Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(5), 1165-1177
Published Online September 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.155106

面向政务服务的知识增强问答系统研究与实现

张洋洋，李　桐*

北京印刷学院信息工程学院，北京

收稿日期：2026年7月17日；录用日期：2026年9月1日；发布日期：2026年9月10日

摘　要

针对政务领域知识检索准确性不足、回答缺乏事实依据以及大语言模型专业知识能力有限等问题，本研究构建了一种融合领域知识图谱与大语言模型的政务智能问答系统(GOV_QA)。首先，对政务服务数据进行知识建模与结构化组织，构建政务领域知识图谱；其次，提出分层知识召回方法，通过实体匹配构建局部子图，将图谱信息转化为知识单元并筛选高相关知识；最后，将召回知识与用户问题共同输入大语言模型，约束回答生成。实验结果表明，GOV_QA的BERTScore精确率、召回率和F1值分别达到0.7600、0.8557和0.8025，主观评价平均得分为4.49分，其中76%的回答获得5分评价，整体性能优于Doubao、DeepSeek和Qwen等基准模型，为政务智能问答提供高可靠的技术方案。

关键词

政务问答系统，知识图谱，大语言模型，检索增强生成

Research and Realization of Knowledge-Enhanced Q&A System for Government Services

Yangyang Zhang, Tong Li*

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: July 17, 2026; accepted: September 1, 2026; published: September 10, 2026

Abstract

In view of the insufficient accuracy of knowledge retrieval, the lack of factual support for answers, and the limited domain knowledge of large language models in government services, this study

*通讯作者。

文章引用: 张洋洋, 李桐. 面向政务服务的知识增强问答系统研究与实现[J]. 人工智能与机器人研究, 2026, 15(5): 1165-1177. DOI: 10.12677/airr.2026.155106

develops an intelligent question answering system named GOV_QA, which integrates a domain knowledge graph with a large language model. First, government service data are modeled and structurally organized to construct a government service knowledge graph. Second, a hierarchical knowledge retrieval method is proposed to construct local subgraphs through entity matching, transform graph information into knowledge units, and select highly relevant knowledge. Finally, the retrieved knowledge and user questions are jointly input into the large language model to constrain answer generation. Experimental results show that GOV_QA achieves BERTScore precision, recall, and F1 values of 0.7600, 0.8557, and 0.8025, respectively. The average subjective evaluation score is 4.49, with 76% of the answers receiving a score of 5. The overall performance is superior to that of the Doubao, DeepSeek, and Qwen baseline models, providing a reliable technical solution for intelligent question answering in government services.

Keywords

Government Q&A System, Knowledge Graph, Large Language Model, Retrieval-Augmented Generation

1. 引言

近年来，我国积极推进数字政府和数字化国家建设，加快政务服务数字化转型，重点推动政务服务由传统人工窗口和电话咨询向在线化、智能化与精准化方向升级[1]。随着政务服务事项持续增加和数字化转型不断深入，政务数据来源更加多元、更新速度加快，传统依赖人工咨询与窗口办理的服务模式逐渐显现不足[2] [3]。主要表现为：政策信息分散、部门间数据衔接不畅，群众难以及时获取准确办事指引，工作人员重复答复压力较大，复杂事项仍依赖经验判断，导致咨询效率偏低、服务成本上升[4]，这些问题影响了群众办事体验和政策落实效果。因此，亟需构建高效可信的人机交互服务体系，推动政务服务向智能化、精准化方向升级。近年来，大语言模型[5] (Large Language Model, LLM)的快速发展为上述问题提供了新的技术支撑。凭借其强大的自然语言理解能力、逻辑推理和数据处理能力[6] [7]，其能够更好地解决人工检索效率低、响应不及时等业务痛点，加速政务服务领域数智化升级[8]。但通用 LLM 对政务服务领域专业知识掌握不足，生成过程中仍可能出现事实偏差与内容“幻觉”。知识图谱[9] (Knowledge Graph, KG)作为一种结构化知识表示方式，能够为 LLM 提供显式的政务知识支持，增强模型回答的可解释性、可追溯性与可信性[10]。将知识图谱与 LLM 相结合，可有效弥补通用 LLM 在领域知识准确性、专业适配性等方面的不足，为政务服务智能化与精准化提供技术支撑。

政务服务领域涉及行政审批、社会保障和公共事务等多个业务领域，其事项类型多样、政策关联复杂，对人员的业务知识与实践经验提出了较高要求。目前，以人工查询办事指南为主的信息检索方式，难以应对海量政务数据与复杂咨询问题的处理需求。为解决传统政务信息查询中存在的信息分散、人工检索方式效率低下等问题，程序等[11]构建了基于知识图谱的政务问答机器人，通过对多源异构政务数据进行知识抽取、融合与存储，实现了知识图谱在信息检索中的应用。然而，基于知识图谱的信息检索与问答方法仍存在交互效率低、泛化能力差等问题[12]。

LLM 具有较强的自然语言理解与人机交互能力。王昀等[13]和董昌其等[14]探讨了 LLM 在政务服务领域的应用，展现出其在政务问答与信息处理方面的应用潜力，但实际应用中仍面临知识不足和“幻觉”

等挑战。沈思等[15]采用检索增强生成方法提升了 LLM 在政策问答中的准确性，但文本检索过程中可能引入无关信息，影响回答质量。孙雨生等[16]通过知识图谱与 LLM 融合开展政策知识问答，改善了政策知识组织与查询困难等问题。然而，该方法主要面向政策文本，尚未深入政务事项办理场景，如申请材料、办理条件与政策依据之间的精准关联。因此，现有知识图谱与 LLM 的融合方法在政务服务场景中的应用仍不够深入，难以满足实际政务服务需求。

针对知识图谱与 LLM 在政务服务领域中交互效率低、领域适配性差及复杂场景支持不足等问题，本研究面向政务服务问答场景构建政务知识增强问答系统(Government Knowledge Graph-Enhanced Question Answering System, GOV_QA)，系统核心围绕以下三个方面展开：

1) 构建涵盖政务事项、政务部门、办理地点、申请材料和政策文件的政务知识图谱，实现多源政务知识的结构化组织。

2) 提出分层知识召回方法，通过实体匹配构建局部子图，将图谱信息转化为知识单元并筛选 Top-M 相关知识，为回答生成提供可靠依据。

3) 实现知识图谱与大语言模型融合的政务问答系统，并通过客观指标与主观评价验证系统的准确性和可靠性。

2. 系统设计

政务服务场景呈现事项类型繁多、业务规则复杂和知识更新频繁等特点。不同政务事项在办理条件、申请材料和政策依据等方面存在较大差异，且部分事项涉及多个部门及关联政策，知识之间具有较强的关联性。同时，政务数据来源于办事指南、政策文件、部门公告和业务系统等多种渠道，存在数据格式不统一、字段内容分散及实体名称不规范等问题，不利于政务知识的统一表征与有效检索。此外，用户咨询通常包含口语化表达、事项简称和多意图组合等情况，传统关键词匹配和固定问答方式难以准确理解用户需求，迫切需要数智化技术实现政务知识的结构化管理与准确检索。

因此，本研究构建了政务知识增强问答系统 GOV_QA。该系统由数据获取与处理、知识图谱构建、知识检索增强和生成回复四个部分组成，其功能分别为政务数据采集与规范化处理、政务知识图谱构建、知识图谱与 LLM 融合以及政务问答交互。系统总体框架如图 1 所示。

1) 数据获取与处理

数据获取与处理部分负责政务服务数据的采集、解析与规范化。本研究以北京市政务服务网办事指南为数据来源，通过自动化程序获取事项索引及详情页面，提取事项名称、申请材料、办理地点和政策依据等字段，形成包含 1258 项政务事项的原始数据集。针对动态加载、字段缺失、内容重复、格式不统一和多值字段层级混乱等问题，依次进行页面解析、字段映射、文本清洗、空值处理、数据去重和结构化转换，并开展完整性与一致性校验，最终生成统一的 JSON 数据，为后续分析提供数据支撑。

2) 知识图谱构建

知识图谱构建部分主要负责政务知识的结构化组织与关联表达。本研究结合政务服务业务特点，采用自顶向下与自底向上相结合的方式进行知识建模，定义五类实体和七类关系。利用大语言模型从半结构化数据中抽取实体、属性和关系，并通过实体名称统一、重复实体合并、属性补全和关系一致性校验完成知识融合。采用 Neo4j 图数据库进行存储，实现政务知识图谱的系统化构建。

3) 知识检索增强

知识检索增强部分采用分层知识召回方法。系统首先结合实体词典完成候选实体召回；随后利用大语言模型选择目标实体并识别咨询意图，再围绕目标实体查询其属性、关系和相邻节点，构建局部知识子图；最后，将图谱信息转化为自然语言知识单元，经过去重和相关性排序，筛选高相关知识，为回答

生成提供准确、精简的知识依据。

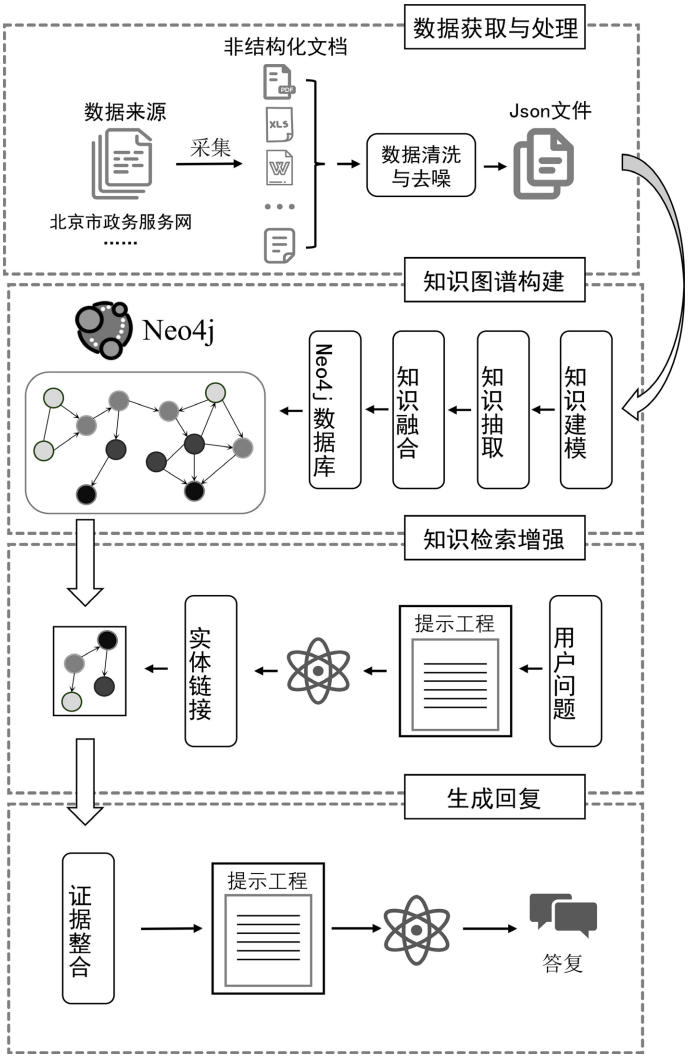


Figure 1. Overall architecture of the GOV_QA system
图 1. GOV_QA 系统总体架构

4) 回答生成

回答生成部分将召回得到的相关知识单元作为证据，与用户问题共同构建提示词并输入大语言模型。生成过程中，系统要求模型优先依据召回知识组织答案，对办理条件、申请材料、办理流程等信息进行归纳表达，从而减少无关内容和事实偏差，生成准确、连贯且符合用户需求的自然语言回答。

3. 系统关键技术

3.1. 政务知识图谱构建

3.1.1. 知识建模

知识建模是政务知识图谱构建的基础，其核心任务是确定图谱模式层中的实体类型、属性集合及关系模式。常用的知识建模方式包括自顶向下建模、自底向上建模和混合建模[17]。针对政务服务领域事项

类别繁多且数据复杂等特征，本研究采用自顶向下与自底向上相结合的混合知识建模方法。

在自顶向下建模过程中，结合群众办事咨询需求和政务服务业务流程，抽象出政务事项、政务部门、办理地点、申请材料和政策文件五类核心实体。其中，政务事项是知识图谱的中心实体，用于承载办事指南中的主要业务信息；政务部门表示事项的实施主体和管理机构；办理地点表示政务事项的线下办理场所及服务网点；申请材料表示办理事项时需要提交的证明或文件；政策文件表示事项办理所依据的法律法规和政策规范。在自底向上建模过程中，根据原始政务数据中的实际字段及内容，对实体属性和关联关系进行补充与调整。从而提高图谱模式与实际政务数据之间的一致性。根据建模结果，构建了以政务事项为核心的政务知识图谱模式层，其结构如图 2 所示。

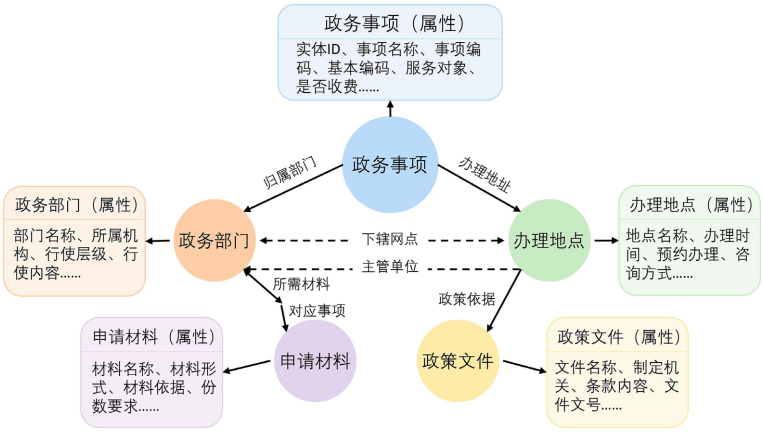


Figure 2. Government knowledge graph
图 2. 政务知识图谱

在政务知识图谱建模过程中，共定义了五类实体、七种关系以及七十四项属性。具体实体类型和关系见表 1、表 2。本研究采用结构化三元组形式表示政务服务知识图谱，定义如下：

$$KG=(E,R,S) \tag{1}$$

$$E=\{(e_1,e_2,\cdots,e_n),(A_1,A_2,\cdots,A_n)\} \tag{2}$$

$$R=(R_1,R_2,\cdots,R_n) \tag{3}$$

$$S=E\times R\times E \tag{4}$$

式(1)~(4)中，KG 表示政务服务知识图谱；E 表示带有属性的实体集合；R 为关系类型集合；S 为多个实体及关系组成的三元组集合；e_n 为第 n 个实体；A_n 为第 n 个实体的属性集合。

Table 1. Knowledge graph entity types
表 1. 知识图谱实体类型

实体类型	中文含义
Task	政务事项
Department	政务部门
Location	办理地点
Material	申请材料
Statute	政策文件

Table 2. Example table of relationship types in knowledge graph
表 2. 知识图谱关系类型示例表

主体	关系	客体
Task	归属部门	Department
Task	办理地址	Location
Task	所需材料	Material
Task	政策依据	Statute
Department	下辖网点	Location
Location	主管单位	Department
Material	对应事项	Task

3.1.2. 知识抽取

知识抽取的核心目标是从半结构化文件中提取实体、关系和属性特征，从而构建结构化知识图谱。本研究采用 LLM 自动抽取的知识抽取方法。通过设计政务领域结构化提示模板，约束模型仅依据原始文本提取信息，未出现的字段统一设置为空值，以减少知识补充和事实偏差。模型生成温度设置为 0，并限定输出格式为 JSON 对象，从而提高抽取结果的稳定性和规范性。

根据政务办事指南的内容特点，将抽取结果划分为基本信息、申请材料等模块。其中，基本信息包括事项名称、办理时间和办理地点等属性；申请材料包括材料名称、材料形式和受理标准等属性；最终，每个政务事项被转换为结构统一的 JSON 文件，为后续实体、属性和关系生成提供数据基础。

3.1.3. 知识融合

针对不同政务事项中存在的实体重复、字段格式不一致和属性缺失等问题，采用基于规则的知识融合方法。首先，压缩连续空白字符，将列表、字典和布尔值转换为统一的文本格式，清除办理地点中的详情提示、无效链接和 URL，并将其拆分为地点名称与办理地址。实体对齐仅在相同类型内进行，政务事项和政务部门分别以规范化事项名称、部门名称为对齐键，办理地点以地点名称和办理地址为组合键，申请材料和政策文件分别以名称和对应事项为组合键。对齐键相同的记录分配相同实体 ID 并合并，同名地点地址不同，或同名材料、文件对应事项不同时，保留为独立实体。

其次，按照字段类别解决属性冲突。同一实体首次写入的普通属性作为基础值，后续记录不直接覆盖，仅合并来源文件，办理地点的对应事项和主管部门采用去重并集。预约状态为空、“无”或“无需预约”，但办理时间中含预约要求时，统一校正为“部分需要预约”并保留预约说明。例如，多个政务事项均包含名称和地址相同的“北京市政务服务中心”，系统将其合并为同一办理地点实体，并对对应事项、主管部门进行去重汇总。对于收费属性，明确为“不收费”时清空与之冲突的收费项目、标准、金额、单位、依据及减免信息，明确为“收费”时过滤全空明细，并对各项收费信息进行汇总和去重。

最后，对实体关系进行融合与一致性校验。以头实体 ID、关系类型和尾实体 ID 作为关系去重键，重复关系仅保留一条，并合并其来源事项和来源文件。关系生成前过滤头尾实体为空的记录，导入时检查端点 ID 能否匹配相应类型的实体，并对重复关系、端点缺失、地点重复和收费字段冲突进行校验，最终形成规范化的实体与关系数据。

3.1.4. 知识存储

采用 Neo4j 图数据库对三元组信息进行存储。Neo4j 采用图结构组织复杂关联知识，并依托索引与缓存机制提高数据存储和查询效率，可实现大规模政务知识的关联存储、快速检索与可视化管理[18]。最终构建得到政务事项节点 1258 个、政务部门节点 55 个、办理地点节点 185 个、申请材料节点 10,094 个、政策文件节点 2853 个，共计 14,445 个实体节点和 31,144 条关系边，形成了以政务事项为核心，关联部

门、地点、材料和政策文件的政务知识图谱。Neo4j 图数据库的部分样例如图 3 所示。

Figure 3. Partial samples of Neo4j
图 3. Neo4j 部分样例

3.2. 分层知识召回方法

在政务问答场景中，用户问题通常存在表达口语化、事项名称不完整以及多个信息需求混合等情况。若直接基于原始问题进行图谱检索，容易召回无关节点，降低检索结果的准确性，同时增加大语言模型的上下文处理负担。针对上述问题，本研究提出分层知识召回方法，将知识召回划分为问题解析与实体定位、局部图谱知识组织和知识相关性筛选三个层次，逐层缩小候选知识范围，具体方法如图 4 所示。

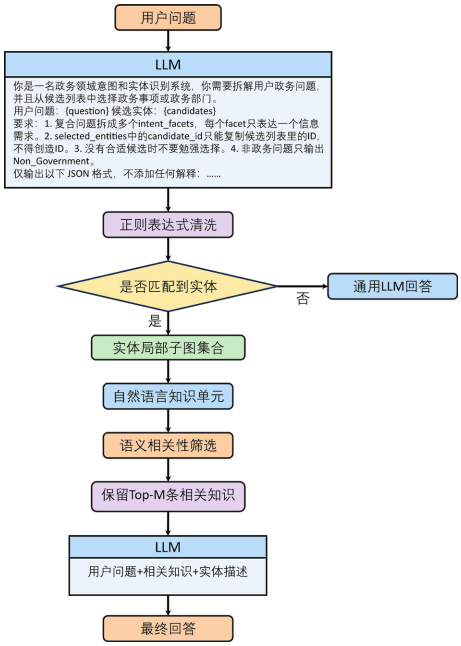


Figure 4. Knowledge recall method
图 4. 知识召回方法

在问题解析与实体定位层，首先对用户输入进行规范化处理，消除异常符号和冗余空格，随后利用统一实体词典，采用实体名称精确匹配、别名匹配、子串匹配及字符二元组 BM25 检索，从政务事项、政务部门实体类型中形成候选集合。获得候选实体后，通过结构化提示词引导 LLM 选择目标实体、识别咨询意图，并将复合问题拆分为单意图子问题。提示词限定模型只能选择候选列表中已有的实体 ID，并以 JSON 格式输出意图类型、目标实体及置信度。系统对模型输出进行清洗和实体 ID 校验，得到链接实体集合 E_q，首轮实体链接面向政务事项和政务部门，申请材料、办理地点和政策文件则通过关联关系进一步召回。非政务问题或未能匹配到图谱实体时，转入通用 LLM 问答分支。

在局部图谱知识组织层，系统以链接实体集合 E_q 中的实体 e 为中心，通过参数化 Cypher 语句查询其属性、关联关系及相邻节点，构建局部子图 $KG_e=(E_e,R_e,S_e)$。各实体对应的局部子图进一步合并为与用户问题相关的子图集合。

本研究采用意图感知的规则模板法将局部子图转换为自然语言知识单元。首先，根据实体类型统一关系方向，如将“申请材料－对应事项－政务事项”归一为“政务事项－所需材料－申请材料”；随后，依据问题意图选取非空字段。实体属性采用“{实体名称}：{字段 1}：{值 1}；{字段 2}：{值 2}”模板，关系采用“办理{事项名称}需要提交{材料名称}”“{事项名称}的政策依据包括{文件名称}”等模板，并附加相关限定属性；最后，清除空值、重复字段和重复事实，形成候选知识单元集合 $U_q=\{u_1,u_2,\cdots,u_n\}$。

在知识相关性筛选层，采用基于字符二元组的 BM25 算法计算意图子问题与各知识单元之间的相关性，其评分函数表示为：

$$S(q_j,u_i)=\sum_{t\in q_j} IDF(t)\frac{f(t,u_i)(k_1+1)}{f(t,u_i)+k_1\left(1-b+b\frac{|u_i|}{avgdl}\right)} \tag{5}$$

式(5)中，q_j 表示第 j 个意图子问题，u_i 表示第 i 个候选知识单元，$f(t,u_i)$ 表示字符二元组 t 在知识单元 u_i 中的出现频次，$|u_i|$ 表示知识单元长度，$avgdl$ 表示候选知识单元的平均长度，$IDF(t)$ 表示字符二元组 t 的逆文档频率。

对于每类问题意图，系统首先保留相关性最高的 7 个知识单元，并优先选取各意图排名首位的知识，保证复合问题中的不同信息需求均获得知识覆盖。其余知识单元统一排序，最终选取前 28 个知识单元构成问题证据集合。

最后，将用户问题、意图信息、链接实体、政务事项属性及筛选后的知识单元共同写入证据提示模板，由 LLM 依据图谱证据生成答案。

4. 实验与结果分析

4.1. 实验条件

实验的硬件环境为：CPU 为 Intel(R) Xeon(R)处理器，GPU 为 NVIDIA GeForce RTX 4090 (24 GB)，内存容量为 45 GB，操作系统为 Ubuntu 22.04 LTS 64 bit，深度学习框架为 PyTorch 2.1.0 (基于 CUDA 12.1)。在系统实现过程中，用户问题分析与最终答案生成均调用 DeepSeek-V4-Pro 大语言模型完成。

4.2. 数据集

基于北京市政务服务网中获得的政务数据，构建政务知识问答测试数据集。根据问题的语义表达和查询难度，将测试数据划分为简单事实、复杂多关系、模糊意图与口语化、边界抗压与防幻觉查询共四类，共包含 100 组政务问答数据。具体类别及数量如表 3 所示。

Table 3. Dataset categories and quantities
表 3. 数据集类别和数量

类别	问题特点	数量
简单事实查询	单一属性检索	25
复杂多关系查询	多属性与多关系联合查询	25
模糊意图与口语化	口语化表达识别	25
边界抗压与防幻觉	错误前提纠正	25

4.3. 知识图谱构建质量评估

本研究重点从实体抽取和关系抽取两个方面评估知识图谱构建中的知识抽取质量。从 1258 项政务事项中分层抽取 50 项作为评价样本，并以原始办事指南为依据，通过人工标注构建评价标准，共标注 354 个实体和 324 条关系。实体抽取结果仅在实体类型和规范化实体名称均与人工标准一致时判定为正确，关系抽取结果仅在头实体类型、头实体名称、关系类型、尾实体类型和尾实体名称均一致时判定为正确。

采用精确率(Precision)、召回率(Recall)和 F1 值评价知识抽取性能。其中，精确率表示正确抽取结果占模型全部抽取结果的比例，召回率表示被模型正确抽取的标准结果占人工标准全部结果的比例，F1 值为精确率和召回率的调和平均值。实体和关系抽取的评价结果见表 4。

Table 4. Knowledge extraction performance evaluation results
表 4. 知识抽取性能评价结果

抽取对象	Pre	Rec	F1
实体抽取	0.9124	0.9202	0.9163
关系抽取	0.8933	0.9043	0.8988

结果表明，实体抽取的精确率、召回率和 F1 值分别为 0.9124、0.9202 和 0.9163，关系抽取的三项指标分别为 0.8933、0.9043 和 0.8988。错误主要来源于同名实体的区分以及原文中隐含关系的识别。总体来看，知识抽取结果能够为后续知识融合、图谱检索和回答生成提供较为可靠的数据基础。

4.4. 系统评价指标

4.4.1. 客观评价指标

传统指标(如 BLEU、ROUGE 等)基于词汇重叠的评价指标难以准确衡量政务问答中表述不同但语义一致的回答。本研究采用 BERTScore [19]作为客观评价指标对模型性能进行评估。通过分别计算生成文本和参考文本在预训练模型 BERT 隐藏层表征的余弦相似度，得出精确率 Precision (Pre)、召回率 Recall (Rec)和调和平均值(F1)。

$$\mathrm{Pre}=\frac{1}{|X|}\sum_{x\in X}\max_{y\in Y}S(x,y) \tag{6}$$

$$\mathrm{Rec}=\frac{1}{|Y|}\sum_{y\in Y}\max_{x\in X}S(x,y) \tag{7}$$

$$S(x,y)=\frac{x\cdot y}{\|x\|\|y\|} \tag{8}$$

$$\mathrm{F1}=2\frac{\mathrm{Pre}\cdot\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{9}$$

式(6)~(9)中，x 和 y 分别为生成文本和参考文本中 token 的 BERT 嵌入向量；X 和 Y 分别为生成文本和参考文本中 token 向量的集合。

4.4.2. 主观评价指标

为了进一步验证系统的回答效用及实用价值，本研究联合多名政务知识专家拟定了评价标准，具体如表 5 所示，并对测试题进行主观评价打分。根据是否生成回答、是否理解用户问句、回答是否包含关键信息、内容是否完整、结果是否正确设计了五个评价维度，对各模型回答结果进行评分，分析其在实际工作的实用性。

Table 5. Scoring standards
表 5. 评分标准

评分标准	分值
未生成有效回答。	0
生成了回答，但未理解问题，内容与问题不符。	1
生成了回答，理解了问题，但回答结果错误。	2
生成了回答，理解了问题，但缺少关键信息，回答结果正确。	3
生成了回答，理解了问题，包含关键信息，回答结果正确，但内容不够完整。	4
生成了回答，理解了问题，关键信息完整且表述完善，回答结果正确，内容完整。	5

4.5. 实验结果分析

本研究通过严谨的对比实验对 GOV_QA 系统的性能进行全面评估。本研究选择了三个通用大模型基准系统：1) Doubao-Seed-Evolving；2) DeepSeek-V4-Pro；3) Qwen3.7-Plus。基于构建的测试集，对系统整体性能进行了系统性评估。

4.5.1. 客观评价结果

BERTScore 评价结果见表 6。GOV_QA 在精确率、召回率和 F1 值上均取得最高得分，分别达到 0.7600、0.8557 和 0.8025。与各项指标表现最优的通用模型相比，GOV_QA 的精确率、召回率和 F1 值分别提高 0.1193、0.1381 和 0.1272。其中，F1 值较 3 种基准模型的平均得分提高约 21.6%。较高的精确率表明，GOV_QA 生成的回答与标准答案具有更高的事实一致性；召回率的提升说明回答能够覆盖更多与问题相关的关键信息。实验结果表明，基于知识图谱的信息检索增强方法能够从知识图谱中获取与用户问题相关的知识证据，减少无关信息干扰，从而提高回答的准确性和完整性。

Table 6. BERTScore scoring results
表 6. BERTScore 得分结果

模型	Pre	Rec	F1
Doubao-Seed-Evolving	0.6149	0.6884	0.6483
DeepSeek-V4-Pro	0.6060	0.7176	0.6560
Qwen3.7-Plus	0.6407	0.7163	0.6753
GOV_QA	0.7600	0.8557	0.8025

4.5.2. 主观评价结果

主观评价结果如图 5 所示。GOV_QA 获得 5 分评价的样本占比达到 76%，4 分和 5 分样本合计占 81%；低分样本(0~2 分)仅占 7%，其中不存在未生成有效回答的 0 分样本。相比之下，Doubao、DeepSeek 和 Qwen 获得 5 分评价的样本占比分别为 8%、15%和 17%，低分样本占比分别为 51%、47%和 47%。从平均得分来看，Doubao、DeepSeek 和 Qwen 分别为 2.62、2.89 和 2.90 分，GOV_QA 达到 4.49 分，较表现最好的通用模型提高 1.59 分。说明通用模型虽然能够理解部分政务问题，但回答中仍容易出现关键信息缺失或内容不完整等情况。GOV_QA 通过引入知识图谱证据约束回答范围，使模型能够更准确地获取信息，提高了回答的事实可靠性与实际可用性。

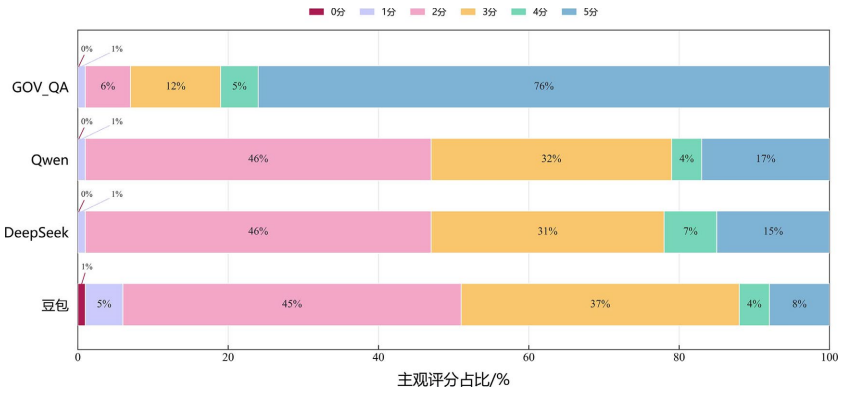
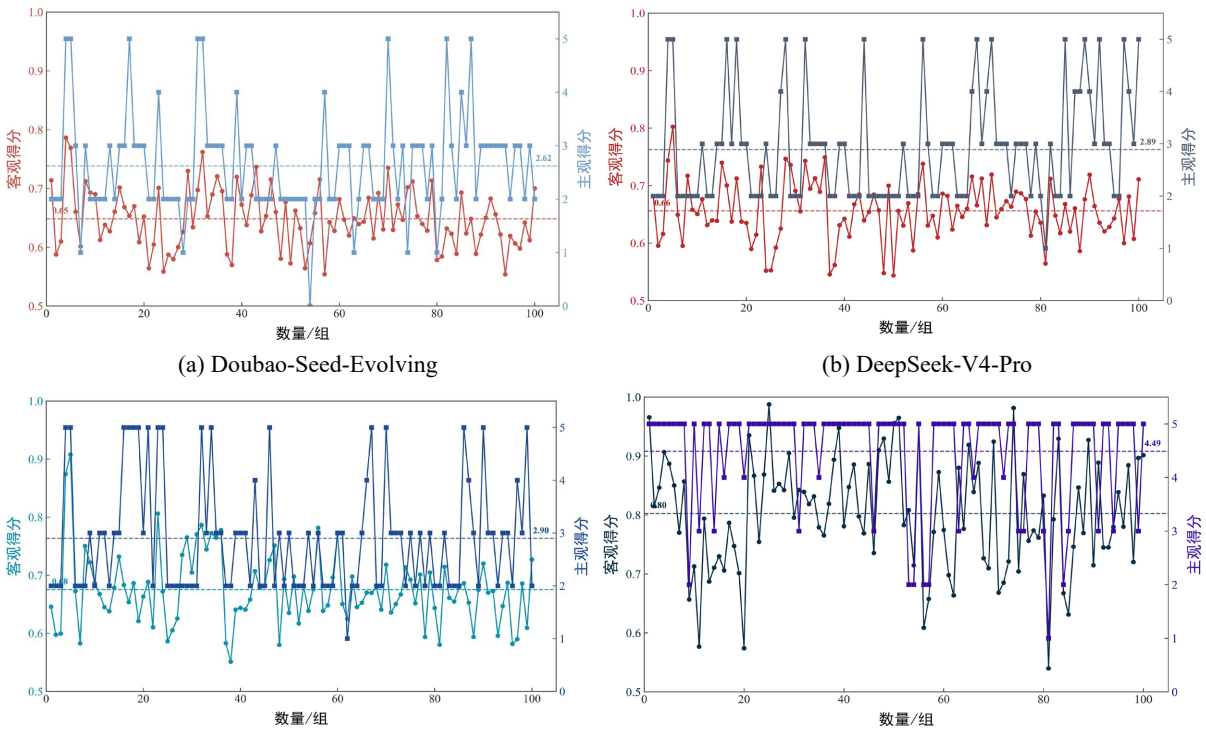


Figure 5. Subjective evaluation results
图 5. 主观评价结果



(a) Doubao-Seed-Evolving

(b) DeepSeek-V4-Pro

(c) Qwen3.7-Plus

(d) GOV_QA

Figure 6. Comprehensive comparative analysis results
图 6. 综合对比分析结果

4.5.3. 综合对比分析

各模型逐题客观得分与主观得分的综合对比结果如图 6 所示。Doubao、DeepSeek 和 Qwen 的 BERTScore F1 均值分别为 0.6483、0.6560 和 0.6753，主观评价均值分别为 2.62、2.89 和 2.90 分，整体表现较为接近。GOV_QA 的 F1 均值达到 0.8025，主观评价均值达到 4.49 分，在多数测试样本上均高于 3 种通用模型，且主客观评价总体呈现一致的变化趋势。GOV_QA 仍有少量样本获得 1~3 分评价，主要集中在复合多意图查询和错误前提纠正等问题上，说明系统在复杂实体匹配、知识覆盖范围和边界问题处理方面仍有改进空间。总体而言，实验结果验证了知识图谱与大语言模型的互补作用，以及分层知识召回方法在提升政务问答准确性、完整性和可靠性方面的有效性。

5. 总结

本研究针对政务问答中知识检索准确性不足、回答缺乏事实依据等问题，构建了融合政务知识图谱与大语言模型的 GOV_QA 系统，并提出分层知识召回方法，通过实体匹配、局部子图构建和相关知识筛选，为模型生成回答提供知识支撑。实验结果表明，GOV_QA 的 BERTScore 精确率、召回率和 F1 值分别达到 0.7600、0.8557 和 0.8025，主观评价平均得分为 4.49 分，整体性能优于通用大语言模型，验证了所提方法的有效性。未来将进一步扩充知识图谱的覆盖范围，提升系统对复合意图、模糊表达及知识边界问题的处理能力。

参考文献

[1] 杨迪丹. 新质生产力背景下的大模型与数据要素在政务服务问答图谱中的协同作用研究[J]. 中国信息界, 2025(1): 155-157.

[2] 徐晓林, 明承瀚, 陈涛. 数字政府环境下政务服务数据共享研究[J]. 行政论坛, 2018, 25(1):50-59.

[3] 许鹿, 黄未. 资产专用性: 政府跨部门数据共享困境的形成缘由[J]. 东岳论丛, 2021, 42(8): 126-135.

[4] 李晓方, 王友奎, 孟庆国. 政务服务智能化: 典型场景、价值质询和治理回应[J]. 电子政务, 2020(2): 2-10.

[5] 赵军, 曹鹏飞. 大语言模型——人工智能发展史上的里程碑[J]. 新兴科学和技术趋势, 2023, 2(1): 80-88.

[6] Brown, T.B., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners. 2020 *Advances in Neural Information Processing Systems* 33, Online, 6-12 December 2020, 1877-1901.

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., *et al*. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 24824-24837. https://doi.org/10.52202/068431-1800

[8] 汪波, 牛朝文. 从 ChatGPT 到 GovGPT: 生成式人工智能驱动的政务服务生态系统构建[J]. 电子政务, 2023(9): 25-38.

[9] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. and Wu, X. (2024) Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, **36**, 3580-3599. https://doi.org/10.1109/tkde.2024.3352100

[10] Ji, S., Pan, S., Cambria, E., Marttinen, P. and Yu, P.S. (2022) A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, **33**, 494-514. https://doi.org/10.1109/tnnls.2021.3070843

[11] 程序, 谭太龙, 王苗苗. PKS 体系下基于知识图谱的政务问答机器人研究[J]. 电子技术应用, 2023, 49(4): 128-132.

[12] Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X. and Wen, J. (2023) Complex Knowledge Base Question Answering: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 11196-11215. https://doi.org/10.1109/tkde.2022.3223858

[13] 王昀, 胡珉, 塔娜, 等. 大语言模型及其在政务领域的应用[J]. 清华大学学报(自然科学版), 2024, 64(4): 649-658.

[14] 董昌其, 李大宇, 米加宁. 大模型嵌入政务服务: 能力边界、协同治理与发展路径——基于地方政府大规模部署 DeepSeek 的观察[J]. 电子政务, 2025(8): 13-21.

[15] 沈思, 冯暑阳, 吴娜, 等. 融合大语言模型的政策文本检索增强生成研究[J]. 数据分析与知识发现, 2025, 9(9): 37-48.

[16] 孙雨生, 曾俊皓, 陈思妤, 等. 知识图谱增强的政策大模型知识问答系统构建研究[J]. 图书馆学研究, 2025(5): 66-75.

[17] Mao, S., Zhao, Y., Chen, J., Wang, B. and Tang, Y. (2020) Development of Process Safety Knowledge Graph: A Case Study on Delayed Coking Process. *Computers & Chemical Engineering*, **143**, Article 107094. https://doi.org/10.1016/j.compchemeng.2020.107094

[18] 沈建伟, 陈佳雯, 陈汉林, 等. 基于元数据语义增强的数据集知识图谱构建与应用[J]. 计算机科学, 2026, 53(S1): 62-71.

[19] Zhang, T.Y., Kishore, V., Wu, F., *et al*. (2020) BERTScore: Evaluating Text Generation with BERT. 2020 *International Conference on Learning Representations*, Addis Ababa, 26 April-1 May 2020, 1-41.