert-annotated question-answer pairs, evaluated across four mainstream open-source LLMs with varying parameter scales, model types, and bilingual capabilities. The results indicate that, for models with medium or large scales, the proposed method significantly outperforms traditional retrieval-augmented generation approaches in answer accuracy, coverage, and expression quality. Specifically, on DeepSeek-R1-70B, the bilingual evaluation understudy (BLEU) score improved by 27.97% and the simulated human preference score increased by 7.99%; on Qwen-2.5-32B, BLEU improved by 54.67% and the simulated score increased by 8.58%. The proposed method is applicable not only to aviation standards, but also to other structured document scenarios in the aviation field, including airworthiness regulations and maintenance manuals, as well as to domains such as law and healthcare that require high answer quality, reliability, and traceability. It offers a versatile technical framework and implementation approach for the construction of domain-specific question-answering systems.

Key words: large language model (LLM); retrieval-augmented generation (RAG); aviation standards; vectorized knowledge base; iterative mechanism

航空标准[1]是航空工业领域内设计、制造、维修和管理过程的重要技术依据。随着航空工业数字化和智能化转型的推进,以中华人民共和国国家标准、国家军用标准、航空行业标准以及中国航空工业集团公司标准等为代表的大量航空标准文档持续发布与修订,形成了规模庞大、格式多样的数据[2]。航空标准文档包含精确而严格的章节结构、专业术语和规范要求,标准间还存在大量的跨文档关联关系,给标准数据的知识获取与有效检索带来了较大的挑战[3]。当前常见的航空标准知识获取方法,如关键词匹配、简单语义检索[4]等难以满足用户在复杂语义环境下的知识应用需求。同时,现有检索系统大多停留在检索、展示和规则化问答的阶段,难以支持用户通过自然语言实现高效柔性的人机交互。

近年来,基于大语言模型(Large language models, LLMs)[5]的检索增强生成(Retrieval-augmented generation, RAG)方法[6]为解决知识密集型任务提供了新思路。然而,现有的通用RAG框架大多采用单轮检索与单步生成的模式,未考虑用户复杂查询[7]的多步检索与精细拆解需求,难以实现针对航空标准这种结构严谨、语义精细文档的知识获取。

在航空标准问答场景中,用户查询常呈现出两类复杂性特征(图1)。一类为多知识点聚合型问题,需定位来自不同条款的多个知识点,并进行逻辑整合。例如“时码输入接口与数据实时采集接口的设计要求分别是什么?”,应将其拆分为“时码输入接口的设计要求是什么?”“数据实时采集接口的设计要求是什么?”等子问题,并行检索后整合分析。另一类为语义递进型问题,初步检索结果可能仍无法满足回答要求,应自动生成补充性子问题继续检索,实现更深入的答案生成。例如,用户询问“电源系统在主电源故障时如何处理?”,初步检索

多知识点聚合型问题
• 时码输入接口与数据实时采集接口的设计要求分别是什么? • 螺栓的密封形式和铆接的密封形式有什么不同? • 反辐射无人机发射车的俯仰调节精度指标应该是多少?应该怎么检验? • ……
语义递进型问题
• 电源系统在主电源故障时如何处理? • 标准“Q/AVIC XXXX—2018”引用的文件有哪些?给出每条引用标准的适用范围。 • 直升机蒸发循环系统蒸发器产品的封存包装应该符合什么规定?给出具体规定的信息。 • ……

图1 复杂用户查询问题示意

Fig.1 Illustration of complex user query problems

结果可能提到“可自动切换至备用电源”，但进一步应发现缺少“如何判定主电源故障？”等细节。若使用传统RAG方法，只能获取表层答案，导致回答缺乏深度与语义闭环。

此外，传统RAG方法在知识库构建方面亦存在适配性不足的问题。通用知识库构建大多使用基于固定窗口、语义或章节等切分方式，忽略了标准文档所特有的严谨章节结构与语境组织方式，导致知识片段在召回过程中上下文脱节、路径信息缺失。

为克服以上问题，本文提出一种面向航空标准的大模型迭代检索增强生成方法，主要贡献如下：(1)通过大模型驱动的主动问题拆解和迭代式知识召回机制，实现了航空标准复杂问题的精准检索和高质量语义推理生成；(2)设计了结构路径感知的知识单元拆分机制，通过显式记录知识单元的章节路径、标题链及上下文语义信息，提高了航空标准知识单元的语义独立性与检索召回精度；(3)构建了包含7 459份航空标准文档和500个专家标注问题的数据集，开展多种开源大模型的对比实验，结果表明本文方法在检索召回率、回答准确性等指标上具有显著优势，验证了其在航空标准问答任务中的实用性与优越性。

此外，本文所提方法具有良好的通用性，可推广至适航规章、维修手册等航空结构化文档场景，以及法律、医疗等知识高度结构化、对回答效果与可溯源性要求较高的复杂问答任务中，为构建面向垂直领域的大模型增强问答系统提供通用的方法框架与实践路径。

1　相关工作

随着大模型在自然语言理解生成领域的突破，面向结构化领域知识的问答系统逐步由基于关键词检索、规则化回答的模式，演进为融合外部知识调用与语言生成能力的RAG范式[8]。典型实现包括Facebook提出的原始RAG框架[6]、探索语言模型的预训练阶段引入检索机制的检索增强语言模型(Retrieval-augmented language model, REALM)[9]以及采用多段融合输入提升生成质量的Fusion-in-Decoder[10]等。

然而，现有RAG方法多数采用静态单轮检索策略，缺乏对用户复杂查询意图的精细理解与逐步建模。为解决该问题，研究者逐步探索在RAG框架中引入调度机制，实现复杂任务的分阶段处理与多轮语义闭环。例如，Toolformer[11]通过无监督训练方式，使语言模型能够判断是否需要插入接口调用，从而动态触发工具使用；Self-RAG[12]提出引导大模型自动评估是否需要检索，并对生成的响应进行自我评估。

另一方面，航空标准文档通常具有条款清晰、结构层级明确的特点，通用的滑窗切分、自然段落切分或语义聚类等RAG知识库构建方法[13-15]难以保留原有语境与结构线索。同时，现有通用语义检索方法多基于FAISS[6]等向量库构建索引，或使用Elasticsearch[16]等检索引擎构建混合召回通道，但在标准文档场景中仍存在显著局限。首先，标准条款表述相近，若使用通用向量模型，易导致语义混淆与条款错位；其次，用户查询常含结构性意图，现有方法可能难以对齐上下文信息，影响大模型对检索结果的理解与生成质量。

因此，构建具备结构路径感知能力的标准知识库，且引入大模型驱动的迭代检索机制，是解决当前标准问答系统精度不足、理解浅表等问题，并推动垂直领域大模型[17-18]系统性建设的可行路径。

2　本文方法

本文提出了一种面向航空标准的大模型迭代检索增强生成方法。该方法包括以下两个关键模块：(1)在知识建模阶段，提出结构路径感知的知识单元划分与建模机制，结合标准文档的章节结构与标题链构建支持语境追溯的知识库，并构建基于关键词与语义融合的知识检索机制；(2)在问题理解阶段，利用大语言模型，实现“提出子问题-执行检索-筛选信息-自主判断是否继续”的动态调度与闭环迭代机制，有效保障复杂查询中信息覆盖度与回答精度。

2.1　基于结构路径感知的标准向量知识库构建与检索机制

航空标准文档具有高度结构化与专业化的文本特征，其内容通常按“章-节-条-项”等层级进行组织，形成严格的编号体系及标题链条。现有大多数通用文本处理方法主要基于自然段落、固定长度窗口和内容关联性进行内容切分，忽略了文档内部的层级语义与结构路径信息，导致切分结果在后续检索与问答任务中存在语义割裂、结构缺失等问题。为此，本文方法提出一种结构路径感知的知识单元拆分机制，面向航空标准文档构建支持上下文追溯、路径聚合和语义还原的结构化知识块。通过对标准的章节路径、标题链及父级语义等信息进行显式建模，为大模型问答提供高质量知识支撑，此机制总体架构如图2所示。

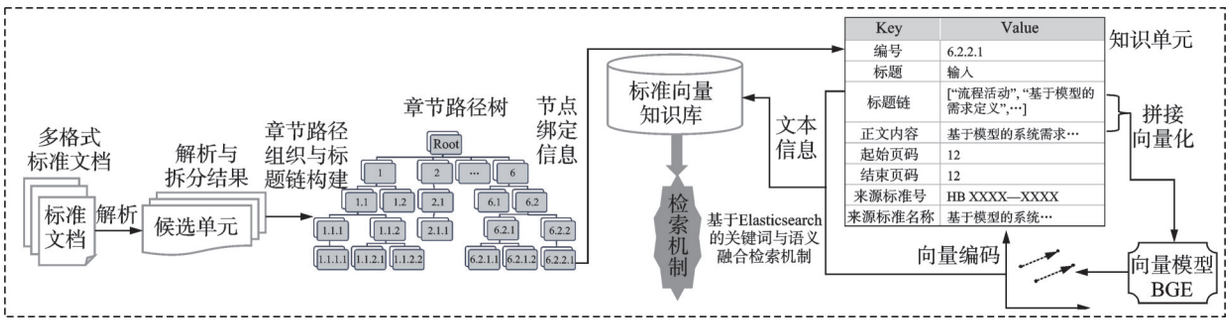

图2 基于结构路径感知的标准向量知识库构建与检索机制架构图

Fig.2 Architecture of the structure-path-aware vector knowledge base construction and retrieval mechanism

2.1.1 多格式标准文档解析与拆分机制

航空标准作为规范化管理的重要载体,其文档来源长期存在格式异构、结构复杂等问题[3]。不同单位、历史版本及发布渠道形成了多种存储形态的标准文档。首先构建一套多格式标准文档解析机制,作为后续章节识别与知识单元构建的输入基础。对于PDF与Word等主流文档格式,基于段落缩进、标题模板匹配、字体样式和页面坐标等特征提取文本,恢复逻辑结构;对于HTML、XML类文档,通过解析器提取嵌套标签及路径,保留结构信息;对于扫描型文件,引入基于深度学习的光学字符识别(Optical character recognition, OCR)模型进行文字还原[19],并采用段落合并、换行归一手段恢复语义结构。完成文档解析后,获得一系列带编号与标题的章节候选单元,作为解析与章节拆分的结果。每个章节将作为后续章节路径树的1个节点,并作为大模型检索的知识库中1个知识块单元。

2.1.2 章节路径组织与标题链构建机制

为了实现结构化知识的语义定位、路径导航与上下文建模,提出标题链构建与章节路径组织机制,将每个解析拆分后的章节嵌入到标准文档的结构体系中,形成结构连续、语义可还原的知识单元体系。标准文档中的编号体系通常体现出明确层级关系,其结构路径编号形式如“3”“3.2”“3.2.1”“3.2.1.4”,天然构成1棵文档章节路径树[20]。此层级结构反映了不同章节之间的语义关联。编号路径中的上层条目往往包含对下层内容的定义、背景或前提信息,在后续检索或生成任务中具有重要的语境参考价值。

章节路径组织通过章节路径树的构建实现。对于标准文档D,将所有具备结构编号的章节组织为1棵有序的章节路径树T_D。该路径树中的每个节点表示1个结构编号单元,每条边表示编号间的层级从属关系。设文档D中结构编号集合为$N_D=s_1, s_2, \cdots, s_n$,每个结构编号$s_i$表示为1个由若干正整数组成的编号序列

$$s_i=(l_1, l_2, \cdots, l_k) \tag{1}$$

式中:k为结构的层级深度,l_j表示该编号在第j层的位置编号。该定义的设计用于构建章节层级的形式化表示,便于在后续路径树建模与层级追踪中实现结构映射。例如,“3.2.1.4”这一章节编号序列应为(3,2,1,4)。路径树的构建过程如算法1所示。

算法1 标准文档章节路径树构建算法

输入:所有结构编号组成的集合$N_D=s_1, s_2, \cdots, s_n$,其中$s_i$为层级编号序列,如(3,2,1,4)

输出:构建完成的章节路径树T_D

(1) 初始化根节点root,令路径树$T_D \leftarrow \{\text{root}\}$

(2) for $i=1$ to n do

(3) 取结构编号$s_i=(l_1, l_2, \cdots, l_k)$

(4) 当前节点指针curr$\leftarrow$root

(5) for $j=1$ to k do

(6) prefix$\leftarrow(l_1, l_2, \cdots, l_j)$

(7) if prefix不在curr的子节点中then

(8) 创建节点node(prefix),并将其加入curr的子节点集合

(9) end if

(10) curr$\leftarrow$node(prefix)

(11) end for

(12) end for

(13) 返回路径树T_D

(14) 结束

算法1基于层次化树结构构建原理设计,定义章节路径树T_D以形式化表示章节间的隶属关系。通过逐层解析编号序列s_i,实现章节结构的层级还原与路径索引构建。在算法1中,对于每一个编号s_i,第(3)行将其解析为层级编号序列形式$(l_1, l_2, \cdots, l_k)$。第5~11行为该章节编号的每一级前缀路径构建树结构节点。第6行生成当前层级编号的路径前缀$(l_1, l_2, \cdots, l_j)$。第7~8行判断当前路径是否已在树结构中出现,若不存在则自动补全

路径缺失的层级。最终，第13行输出路径树 T_D，作为后续标题链生成、知识单元组织与检索的基础结构。

在 T_D 构建完成后，进一步引入标题链机制，记录每个节点在文档中的语义定位路径，为大模型提供结构上下文提示词、路径回溯能力。设 $s_i=(l_1,l_2,\cdots,l_k)$ 为某结构编号路径，其对应路径树中存在编号节点链 $(n_1,n_2,\cdots,n_k)$，其中 n_j 为层级 j 的路径编号，绑定有对应标题 t_j。则该结构单元的标题链可表示为

$$p_i=(t_1,t_2,\cdots,t_k) \tag{2}$$

该定义基于章节路径树的层次结构设计，以章节标题的逐级连接形式刻画文档内容的语义定位路径，从而在结构层面实现章节间的语义追溯与上下文约束。例如，当模型需要定位“6.2.2.1”章节内容时，$s_i=(6,2,2,1)$，路径树中对应编号节点链为 $(6,6.2,6.2.2,6.2.2.1)$。假设章节编号“6”“6.2”“6.2.2”“6.2.2.1”分别对应的章节标题为“流程活动”“基于模型的需求定义”“基于模型的系统需求定义”以及“输入”，则“6.2.2.1”章节的标题链可表示为：$p_i=$（流程活动，基于模型的需求定义，基于模型的系统需求定义，输入）。

同时，在路径树构建过程，以章节标题为牵引，设置节点内容绑定机制。对于每一个标准章节编号 s_i，在其插入路径树 T_D 的同时，将其对应的章节标题 t_i、标题链 p_i、正文内容 c_i、此章节起始页码 g_i、此章节结束页码 e_i、来源标准号 h_i 以及来源标准名称 m_i 作为字段附加绑定至该编号节点，即完成如下结构

$$\mathrm{Node}(s_i)=(s_i,t_i,p_i,c_i,g_i,e_i,h_i,m_i) \tag{3}$$

该结构用于对文档章节的多源信息进行统一封装，以实现结构化节点的语义可追溯、内容可定位与元信息可关联。通过该定义，可在路径树节点层面建立从编号到语义内容的多维映射，为后续的向量化表达与知识检索提供数据基础。各字段将在大模型召回阶段按照预定义的提示词格式进行拼接，用于构造结构化的上下文输入，确保模型在理解与生成过程中具备清晰的语义路径与来源标注。

2.1.3　标准向量知识库构建机制

使用2.1.2节所述机制形成标准文本知识库后，本文方法进一步对标准文本知识库中的各知识单元进行语义向量化处理，并构建标准向量知识库，以支撑后续的大模型迭代式语义检索机制。基础向量模型选用北京智源通用语义向量模型（BAAI general embedding，BGE）[15]。

为提升模型在航空领域语境下的语义对齐能力与专业表达适应性，本文基于bge-large-zh-v1.5模型开展领域化微调。微调数据由人工标注获得，数据来源于航空科研报告与标准文档，标注遵循“短问题-长文本参考知识片段”配对原则，即从语义完整的章节中，设计可由该章节完整回答的自然语言问题。主要目的是明确刻画用户查询意图与文本内容之间的语义关联关系，以此提升向量模型在标准问答任务中的适应性与准确性。共形成约9 000对样本，经复核以确保标注一致性与语义准确性。

为增强模型的判别能力，训练前对原始样本进行困难负样本挖掘，采样范围设定为[2,200]，以生成语义相近的负例；随后采用对比学习目标函数进行模型优化，学习率设为 1×10^{-5}，训练轮数为5，批次大小为2，查询与段落最大长度分别为128和512，训练组大小为11。训练后，利用模型融合方法，将基础模型与微调模型按权重0.5∶0.5进行参数融合，以兼顾通用语义能力与航空领域语义表达能力。

在构建标准向量知识库时，将标准文本知识库中每个知识单元的标题链与正文内容拼接为统一语义输入，通过微调后的向量模型生成定长语义向量，并将编码结果与其对应的所有元信息（包括编号、标题、标题链、正文内容、起始页码、结束页码、来源标准号和来源标准名称）共同存储于标准向量知识库中。标准向量知识库基于开源搜索引擎软件Elasticsearch构建，支撑语义相关性驱动下的大规模快速召回。

2.1.4　基于Elasticsearch的关键词与语义融合检索机制

为增强标准向量知识库对复杂查询任务的响应能力，本文方法在构建完成的知识库基础上，设计了一套融合关键词检索与语义检索的多通道召回机制。该机制基于Elasticsearch[16]搜索引擎框架，联合利用文本倒排索引与语义向量检索能力，实现了对标准知识中关键信息的全面、精准召回。

在文本索引方面，针对知识单元的标题、标题链、来源标准号、来源标准名称与正文内容等字段建立多级索引通道。对于各文本字段，采用基于分词的关键词索引方法。在语义检索方面，引入高维向量检索机制。用户输入的问题首先被微调后的BGE模型编码为查询向量，与知识库中所有向量计算语义相似度。为提升综合排序质量，融合向量相似度得分与文本字段匹配得分，构建加权打分函数，设计为

$$\text{score}_{\text{final}}=\frac{1+\cos(\boldsymbol{q},\boldsymbol{d})}{2}+\lambda\cdot\text{sigmoid}(\text{score}_{\text{word}},a,b) \tag{4}$$

该函数由语义相似度部分、关键词匹配两部分组成,λ为关键词部分的权重。其中,排序函数整体为本文设计,用于适配航空标准文本中语义一致性与关键词精确匹配并存的检索需求。语义相似度部分采用余弦相似度[16]进行建模,并归一化至区间[0,1],其中$\boldsymbol{q}$表示查询问题的语义向量,$\boldsymbol{d}$表示知识单元的语义向量,其定义如下

$$\cos(\boldsymbol{q},\boldsymbol{d})=\frac{\boldsymbol{q}\cdot\boldsymbol{d}}{\|\boldsymbol{q}\|\cdot\|\boldsymbol{d}\|} \tag{5}$$

关键词匹配部分引入参数化sigmoid函数[14]对其得分进行平滑压缩,具体表达式为

$$\text{sigmoid}(\text{score}_{\text{word}},a,b)=\frac{1}{1+\mathrm{e}^{-a(\text{score}_{\text{word}}-b)}} \tag{6}$$

式中:参数a控制函数斜率,b控制激活拐点位置,$\text{score}_{\text{word}}$表示基于关键词索引机制所获得的文本相关性得分,由知识单元中多个结构化字段(如标题、正文、标准号等)在倒排检索中累积计算得到

$$\text{score}_{\text{word}}=\sum_{i=1}^{n}\beta_i\cdot\text{BM25}(\boldsymbol{q},d_i) \tag{7}$$

式中:$\text{BM25}(q,d_i)$为最佳匹配25算法(Best matching 25, BM25)[21]在第i个文本字段内容d_i上对查询$\boldsymbol{q}$的匹配得分,β_i表示各字段的权重系数。本文方法采用$a=5,b=4,\lambda=1.5$的配置,在多个检索任务中表现出较强的排序稳定性与召回质量。

2.2 大模型驱动的自动迭代检索与生成机制

在复杂的航空标准问答任务中,用户问题往往具有复合性与层次性,可能同时涉及多个标准或条款的知识点,也可能需要依赖初步回答进一步挖掘。为此,本文方法构建了大模型驱动的自动迭代检索与生成机制,依托大模型的理解能力开展子问题拆解与深层意图识别,结合多轮检索与动态调度策略,形成自动迭代式[22]问答流程,提升系统在复杂任务中的语义覆盖、逻辑一致性与生成质量。总体架构如图3所示。虚线表示大模型语义调用过程。

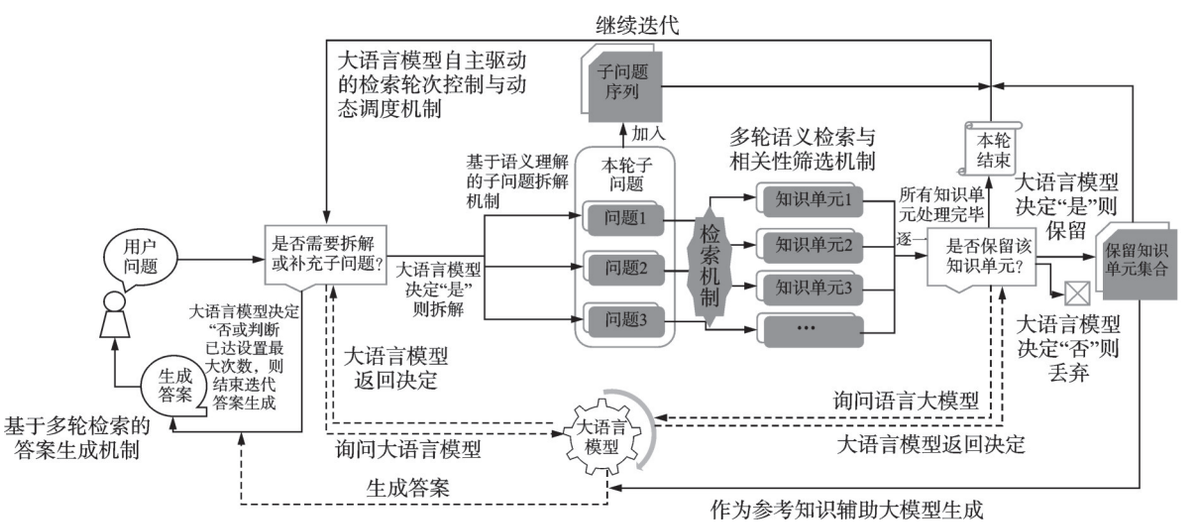

图3 大语言模型驱动的自动迭代检索与生成机制架构图

Fig.3 Architecture of the LLM-driven automatic iterative retrieval and generation mechanism

2.2.1 基于语义理解的子问题拆解机制

在面对涉及多个知识点或逻辑层级较深的航空标准问题时,传统RAG方法难以精准对齐所有潜在意图,容易出现知识覆盖不全或答案逻辑较浅等问题。为此,本文方法设计了一种基于提示工程的大模型子问题拆解机制,通过提示指令引导大模型自动将原始复杂问题拆解为结构清晰、语义聚焦的若干子问题。

该机制采用少量示例驱动的提示模板,引导大模型将原始查询拆分为最多3个子问题,以标准化的列表结构返回。提示内容中提供明确的拆解目标、输出格式要求以及典型航空标准问题的示例输入与拆分结果。对于语义结构简单、不需要进一步分解的问题,支持保留原始问题作为唯一子问题,保障通用性与鲁棒性。生成结果在结构上表现为具有独立表达能力的子问题序列,每个子问题均可单独触发检索流程,并具备原始查询的上下文语义与限定条件。

2.2.2 多轮语义检索与相关性筛选机制

完成子问题拆解后,本文方法针对每一个子问题分别自动执行语义检索流程。对于每个子问题,调用2.1.4节所述基于Elasticsearch的关键词与语义融合检索接口,从标准向量知识库中筛选,形成候选列表。随后,引入基于大模型的相关性判别机

制，对每条候选知识单元进行逐一判断，以决定是否保留用于后续辅助大模型生成结果。设用户查询为x，候选单元为z_k，则是否保留该段落由语言模型判别函数$\mathcal{L}$给出

$$\delta_k = \mathcal{L}(x, z_k) \tag{8}$$

式中：$\delta_k \in \{0,1\}$表示知识单元z_k是否被判定为与问题x存在直接知识支撑关系。仅当$\delta_k = 1$时，该知识单元才会被纳入最终用于生成的知识单元集合。该判别函数为本文提出的自定义二分类策略，使用大模型的理解能力实现动态判别。若Z表示候选知识单元集合，则最终使用的知识单元集合Z'表示为

$$Z' = \{z_k | \delta_k = 1, z_k \in Z\} \tag{9}$$

Z'表示通过本文判别机制筛选后的有效知识集合，其构成了后续生成阶段的核心输入。该判别过程采用严格控制策略，避免无效冗余段落对生成环节产生干扰。判别标准不仅考虑知识单元与问题之间的语义关联性，还关注是否包含支撑问题回答的高价值内容。

2.2.3　大模型自主驱动的检索轮次控制与动态调度机制

本文方法首先通过2.2.1节中的大模型提示引导机制，将原始复杂问题拆解为结构清晰、语义聚焦的子问题序列；随后在2.2.2节中，依次对每个子问题执行语义检索，并利用大模型判断候选段落的实际支撑性。然而，首轮检索结果往往无法完全覆盖用户查询所需的全部知识点，尤其在面对涉及多跳逻辑或语义隐含关系的问题时，仅依靠初轮子问题生成与静态检索可能导致答案不完整。为此，进一步提出大模型自主反思机制，每轮检索结束后综合分析当前获取的信息，判断是否仍存在内容空缺或逻辑跳跃，动态生成新的补充性子问题，驱动系统进入下一轮检索，这也是本文方法提出“迭代”检索增强生成的内涵所在。

假设已完成第t轮次迭代，判断是否进入第$t+1$轮次。完成第t轮的所有子问题检索与筛选后，将原始问题x、已处理的子问题序列$\{x_1, x_2, \cdots, x_m\}$以及第$t$轮保留知识单元集合$Z'^{(t)}$输入至大模型中，触发反思式判断过程。大模型分析已获信息与问题意图之间的语义差距，若认为当前信息存在缺口，自动给出新的补充性子问题集合，作为下一轮检索的输入；否则迭代终止，进入大模型回答阶段。该判定过程可形式化为

$$X^{(t+1)} = \Psi(x, \{x_1, x_2, \cdots, x_m\}, Z'^{(t)}) \tag{10}$$

式中：$X^{(t+1)}$表示由大模型生成的第$t+1$轮新子问题集合；$\Psi(\cdot)$为本文设计并定义的大模型反思函数。若$X^{(t+1)} \neq \varnothing$，则进入下一轮迭代的子问题处理流程；若$X^{(t+1)} = \varnothing$，或达到设定的最大轮次数即$t = T_{\max}$，则终止迭代，进入答案生成阶段。为避免过度搜索与语义漂移，在提示词设计中约束新的子问题生成范围必须服务于原始问题。同时，通过限制每轮新增子问题数量与总轮次数，实现对迭代路径的精细控制。

2.2.4　基于多轮检索的答案生成机制

当大模型判断自主迭代终止时，进入答案生成阶段，对保留的所有高相关知识片段进行聚合分析，结合原始查询，生成回答结果。将原始查询、拆解后的子问题序列以及所有经过筛选保留的参考知识单元共同注入大模型，作为其构建答案的输入语境。假设最终迭代n轮次，整个答案生成流程可表示为对原始查询x、子问题序列$\{x_1, x_2, \cdots, x_n\}$与对应参考片段集合$Z'^{(n)}$的融合处理，有

$$\text{Answer} = \text{Generate}(x, \{x_1, x_2, \cdots, x_n\}, Z'^{(n)}) \tag{11}$$

式中：Answer为输出的答案文本，Generate$(\cdot)$为本文设计的基于提示工程与语义约束融合的生成函数，用于综合原问题、各轮子问题及其对应知识片段，在保证逻辑一致性与语义完整性的前提下生成回答。

3　实验与分析

3.1　实验设置

为系统验证本文所提方法的有效性与适应性，设计并开展一系列基准对比实验，旨在量化分析本文方法在提升答案准确性、知识召回能力与语义生成质量方面的综合表现。

3.1.1　数据集与知识库构建

选取7 459份航空标准文档，包含4 552份中国航空工业集团公司标准Q/AVIC、2 907份航空行业标准HB，具有典型的结构复杂与语境严谨特征。数据集规模情况如表1所示。采用2.1节所述机制构建基于结构路径感知的标准向量知识库，共形成390 061个标准知识单元，覆盖全部7 459份文档。为对比验证本文方法所提知识库构建机制的有效性，同时构建基于固定窗口切分的对照组，以500个字符为窗口进行滑动切分，不考虑文档结构与语境划分，形成178 432个知识单元，并采用相同的向量模型进行编码。

在问答数据集方面，本文构建了一个由航空标准领域专家标注的数据集，共包含500个真实问题及其参考答案。所有问题均来源于上述7 459份航空标准文档，确保每个问题在知识库中都存在明确的答案依据，用于评估系统对真实标准知识的召回

与应用能力。不同于开放式问答任务,本研究强调“参考知识支撑回答”的场景,要求生成答案时充分依赖已知标准片段,而非凭空发挥。每条样本均由专家标注其参考答案所依据的具体知识段落列表,作为检索覆盖率的评估依据。数据集主要包含3类问题:单一来源问题,仅需引用一个标准条款可回答;多来源型问题,涉及跨章节或文档的知识聚合;深度型问题,初步检索结果不足以回答问题,需要分多轮逐步深入补充信息。

表1 数据集规模统计

Table 1 Statistics of dataset scale

标准类型	标准数量	标注问题数量
航空行业标准HB	2 907	265
中国航空工业集团公司标准Q/AVIC	4 552	235
总计	7 459	500

3.1.2 实验对比方法设置

本文方法的核心创新主要体现在2个模块:模块1为基于结构路径感知的标准向量知识库构建与检索机制,有效提升了标准知识的语义表达质量与结构可追溯性;模块2为大模型驱动的自动迭代检索与生成机制,增强了复杂问答任务中的知识覆盖能力与语义生成质量。为系统评估该方法的整体性能及上述2个关键模块的贡献,本文设计了5组对比实验,如表2所示。

方法1(本文方法完整方案):同时使用基于结构路径感知的标准向量知识库(模块1)和大模型驱动的自动子问题拆解与多轮迭代检索生成机制(模块2),为本文提出的完整的方法路径。

方法2(替换模块2):保留使用基于结构路径感知的标准向量知识库(模块1),模块2替换为传统RAG方法,即直接以原始问题进行一次性语义检索并辅助大模型生成答案,用于验证模块2的贡献。

方法3(替换模块1):保留使用大模型驱动的自动子问题拆解与多轮迭代检索生成机制(模块2),模块1替换为基于固定窗口切分的对照组,用于验证模块1的贡献。

方法4(同时替换模块1与模块2):同时替换使用基于固定窗口切分的对照组和传统RAG方法。此方法将本文方法完整方案中的模块1、模块2皆替换为传统方法,用于验证模块1、模块2的整体贡献。

方法5(无检索对照组):不使用任何知识库与召回机制,直接将问题输入大模型生成答案,完全依赖模型本身的知识完成问答,作为基础对照组。

表2 实验对比方法设置

Table 2 Experimental settings of comparative methods

方法	知识库类型	检索方式	模块1	模块2	方法说明
1	路径感知向量知识库	自动迭代检索机制	√	√	完整方案
2	路径感知向量知识库	传统RAG	√		验证模块2的贡献
3	滑窗对照知识库	自动迭代检索机制		√	验证模块1的贡献
4	滑窗对照知识库	传统RAG			验证模块1、2的整体贡献
5	无	无			基础对照组

3.1.3 模型配置与评估指标

为系统评估本文方法在不同大语言模型上的适配效果与稳定性表现,本文选取4种主流大语言模型作为问答生成引擎,具体包括:DeepSeek-R1-70B[23]、Qwen-2.5-32B(通义千问)[24]、GLM-4-9B(智谱GLM)[25]以及LLaMA-2-7B[26]。所选模型覆盖从7B至70B的典型参数规模,既包含在中文理解与生成方面表现优异的国产模型,也包括通用性能强的国际模型,验证此方法在不同模型能力边界下的适应性。

上述模型均在本地服务器环境中部署运行,采用统一的输入格式、提示模板与生成配置。由于LLaMA-2-7B对中文任务的原生适配能力存在差异,在其实验中额外添加了“请使用中文回答以下问题”等指令性提示。其余模型均使用完全一致的提示结构与问题输入形式。此外,考虑到航空标准领域的问答任务通常要求回答内容精准、简洁且不发散,避免生成与问题无关的冗余描述,提示模板中进一步加入了“请简明回答,不要过度延展内容”等约束性指令,以引导模型输出更符合领域专家的实际回答习惯与业务需求。

在评估体系方面,本文结合自动化评估指标与模拟主观评分指标,从多个维度对生成结果进行全面分析。自动化指标方面,选取如下3类主流生成质量指标:BLEU[27],用于衡量词级n-gram匹配精度,本文采用BLEU-1与BLEU-2;面向召回的自动摘要评价指标(Recall-oriented understudy for gisting evaluation, ROUGE)[28],衡量生成文本与参考答案之间的重叠程度,选取ROUGE-1-F、ROUGE-2-F和ROUGE-L-F作为主评估指标;基于BERT的语义相似度指标BERTScore[29],用于

计算生成结果与参考答案在深层语义表示上的相似度，采用BERTScore-F值进行展示。

主观评分方面，采用通用大模型GPT-4o[30]构建结构化模拟评分机制，对模型生成答案从5个维度进行10分制评价，包括准确性、覆盖度、条理性、表达质量和专业性，并取平均值作为主观评分得分。评分过程基于统一设计的提示词模板执行，以确保评分的可复现性与一致性。

3.2　实验结果与分析

3.2.1　对比实验结果

为全面评估本文方法的整体性能与各模块贡献，本文在4种主流大语言模型上分别进行了5组对比实验。实验结果如表3～6所示，每个评分指标下表现最优的结果已用加粗体标注。

表3　DeepSeek-R1-70B实验结果

Table 3　Experimental results of DeepSeek-R1-70B

方法	BLEU-1	BLEU-2	ROUGE-1-F	ROUGE-2-F	ROUGE-L-F	BERTScore-F	主观评分
1(本文方法)	**0.378 4**	**0.290 1**	**0.492 0**	**0.257 9**	**0.379 5**	**0.768 4**	**8.51**
2(替换模块2)	0.335 2	0.247 9	0.452 8	0.210 2	0.321 0	0.754 4	8.16
3(替换模块1)	0.331 6	0.238 9	0.403 7	0.179 4	0.299 9	0.734 2	7.50
4(替换模块1、2)	0.304 8	0.220 1	0.400 7	0.174 0	0.286 2	0.735 2	7.88
5(无检索对照组)	0.201 4	0.127 3	0.240 7	0.060 3	0.170 6	0.679 4	6.66

表4　Qwen-2.5-32B实验结果

Table 4　Experimental results of Qwen-2. 5-32B

方法	BLEU-1	BLEU-2	ROUGE-1-F	ROUGE-2-F	ROUGE-L-F	BERTScore-F	主观评分
1(本文方法)	**0.426 2**	**0.339 9**	**0.587 8**	**0.347 5**	**0.469 5**	**0.800 3**	**8.47**
2(替换模块2)	0.359 4	0.273 5	0.490 1	0.243 5	0.344 4	0.768 2	8.23
3(替换模块1)	0.375 5	0.291 0	0.509 8	0.283 3	0.387 5	0.768 5	7.50
4(替换模块1、2)	0.288 5	0.210 3	0.394 2	0.169 4	0.263 9	0.732 6	7.80
5(无检索对照组)	0.205 2	0.127 4	0.262 9	0.067 8	0.170 1	0.685 7	6.66

表5　GLM-4-9B实验结果

Table 5　Experimental results of GLM-4-9B

方法	BLEU-1	BLEU-2	ROUGE-1-F	ROUGE-2-F	ROUGE-L-F	BERTScore-F	主观评分
1(本文方法)	0.358 6	0.284 3	**0.542 7**	**0.323 0**	**0.427 3**	0.772 0	7.95
2(替换模块2)	**0.423 1**	**0.327 2**	0.540 2	0.291 4	0.409 6	**0.783 3**	**8.22**
3(替换模块1)	0.320 7	0.247 4	0.459 4	0.246 4	0.348 5	0.737 8	7.08
4(替换模块1、2)	0.353 8	0.259 6	0.441 0	0.212 1	0.323 7	0.742 0	7.82
5(无检索对照组)	0.196 0	0.111 7	0.231 8	0.054 7	0.153 0	0.666 2	6.77

表6　LLaMA-2-7B实验结果

Table 6　Experimental results of LLaMA-2-7B

方法	BLEU-1	BLEU-2	ROUGE-1-F	ROUGE-2-F	ROUGE-L-F	BERTScore-F	主观评分
1(本文方法)	0.084 5	0.034 7	0.050 5	0.015 7	0.036 1	0.514 5	4.43
2(替换模块2)	**0.181 4**	**0.119 0**	**0.254 2**	**0.112 3**	**0.161 9**	**0.627 4**	**6.66**
3(替换模块1)	0.084 0	0.036 3	0.051 2	0.013 6	0.036 7	0.515 9	4.29
4(替换模块1、2)	0.115 1	0.064 0	0.146 2	0.048 1	0.088 8	0.576 9	5.50
5(无检索对照组)	0.078 8	0.038 6	0.082 0	0.020 8	0.049 2	0.537 8	4.99

同时，为全面评估不同大模型在各方法下的表现，进一步在5种实验方法上，对4种大模型在各项评估指标上进行了横向对比分析，结果如图4所示。

综合分析结果如下：

(1)本文方法整体优势显著，完整方案在多数模型中表现最优。在DeepSeek-R1-70B与Qwen-2.5-32B上，方法1在所有指标上均取得最优成绩，显著优于其他对照方法。例如，相较于传统方法(方法4)，DeepSeek-R1-70B使用本文方法后BLEU平均提升27.97%，主观评分提升7.99%；Qwen-2.5-32B使用后BLEU平均提升54.67%，主

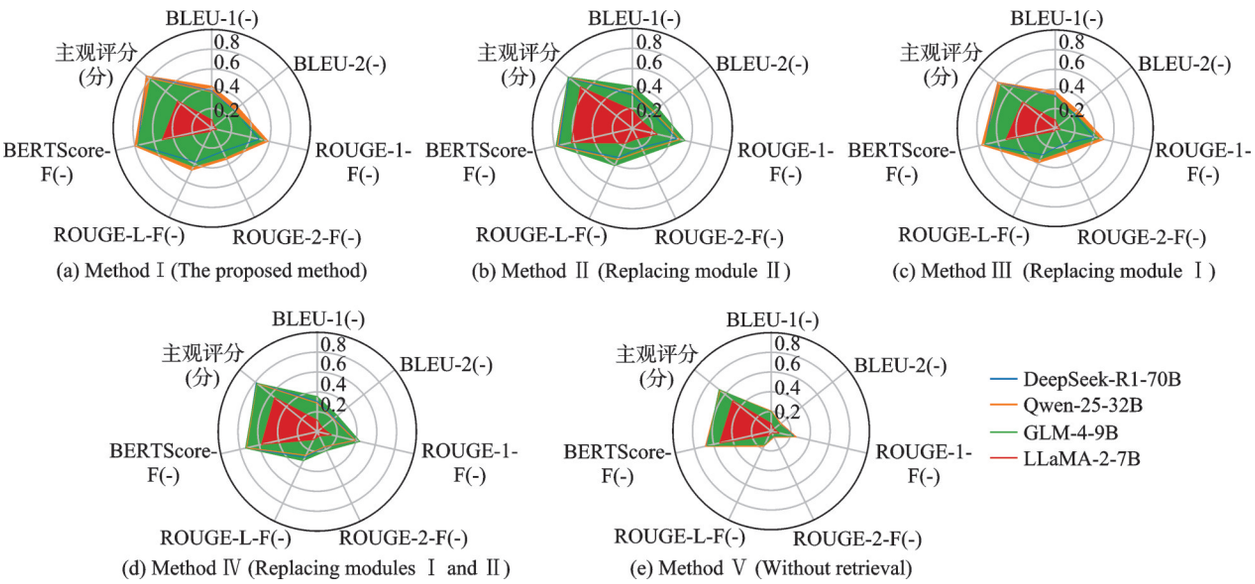

图4 不同方法下大语言模型性能综合对比图

Fig.4 Comprehensive performance comparison of LLMs under different methods

观评分提升8.58%;GLM-4-9B使用后BLEU平均提升5.43%。这表明在具备强大语义理解与推理能力的大模型支撑下,本文方法能够显著提升复杂问答的质量与准确性。

(2)模块1更具通用性,模块2依赖模型能力。在多个模型中,方法2效果优于方法3。特别是在参数量较小的模型(GLM-4-9B与LLaMA-2-7B)上,这一趋势更为显著。这可能显示出,模块1提供的路径感知向量知识库构建方法具备较强的迁移性与稳定性,而模块2自动迭代检索机制的效果发挥则更加依赖大模型本身的能力边界,适用于高能力模型精细任务调优。

(3)小模型虽不适合本文方法的复杂生成路径,但仍可采用模块1提升效果。在GLM-4-9B的实验中,本文方法并未取得绝对最优,这一点在非中文原生模型LLaMA-2-7B上表现更加明显。这表明对于小规模模型而言,本文方法中的拆解与迭代机制可能带来额外推理负担,增加生成逻辑复杂度,削弱回答的条理性与准确性。因此,本文方法可能更适用具备中大型参数规模的大模型,在资源受限的场景下,应对其完整流程的适用性予以评估与适度裁剪。尽管如此,方法2在各类模型中始终展现出稳定显著的性能提升作用,即便在小模型配置下,保留模块1的方案仍优于完全使用传统方法的基线,这进一步印证了本文方法提出的高质量结构化知识库构建方法对于语言模型问答能力的普遍增益,具有良好的通用性与迁移能力。

(4)无检索对照组(方法5)表现始终最差。这一结果说明标准问答任务对知识支撑的刚性需求,大模型自身知识不足以覆盖行业语料,检索增强生成机制是复杂问答系统构建的保障。

(5)Qwen-2.5-32B在此场景中总体表现优于DeepSeek-R1-70B,可能对本文方法思路具有更高适配性。尽管DeepSeek-R1-70B参数量更大,但Qwen-2.5-32B在方法1、方法2、方法3等多个任务设置中表现接近甚至更优,表明模型参数与本文方法适配性之间存在非线性关系。Qwen-2.5-32B擅长中文结构问答,具备“命中参考-快速重述”的风格;而DeepSeek-R1-70B虽具有更大规模与英文场景泛化能力,但在处理高结构性中文任务时仍存在轻微语义漂移与输出波动,易出现答非所问或偏离参考的现象,在此评分机制下可能不占优势。

(6)GLM-4-9B在方法4、方法2上表现突出,某些指标甚至优于参数量更大的DeepSeek-R1-70B和Qwen-2.5-32B,这可能表示GLM-4-9B风格偏向于强依赖知识注入,更注重参考所提供的材料,在本文设置的问答数据集与评测方法上具有一定优势。

3.2.2 案例分析

为直观验证本文方法在复杂问答任务中的表现,进行典型案例分析。选取两个案例,分别代表多知识点聚合型问题与语义递进型问题。基于Qwen-2.5-32B模型,保持3.1.2节设计的5种实验设置,生成对比结果,分析不同实验设置对模型回答完整性、条款可追溯性及语义连贯性的影响。

(1)多知识点聚合型问答案例。选择问题为“时码输入接口与数据实时采集接口的设计要求分别是什么?”。对比实验结果如表7所示。

(2)语义递进型问答案例。选择问题为“电源系统在主电源故障时如何处理?”。对比实验结果如表8所示。

表 7　多知识点聚合型问答案例分析

Table 7　Case analysis of multi-knowledge-point aggregated question answering

实验方法	模型生成结果	特征说明
方法 1 （本文方法）	根据 Q_AVIC XXXXX-XXXX 第 8.3.3 和 8.3.4 节，时码输入接口的设计要求是为光纤通道（Fibre channel，FC）数据提供时间戳，可选择遵循 XXXX-X 或 XXXX-XXXX 标准。数据实时采集接口的设计要求是考虑接口带宽和实时性要求，可采用千兆以太网接口	同时召回多条知识，信息完整，条理清晰，且信息可追溯
方法 2 （替换模块 2）	针对您的提问，《XXXXX》标准中提供了关于“数据实时采集接口”的相关设计要求，指出其设计应考虑接口带宽和实时性需求，并建议可采用千兆以太网接口。然而，在提供的信息范围内并无明确针对“时码输入接口”具体的设计要求的直接描述。如需了解时码输入接口的具体标准细节，可能需要参考其他相关的航空标准化文件或补充相关信息	无迭代机制导致召回范围不足。仅覆盖局部知识点
方法 3 （替换模块 1）	时码输入接口应遵循 XXXX-X 或 XXXX-XXXX 标准，为 FC 数据提供时间戳。数据实时采集接口设计应考虑接口带宽和实时性要求，可采用千兆以太网接口。	缺乏结构路径支撑，可追溯性有限
方法 4 （替换模块 1、2）	根据《XXXXX》，关于时码输入接口，它应为 FC 数据提供时间戳，并可选择遵循 XXXX-X 或 XXXX-XXXX 标准。对于数据实时采集的具体接口设计要求没有直接提及，但可以推测，在设计中需要确保该接口具备能够接收和处理实时数据流的能力，具体实现细节可能需参考相关的技术文档或进一步咨询标准制定者来明确	信息零散，逻辑弱。含推测性描述
方法 5 （无检索对照组）	时码输入接口主要设计要求是确保时间信号的精确传递，通常需要支持国际标准的时间格式和协议，并保证低延迟和高精度；而数据实时采集接口则着重于高效准确地收集各种传感器或系统产生的数据流，其设计需考虑高速传输、数据完整性和同步性等方面的标准。两者都必须满足相关行业的安全与性能标准	缺乏领域知识支撑。属于通用性描述

表 8　语义递进型问答案例分析

Table 8　Case analysis of semantic-progressive question answering

实验方法	模型生成结果	特征说明
方法 1 （本文方法）	当飞机的主电源发生故障时，应能够自动地切换至备用电源继续工作（参照 Q_AVIC XXXXX-XXXX 要求）。备用电源可以是来自飞机自身的其他可用电源或者是专用蓄电池。这样的设计确保了在任何单个主电源失效的情况下系统依然能够正常运行，且不造成危险，也不会损害其他主电源向重要负载设备的供电能力（参照 HB XXXX-XXXX（XXXX）的要求）。备用电源应具备足够容量以支持应急系统的启动和工作，并确保在飞机安全着陆前提供所需电力，同时在飞行中能够检查其状态且不应因为这样的检查导致其电能降低。此外，电源的连接设计还需防止异常放电现象影响到整个系统的可靠性（参照 HB_Z XXX-XXXX（XXXX）中的规定）。总之，备用电源系统的设计和管理确保了当主电源出现故障时，关键设备仍可以持续获得电力支持	回答完整、逻辑层次清晰，条款引用准确
方法 2 （替换模块 2）	在民用飞机电气系统的飞行试验中，当发生主电源（包括 APU 发电机）故障时，会进行一系列的验证和考核来确保系统的可靠性和安全性。这主要包括验证通道转换功能、测试稳态与瞬态供电特性、评估不同气象条件下各飞行阶段以及工作构型下的系统表现等。因此，在主电源故障时，飞机电气系统会通过预设的机制进行自动或手动切换至备用电源，并确保所有关键操作能够继续正常执行以保障飞行安全	无法整合逻辑，缺乏语义递进性
方法 3 （替换模块 1）	可以采用多个辅助动力装置或在其上装设多个发电机作为备选方案；在非正常工作状态下，即供电系统在规定的电源故障模式下运行时，负载控制器应自动切断非重要负载。如果可用的话，飞行中可通过由辅助动力装置驱动的发电机增强应急供电系统的能力。此回答基于 HB_Z XXX-XXXX（XXXX）标准中的信息给出，并不包含更多未提及的技术细节或步骤	生成某些关键描述，但缺乏条款引用与上下文衔接，语义链条中断
方法 4 （替换模块 1、2）	材料并未直接描述当主电源发生故障时的具体处理措施，但是可以参考标准中关于故障检测和指示的要求。当出现某个电源故障时，比如输出欠压（A）或其他内部故障（E），需要有相应的故障信号来提示这一情况，以便采取进一步的维护或应急操作。如果电源是并联运行，则还需要增加一个故障信号来识别哪个具体电源出现了问题。标准中并没有详细说明主电源故障后备用或冗余系统的直接处理步骤。在实际应用中，一般会有冗余设计或者备用电源系统，当检测到主电源出现故障时，系统会自动切换至备用电源继续供电，以确保机载电子设备能够正常工作不受影响。具体的切换逻辑和措施应根据具体机型的设计规范来确定	知识分散且缺乏逻辑组织，输出存在主观推测与语义跳跃
方法 5 （无检索对照组）	电源系统的冗余设计确保了在主电源故障时有备用电源立即接管供电任务，常见的备用方案包括使用不间断电源（UPS）、电池组或辅助发电机等，以保障飞行安全和关键系统的正常运行	依赖模型内部知识生成，缺乏标准支撑和技术细节

实验表明,本文方法表现出结构化理解与语义推理优势。多知识点聚合型任务中,方法1实现了跨条款的多源知识联合召回与语义整合,生成内容充分、层次分明和条款可追溯的回答;语义递进型任务中,该方法能够沿任务语义链展开推理,逻辑清晰,条款引用准确。相比之下,方法2在知识覆盖与逻辑递进方面均存在不足;方法3虽能生成主要内容,但缺乏条款关联与上下文衔接;方法4的输出内容零散,语义连贯性较弱;方法5主要依赖模型内部知识,缺乏领域针对性与标准支撑。

3.2.3 参考知识命中率分析

为进一步评估本文方法在检索和生成联动过程中的知识召回效果,本文提出参考知识命中率指标,用于衡量各模型在完整的方法流程下,最终生成答案时所依赖的检索内容是否覆盖了专家标注的所有参考依据。如3.1.1节所述,本文使用的问答数据集由航空标准领域专家标注而成,共包含500条真实问题及其对应参考答案。标注过程中,每条问答样本除包含专家答案外,专家同时记录了支撑该答案的参考知识片段列表,即形成该答案所依赖的一个或多个标准章节知识单元。对于每个问题,计算本文方法最终参考的知识单元列表与专家参考列表的重合程度,定义为该问题的参考知识命中率

$$h_i=\frac{n_{\text{hit}}^{(i)}}{n_{\text{ref}}^{(i)}} \tag{12}$$

式中:$n_{\text{ref}}^{(i)}$表示第i个问题对应的专家参考片段数量;$n_{\text{hit}}^{(i)}$为本文方法最终选择参考的片段中成功命中的数量。其中,当本文方法参考的片段在内容上完全匹配或语义相似度超过0.90时,视为命中。总体命中率为所有问题命中率的平均值:$H=\frac{1}{N}\sum_{i=1}^{N}h_i$,其中$N$为问题总数。参考知识命中率在各模型中的统计结果如图5所示。

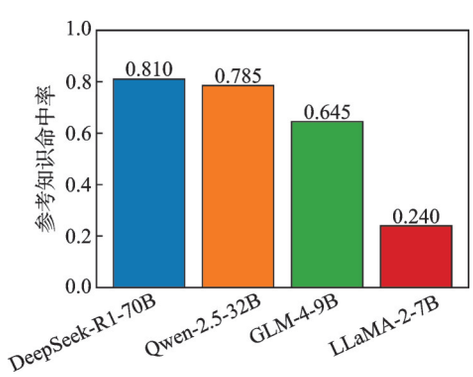

图5 参考命中率实验结果

Fig.5 Experimental results of reference hit rate

DeepSeek-R1-70B与Qwen-2.5-32B均在参考知识命中率上取得较高分值,说明它们在执行本文方法中多轮子问题拆解与检索时,能够较好地命中专家标注的重要参考段落。特别是Qwen-2.5-32B,在参数量明显低于DeepSeek的情况下,仍能实现接近水平的召回覆盖,展现出其在中文任务下对结构化检索结果的较强适配能力。相比之下,LLaMA-2-7B的参考命中率最低,反映出小模型、非中文原生模型在多轮迭代机制下的指令解析能力不足,难以准确关联检索结果与生成目标,导致知识引用片段与任务核心脱节。

4 结 论

本文面向航空标准的智能问答场景,提出了一种面向航空标准的大模型迭代检索增强生成方法,通过引入基于结构路径感知的标准向量知识库构建和检索机制与大模型驱动的自动迭代检索与生成机制,优化了标准知识单元的结构可追溯性与语义独立性,增强了检索召回的精度与稳定性,实现了航空标准复杂问题的精准检索和高质量语义推理生成,旨在解决多知识点聚合型问题、语义递进型问题等典型挑战。本文构建了包含7 459份航空标准文档的数据集与500条专家标注问答对,设计5种对比实验,在DeepSeek-R1-70B、Qwen-2.5-32B、GLM-4-9B与LLaMA-2-7B这4类主流大语言模型上全面评估本文方法的性能表现。实验结果表明,该方法在多个生成质量指标上相较传统方案取得显著提升,特别是在中高参数模型上展现出最优性能。该方法具有较强的通用性,可推广至结构严谨、语境明确和回答容错率低的其他领域应用场景,具备较强的跨任务迁移潜力,为复杂技术文档下的智能问答提供了新的实现思路。

为更全面地评估所提方法的适用性,进一步分析其局限性与改进方向。一方面,迭代检索与生成机制在提升回答深度与质量的同时带来了额外的推理开销,使系统整体响应时间受限于迭代次数与模型规模;另一方面,当采用中小参数模型时,模块2的语义生成能力相对不足,易出现语义偏移等问题。针对上述不足,后续研究将重点探索动态迭代控制策略与模型轻量化方案,通过自适应地平衡检索深度与推理成本,提升本文方法在复杂场景下的效率与工程可用性。

参考文献

[1] 何瑞恒,郑朔昉,王旭峰,等. 航空装备标准/规范研制的溯源定义方法研究[J]. 标准科学, 2023(7): 31-35.

HE Ruiheng, ZHENG Shuofang, WANG Xufeng, et al. Research on the traceability definition method for the development of aviation equipment standards/specifications[J]. Standard Science, 2023(7): 31-35.

[2] 李翔宇，傅田，潘鑫，等. 标准数字化在航空行业应用探索与实践[J]. 信息技术与标准化，2022(10)：68-72.
LI Xiangyu, FU Tian, PAN Xin, et al. Exploration and practice of standards digitalization in aviation industry[J]. Information Technology & Standardization, 2022(10): 68-72.

[3] 曹平，吴超，潘鑫，等. 航空装备标准数字化应用实践与展望[J]. 信息技术与标准化，2024(8)：56-61.
CAO Ping, WU Chao, PAN Xin, et al. Application and prospect of standard digitization of aviation equipment[J]. Information Technology & Standardization, 2024(8):56-61.

[4] HUANG J T, SHARMA A, SUN S, et al. Embedding-based retrieval in facebook search[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: Association for Computing Machinery, 2020: 2553-2561.

[5] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[6] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances In Neural Information Processing Systems, 2020, 33: 9459-9474.

[7] CRESWELL A, SHANAHAN M, HIGGINS I. Selection-inference: Exploiting large language models for interpretable logical reasoning[EB/OL].(2022-05-19). https://arxiv.org/abs/2205.09712.

[8] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: A survey[EB/OL]. (2023-12-18). https://arxiv. org/abs/2312.10997.

[9] GUU K, LEE K, TUNG Z, et al. REALM: Retrieval-augmented language model pre-training [C]//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: Journal of Machine Learning Research, 2020: 3929-3938.

[10] IZACARD G, GRAVE E. Leveraging passage retrieval with generative models for open domain question answering[EB/OL]. (2020-07-02). https://arxiv.org/abs/2007.01282.

[11] SCHICK T, DWIVEDI-YU J, DESSÌ R, et al. Toolformer: Language models can teach themselves to use tools[J]. Advances in Neural Information Processing Systems, 2023, 36: 68539-68551.

[12] ASAI A, WU Z, WANG Y, et al. Self-RAG: Learning to retrieve, generate, and critique through self-reflection[C]//Proceedings of the Twelfth International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2023: 1-30.

[13] 田永林，王雨桐，王兴霞，等. 从RAG到SAGE：现状与展望[J]. 自动化学报，2025，51：1-25.
TIAN Yonglin, WANG Yutong, WANG Xingxia, et al. From retrieval-augmented generation to SAGE: The state of the art and prospects[J]. Acta Automatica Sinica, 2025, 51: 1-25.

[14] KISS C, NAGY M, SZILÁGYI P. Max-min semantic chunking of documents for RAG application [J]. Discover Computing, 2025, 28(1): 117.

[15] XIAO S, LIU Z, ZHANG P, et al. C-pack: Packed resources for general chinese embeddings[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC, USA: Association for Computing Machinery, 2024: 641-649.

[16] CHEN J, BAO R, ZHENG H, et al. Optimizing Retrieval-augmented generation with elasticsearch for enhanced question-answering systems[EB/OL]. (2024-10-18). https://arxiv.org/abs/2410.14167.

[17] 高志强，沈佳楠，姬纬通，等. 大模型技术的军事应用综述[J]. 南京航空航天大学学报，2024，56(5)：801-814.
GAO Zhiqiang, SHEN Jianan, JI Weitong, et al. Review of military applications of foundation model technology[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2024, 56(5): 801-814.

[18] 邹通，丁学良，戴瀚苏，等. 基于大语言模型的集群协同决策[J]. 南京航空航天大学学报(自然科学版)，2025，57(6)：1061-1071.
ZOU Tong, DING Xueliang, DAI Hansu, et al. Swarm cooperation decision-making based on large language models[J]. Journal of Nanjing University of Aeronautics & Astronautics (Natural Science Edition), 2025, 57(6): 1061-1071.

[19] 李鸿亮，刘禹良，廖文辉，等. 大模型时代的光学文字识别：现状及展望[J]. 中国图象图形学报，2025，30(6)：2023-2050.
LI Hongliang, LIU Yuliang, LIAO Wenhui, et al. OCR in the era of large models: Current status and prospects[J]. Journal of Image and Graphics, 2025, 30(6): 2023-2050.

[20] 刘俊伟，钱昱辰，周泽宇，等. 基于大模型和检索增强生成技术的军事知识问答系统研究与应用[C]//第十三届中国指挥控制大会. 杭州：中国指挥与控制学

会,智元研究院有限公司, 2025: 598-604.
LIU Junwei, QIAN Yuchen, ZHOU Zeyu, et al. Research and application of military knowledge question-answering systems based on large models and retrieval-augmented generation technologies[C]//Proceedings of the 13th China Conference on Command and Control. Hangzhou: Chinese Institute of Command and Control, Hangzhou Zhiyuan Research Institute Co., Ltd., 2025: 598-604.

[21] ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

[22] GUAN X, ZENG J, MENG F, et al. DeepRAG: Thinking to retrieval step by step for large language models[EB/OL]. (2025-02-03). https://arxiv.org/abs/2502.01142.

[23] GUO D, YANG D, ZHANG H, et al. Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning[EB/OL]. (2025-01-22). https://arxiv.org/abs/2501.12948.

[24] YANG A, YANG B, ZHANG B, et al. Qwen-2.5 technical report[EB/OL]. (2024-12-19). https://arxiv.org/abs/2412.15115.

[25] ZENG A, XU B, WANG B, et al. CHATGLM: A family of large language models from GLM-130b to GLM-4 all tools[EB/OL]. (2024-06-18). https://arxiv.org/abs/2406.12793.

[26] TOUVRON H, MARTIN L, STONE K, et al. LLaMA 2: Open foundation and fine-tuned chat models[EB/OL]. (2023-07-18). https://arxiv.org/abs/2307.09288.

[27] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2002: 311-318.

[28] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of in Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004: 74-81.

[29] ZHANG T, KISHORE V, WU F, et al. Bertscore: Evaluating text generation with bert[EB/OL].(2019-04-21).https://arxiv.org/abs/1904.09675.

[30] HURST A, LERER A, GOUCHER A P, et al. GPT-4o system card[EB/OL].(2024-10-25).https://arxiv.org/abs/2410.21276.

(编辑:刘彦东)