·专题报道——人工智能在公共卫生中的应用·

# 智能问答模型构建以及在公共卫生领域应用：基于检索增强生成技术

刘婕，郝舒欣，刘博涵，刘悦

中国疾病预防控制中心环境与人群健康重点实验室/中国疾病预防控制中心环境与健康相关产品安全所，北京 100021
通信作者：刘悦，E-mail：liuyue@nieh.chinacdc.cn

**【摘　要】目的** 通过构建基于检索增强生成（RAG）技术的智能问答模型，辅助公共卫生领域的技术人员从规范、指南、手册、文献等技术文档中快速、准确识别并提取关键信息，提升知识获取效率与规范应用能力。**方法** 基于 RAG 技术，集成自动化文本预处理机制，采用 LangChain 框架，将带有阈值管理的信息检索模块与历史对话记录支持的回答生成模块整合，构建基于大语言模型的检索增强智能问答模型（CURA-LLM）。**结果** 在操作规范、检测流程与法律规范 3 个公共卫生领域典型数据集上 CURA-LLM 模型表现优异，F1 评分分别为 0.941、0.891 和 0.947，余弦相似度分别为 0.968、0.929 和 0.963。与 3 种现有领域问答模型相比，CURA-LLM 模型在检索相关性和回答准确性方面更具有优势。**结论** CURA-LLM 模型为公共卫生技术文档的智能化解读与知识快速总结与提取提供了有效支撑。辅助专业人员在突发公共卫生事件响应、日常管理、培训与教学等场景中快速响应与规范执行，具有良好的应用前景。

**【关键词】** 检索增强生成（RAG）；大语言模型；智能问答；公共卫生

# Development of an intelligent question-answering model and its application in public health: based on RAG technology

LIU Jie, HAO Shuxin, LIU Bohan, LIU Yue (*China CDC Key Laboratory of Environment and Population Health*, *National Institute of Environmental Health*, *Chinese Center for Disease Control and Prevention*, *Beijing* 100021, *China*)
Corresponding author: LIU Yue, E-mail: liuyue@nieh.chinacdc.cn

**【Abstract】 Objective** To develop an intelligent question-answering model based on retrieval-augmented generation (RAG) to assist public health professionals in efficiently and accurately extracting key information from technical documents such as regulations, guidelines, manuals, and literature, thereby improving knowledge acquisition efficiency and standardized application capabilities. **Methods** On the basis of RAG, this study integrates an automated text preprocessing mechanism and employs the LangChain framework to construct a context-aware unified retrieval-augmented question-answering model powered by large language models (CURA-LLM). This model combines a threshold-managed information retrieval module with a response generation module supported by historical dialogue records. **Results** CURA-LLM demonstrated outstanding performance on three representative public health datasets related to operational guidelines, testing procedures, and legal regulations, with the F1 scores of 0.941, 0.891, and 0.947, and the cosine similarities of 0.968, 0.929, and 0.963, respectively. Moreover, CURA-LLM outperformed three existing domain-specific question-answering models in terms of retrieval relevance and answer accuracy. **Conclusions** CURA-LLM provides effective support for the intelligent interpretation and rapid extraction of knowledge from public health technical documents. It can assist personnel in quickly responding and standardizing execution in scenarios such as emergency responses, routine management, and professional training, offering strong potential for practical application.

**【Keywords】** retrieval-augmented generation; large language model; intelligent question answering; public health

公共卫生领域知识更新迅速，各类标准、指南和技术文件不断发布，对基层专业人员在短时间内高效获取、理解并规范应用专业知识提出了更高要求[1]。同时，传统培训模式受制于师资和资源的限制，难以满足专业人员对复杂技术文件的深度理解和高效应用的需求。随着人工智能技术的发展，检索增强生成（retrieval-augmented generation, RAG）技术已成为智能问答系统发展的重要方向，已扩展至医疗、法律等特定领域[2–3]。然而，现有 RAG 系统在应对高度结构化、术语密集的技术文件时，往往缺乏针对性设计，尤其在文档预处理标准化、多轮对话中上下文动态管理与调用以及低相关性检索结果的过滤方面仍存在局限。这限制了其在公共卫生等领域复杂问答任

开放获取：CC BY-NC-ND 4.0 DOI：10.11847/zgggws1148202
第一作者：刘婕（1988—），助理研究员，硕士，研究方向：信息技术应用、环境与健康。
收稿日期：2025-06-26 修回日期：2025-08-11 录用日期：2025-09-30
利益冲突：不存在 伦理审查：不需要 出版授权：全体作者已与编辑部签署作者贡献声明及版权转让协议

务中的有效性。针对上述问题，本研究提出检索增强智能问答模型（Context-aware Unified Retrieval-Augmented with LLM, CURA-LLM），在检索前引入文档结构感知的预处理模块，并通过阈值筛选与历史对话管理模块，为构建质量可靠的本地化智能问答系统提供新思路。

## 1 材料与方法

1.1 数据来源　选取《洪涝灾区环境卫生处置与预防性消毒指引（2023 版）》[4]（简称“消毒指引”），《WST799—2022 污水中新型冠状病毒富集浓缩和核酸检测方法标准》[5]（简称“检测标准”）和《中华人民共和国传染病防治法》[6]（简称“传染病防治法”）3 份技术文档作为测试文件，分别涵盖操作规范、检测流程与法律法规，覆盖不同公共卫生应用场景，支持多类型任务的模型能力验证。

1.2 方法

1.2.1 模型框架（图 1）　模型框架由文本处理、信息检索和答案生成 3 个核心模块组成，协同实现从数据预处理到多轮问答的完整流程，见图 1。文件处理模块对文件进行表格转化、清洗和格式化，为信息检索模块提供高质量输入数据；信息检索模块通过语义向量化和阈值筛选功能，筛选高相关性内容；答案生成模块基于大语言模型结合提示工程（prompt engineering）与对话记忆，生成精准且连贯的回答，同时将相关知识单元作为答案参考同时输出。基于 LangChain 框架，将信息检索与答案生成相结合，为模型的语义理解、任务适配性及未来功能拓展提供支持。

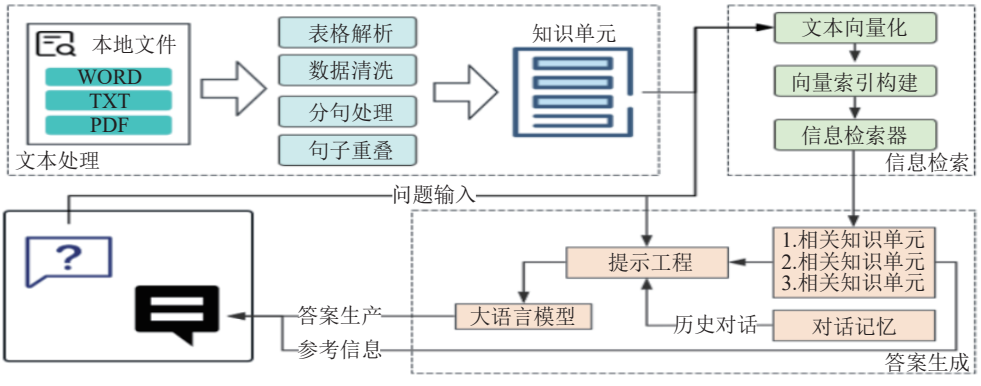



**图 1**　CURA-LLM 模型框架

**Fig. 1**　Framework of CURA-LLM

1.2.1.1 开放式文本处理　为提高输入文本质量并适应后续处理需求，设计高效自动化文件预处理流程。针对文件中表格，采用逐行解析，将表头与内容逐列映射为“字段–值”形式，并转换为自然语言描述后替换表格内容，确保结构化数据的语义完整性。文本部分清理噪声信息并统一编码格式为 UTF-8。针对技术文档中标点符号不完整且句子较长的问题，采用正则表达式的分句方法，灵活识别句子结束标志。将若干句子组合为段落，段落长度限定为 500 字符内，通过段落首尾两句的重叠维持上下文关联，确保模型输入的高质量和一致性[7]。

1.2.1.2 信息检索器构建　基于向量化检索策略实现高效的信息检索。采用中文嵌入模型将知识单元和用户问题分别转换为高维向量，利用 FAISS（Facebook AI Similarity Search）构建文档向量存储结构。通过计算用户问题向量与知识单元向量的余弦相似度获得初步检索结果，相似度为：

$$\text{cosine_similarity}(Q,K)=\frac{Q\cdot K}{\|Q\|\|K\|}$$

其中：$Q\cdot K$ 为用户问题向量与知识单元向量的点积，$\|Q\|$为用户问题向量的欧几里得长度，$\|K\|$为知识单元向量的欧几里得长度。

为确保检索结果的高相关性与可控性，将相似度结果进行排序，过滤相似度低于 0.5 的知识单元，保留相似度最高的 5 个知识单元。该信息检索过程以模块化接口的形式实现，便于后续针对不同数据规模或业务需求进行扩展与迭代。

1.2.1.3 答案生成模块构建（表 1）　基于 LangChain 框架整合检索模块输出的知识单元、用户问题、历史对话记录及提示内容，输入至智能问答模型（LLM）进而生成准确回答。提示工程（prompt engineering）引导 LLM 综合多轮对话与检索结果，生成符合上下文的回答，本研究 Prompt 如表 1 所示。利用会话缓冲存储器记录并维护历史问答内容，使对话能够参考先前的交互上下文，从而生成更加精准的回答。模型设定将搜索到的知识单元与回答同时输出，便于用户查询原文信息。此外，当检索结果为空时，模型将及时提示用户以帮助用户优化输入，提高交互体验和系统的智能响应能力。

**表 1** 提示工程示例
**Table 1** Sample prompts for prompt engineering

| 提示 | 提示文本 |
|---|---|
| LLM 答案生成提示 | 请阅读并整合检索到的相关信息和之前的对话内容，全面回答用户问题，回答时避免重复用户问题，使用原文回答。<br>用户问题:{question}<br>以下是之前的对话内容:<br>{chat_history}<br>以下是检索到的相关信息:<br>{context} |

1.2.2 模型评价

1.2.2.1 评价指标 为全面验证模型性能，本研究通过基于变换器的双向编码器表征法(bidirectional encoder representations from Transformer, BERT)比较生成文本与参考文本之间的嵌入表征。BERT 评分包括精确率(Precision)、召回率(Recall)和 F1 评分，综合反映回答的质量[8]。通过计算生成回答与参考原文词向量之间的余弦相似度，量化语义匹配程度。

1.2.2.2 消融实验 为验证所提出的相似度阈值管理机制对模型性能的影响，设计消融实验。实验在相同测试集下，分别比较设置相似度阈值(0.5)与取消该阈值条件下的性能差异。通过量化不同策略对回答质量的影响，评估阈值机制在过滤低相关片段、提升语义匹配度与回答准确性方面的有效性。

1.2.2.3 多轮对话定量分析 通过 10 组具备上下文依赖的多轮对话任务测试模型历史记忆能力。每组对话包含 5 轮递进式提问，回答需调用历史内容，问题均来源于输入文档，参考答案为匹配的原文句。通过评价指标评估各轮回答，量化语义保持与记忆调用效果。

1.2.2.4 基准方法(表 2) 本研究模型 CURA-LLM 与 3 种领域问答模型进行对比，以评估模型性能。每个模型回答了测试文档中提取的 50 个典型问题，涵盖技术细节回答等多种任务场景，问题均来源于输入文档，参考答案为匹配的原文句。对于典型问题，以实例方式进行展示。

**表 2** 模型说明
**Table 2** Overview of models

| 模型 | 英文名 | 模型说明 | 是否依赖外部知识 | 检索方式 |
|---|---|---|---|---|
| DPLM | Direct Prompt Learning Method | 直接提示学习法 | 否 | 无 |
| DPR-LLM | Dense Passage Retrieval with LLM | 密集检索增强 | 是 | 语义密集检索 |
| BM25 | Okapi Best Matching 25 | 稀疏检索增强 | 是 | 关键词稀疏检索 |
| CURA-LLM | Context-aware Unified Retrieval-Augmented with LLM | 阈值历史对话检索增强 | 是 | 语义密集检索，设置阈值检索并与历史对话结合 |

1.3 实验环境 实验使用通义千问对话模型(Qwen 2.5-7B)，其在语义表达和逻辑推理方面具有较强能力和稳定性[9]。嵌入模型选择中文文本向量化模型 shibing624/text2vec-base-chinese[10]。该模型经过大量的中文语料训练，能够捕捉文本中的语义信息。实验计算机配备 Apple M4 Max 芯片，基于 macOS 15.1.1 操作系统，结合 Python 3.12.7 软件和 PyTorch 2.0 软件进行开发与执行。

## 2 结 果

2.1 文档处理结果(表 3) 本研究构建的自动化文本处理流程实现了对技术文档中表格和文本的解析与清洗。对“消毒指引”、“检测标准”与“传染病防治法”3 份文档完成预处理。每个知识单元的字符均值均低于中文嵌入模型的最大处理字符数 512。

**表 3** 文档预处理
**Table 3** Document preprocessing

| 文档 | 总字符数 | 句子数 | 句子字符均值 | 整合段落数 | 知识单元字符均值 |
|---|---|---|---|---|---|
| 消毒指引 | 6 146 | 157 | 38.516 | 15 | 465.067 |
| 检测标准 | 7 885 | 153 | 50.732 | 22 | 452.273 |
| 传染病防治法 | 22 672 | 304 | 73.789 | 76 | 496.776 |

2.2 消融实验(表 4) 为评估相似度阈值管理模块对问答性能的影响，在保持其他组件一致的前提下，构建对照模型，即移除相似度阈值筛选，仅基于向量数据库对所有预处理段落进行检索并生成答案，记作 CURA-LLM 去阈值。与 CURA-LLM 去阈值相比，CURA-LLM 在 3 个数据集上均表现出更高的精确率、F1 值及余弦相似度，同时平均应答时间更短，显示出更优的回答质量与生成效率。结果表明，相似度阈值管理机制在提升回答准确性、过滤冗余信息干扰方面具有一定应用价值。

**表 4** 消融实验
**Table 4** Ablation experiments

| 文档 | 模型 | 精确率 | 召回率 | F1 | 余弦相似度 | 平均应答时间/s |
|---|---|---|---|---|---|---|
| 消毒指引 | CURA-LLM去阈值 | 0.881 | 0.934 | 0.905 | 0.931 | 6.881 |
| | CURA-LLM | 0.928 | 0.956 | 0.941 | 0.968 | 4.909 |
| 检测标准 | CURA-LLM去阈值 | 0.811 | 0.887 | 0.845 | 0.903 | 6.292 |
| | CURA-LLM | 0.878 | 0.908 | 0.891 | 0.929 | 3.911 |
| 传染病防治法 | CURA-LLM去阈值 | 0.879 | 0.932 | 0.902 | 0.928 | 4.889 |
| | CURA-LLM | 0.941 | 0.953 | 0.947 | 0.963 | 3.021 |

## 2.3 模型评价

2.3.1 模型性能结果(表 5) 去除多轮对话任务，评估 CURA-LLM 与 3 种基准模型在 3 份文档上的问答性能。CURA-LLM 在所有文档中均取得最高的 F1 评分和余弦相似度，F1 评分分别为 0.941、0.891 和 0.947；余弦相似度分别为 0.968、0.929 和 0.963。表明 CURA-LLM 在回答准确性与语义相关性方面均表现稳定，能够有效捕获用户需求，生成逻辑连贯、内容贴切的回答，具备良好的综合性能。在实验室操作、灾后环境消毒、疫情防控等场景下，能够帮助现场人员高效获取关键操作标准与防控要点，辅助提升应急响应效率和技术规范性落实。模型在不同类型文档中的表现存在差异，这可能与语义组织方式有关。“传染病防治法”其法律文本结构规范、语义边界清晰，有助于模型准确提取答案，模型在该文档上的表现最优。“消毒指引”文本以场景叙述为主，语义关联紧密，模型性能同样较高。“检测标准”偏向技术细节，结构复杂且语义连贯性较低。相比之下，BM25 在该文档中表现相对更好，说明使用关键词匹配策略对结构化内容具有一定优势。

**表 5** 模型性能结果
**Table 5** Model performance results

| 文档 | 模型 | 精确率 | 召回率 | F1 | 余弦相似度 | 平均应答时间/s |
|---|---|---|---|---|---|---|
| 消毒指引 | DPLM | 0.587 | 0.679 | 0.629 | 0.763 | 4.539 |
| | DPR-LLM | 0.852 | 0.886 | 0.867 | 0.912 | 2.78 |
| | BM25 | 0.641 | 0.688 | 0.663 | 0.76 | 2.534 |
| | CURA-LLM | 0.928 | 0.956 | 0.941 | 0.968 | 4.909 |
| 检测标准 | DPLM | 0.549 | 0.683 | 0.607 | 0.738 | 4.691 |
| | DPR-LLM | 0.813 | 0.858 | 0.833 | 0.889 | 2.25 |
| | BM25 | 0.642 | 0.715 | 0.674 | 0.769 | 2.391 |
| | CURA-LLM | 0.878 | 0.908 | 0.891 | 0.929 | 3.911 |
| 传染病防治法 | DPLM | 0.602 | 0.752 | 0.668 | 0.815 | 4.352 |
| | DPR-LLM | 0.902 | 0.919 | 0.907 | 0.933 | 1.643 |
| | BM25 | 0.644 | 0.719 | 0.679 | 0.775 | 2.107 |
| | CURA-LLM | 0.941 | 0.953 | 0.947 | 0.963 | 3.021 |

2.3.2 CURA-LLM 模型多轮问答任务结果(表 6) 结果显示第 1 轮问答(Q1)F1 值为 0.943，平均应答时间为 2.534 s；至第 5 轮(Q5)，F1 值下降至 0.782，应答时间上升至 5.583 s，呈现性能衰减趋势。结果表明，模型在对话初期具备良好的上下文理解与应答能力，但随着轮次推进，对历史信息的保持能力减弱，导致回答准确性降低。同时，性能下降并非均匀递减，个别轮次如 Q3 与 Q4 波动较大，提示问答质量不仅受轮次影响，还可能与问题对历史语境的依赖强度及语义复杂性密切相关。

**表 6** 多轮问答任务结果
**Table 6** Results on the multi-turn question answering task

| 对话轮次 | 精确率 | 召回率 | F1 | 余弦相似度 | 平均应答时间/s |
|---|---|---|---|---|---|
| Q1 | 0.946 | 0.941 | 0.943 | 0.96 | 2.534 |
| Q2 | 0.93 | 0.939 | 0.934 | 0.953 | 3.18 |
| Q3 | 0.849 | 0.88 | 0.863 | 0.92 | 4.511 |
| Q4 | 0.796 | 0.867 | 0.824 | 0.841 | 4.777 |
| Q5 | 0.787 | 0.783 | 0.782 | 0.806 | 5.583 |

### 2.4 模型回答实例分析

2.4.1 模型回答实例(表 7) 结果显示，CURA-LLM 能够精准捕捉对话中的语义信息，通过阈值检索结果生成更符合标准答案的回答。相比之下，DPLM 缺乏检索能力，回答多为泛化内容；BM25 在复杂语义场景中对关键信息的覆盖率较低，DPR-LLM 在细节提取方面略显不足。回答实例进一步验证了相较对比模型，CURA-LLM 在处理涉及操作流程、疫情防控等与公共卫生实践密切相关内容时更具有优势；能帮助现场工作人员及时定位所需知识，强化标准执行能力。

**表 7** 问答模型实例
**Table 7** Answer examples of QA models

| 模型 | 消毒指引 | 检测标准 | 传染病防治法 |
|---|---|---|---|
| 问题 | 动物尸体深埋的场所有什么要求? | 聚乙二醇沉淀法实验方法是什么? | 国家对患有特定传染病的困难人群提供哪些医疗保障措施? |
| DPLM | .....深度要求：通常建议将动物尸体深埋至地下>2 m，确保不会被挖掘到地表或再次暴露..... | .....这一过程会导致蛋白质之间的相互作用减弱或受到抑制，从而有助于它们从溶液中析出..... | 1. 免费或低价治疗.....<br>2. 医疗救助.....<br>3. 基本医疗保险覆盖.....<br>4. 专项救助基金..... |
| DPR-LLM | .....挖土坑深度需>2 m.....再用漂白粉按每平方米20～40 g的比例撒盖于动物尸体上..... | .....收集形成的沉淀用于后续的核酸提取和实时荧光RT-PCR检测..... | 国家对患有特定传染病的困难人群实行医疗救助，减免医疗费用，并且在患者、疑似患者的治疗费用方面..... |
| BM25 | .....处理动物尸体时应遵循当地环保部门的规定，并确保安全、卫生地处置，避免对环境和公共卫生造成影响。..... | .....并未包含关于聚乙二醇沉淀法的具体实验方法说明..... | 给出的相关信息中，并未直接提及国家为患有特定传染病的困难人群提供的具体医疗保障措施..... |
| CURA-LLM | .....挖土坑深>2 m.....再用漂白粉按20～40 g/m² 撒盖于动物尸体，..... | .....收集病毒颗粒多聚体形成的沉淀，用于后续核酸提取和实时荧光RT-PCR检测。 | 有特定传染病的困难人群实行医疗救助，减免医疗费用。对于符合国家规定的传染病医疗费用，基本医疗保险按照规定予以支付。患者、疑似患者的治疗甲类传染病..... |
| 原文答案 | .....挖土坑深>2 m，.....再用漂白粉按20～40 g/m²撒盖于动物尸体上..... | .....收集病毒颗粒多聚体形成的沉淀，用于后续核酸提取和实时荧光RT-PCR检测。 | 国家对患有特定传染病的困难人群实行医疗救助，减免医疗费用。同时鼓励商业保险机构开发传染病防治相关保险产品，并规定符合国家规定的传染病医疗费用由基本医疗保险按照规定予以支付，对于患者、疑似患者的医疗费用..... |

2.4.2 答案错误类型及成因分析(表 8) 为进一步评估 CURA-LLM 模型表现，选取对话任务中典型错误案例进行分析，识别其表现不佳的具体原因及共性特征。问题包括：模型未能准确把握提问意图，导致未定位至关键条款或概念，回答内容泛化、模糊；知识单元划分策略将连贯信息切分至多个片段，影响向量召回对完整内容的覆盖，导致生成回答信息缺失；此外，尽管设置了“使用原文回答”的提示工程策略，但在部分任务中模型仍存在语言重写或复述现象，提示词对生成结果的约束能力有限，影响模型在任务中的合规性和精确性。

**表 8** 模型错误案例
**Table 8** Error cases of CURA-LLM answers

| 问题 | 模型回答 | 原文答案 |
|---|---|---|
| 医疗机构感染包括哪些情况? | 根据提供的信息，文中并没有直接提及“医疗机构感染”的具体情形，但文中涉及了传染病预防控制的相关措施，例如对甲类、乙类和丙类传染病患者的隔离治疗、医学观察等。这些措施可以视为在医疗机构中防控传染病的措施之一。如果需要了解具体的医疗机构内感染情况，请提供更详细的信息或具体询问相关条款。 | 医疗机构感染，是指在医疗机构内获得的感染，包括在医疗机构内发生的感染和在医疗机构内获得、离开医疗机构后发生的感染，但不包括进入医疗机构前已开始或者已处于潜伏期的感染。医疗机构工作人员在医疗机构内获得的感染也属医疗机构感染。 |
| 铝盐混凝沉淀法需要哪些试剂? | 铝盐混凝沉淀法需要以下试剂：<br>.....7.2.2.1.4 盐酸溶液[c(HCl)=1 mol/L]：移取43 mL盐酸(分析纯，ρ=1.19 g/mL)于烧杯中，加入适量水，用玻璃棒搅拌混匀后转移至500 mL容量瓶中，并定容至刻度。 | 7.2.2.1试剂<br>.....<br>7.2.2.1.4盐酸溶液[c(HCl)=1 mol/L]：移取43 mL盐酸(分析纯，ρ=1.19 g/mL)于烧杯中，加入适量水，用玻璃棒搅拌混匀后转移至500 mL容量瓶中，并定容至刻度。<br>7.2.2.1.5乙二胺四乙酸二钠二水合物($C_{10}H_{14}N_2O_8Na_2·2H_2O$)：纯度≥98%。 |

**续表 8**
**Table 8** Continued

| 问题 | 模型回答 | 原文答案 |
|---|---|---|
| 洪灾之后农村怎么处理粪便? | 洪灾之后农村处理粪便的方法如下：<br>1. 对于可清掏的贮粪构筑物内的粪便，应将其清掏后加等量生石灰或按1 kg/m²撒布有效氯含量>20%的漂白粉混匀消毒后清除。<br>2. 对于不可清掏的贮粪构筑物内的粪便，则需按体积比10 : 1加漂白粉进行消毒处理，混合2 h后再清除。处理后的粪便应符合GB 7959—2012《粪便无害化卫生要求》的要求。<br>3. 对于废弃池坑的处理，可用生石灰或漂白粉按1 kg/m²撒布、浇湿后，用农田土壤或建筑渣土填平、压实。 | 粪便处理：应将可清掏贮粪构筑物内的粪便清掏后加等量生石灰或按1 kg/m²撒布有效氯含量>20%的漂白粉混匀消毒后清除；<br>不可清掏贮粪构筑物内的粪便按10 : 1(体积比)加漂白粉进行消毒处理，混合2 h后清除。处理后的粪便应符合GB 7959—2012《粪便无害化卫生要求》的要求。 |

2.4.3 CURA-LLM 回答效果与空检索处理(表 9)结果表明，CURA-LLM 模型可通过显示相关段落和相似度，帮助用户快速定位答案来源，提升信息获取的透明性和信任度。同时，当问题无法匹配到相关内容时，模型提供明确的提示，引导用户重新输入更具体的查询问题，从而改善用户体验和交互效果。智能反馈机制适用于突发公共卫生事件中面对新问题或模糊提问的场景，有助于提高沟通效率和应对灵活性，增强模型在动态场景下的实用性和友好性。

**表 9** CURA-LLM 模型回答实例及检索结果
**Table 9** Answer examples and retrieval results of CURA-LLM

| 问题 | 病毒灭活需要注意什么? | 南方地区面临低温寒潮应该怎么办? |
|---|---|---|
| 模型回答 | 病毒灭活时应将样本充分混匀后置于60 ℃水浴30 min。注：水浴锅水位应高于样本液面，确保容器内样本达到目标温度。 | 抱歉，输入文件中未能找到与您的问题直接相关的内容。请重新输入。 |
| 相关段落及相似度 | 段落 1：(相似度: 0.570) 样本灭活、检测等操作应在生物安全二级及以上实验室……<br>段落 2：(相似度: 0.551) ……将样本充分混匀后置于60 ℃水浴30 min。注：水浴锅水位应高于……<br>段落 3：(相似度: 0.507) 污水中新型冠状病毒富集浓缩…… | 无相关段落信息。 |

## 3 讨　论

本研究提出的 CURA-LLM 模型基于 RAG 框架，融合结构化文本预处理、相似度阈值筛选与历史语境保持机制，面向公共卫生领域技术文档的问答需求，构建了增强型智能问答服务。在三类代表性文档上的实证结果表明，模型在回答准确性和信息调用效率方面均优于现有基础模型，验证了其在专业场景中的有效性和可行性。

在文档结构感知与语义保持方面，CURA-LLM 通过自动化文本清理、表格转化、段落级切分和基础向量化策略，有效提升了模型对不同结构密集型专业文档结构的适应能力。这种轻量化结构增强路径与复杂图模型建模方法或通用分词工具相比更具稳定性，更适用于实际部署[11-12]。模型中设置的相似度阈值筛选机制，有效过滤了语义相关性差的干扰内容，优化了生成质量[13]；在多轮问答任务中，模型亦表现出一定程度的语境保持能力。

CURA-LLM 对公共卫生业务流程与体系化运行具备良好适应能力。可广泛应用于培训、继续教育、日常技术指导等场景；缓解基层培训资源有限、知识更新滞后的问题，可提升基层人员的执行能力与专业素养。在地方病防治、慢性病管理、饮用水监测、职业卫生监督等常规工作中，可为基层提供即时、可靠的技术支持，推动标准的精准落实，保障执行一致。在突发公共卫生事件中，有助于快速理解规程与技术要点，提升应急响应效率与规范化水平。

尽管模型在实验任务中表现稳定，但仍存在一定局限。当前模型仅支持单文档处理，尚不具备跨文档、多源异构信息整合能力；对于结构松散、语义分布不均的文本仍可能导致上下文割裂与信息遗漏；生成内容偶有偏离原文的情况。提示其在高合规性任务中的内容控制仍待优化。未来研究可采取基于语义嵌入的自适应切分等方法[14]，引入片段级答案校正与监督机制，通过提示工程在生成阶段增强内容对齐性，实现更高的回答准确性[15]。借鉴对话管理机制[16]，通过动态记忆增强对话连续性处理能力。同时应围绕交互优化与系统集成等方向深化探索，助力智慧公共卫生体系建设。

CURA-LLM 模型为公共卫生技术文件的高效利用提供新路径，在知识获取、基层规范执行与应急响应等方面展现出良好潜力与应用推广价值。

**数据可用性声明**　本分析中使用的所有输入数据

的引用和元数据均可从公开数据库下载：(1)《洪涝灾区环境卫生处置与预防性消毒指引(2023 版)》：来源于国家疾控局官方网站(网站地址：https://www.ndcpa.gov.cn/jbkzzx/yqxxxw/common/content/content_1696053564394565632.html)；(2)《WST 799—2022 污水中新型冠状病毒富集浓缩和核酸检测方法标准》：来源于中华人民共和国国家卫生健康委员会官方网站(网站地址：https://www.nhc.gov.cn/fzs/c100048/202204/f7a4d18c309c4813b9e1b8a7e4faff42.shtml)；(3)《中华人民共和国传染病防治法》：来源于中国人大网(网站地址：http://www.npc.gov.cn/npc/c2/c30834/202504/t20250430_445085.html)。本研究使用的相关数据若因平台更新、调整、关闭，数据无法获取，作者不再单独提供相应数据，敬请理解。

**AIGC 使用声明** 在本研究的准备过程中，作者使用了通义千问对话模型(Qwen 2.5-7B)，中文文本向量化模型(shibing624/text2vec-base-chinese)和通义千问服务来辅助完成模型构建，用于文本语义生成、问答内容生成、语义嵌入向量提取与文章前期资料收集等任务。使用这些工具/服务的时间为 2024 年 12 月—2025 年 6 月，并通过设计提示词对 Qwen 2.5-7B 进行指令交互；调用向量化模型生成文本嵌入以支持相似度计算与语义召回生成相关回答文本等。使用此工具/服务后，作者根据需要对内容进行了必要的审查和编辑，并对出版物的内容承担全部责任。

## 参考文献


[1] 赵俊贤, 杨楠, 刘辉, 等. 我国卫生健康标准现状与发展 [J]. 协和医学杂志, 2023, 14(6): 1325-1329.

[2] Liu S R, McCoy A B, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines[J]. Journal of the American Medical Informatics Association, 2025, 32(4): 605-615.

[3] Arslan M, Ghanem H, Munawar S, et al. A survey on RAG with LLMs[J]. Procedia Computer Science, 2024, 246: 3781-3790.

[4] 国家疾病预防控制局. 国家疾控局发布洪涝灾区环境卫生处置与预防性消毒指引 (2023 版)[EB/OL]. (2023-08-04)[2023-12-20]. https://www.ndcpa.gov.cn/jbkzzx/yqxxxw/common/content/content_1696053564394565632.html.

[5] 中华人民共和国国家卫生健康委员会. 关于发布推荐性卫生行业标准《污水中新型冠状病毒富集浓缩和核酸检测方法标准》的通告[EB/OL].(2022-04-06)[2023-12-20]. https://www.nhc.gov.cn/fzs/c100048/202204/f7a4d18c309c4813b9e1b8a7e4faff42.shtml.

[6] 中国人大网. 中华人民共和国传染病防治法 [EB/OL]. (2025-04-30)[2025-06-15]. http://www.npc.gov.cn/npc/c2/c30834/202504/t20250430_445085.html.

[7] 齐思洋, 胡慧云, 李洪冰, 等. 融合大语言模型的领域问答系统构建方法 [J]. 北京邮电大学学报, 2024, 47(4): 50-56.

[8] Al-Besher A, Kumar K, Sangeetha M, et al. BERT for conversational question answering systems using semantic similarity estimation[J]. Computers, Materials and Continua, 2022, 70(3): 4763-4780.

[9] Bai J Z, Bai S, Chu Y F, et al. Qwen technical report[EB/OL]. (2023-09-28)[2024-12-20]. https://arxiv.org/pdf/2309.16609.

[10] Xu M. Text2vec: text to vector toolkit (Version 1.1.2) [Computer Software][EB/OL]. (2023-09-20)[2024-12-20]. https://github.com/shibing624/text2vec.

[11] Qu R Y, Tu R X, Bao F. Is semantic chunking worth the computational cost?[C]. Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, New Mexico: Association for Computational Linguistics, 2025: 2155-2177.

[12] Dehal R S, Sharma M, Rajabi E. Knowledge graphs and their reciprocal relationship with large language models[J]. Machine Learning and Knowledge Extraction, 2025, 7(2): 38

[13] Rossi N, Lin J X, Liu F, et al. Relevance filtering for embedding-based retrieval[C]. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: Association for Computing Machinery, 2024: 4828-4835.

[14] Kiss C, Nagy M, Szilágyi P. Max–Min semantic chunking of documents for RAG application[J]. Discover Computing, 2025, 28(1): 117.

[15] Wang L, Bi W S, Zhao S L, et al. Investigating the impact of prompt engineering on the performance of large language models for standardizing obstetric diagnosis text: comparative study[J]. JMIR Formative Research, 2024, 8: e53216.

[16] Wang Q, Fu Y, Cao Y, et al. Recursively summarizing enables long-term dialogue memory in large language models[J]. Neurocomputing, 2025, 639: 130193.