软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(4):1740−1758 [doi: 10.13328/j.cnki.jos.007435] [CSTR: 32375.14.jos.007435]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大语言模型驱动的可信政务问答技术*

王骞玥[1], 胡晋武[1,2], 王宇丰[1,2], 胡 宇[2], 高浩然[1], 邱舟强[3], 谭明奎[1]

[1](华南理工大学 软件学院, 广东 广州 510006)

[2](人工智能与数字经济广东省实验室 (广州), 广东 广州 510335)

[3](广东省科技创新监测研究中心, 广东 广州 510033)

通信作者: 谭明奎, E-mail: mingkuitan@scut.edu.cn

摘 要: 政务问答系统能实时处理政务咨询, 在降低人工咨询压力的同时提高企业和群众的办事效率. 政务问答系统的服务场景多样且重视回答表述的准确规范, 现有方法或基于预设知识库产生回答, 或基于规模有限的语言模型生成回答, 均无法在多服务场景下有效理解咨询并生成准确且可解释的可信回答. 为此, 提出一种基于大语言模型的政务问答技术以实现可信政务回答. 所提方法以政务大语言模型为内容理解和生成的核心模块, 并由分析引导模块和领域知识库模块辅助. 政务大语言模型生成咨询回答时参考分析引导模块提供的咨询分析结果和领域知识库模块提供的咨询相关领域知识, 并针对咨询生成内容表述与事实一致的准确回答. 生成回答时参考的信息可作为回答依据提升回答的可解释性. 为构建方法涉及的相关模块并测试其有效性, 收集并整理了一个包含多层次多粒度政务公开信息的综合性数据集, 其中包含 1 901 篇文档和 10 503 条问答对数据. 最后, 通过实验分析验证了基于该方法实现的原型系统能在多服务场景下针对用户咨询生成表述准确且可解释的可信咨询回答.

关键词: 智慧政务; 政务问答系统; 大语言模型; 知识增强; 意图识别

中图法分类号: TP391

中文引用格式: 王骞玥, 胡晋武, 王宇丰, 胡宇, 高浩然, 邱舟强, 谭明奎. 大语言模型驱动的可信政务问答技术. 软件学报, 2026, 37(4): 1740–1758. http://www.jos.org.cn/1000-9825/7435.htm

英文引用格式: Wang QY, Hu JW, Wang YF, Hu Y, Gao HR, Qiu ZQ, Tan MK. Trustworthy Government Q&A Technology Based on Large Language Model. Ruan Jian Xue Bao/Journal of Software, 2026, 37(4): 1740–1758 (in Chinese). http://www.jos.org.cn/1000-9825/7435.htm

Trustworthy Government Q&A Technology Based on Large Language Model

WANG Qian-Yue[1], HU Jin-Wu[1,2], WANG Yu-Feng[1,2], HU Yu[2], GAO Hao-Ran[1], QIU Zhou-Qiang[3], TAN Ming-Kui[1]

[1](School of Software Engineering, South China University of Technology, Guangzhou 510006, China)

[2](Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou 510335, China)

[3](Guangdong Science and Technology Innovation Monitoring and Research Center, Guangzhou 510033, China)

Abstract: The government Q&A system can handle user queries in real-time, improving the efficiency of businesses and the public, while reducing the pressure of manual consultation. However, the service scenarios of the government Q&A system are diverse and require accurate and standardized expression of answers. Existing methods, which either utilize preset knowledge bases to generate answers or language models with limited scale, are unable to effectively understand consultations and generate trustworthy answers that are accurate and interpretable across multiple service scenarios. Therefore, this study proposes a government Q&A system based on a large language model to provide trustworthy government responses. The method employs a large language model specific to government service as the core module for content understanding and answer generation, assisted by an analysis guidance module and a domain knowledge base

* 基金项目: 国家自然科学基金 (62072190)
王骞玥和胡晋武为共同第一作者.
收稿时间: 2024-07-06; 修改时间: 2024-11-16, 2025-03-05; 采用时间: 2025-03-25; jos 在线出版时间: 2025-10-29
CNKI 网络首发时间: 2025-10-30

module. When generating answers, the large language model references the consulting analysis results provided by the analysis guidance module and the domain knowledge offered by the domain knowledge base module to produce answers that are accurate and consistent with the facts. The reference information during answer generation serves as a foundation to enhance the interpretability of the answers. A comprehensive dataset, containing multi-level and multi-granularity government public information, is collected and organized to construct the modules involved in the method and to test their effectiveness. This dataset includes 1 901 documents and 10 503 question-answer pairs. Finally, experiments verify that the prototype system, implemented based on the proposed method, can generate accurate and interpretable answers for user inquiries in multiple service scenarios, proving the effectiveness of each module in the system.

Key words: intelligent government; government Q&A system; large language model (LLM); knowledge enhancement; intention recognition

随着政务服务网络化和数字政府建设, 我国在线政务服务咨询呈现出用户数量增大、咨询话题增广、咨询内容复杂化的趋势. 2023 年, 全国政务咨询超 6.2 万条 (https://doi.org/10.28655/n.cnki.nrmrb.2024.015440). 政务问答系统能提供全时段、全面且实时的政务咨询服务, 能在便利企业及群众办事的同时减轻人工咨询服务压力, 因此其在改善政务咨询服务中具有价值和潜力.

政务问答系统作为政府机构重要的信息发布窗口, 要求其回答可信. 然而, 相较于其他知识更新周期长, 使用人群相对固定的可信问答系统, 如医疗问答系统[1]和客服问答系统[2], 政务问答系统在政策、案例和规则上均有独特规律和可用性要求 (如 GB/T 44230-2024 政务信息基本要求). 具体来说, 政策更新快要求政务问答系统的回答有时效性; 用户核心咨询意图模糊要求政务问答系统能有效理解非政务语言组织规范的用户咨询; 政务咨询回答有特定规则, 要求政务问答系统产生的回答内容规范严谨.

目前政务问答系统的构建方法主要有两类. 一类基于专家预设的问答库[3]提供咨询回答. 此类方法从问答库中匹配与咨询最接近的问题并将问题对应的回答作为系统对咨询的回答. 此类问答系统给出的回答虽然内容组织规范, 但其针对性受限于咨询语言组织的规范性和预设问答对的完备性. 因此, 这类问答系统在面对非规范咨询时常有“答非所问”情况[4], 如图 1 的左侧模块示例, 面对用户提出的非规范咨询时, 系统未能提供教育资金补贴的内容. 另一类政务问答系统基于规模有限的语言模型[5]. 此类方法能针对用户咨询生成对应回答, 提升回答的针对性. 但此类系统生成回答的过程是黑盒的, 且缺乏依据, 导致回答内容的准确性和可解释性不足. 如图 1 中间模块示例, 系统虽能针对咨询生成回答, 但内容无据可依, 表述正确性无法验证, 导致回答可信度降低, 且此类系统在多服务场景下对咨询的理解受限于咨询语言组织的规范性[5]. 因此, 政务问答系统在多服务场景下针对用户咨询生成准确且可解释的可信回答的能力仍有待提升.

针对前述问题, 本文提出一种基于大语言模型的可信政务问答技术, 如图 1 右侧模块示例. 本文方法将政务大语言模型作为系统核心的内容理解和生成模块, 其具备理解简单规范政务咨询和生成表述符合政务咨询规范的回答的能力. 本文方法有效解决前人方法对非规范化咨询理解不足的问题, 加强大语言模型在多服务场景下对咨询内容和咨询意图的有效理解, 保障多场景适应能力. 同时, 方法中包含可低成本更新的领域知识库, 其能根据前述对咨询的分析结果提供咨询相关领域知识. 保障多服务场景下领域知识检索的有效性. 政务大语言模型生成回答时综合参考分析结果和提供的咨询相关领域知识, 确保在有效理解多服务场景中用户咨询的前提下, 针对用户咨询生成表述与事实一致的准确回答. 同时, 参考信息可作为回答依据, 提升回答的可解释性. 通过上述分析和生成过程, 方法实现了多服务场景下针对用户咨询生成准确且可解释的可信咨询回答[4]. 相较前人提出的政务问答方法, 本文方法兼顾回答的针对性和表述的正确性. 同时, 本文方法通过不同模块的协同合作, 实现在多服务场景下生成准确且可解释的可信回答. 本文的主要贡献如下.

(1) 针对当前政务问答系统生成回答可信度不足的问题, 本文提出了基于政务大语言模型的可信政务问答技术. 该方法通过加强政务大语言模型在多服务场景中对咨询的理解并在生成过程中提供咨询相关领域知识, 实现在多服务场景中对用户咨询生成表述准确规范且可解释的可信回答.

(2) 针对政务公开信息数据集缺乏问题, 本文从多源官方信息公开平台上收集并整理了一个包含多层次多粒度政务公开信息的综合性数据集 (https://doi.org/10.57760/sciencedb.24847). 该数据集包含 1 901 篇文档和 10 503 条问答对, 用于微调政务大语言模型和构建领域知识库.

(3) 多服务场景下的方法对比实验结果和生成案例分析证明, 基于本文方法构建的原型系统能在不同服务场

景下针对用户咨询生成内容表述准确规范且可解释的可信回答. 同时, 各服务场景下的消融实验结果证明了系统各组成模块的有效性.

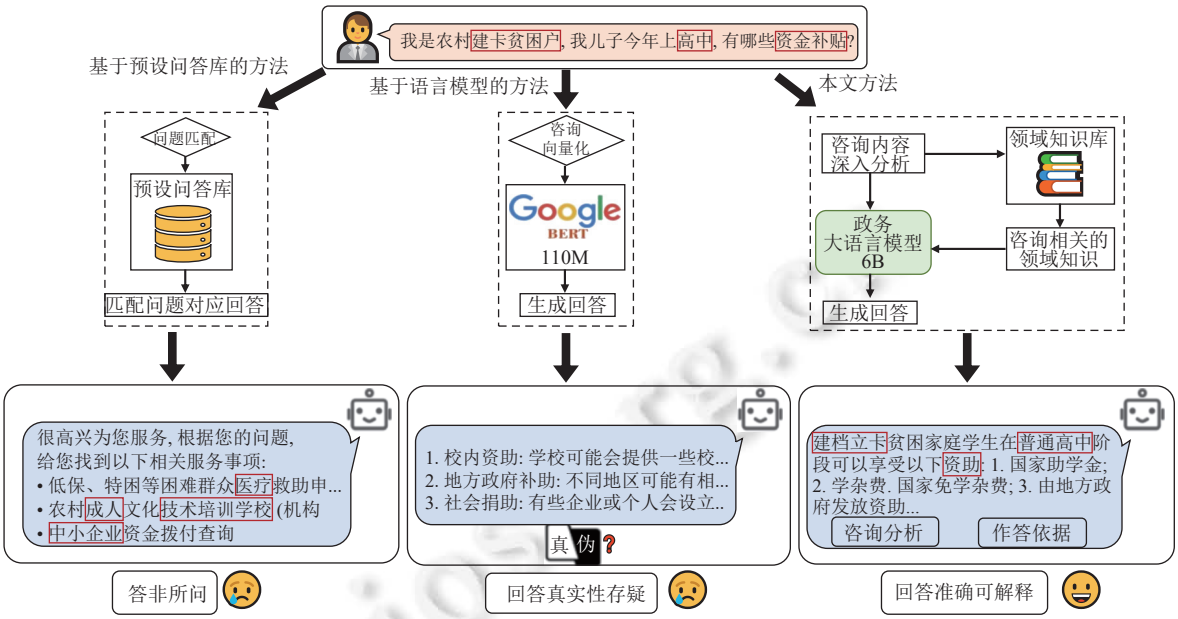

图 1 各类政务问答系统回答效果对比

本文第 1 节介绍相关工作. 第 2 节进一步阐述本文解决的问题并概述本文提出的基于政务大语言模型的可信政务问答技术. 第 3 节详细介绍可信政务问答技术. 第 4 节介绍政务公开信息数据集的收集整理细节. 第 5 节通过定量实验和生成案例分析证明基于本文方法实现的原型系统及其组成模块的有效性. 第 6 节进一步讨论了方法在多服务场景中的可用性, 总结了方法的局限性和改进方向.

1 相关工作

1.1 政务问答技术

政务问答系统是问答系统在政务领域的应用[6]. 王芳等人[7]对我国 30 个省级政府网站搭载的政务问答系统进行了实际测评, 认为当前国内政务问答系统普遍缺乏对咨询意图的深入理解和场景化服务能力. 目前, 政务问答系统的构建技术主要分两类.

一类工作关注政务问答系统对咨询的有效理解. Liang 等人[8]采用了基于三向决策 (three-way decision, TWD) 的两阶段意图识别方法来提高政务问答系统对不常见咨询的理解能力. Gao 等人[3]以大语言模型作为系统理解咨询的工具, 并将其集成到现有的政务问答系统中. 耿云飞[9]提出一种基于 BERT 和自注意力机制的意图识别方法, 以提升政务问答系统意图识别准确率. 此类工作虽然一定程度上提升了系统对用户咨询的理解程度, 但仍依赖预设问答库检索匹配产生回答. 虽然保障了回答表述的正确性, 但咨询针对性不足, 常出现“答非所问”现象. 此类方法产生的回答准确性不足, 无法满足政务问答系统对可信回答的要求.

另一类工作关注政务问答系统回答的咨询针对性. 这类方法基于语言模型理解咨询并生成用户咨询回答. 程序等人[5]在 PKS 体系下设计实现了基于知识图谱和 BERT 的政务问答系统, 该系统依托 BERT 实现咨询理解和回复生成, 并基于知识图谱检索结果生成针对性回复. Fang 等人[10]使用 Qwen 模型参考现有的政务问答库中关联咨询的问答对生成针对用户咨询的回答. 此类方法虽然能提升回答的针对性, 但没有考虑政务问答系统对场景化服务的要求. 因此, 此类方法的回答效果受限于咨询的语言组织是否符合领域规范. 同时, 此类系统回答由黑盒语

言模型产生且回答依据不足, 导致回答的可解释性不足.

综上所述, 目前政务问答系统的构建技术仍无法使政务问答系统生成可信回答. 在本文方法中, 分析引导模块的多步分析可以提升方法对咨询的理解, 增强回答的针对性. 同时, 本文方法还可在多服务场景下生成准确且可解释的回答. 通过对咨询进行多维分析和参考材料检索, 保障在多服务场景下, 政务大语言模型能在理解咨询的前提下生成表述正确的准确回答. 生成回答时参考的内容可作为回答依据, 提升回答的可解释性.

1.2 可信问答技术

Liu 等人[11]从 7 个维度定义大语言模型回答的可信, 即可靠性 (reliability)、安全性 (safety)、公平性 (fairness)、抵抗滥用 (resistance to misuse)、解释性和推理 (explainability & reasoning)、遵循社会规范 (social norm) 和稳健性 (robustness), 并指出通常情况下, 回答的可信度与回答实用性正相关. 本文从政务问答系统生成回答的可靠性、可解释性和稳定性这 3 方面提升回答可信度. 以下将分别介绍关注大语言模型这 3 种特性的相关工作.

回答的可靠性, 是对信息、事实和结果的准确表示[11]. 研究人员主要从控制大语言模型推理阶段的行为来加强回答的可靠性. Dhuliawala 等人[12]提出了 CoV (chain of verification) 框架通过大语言模型自证回答正确的方式加强最终回答表述的正确性. Wei 等人[13]提出思维链 (chain of thought, CoT) 通过大语言模型自主分析并回答的方式, 增加回答的准确性. Wen 等人[14]提出的 MindMap 方法使得大语言模型能够生成思维导图, 基于思维导图进行推理并最终总结得出回答.

回答的稳健性体现系统在各种情况下保持其性能水平的能力[11]. 研究人员在推理阶段主要基于提示学习, 通过加强大语言模型对同一事物不同表述的对齐能力, 以提升其回答的稳健性. Jiang 等人[15]证明了使用合适的提示加强回答约束, 有效增强语言模型对同一事物不同表述的对齐能力. Zhou 等人[16]还提出了根据问题难度使用不同提示的方法, 增强回答稳健性.

回答的可解释性, 用于说明生成回答是否有理有据[11]. 为使回答有据可依, 常用的方法是通过知识库提供语言模型生成时的回答参考并将其作为生成回答依据. 知识库的形式可以是知识图谱[14]或向量数据[17]. Guu 等人[18], 将这类在大语言模型推理时根据外接知识库提供参考的方法, 总结为检索增强语言建模 (retrieval-augmented language modeling, RALM).

本文方法综合考虑回答的可靠性、稳健性和可解释性, 通过加强这 3 方面的特性来提升回答的可信度. 首先, 领域知识库在政务大语言模型生成回答时提供咨询相关领域的知识, 这使得政务大语言模型能够参考领域知识的语言组织规范和事实, 生成表述规范且与事实一致的可靠回答. 其次, 该方法对咨询内容进行多步主动分析, 通过提取并转述咨询关键信息的方式, 加强了对不同语言组织方式下咨询内容的理解, 保障多服务场景下政务大语言模型生成回答的针对性. 领域知识库模块提供的咨询相关领域知识, 结合咨询分析结果, 共同作为政务大语言模型针对咨询生成回答时的参考材料, 使得政务大语言模型能够针对咨询生成对应表述正确的准确回答, 保障了回答的可靠性. 同时, 参考材料 (包括分析结果和检索知识) 作为回答的依据, 可提升回答的可解释性. 分析过程中对非规范咨询的关键信息进行转述, 保障了政务大语言模型和领域知识库模块在多服务场景下的稳定性, 实现了在多服务场景下生成准确且可解释的回答, 保障了回答的稳健性.

2 问题定义及方法概述

2.1 问题定义

政务问答系统既能有效缓解政务咨询专员的工作压力, 又能提供全天候、多渠道、全方位的智能政务服务. 政务问答系统 S 对于用户咨询内容 X, 会给出回复 A. 整个过程的形式化表示为:

$$A = S(X) \tag{1}$$

由于政务问答系统面对的咨询对象不是特定群体, 因此咨询的语言组织规范性无法保证, 即输入内容的分布与政务大语言模型的训练数据存在偏差[19]. 记输入内容为 $\{X\}_{i=1}^{n} \sim P^n$, 领域模型的训练数据为 $\{Y\}_{i=1}^{n} \sim Q^n$. 其中 P^n 和 Q^n 表示不同分布. 大语言模型独立理解训练分布外数据的能力有限, 因此容易出现答非所问的现象.

对于政务问答系统 S, 用户咨询 X 和咨询对应的参考回答记为 $A_{\text{ground\_truth}}$. 系统产生的回答 $A_{\text{prediction}}$ 形式化表示如公式 (2) 所示. 本文的目标是获得优化后的政务问答系统 S_{opt}, 使得 $A_{\text{prediction}}$ 在语义和表述规范上更接近 $A_{\text{ground\_truth}}$. 其形式化表达如公式 (3) 所示.

$$A_{\text{prediction}} = S_{\text{opt}}(X) \tag{2}$$

$$\underset{S_{\text{opt}}}{\operatorname{argmin}}\ dist(A_{\text{prediction}}, A_{\text{ground\_truth}}) \tag{3}$$

2.2 可信政务问答技术概述

本文提出的可信政务问答技术如图 2 所示. 该方法主要包含 3 个模块: 政务大语言模型、分析引导模块和领域知识库模块. 政务大语言模型是系统的核心, 负责咨询内容的理解和回答的生成. 分析引导模块通过指令与政务大语言模型交互, 实现对咨询内容的逐步分析. 领域知识库模块由领域知识库和基于关键词的知识检索方法组成, 为政务大语言模型提供咨询相关的领域知识.

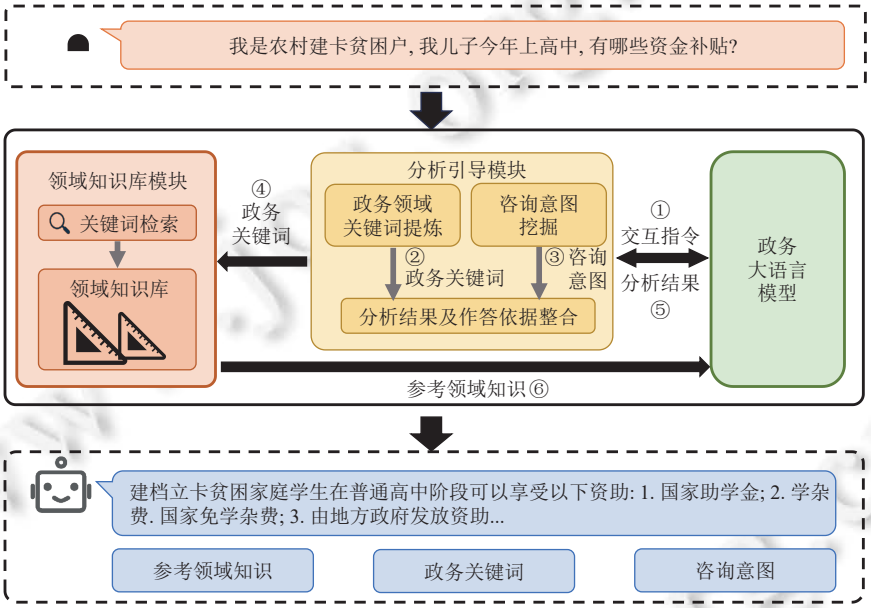

图 2 可信政务问答技术示意图

对于用户提出的政务咨询, 分析引导模块首先根据用户咨询构建交互指令. 通过指令与政务大语言模型多步交互, 提取并规范咨询内容中涉及的领域关键词和咨询意图, 提升对咨询内容的理解. 同时, 分析结果作为政务大语言模型生成回答时的参考, 确保了在多服务场景下生成回答的针对性. 领域知识库模块根据前述分析结果 (即咨询相关领域关键词) 检索相应领域知识, 提供给大语言模型作为生成回答的参考材料, 保障了生成回答表述的正确性. 政务大语言模型结合分析和检索结果, 实现在理解政务咨询的前提下针对咨询生成内容表述正确的准确回答. 同时, 回答生成时的参考材料作为回答依据, 增强回答的可解释性. 综上, 方法基于预设的分析和回答步骤生成回答, 提升了回答过程的透明度, 用户在获得咨询回答的同时, 还可以参考各中间步骤产生的分析和检索结果, 判断回答的可用性.

3 基于大语言模型的可信政务问答技术

3.1 政务大语言模型

通识大语言模型通常无法独立支撑特定领域对话[20], 表现为无法有效理解领域提问并生成符合领域规范的

回答. 相较于医疗问答更偏向于症结判断[1], 客服问答注重解决问题和客户满意度[2], 政务问答着重提供解决方式或办事流程并要求回答表述的规范和严谨性, 符合政务信息系统基本要求 (GB/T 44230-2024).

因此, 本文构建了政务大语言模型, 该模型具备政务背景知识, 能理解简单的咨询问题并生成符合政务语言组织规范的回答. 政务大语言模型是本文系统内容理解和回答生成的核心支撑模块. 本节将说明政务大语言模型的构建细节.

政务大语言模型由开源通识大语言模型 ChatGLM2-6B[21]通过指令微调构建. 本文采用低秩自适应 (low rank adaptation, LoRA) 方法[22]来降低微调成本并保证微调效果. LoRA 方法在大语言模型外新增旁路, 并通过调整旁路中的矩阵权重实现高效参数微调. 旁路由降维矩阵 A 和升维矩阵 B 组成. 训练前, 用全 0 矩阵初始化降维矩阵 A, 用随机高斯分布初始化升维矩阵 B, 该操作能保证训练开始时, 旁路矩阵的作用等价于一个全 0 矩阵, 不改变整个模型的输出. 训练时, 针对输入特征向量, 旁路先降低输入特征向量的维度再对其升维, 以模拟模型参数的更新, 实现低秩近似.

LoRA 方法的优势在于只调整较小的旁路矩阵参数, 可以显著减少训练时的计算资源和时间. 同时, LoRA 方法能够在不牺牲太多语言模型性能的情况下, 有效地对模型进行定制化调整, 使其更好地适应特定的任务或特定领域. 微调任务中大语言模型参数可符号化如公式 (4) 所示.

$$W_0+\Delta W \tag{4}$$

LoRA 方法微调时, 参数更新如公式 (5) 所示.

$$W_0+\Delta W=W_0+BA,\ B\in R^{d\times r},\ A\in R^{r\times k} \tag{5}$$

其中, $W_0\in R^{d\times k}$ (d 为上一层输出维度, k 为下一层输入维度) 表示预训练大语言模型的权重参数矩阵, ΔW 表示需要更新的旁路网络, 秩 $r\ll\min(d,k)$. r 的表示如公式 (6) 所示.

$$r=W_0x+\Delta Wx=W_0x+BAx \tag{6}$$

完成训练后, 将 LoRA 更新后网络与基础模型合并以构建政务大语言模型, 加快模型推理速度. 合并后的模型权重可以表示为:

$$W_{\text{merge}}=W_0+A\cdot B \tag{7}$$

本文通过将基础模型与 LoRA 微调部分合并构建的政务大语言模型在保留基础模型本身的语言理解和生成能力的同时, 还具备政务领域的对话能力. 构建的政务大语言模型, 能直接理解简单明确的政务服务咨询提问并针对提问生成符合政务语言组织规范的咨询回复, 是本文系统理解咨询内容和生成规范化回复的关键模块. 关于政务大语言模型微调数据的来源及构建细节在本文的第 4 节详述, 关于政务大语言模型微调构建使用的训练数据量讨论及微调方法的超参数设置在本文的第 5 节详述.

3.2 分析引导模块

相较于医疗问答和客服问答的用户咨询聚焦于病症或产品本身, 政务问答系统的咨询核心需求隐含在用户咨询的表述中[23]. 政务问答系统的服务对象知识结构多样且语言表达的准确性和规范性无法保证, 这增加了咨询的复杂度[4]. 同时, 政务问答系统有两个典型的使用场景: 专家政务信息查询和市民政务服务咨询. 在专家政务信息查询场景中, 系统面向领域内人员, 根据简练规范的提问, 提供政务公开信息的查询服务. 在市民政务服务咨询场景中, 系统面向市民提供政务服务相关的办事建议或答疑.

目前, 政务问答方法主要基于检索或通过小规模语言模型直接给出回答. 然而, 这些方法对咨询的分析深度不足, 且在多服务场景中的非规范咨询进一步阻碍了对咨询内容的有效理解. 咨询理解深度的局限影响了回答的准确性及这些方法在多服务场景中的稳定性, 最终导致“答非所问”的现象[2].

为提升在多服务场景下对咨询内容的有效理解, 如图 3 所示, 政务咨询专员在处理政务咨询时, 通常会遵循特定顺序的思考过程 (A, B, C). 具体而言, 政务咨询专员通过对咨询进行多步分析, 能在面对非规范化语言组织的咨

询时, 从咨询内容中提取并以规范语言转述出政务领域关键词和咨询意图, 确保了对用户咨询的有效理解, 从而缓解多服务场景中用户对系统理解造成的偏差.

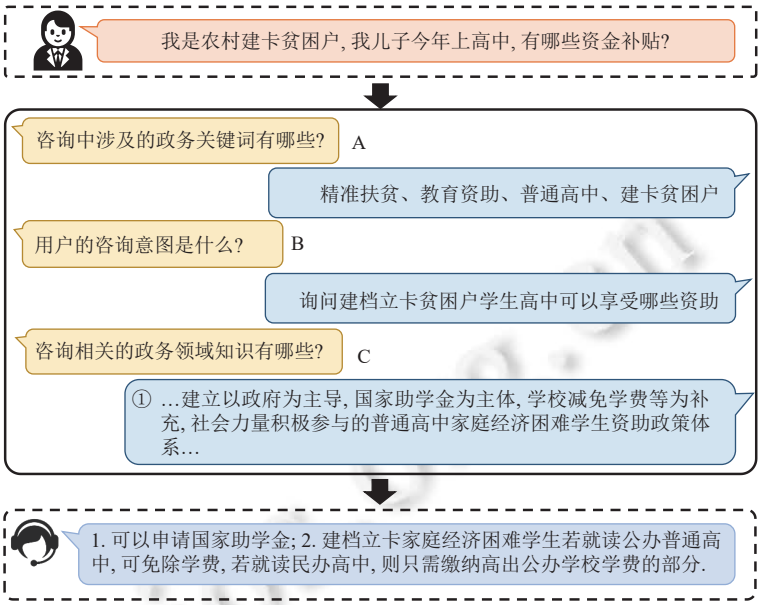

图 3 政务咨询专员面对咨询时的思考过程

受到上述政务咨询专员对咨询分析过程的启发, 方法设计了分析引导模块, 通过与政务大语言模型交互, 引导模型逐步模拟专员的分析过程, 以提取用户咨询中的关键信息 (包括咨询意图和政务领域的关键词). 通过多步预设分析步骤, 可提升对组织规范咨询的有效理解. 多步分析结果作用于政务大语言模型回答生成时, 保障大语言模型对咨询的有效理解, 提升大语言模型生成回答的针对性并为回答的准确性奠定基础, 有效克服了政务问答方法因对咨询理解不足而出现的“答非所问”的现象. 对于不符合政务领域组织规范的咨询, 分析引导模块依托政务大语言模型的语言理解能力和政务语言标准化生成优势对其提取的分析结果进行规范化转述, 使得该模块的输出结果均以领域规范化表达呈现. 因此, 在面对多服务场景下各类语言组织形式的咨询时, 分析引导模块能对提取的政务关键信息进行规范化转述, 保障系统其他模块的稳定运行, 提升了方法对多服务场景的适应性.

考虑到政策和文件的更新与丰富周期短, 分析引导模块需要针对新增服务场景和新增咨询类型及时进行高效适应. 为了保障各分析任务的完成效果, 分析引导模块与具有政务背景知识和语言理解与生成能力的政务大语言模型交互时使用上下文学习策略, 在交互指令中包含任务相关案例, 能进一步确保分析任务顺利完成并提高分析结果的质量. 同时, 在服务场景和咨询类型更新丰富时, 可由领域专家增加典型案例, 实现对更新的快速适应. 以下按照咨询专员对咨询内容的分析顺序, 介绍该模块中指令模板的设计细节.

(1) 咨询相关政务关键词的凝练. 针对该步骤设计一个提示模板 (prompt template), 记作 Keyword Template. 该模板引导政务大语言模型从用户咨询中凝练涉及的政务关键词并在必要时进行规范化转述, 结果以符合领域语言组织规范的表述输出. Keyword Template 的具体内容如图 4(a) 所示.

(2) 用户咨询意图的挖掘. 这一步主要引导模型关注用户的咨询意图. 针对该步骤设计一个提示模板, 并将其记作 Intention Template. 该模板引导政务大语言模型分析用户的咨询目标. Intention Template 的具体内容如图 4(b) 所示.

(3) 分析与检索结果整合. 针对该步骤设计一个提示模板, 并将其记作 Answer Template. 在完成咨询相关的领域知识检索后, 该模板引导大语言模型整合并参考前述分析结果和检索到的领域知识, 就咨询问题生成回答. Answer Template 的具体内容如图 4(c) 所示.

Keyword Template

<指令>你的任务是从用户输入中总结凝练3–5个政务相关关键词.
注意, 严格基于用户输入凝练政务关键词, 关键词简洁并符合政务领域要求, 回答格式为: 关键词1, 关键词2…
以下是一些用于参考的例子:
城市一直都在变得更好 凝练为 稳步高质量发展,
让大家过得安稳保障社会治安 或 让大家过得开心 凝练为 提高人们生活水平,
户籍是湖北, 人目前在浙江 凝练为 异地,
以下是你需要凝练关键词的句子: </指令>
<用户输入>{{query}}</用户输入>

(a) Keyword Template内容

Intention Template

<指令>你的任务是从政务服务角度, 以关键词形式, 根据用户咨询简要总结出用户咨询目标.
注意, 回答只能包含1–2个凝练的咨询目标. 不详细展开!
回答的参考样例: 问政府措施, 问异地办证建议, 问医保报销比例, 了解天津市现代服务业的发展措施等等.
以下是你要总结咨询目标的句子: </指令>
<用户咨询>{{query}}</用户咨询>

(b) Intention Template内容

Answer Template

<指令>你的任务是结合给出的政务背景知识并严格参考提供信息, 详细地回答用户咨询.
提供给你的参考信息有: 问题相关已知信息、咨询目标、咨询涉及的政务领域关键词
注意, 绝对不要出现和咨询内容不相关的信息.
在回答中不要出现'根据已知信息'等类似表述. 以下是提供的参考信息和你要回答的咨询: </指令>
<问题相关已知信息>{{context}}</问题相关已知信息>
<咨询目标>{{intention}}</咨询目标>
<政务领域关键词>{{orders}}</政务领域关键词>
<用户输入>{{query}}</用户输入>

(c) Answer Template内容

图 4　分析引导模块中的指令模板设计细节

3.3　领域知识库模块

不同于其他可信问答系统的背景知识更新周期长, 政务问答系统需要及时响应政策的不断变化与丰富, 在实现回答内容可靠性的同时, 保障回答时效性, 符合政务信息系统基本要求 (GB/T 44230-2024). 然而, 回答由黑盒模型生成, 缺乏可解释性. 因此, 仅依赖大语言模型生成回答, 可靠性无法保证. 同时大语言模型本身的知识更新成本对计算资源和数据资源的要求高, 无法适应政务问答系统对回答时效性的要求.

为解决上述问题, 本文方法包含领域知识库模块. 领域知识库以向量的形式存储收集整理的各类政务文档. 领域知识检索方法基于分析引导模块分析的领域关键词检索咨询相关领域知识. 在政务大语言模型生成回答时, 参考检索到的政务领域知识生成回答, 保障了回答内容表述的可靠性. 该模块提供的咨询相关领域知识与分析模块提供的分析结果可作为大模型生成回答的依据, 保障回答的可解释性以进一步提升回答的可信度.

同时, 领域知识库具有白盒化程度高、维护成本低的优良特性. 政策的变化与丰富可通过对向量知识库的内向量的检索替换和直接扩展实现, 通过向政务大语言模型提供咨询相关的最新政务领域知识, 使得方法生成的咨询回答包含最新信息, 以达到政务问答系统的时效性要求. 以下将详细介绍领域知识库的构建细节和检索方法.

领域知识库以向量形式存储收集整理各类政务文档. 对于一份待存入领域知识库的政务文档, 其将经历以下过程: 首先将文档按照固定长度拆分. 本文设定拆分长度为 250 字, 文段间的重叠长度为 50 字. 然后将拆分文段通过 m3e-base 计算其嵌入表示. 将拆分的文段和对应的嵌入表示存储到领域知识库中, 完成领域知识库的构建. 领域知识库构建具体步骤可以进行如公式 (8) 所示的符号化表示.

$$D \xrightarrow{\text{split}} \{T_i\}_{i=1}^{n} \xrightarrow{\text{embedding}} \{E_i\}_{i=1}^{n} \xrightarrow{\text{store}} K \tag{8}$$

其中, D 表示收集整理的政务公开文档集合. split 操作将上传的文档按照长度进行拆分. T_i 表示拆分的文段, E_i 表示文段的嵌入式表达, store 表示存储拆分文段和对应嵌入式表达, K 表示领域知识库. K 的表示见公式 (9).

$$K = \{(T_i, E_i)\}_{i=1}^{n} \tag{9}$$

政务问答系统的服务场景多样, 不同场景下咨询语言组织的规范性无法保证, 因此, 咨询表述与领域知识库中的内容存在分布偏移[16]. 这会导致基于相似度检索相关内容的成功率降低, 进而降低回答的准确性和可解释性甚至出现“答非所问”的现象[15]. 为克服咨询语言不规范对检索成功率的影响, 领域知识库基于分析引导模块提供的政务关键词进行咨询相关领域知识的检索. 分析结果中的关键词与咨询内容相关且表述符合政务领域语言组织规范, 能保障在语言组织规范性未知的多服务场景中检索到咨询相关领域知识的成功率.

具体来说, 本文通过计算领域知识库中知识与关键词的嵌入式表达的余弦相似度找出关键词相关的领域知识. 对于一组分析得到的政务关键词 $Keyword$ (可按公式 (10) 表示), 首先将关键词以逗号分隔后拼接成一个文本形式的输入 $Input$ (可按公式 (11) 表示). $Input$ 通过嵌入式表达后将得到的输入向量 I, 和知识库中存储文段的嵌入式表达 E_i 进行余弦相似度计算得 Sim_i, 计算方法如公式 (12) 所示. 找出相似度最高的前 k ($k = 3$) 条文本作为政务大语言模型生成回答时参考的咨询相关领域知识. 领域知识库使用的文本数据的来源在第 4 节详述.

$$Keyword = \{k_1, k_2, \ldots, k_m\} \tag{10}$$

$$Input = k_1, k_2, \ldots, k_m \tag{11}$$

$$Sim_i = \cos(\theta) = \frac{I \cdot E_i}{\|I\| \times \|E_i\|} \tag{12}$$

3.4 方法总结

本节以一个政务咨询为例, 总结方法的基本运行流程, 直观展示方法各模块的作用及关联, 处理流程如图 5 所示.

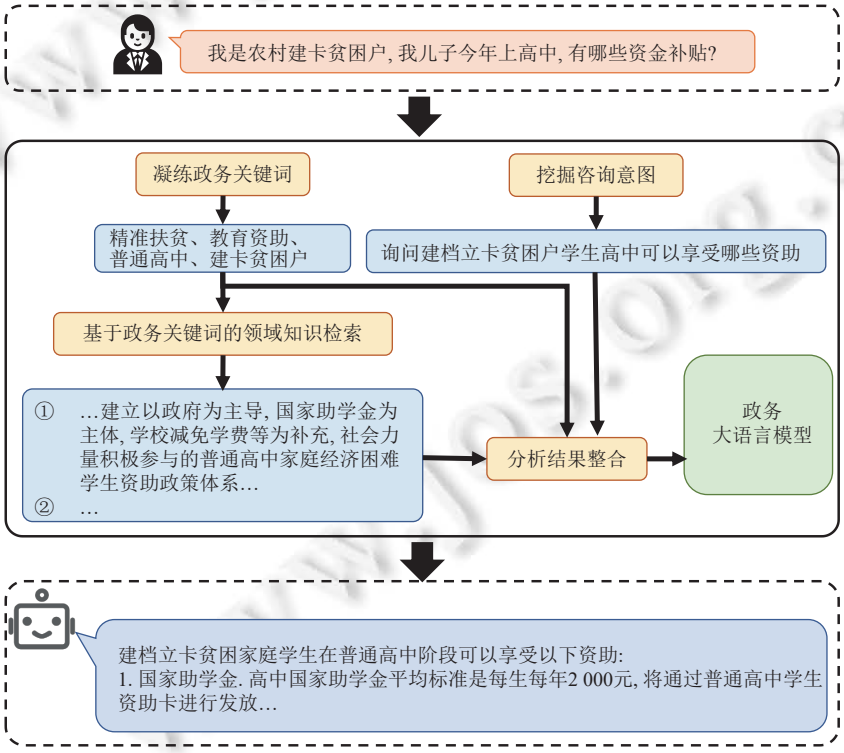

图 5 针对用户咨询的处理流程样例

给定一个咨询输入, 将经历以下步骤得到对应回答.

(1) 关键词提取: 分析引导模块引导政务大语言模型基于用户咨询, 分析咨询涉及的政务领域关键词. 该模块将用户咨询填充至关键词凝练提示模板 Keyword Template 的对应位置构建交互指令. 指令与政务大语言模型交互, 分析出咨询涉及的政务领域关键词——精准扶贫, 教育资助, 普通高中, 建卡贫困户.

(2) 咨询意图分析: 分析引导模块进一步引导政务大语言模型基于用户咨询分析咨询意图. 模块将用户咨询填充至咨询意图挖掘提示模板 Intention Template 的对应位置构建交互指令. 指令与政务大语言模型交互, 分析出咨询意图——询问建档立卡贫困户学生高中可以享受哪些资助.

(3) 领域知识检索: 领域知识库结合 (1) 中分析结果, 检索咨询相关领域知识. 领域知识库计算步骤 (1) 中分析结果的嵌入向量与知识库内容切分文本的嵌入表示之间的相似度, 并找出相似度最高的前 3 条领域知识.

(4) 信息整合与回答生成: 分析引导模块引导政务大语言模型整合步骤 (1)–(3) 产生的结果回答用户咨询. 分析引导模块将步骤 (1)–(3) 结果及用户咨询填充至回答整合提示模板 Answer Template 的对应位置构建交互指令. 指令与政务大语言模型交互, 根据步骤 (1)、(2) 对问题的分析以及提供的咨询相关领域知识, 针对用户咨询生成回答.

4 政务公开信息数据集构建

为构建政务大语言模型和领域知识库, 本文从各官方信息发布网站中收集并整理了一个包含文档和问答对数据的综合性政务公开信息数据集. 该数据集中的大部分数据源自各政府门户网站及多个政务信息公开平台, 部分问答对数据由 ChatGPT 3.5[24]生成, 并经人工筛选精炼得到. 本文将包含文档的子数据集称为政务文档数据集, 包含问答对的子数据集称为政务问答对数据集.

政务文档数据集主要用于构建政务领域知识库, 收集的文档数据来自各大信息发布平台, 包含不同数据类型以及地区和政务主题的多层次多粒度政务信息. 政务问答对数据集用于微调构建政务大语言模型, 其包含两类数据, 一类问答对主要来自各省政务服务指南及其公开的政务问答库; 另一类问答对收集自各级政府官网上选登的民众疑问及解答. 政务公开信息数据集包含 1 901 篇公开政务相关文档和 10 503 条问答对. 数据集的收集及使用概况如图 6 所示.

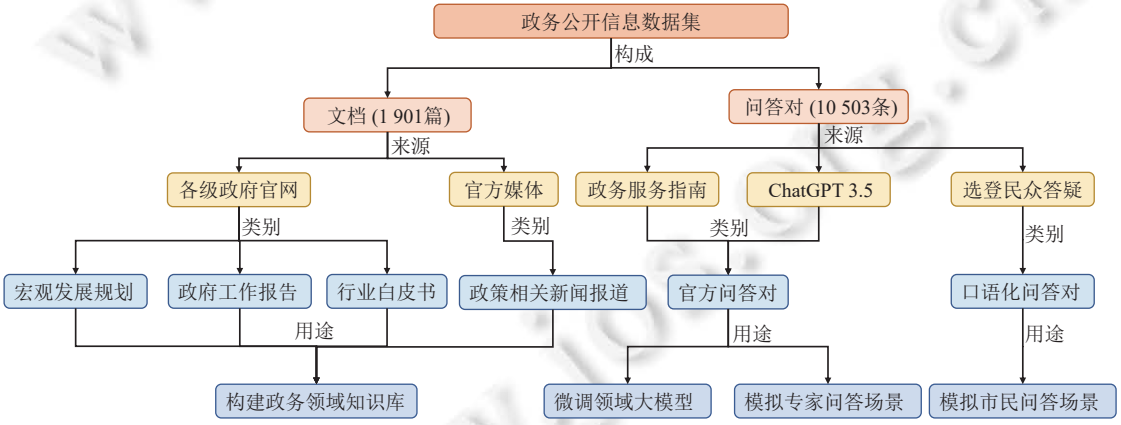

图 6 政务公开数据集收集及使用概况

4.1 政务文档数据集

政务文档数据集中数据类型的详细内容统计情况如图 7 所示, 该数据集收录的各类政务文件, 体现了数据集文档内容的丰富性. 政务文档数据集中政府工作报告覆盖了全国 85% 的省级行政区. 具体而言, 数据集涵盖了东部 (如浙江省)、中部 (如河南省)、西部 (如重庆市) 及少数民族地区 (如新疆维吾尔自治区) 的多个省市, 确保数据集能够反映全国范围内政务公开的差异性和地域特色, 体现了文档数据覆盖地区的多样性.

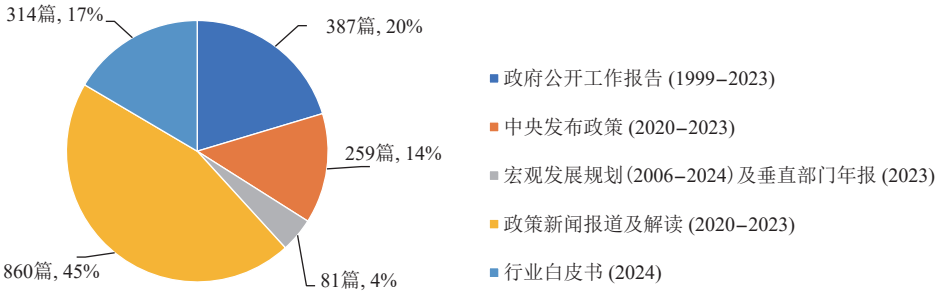

图 7 政务文档数据集统计信息

政务文档数据集中, 文档覆盖政务主题的详细统计情况如图 8 所示. 图 8 的词云统计显示了对政策法规、行政服务和公共服务这 3 大政务服务主题下涉及的政务领域话题进行频次统计的结果, 分别罗列了出现频次前 50 的政务领域话题. 在政策法规类别下, 包含了行政处罚、法治政府、政务服务公开等诸多领域; 在行政服务类别中, 有行政执法、行政审批、知识产权保护等多个细分领域; 在公共服务类别里, 涵盖了国民经济、环境保护、科技创新等丰富的领域. 这些多样化的政务话题领域充分体现了数据集在覆盖不同政务领域上的全面性.

(a) 政策法规 (b) 行政服务 (c) 公共服务

图 8 政务文档数据集的政务领域覆盖

综上, 政务文档数据集覆盖丰富的数据类型、广泛地域和多种政务领域主题, 可多维多层次地全面反映我国政务公开的地域特色和领域多样性.

4.2 政务问答对数据集

政务问答对数据集可细分为两类: 官方问答数据集和口语化问答数据集. 官方问答数据集中的问答对数据主要来源于各省政务服务指南以及公开的政务问答库, 此类问答对由专家编写, 能有效覆盖政务服务咨询的核心内容, 用于领域大语言模型的微调构建和模拟专家查询政务服务信息的使用场景. 此类数据共包含 3 031 条. 口语化问答数据集中的数据源于政府官网上选登的民众疑问解答, 能有效模拟市民政务咨询场景, 包含 5 972 条数据. 为了进一步提升数据质量, 本文对收集到的问答对数据进行了必要的信息增删, 使其在脱离政务服务指南上下文后, 仍能表述清晰.

为扩展问答对数据集在我国政策和发展规划方面知识的覆盖, 帮助领域大语言模型构建对国家发展的宏观认识并进一步规范其语言表达方式, 本文使用 ChatGPT 3.5, 从各级政府公开的工作报告中挖掘问题及相关回答. 本文基于 ChatGPT 的多轮对话能力分批次处理按 2 000 字截断的工作报告, 并通过指令生成基于文档内容的问答对. 使用的提示模板记作 Q&A CollectionTemplate, 如图 9 所示. 对于生成的问答对, 再由人工进行筛选、优化, 以确保问答对的内容质量. 优化后的问答对将被纳入官方问答数据集, 共 1 500 条数据.

Q&A CollectionTemplate

你的任务是基于提供的政务文档内容, 生成一个用户可能提出的问题及其对应的规范回答.
注意:
1. 生成的问题需涵盖具体的政策细节, 确保与文档内容紧密相关;
2. 生成的回答需严格遵循文档表述, 不添加主观解释或推测.
回答的参考样例:
问题: XX补贴的申请条件是什么?
回答: 根据政务文档规定, XX补贴的申请条件包括…
问题: 异地办理身份证的流程是什么?
回答: 根据政务文档要求, 异地办理身份证的流程为…
以下是你要参考的政务文档内容: </指令>
{{document}}

图 9 使用 ChatGPT 构造问答对时使用的提示模板

5 实验分析

5.1 实验数据及资源

本文从政务公开信息数据集中挑选了 1 200 条官方问答对构成初步训练集, 并另选 220 条官方问答对为模型训练时的验证集. 政务问答系统有两个常见的使用场景: 为领域内人士提供政务公开信息查询服务, 以及为市民提供政务服务的办理建议或答疑. 本文称前者为专家问答, 后者为市民问答. 综合考虑其他开源专业性数据集中测试集的规模[25–27]及本文训练数据集规模, 为了验证本文使用的政务大语言模型的领域对话能力, 以及基于前述方法实现的原型系统在专家问答场景下的服务质量, 本文挑选了 220 条官方问答为测试集, 用于模拟专家问答场景. 为了验证原型系统在市民问答场景下的服务质量, 本文挑选 220 条口语化问答为测试集, 用于模拟市民问答场景. 本文代码基于 PyTorch 2.1.2 和 Langchain 0.2.1 实现. 本文所有实验在 4 张 NVIDIA A800 GPU 上完成, 并使用 CUDA 11.3 作为计算架构.

5.2 评价指标及评估参考模型

由于本文使用的问答对数据均包含参考回答, 因此我们比较生成回答和参考回答在划分成不同长度的词组后, 通过词组间的重叠度来评价两者间的一致性, 以衡量生成回答表述的准确性和规范性. 具体来说, 本文采用了 ROUGE (recall oriented understudy for gisting evaluation) 系列[28]和 BLEU (bilingual evaluation understudy) 系列[29]的评估指标. ROUGE 系列指标包含 ROUGE-1, ROUGE-2, ROUGE-L. BLEU 系列指标有 BLEU-1、BLEU-2、BLEU-3、BLEU-4. 同时, 本文使用 BERTScore[30]评估生成回答与参考回答之间的语义相似度, 以评价生成回答的咨询针对性. BERTScore[30]是一种用于衡量文本间语义相似度的指标. 该指标基于 BERT[31]计算两个句子的嵌入式表示, 并通过嵌入式表示之间的余弦相似度评估句子间的相似度. 基于 BERT 计算的嵌入式表示能表达上下文语境, 因此能更准确地衡量表达方式不同的句子间的语义相似度.

为了验证原型系统在各政务咨询服务场景下的生成回答效果, 本文将 ChatGLM2-6B 和 Qwen-72B 在各测试集上的评估结果作为评估参考. ChatGLM2-6B 是构建政务大语言模型的预训练大语言模型, Qwen-72B 的参数量是政务大语言模型的 10 倍以上, 具有更强的语义理解和生成能力. 使用这两个通识大语言模型的评估结果作为对比, 能有效说明原型系统在政务咨询领域的优势. 同时, 本文在各服务场景下针对原型系统各模块进行消融实验, 以说明系统各模块的有效性.

5.3 专家问答场景下实验结果分析

为了评估原型系统在专家问答场景下生成回答的质量, 我们将原型系统及评估参考模型在由官方问答对组成的测试集上进行了评估. 在该测试集上, 我们进行了原型系统的模块消融实验, 以说明在专家问答场景下系统各模块的有效性. 实验结果如表 1 所示.

表 1 系统在专家问答场景下的评估结果 (%)

模型/系统	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ChatGLM2-6B	63.82	24.58	7.76	14.58	17.5	8.51	5.16	3.78
Qwen-72B	67.27	24.86	7.02	16.48	21.88	10.03	5.62	3.81
政务大语言模型	66.43	29.15	10.78	18.17	21.00	11.20	7.28	5.65
政务大语言模型＋领域知识库模块	67.15	28.97	13.06	21.30	23.55	13.46	**9.71**	**8.10**
政务大语言模型＋分析引导模块	68.15	30.35	11.55	21.48	22.42	12.16	8.17	6.31
原型系统	**68.60**	**31.01**	**13.39**	**23.22**	**24.21**	**13.86**	9.70	7.80

(1) 政务大语言模型在政务咨询对话方面展现出优势. 表 1 结果表明, 政务大语言模型在各项评估指标上均明显优于同参数规模的 ChatGLM2-6B. 这说明政务大语言模型的生成结果在语义和用词上与参考回答更接近, 其回答的准确性更高. 然而, 政务大语言模型在 BERTScore 上的评估结果略低于 Qwen-72B. 这是因为政务大语言模型的语义理解受参数规模限制, 其对咨询的理解程度不足, 导致回答的针对性不足. 表 1 结果显示, 政务大语言模型在 ROUGE 和 BLEU 这类基于词重叠度的评估指标上相较 Qwen-72B 有明显优势. 这表明政务大语言模型生成回答与参考回答在用词上更一致, 其生成的回答表述更规范. 综上, 虽然本文使用的政务大语言模型的参数规模有限, 但通过领域数据微调, 在领域对话中能针对咨询生成表述更规范的回答, 更适应政务对话场景.

(2) 专家问答场景中, 原型系统展现出政务咨询服务优势. 在各类评估指标上, 原型系统均有明显优势. 结果如表 1 所示, 原型系统相较于 ChatGLM2-6B 在 BERTScore 上提升了 4.78%, 在 ROUGE-L 上提升了 8.64%. 而相较于 Qwen-72B, 原型系统在 BERTScore 上提升了 1.33%, 在 ROUGE-L 上提升了 6.74%. 这表明相较于通识大语言模型, 原型系统生成的回答在语义和用词上都更接近参考回答, 即其生成的回答具有更高的准确性和规范性. 这些评估结果揭示了原型系统在专家问答场景中的几个关键优势: 首先, 其产生的回答在语义上更贴近参考回答, 这意味着它能够有效理解咨询内容并针对性地产生回答; 其次, 其生成的回答使用更符合领域规范的用词和表达, 表述更规范; 最后, 这些优势共同体现了原型系统在政务信息咨询场景中提供政务咨询服务的能力.

(3) 专家问答场景中分析引导模块能提升系统的回答针对性. 表 1 结果表明, 政务大语言模型结合分析引导模块后, 在 BERTScore 指标上相较于仅使用政务大语言模型提高了 1.72%. 这表明在分析引导模块的辅助下, 政务大语言模型生成的回答和参考回答在语义上更一致, 即生成回答更有问题针对性. 其原因为引导模块能逐步引导政务大语言模型关注咨询输入的不同关键点, 加强了政务大语言模型对咨询内容的理解和对咨询意图的挖掘. 当咨询分析结果作用于回答生成时, 政务大语言模型对咨询内容和咨询意图的理解加深, 从而提升了回答的针对性. 政务大语言模型结合分析引导模块后, 在 ROUGE 系列和 BLEU 系列指标上也有一定提升. 分析其原因, 为该模块辅助后, 政务大语言模型生成回答的针对性提升且大语言模型本身具备一定生成领域规范内容的能力, 因此回答的部分语言组织与参考回答一致, 使得 ROUGE 系列和 BLEU 系列指标有一定提升.

(4) 专家问答场景中领域知识库模块能提升回答内容的规范性. 表 1 结果表明, 使用领域知识库辅助政务大语言模型后, 在 BLEU 系列指标上相较于未使用前有明显提升. 这说明领域知识库辅助政务大语言模型后, 生成的回答在用词上与参考回答更一致, 回答的规范性和准确性得到提升. 分析其原因, 在专家问答场景下, 咨询输入的语言组织符合领域规范, 基于用户咨询检索相关领域知识的成功率较高. 因此, 政务大语言模型在生成回答时能参考咨询相关的领域知识, 能针对用户咨询生成内容表述与事实一致的准确规范的咨询回答.

5.4 系统在市民问答场景下的服务质量评估

为了评估原型系统在市民咨询场景下生成回答的质量, 我们将原型系统及评估参考模型在由口语化问答对组成的测试集上进行评估. 在该测试集上进行了原型系统的模块消融实验, 以说明在市民问答场景下系统各模块的有效性. 实验结果如表 2 所示.

(1) 政务大语言模型的回答质量受限于咨询内容的组织规范性. 表 2 的结果表明, 在市民问答场景下, 政务大语言模型在各评估指标上均优于同参数规模的 ChatGLM2-6B. 但与 Qwen-72B 相比, 政务大语言模型在所有评估

指标上均无明显优势. 这说明在该场景下, 政务大语言模型生成的回答缺乏针对性且在用词上与参考回答的一致性不足, 导致回答的准确性不足. 分析原因, 是由于该场景下的咨询输入与政务大语言模型微调时的输入在语言组织方式上存在明显差异, 致使政务大语言模型无法有效理解咨询内容, 从而导致生成的回答在针对性和用词一致性上与参考回答有一定差异.

表 2 市民问答场景下的评估结果 (%)

模型/系统	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ChatGLM2-6B	63.05	21.32	4.88	12.24	16.60	6.94	3.56	2.31
Qwen-72B	65.61	22.04	5.21	14.87	20.10	8.40	4.24	2.74
政务大语言模型	64.60	23.28	5.76	13.66	17.82	7.99	4.32	2.89
政务大语言模型＋领域知识库模块	64.16	24.64	6.24	15.87	18.77	8.81	5.65	3.79
政务大语言模型＋分析引导模块	66.37	26.36	8.04	18.00	20.97	10.35	6.42	4.79
原型系统	**67.31**	**27.02**	**8.84**	**19.74**	**23.30**	**11.22**	**6.46**	**4.54**

(2) 市民问答场景中的原型系统具有政务咨询服务优势. 表 2 的结果显示, 原型系统在市民问答场景下的各项评估指标均取得最优. 与 ChatGLM2-6B 相比, 原型系统在 BERTScore 上提升了 4.26%, 在 ROUGE 系列指标上的平均提升为 5.72%, 在 BLEU 系列指标上的平均提升为 4.03%. 与 Qwen-72B 相比, 原型系统在 BERTScore 上提升了 1.7%, 在 ROUGE 系列指标上的平均提升为 4.49%, 在 BLEU 系列指标上的平均提升为 2.51%. 实验结果表明, 原型系统生成的回答和参考回答在语义和用词上更一致, 说明生成回答的准确性和规范性较好, 这证明了原型系统在市民问答场景中有政务咨询服务优势.

(3) 市民问答场景下, 分析引导模块有助于提升回答的针对性. 表 2 的结果显示, 政务大语言模型在分析引导模块辅助后在各评估指标上均取得明显提升. 相比于仅使用政务大语言模型, 分析引导模块辅助后的政务大语言模型在 BERTScore 上提升了 1.77%, 在 ROUGE 系列指标上的平均提升为 3.23%, 在 BLEU 系列指标上的平均提升为 2.38%. 其原因为分析引导模块能够引导政务大语言模型逐步分析咨询内容, 将市民口语化咨询内容中的关键信息转述为对应的规范表达, 从而缓解了口语化咨询内容对政务大语言模型语义理解上的阻碍, 提升了政务大语言模型生成回答的咨询针对性. 同时, 由于政务大语言模型具备一定生成规范内容的能力, 所以针对咨询生成的回答在语言组织上与参考回答的一致性得到了提升.

(4) 市民问答场景下, 领域知识库模块需要分析引导模块的辅助以保障其有效性. 在市民问答场景下, 当领域知识库模块单独辅助本文使用的政务大语言模型后, 各评估指标上的提升不明显, 甚至在 BERTScore 上的评估结果反而下降. 这是因为市民问答场景下的市民口语化的咨询输入在语言组织方式上与领域知识库中的内容差别较大, 导致直接使用咨询检索相关领域知识的成功率低. 政务大语言模型在生成回答时能参考的相关知识不足甚至受到无关信息的干扰, 影响了生成回答与参考回答在语义上的一致性[12], 且在用词规范上缺乏参考, 导致在 ROUGE 系列和 BLEU 系列的指标上提升不明显. 使用分析结果检索领域知识后 (评估结果见表 2 最后一行), 相较于未使用分析结果的政务大语言模型, 在 ROUGE 系列和 BLEU 系列的指标上提升明显. 这是因为分析引导模块中由咨询凝练出的政务领域关键词和知识库中的内容在语言组织方式上相似, 进而提升了咨询相关领域知识的检索成功率. 因此, 在该服务场景下, 政务大语言模型在生成内容时能同时参考咨询分析结果和与咨询相关的领域知识, 实现针对咨询生成内容组织规范的准确回答.

(5) 原型系统具有提供可信政务服务优势. 对比表 1 和表 2 结果, 在市民问答场景下, 原型系统和政务大语言模型在各评估指标上的结果均有不同程度的下降, 但原型系统的降低程度较政务大语言模型更缓. 相较于专家问答场景, 原型系统在市民问答场景下 BERTScore 下降 1.29%, 而本文使用的政务大语言模型在市民问答场景下 BERTScore 下降 1.83%. 原型系统在市民问答场景下 ROUGE 系列指标上平均下降 4.26%, 在 BLEU 上平均下降 2.51%. 政务大语言模型在市民问答场景下 ROUGE 系列指标上平均下降 5.13%, 在 BLEU 系列指标上平均下降 3.02%. 这说明, 在各服务场景下原型系统能生成与参考回答在语义和用词规范性上更一致的回答, 即其生成的回答更准确且规范. 综上所述, 原型系统能在不同服务场景下针对用户咨询生成内容组织规范的准确回答并提供回答依据以增强回答的可解释性, 证明原型系统具备提供可信政务咨询服务的优势.

5.5 原型系统问答案例分析

本节分析了系统在专家问答和市民问答场景下, 原型系统的成功案例和失败案例.

在各服务场景下的典型成功案例可以在公开数据集链接 (https://doi.org/10.57760/sciencedb.24847) 中查看. 原型系统的成功案例在不同场景下都能有效提取咨询政务领域关键词并总结用户咨询意图, 进而针对用户咨询产生准确可解释的回答内容. 实际的有效案例说明了验证系统的实用性.

分析发现, 系统在各服务场景下因咨询问题复杂而失效. 在各服务场景下的典型失败案例可以在公开数据集链接 (https://doi.org/10.57760/sciencedb.24847) 中查看. 系统在不同服务场景下, 面对复杂咨询时失效的原因及失效现象如下: 1) 专家问答场景下, 系统面对一类需要综合多条相关信息以形成概括性的回答的咨询问题时失效. 这是由于系统中的大型模型有限的上下文窗口, 本文设计的检索方法仅保留了与政务关键词最相关的前 3 条领域知识, 这限制了对相关问题相关信息的覆盖范围, 从而导致了系统生成回答的内容不全面. 2) 市民问答场景下, 系统在多政务领域关联的咨询时失效. 这是因为系统分析方式的局限性, 未能识别出复杂咨询问题中多个政务话题之间的联系以及由此衍生的用户咨询需求. 系统的分析模块仅能识别部分政务话题并给出部分回答, 导致回答不全面, 未能满足用户的实际咨询需求.

5.6 微调数据量边际效应探究

为了探究 LoRA 方法在本文所探讨的微调场景下需要的微调数据量, 本文从初步训练集中划分构造了包含 900 条、1 000 条、1 100 条以及 1 200 条数据的训练数据集. 将 ChatGLM2-6B 模型分别在上述训练数据集上使用 LoRA 方法进行微调训练, 并在由官方问答对组成的测试集上对各微调后的大语言模型进行评估. 选择在测试集上综合表现最佳的微调结果作为本文所使用的政务大语言模型. 表 3 记录了微调训练时主要超参数的设置情况. 图 10 展示了 ChatGLM2-6B 模型以及经过不同训练量微调后得到的大语言模型在测试集上的评估情况. 其中, “GovGLM (0.9k)”表示由 ChatGLM2-6B 经过 900 条问答对微调得到的政务大语言模型. 图中各评估结果均以百分数形式展示. 实验结果表明, ChatGLM2-6B 在 1 000 条和 1 100 条数据集上的微调结果在各项评估指标上展现出了明显的性能提升. 在综合考虑评估结果后, 选择在 1 100 条问答对上微调后的大语言模型作为本文所采用的政务大语言模型.

表 3 LoRA 微调关键超参数记录表

超参数名	参数值
Learning rate	1.00×10^{-5}
Warmup ratio	5.00×10^{-2}
Weight decay	5.00×10^{-2}
Eval steps	50
LoRA rank	8
per_device_train_batch_size	4
per_device_eval_batch_size	4

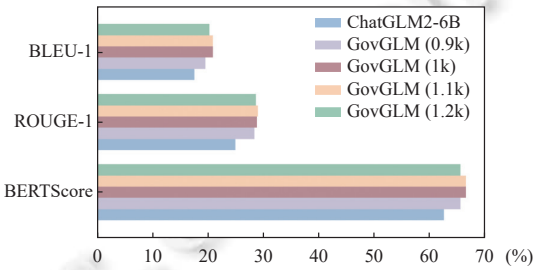

图 10 微调数据量边际效应实验结果

图 10 的实验结果表明, 在该微调场景下 LoRA 微调方法对数据量的要求较低, 只需要 1 000 条左右的微调数据, 就能得到明显超越同参数规模下通识大语言模型的领域对话能力. 实验结果验证了 LoRA 方法参数微调的高效性. 分析其原因, 在 LoRA 方法的训练过程中, 预训练模型的全部参数被冻结, 仅调整旁路网络中的参数. 训练时冻结预训练大语言模型的参数既能有效降低训练成本, 又能一定程度保留大语言模型原有的语义理解和生成能力, 从而降低大语言模型在训练后退化甚至崩溃的概率.

6 相关讨论

6.1 系统多场景可用性分析

本文通过分析引导模块提取并转述咨询涉及的关键政务信息, 确保领域知识库模块和政务大语言模型在不同

场景下的稳定工作, 实现对多服务场景中不同语言组织方式下咨询的有效理解和针对性回答. 第 3.2 节按照使用对象对政务咨询场景进行分类, 并通过实验证明了方法在不同服务场景下的服务优势. 本节将进一步探讨按照政务服务主题分类时, 系统如何适应多样的现实服务场景.

按照政务服务主题分类, 可能涉及政策解释、法律咨询、行政审批等多个场景, 每个场景的知识需求和咨询类型有所不同. 在按照不同政务主题分类的多个政务咨询场景中, 政务大语言模型根据领域知识库提供的领域知识生成回答. 本文方法的领域知识库, 具有地域覆盖广、文本类型多、涵盖主题丰富的特性, 保障了库内知识的完备性. 在不同政务服务场景中, 领域知识库可为政务大语言模型提供对应政务知识, 保障了方法的回答准确性.

综上, 原型系统通过分析引导模块对多语言组织方式下内容的专业化转述, 以及领域知识库对政务服务主题覆盖的完备性和丰富性, 适应不同语言组织方式的用户提出的各类政务服务相关咨询. 因此, 原型系统具备在不同政务服务场景和不同咨询类型的适用性.

6.2 方法局限性分析

目前, 领域知识库的构建方式较为简单, 采用固定长度切分文本并存储. 尽管使用了重叠长度以避免信息丢失, 但仍然可能导致对原始政务知识的语义截断, 从而使检索到的领域知识缺乏咨询所需的细节内容, 极端情况下可能会影响生成回答的详细性和实用性.

考虑到分析引导模块需要对新增服务场景进行快速高效的适应, 为平衡任务完成的稳定性和适应性, 本文的分析引导模块是通过在政务大语言模型上使用上下文学习策略来实现. 虽然第 5.3 节和第 5.4 节的消融实验表明了本文构建的分析引导模块的有效性和重要性, 这也可能导致模块对复杂咨询的解析能力不足, 极端情况下会影响分析结果的完备性.

具有概括性信息需求的咨询, 要求方法整合大量知识并进行精炼总结以回答, 如询问某地区经济发展状况或环保措施. 专家问答场景下的失败案例表明, 在面对此类提问时, 系统检索方法受大语言模型上下文窗口限制, 只取与检索输入最相似的前 3 条知识库知识, 无法完整检索所需的信息, 影响了回答的完备性.

包含多个政务主题和事件场景的咨询, 需要系统抽取咨询涉及的所有政务主题, 发掘它们的正确关联并由此分析用户咨询意图. 市民问答场景下的失败案例属于此类问题. 观察回答结果发现系统的分析引导模块未能完整抽取问题涉及的多个政务主题, 且该模块没有独立步骤分析多主题间的关联导致分析的咨询意图不完备. 最终, 系统在此类问题上产生的回答缺乏完备性且无法满足用户真实意图.

政务问答系统可能接收到的内容有语法错误、语义歧义或信息缺失等表述有误的内容. 分析引导模块依托政务大语言模型, 可基于上下文理解修正简单的错误输入 (如错别字或漏字), 已经具备了一定的错误输入处理能力. 对于深层表述错误的咨询 (如包含语法错误或信息缺失的咨询), 分析引导模块无法自动补齐或修正, 导致对分析结果的偏差, 进而影响回答的针对性和有效性.

系统在实际使用中可能接收包含对抗样本或代码注入的恶意攻击咨询. 本文方法未考虑对恶意攻击的防御, 因此, 在面对恶意攻击咨询时, 分析引导模块和领域知识库模块可能失效, 导致回答失败, 并且无法提供有效的分析和参考结果.

6.3 研究方向展望

为了防止文档存入知识库时出现语义截断的问题, 受到 Luo 等人[32]工作的启发, 我们计划研究基于语义的切分方式, 以保留原始文档的上下文信息, 提升知识库内的信息质量.

研究低数据量微调持续适应技术, 可以使用本文的分析引导模块为分析引导模块中的政务咨询子任务收集微调数据, 为微调构建分析引导模块提供重要基础, 以提升分析引导模块输出的稳定性和效果. 为了在保证新服务场景适应效率的前提下实现分析任务完成的高效稳定, 受到 Transformer-Squared[33]工作的启发, 我们计划研究基于多场景提示组合的任务微调框架, 以保障任务在多服务场景下的完成稳定性和适应效率.

针对高层次信息需求的咨询, 受到 RAPTOR[34]的工作启发, 我们计划设计层次化知识库实现对信息的结构化建模, 以及多信息需求的内容检索与整合方法, 提升面向多信息需求咨询时信息检索的完备性和密集度, 进而保障

回答的完备性.

针对具有复杂政务主题关联的咨询, 研究复杂关联解析技术. 受到 Yu 等人[35]工作的启发, 我们可以尝试将语义关系依存解析的步骤引入分析引导模块中, 以提升分析模块对复杂咨询的理解程度. 同时可以使用依存关系指导领域知识库检索结果中关键信息的整合提炼, 为政务大语言模型提供更凝练聚焦的参考领域知识, 提升方法对复杂咨询的应对能力.

为实现对包含语法错误或关键信息缺失的咨询进行感知和修正, 避免因错误输入导致分析和检索失效, 从而生成错误回答, 我们可以构建基于表述流畅度评估的咨询错误分类器, 通过流畅度评估判断咨询是否存在问题. 为了进一步修正咨询错误以正确回答用户咨询, 受到 Mao 等人[36]工作的启发, 我们计划研究基于多轮对话交互的咨询意图修正技术, 通过多轮问题重写逐步明确用户咨询意图以提升回答的针对性.

为了实现对恶意攻击咨询的拒绝回答, 一种可能的方法是通过设定领域知识库检索时的相似度阈值来判断是否存在相关内容. 如果所有相似度均低于设定阈值, 则认为知识库中不存在与咨询相关的内容, 此时可以通过分析引导模块设计的拒绝回答提示模板, 指示政务大语言模型拒绝回答用户的咨询.

7 总 结

为了提升当前政务问答系统生成可信咨询回答的能力, 本文提出基于大语言模型的可信政务问答技术, 通过加强系统在多服务场景下生成准确且可解释的咨询回答来实现可信咨询. 该方法首先通过分析引导模块将咨询中的关键信息转述为对应的规范化表达, 从而加强政务大语言模型对多服务场景下的非规范咨询的有效理解. 然后, 领域知识库模块根据分析结果检索咨询相关领域知识, 确保多服务场景下检索效果不受咨询语言组织规范性的影响, 进而保障检索成功率. 通过上述两个辅助模块的协同, 政务大语言模型在多服务场景下参考咨询分析结果和咨询相关领域知识, 针对用户咨询生成准确且可解释的可信回答. 为构建领域知识库和微调政务大语言模型, 本文收集整理了一个包含 1 901 篇政务文档和 10 503 条政务问答对的政务公开信息数据集, 通过实验和生成案例分析验证了基于本文方法实现的原型系统生成可信回答的能力. 实验结果表明, 原型系统在多种服务场景下均能针对用户咨询生成准确且可解释的回答, 具备提供可信政务咨询服务的能力. 最后, 本文从系统在两个典型政务服务场景下的失败案例分析出发, 分析了系统方法的局限性并展望了未来研究方向, 为后续工作提供启发.

References

[1] Li YX, Li ZH, Zhang K, Dan RL, Jiang S, Zhang Y. ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus, 2023, 15(6): e40895. [doi: 10.7759/cureus.40895]

[2] Vassilakopoulou P, Haug A, Salvesen LM, Pappas IO, Agerfalk P. Developing human/AI interactions for chat-based customer services: Lessons learned from the norwegian government. European Journal of Information Systems, 2023, 32(1): 10–22. [doi: 10.1080/0960085X.2022.2096490]

[3] Gao SS, Gao L, Li Q, Xu JJ. Application of large language model in intelligent Q&A of digital government. In: Proc. of the 2nd Int'l Conf. on Networks, Communications and Information Technology. Qinghai: ACM, 2023. 24–27. [doi: 10.1145/3605801.3605806]

[4] Wang YK, Zhang N, Zhao XJ. Intelligent Q&A robots in government services: Current status, mechanism, and key support. E-Government, 2020(2): 34–45 (in Chinese). [doi: 10.16582/j.cnki.dzzw.2020.02.004]

[5] Cheng X, Tan TL, Wang MM. Research on government question answering robot based on knowledge graph in PKS system. Application of Electronic Technique, 2023, 49(4): 128–132 (in Chinese with English abstract). [doi: 10.16157/j.issn.0258-7998.223038]

[6] Fan YF, Zou BW, Xu QT, Li ZF, Hong Y. Survey on commonsense question answering. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 236–265 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6913.htm [doi: 10.13328/j.cnki.jos.006913]

[7] Wang F, Wei ZH, Lian ZX. Developing an evaluation index system and its testing questions for government intelligent Q&A systems. Document, Informaiton & Knowledge, 2023, 40(6): 98–111 (in Chinese with English abstract). [doi: 10.13366/j.dik.2023.06.098]

[8] Liang DC, Wu YQ, Duan WY. Multiple granularity user intention fairness recognition of intelligent government Q & A system via three-way decision. Information Sciences, 2023, 631: 305–326. [doi: 10.1016/j.ins.2023.02.070]

[9] Geng YF. Research and implementation of task-oriented dialogue system for government affairs [MS. Thesis]. Changsha: South-central Minzu University, 2022 (in Chinese with English abstract).

[10] Fang KY, Xu KW. Automating government response to citizens' questions: A large language model-based question-answering guidance generation system. In: Proc. of the 3rd Int'l Conf. on Digital Society and Intelligent Systems. Chengdu: IEEE, 2023. 386–389. [doi: 10.1109/DSInS60115.2023.10455136]

[11] Liu Y, Yao YS, Ton JF, Zhang XY, Guo RC, Cheng H, Klochkov Y, Taufiq MF, Li H. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. arXiv:2308.05374, 2024.

[12] Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, Weston J. Chain-of-verification reduces hallucination in large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: ACL, 2024. 3563–3578. [doi: 10.18653/v1/2024.findings-acl.212]

[13] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1800.

[14] Wen YL, Wang ZF, Sun JM. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 10370–10388. [doi: 10.18653/v1/2024.acl-long.558]

[15] Jiang ZB, Araki J, Ding HB, Neubig G. How can we know when language models know? On the calibration of language models for question answering. Trans. of the Association for Computational Linguistics, 2021, 9: 962–977. [doi: 10.1162/tacl_a_00407]

[16] Zhou CT, Liu PF, Xu PX, Iyer S, Sun J, Mao YN, Ma XZ, Efrat A, Yu P, Yu LL, Zhang SS, Ghosh G, Lewis M, Zettlemoyer L, Levy O. LIMA: Less is more for alignment. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2400.

[17] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 793.

[18] Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: Retrieval-augmented language model pre-training. In: Proc. of the 37th Int'l Conf. on Machine Learning. arXiv:2002.08909v1, 2020.

[19] Zhang SH, Song YL, Yang JH, Li YQ, Han B, Tan MK. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[20] Ma JK, Wang YS, Li G, Mei H. Influence of acquirer's participation on project performance: An analysis on e-government projects. Ruan Jian Xue Bao/Journal of Software, 2012, 23(10): 2679–2694 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4244.htm [doi: 10.3724/SP.J.1001.2012.04244]

[21] Zeng AH, Liu X, Du ZX, Wang ZH, Lai HY, Ding M, Yang ZY, Xu YF, Zheng WD, Xia X, Tam WL, Ma ZX, Xue YF, Zhai JD, Chen WG, Liu ZY, Zhang P, Dong YX, Tang J. GLM-130B: An open bilingual pre-trained model. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023. 320–335.

[22] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[23] Tian XH. Algorithm integration, information-driven empowerment, and capacity building in grassroots government governance. CASS Journal of Political Science, 2025(1): 119–134, 190 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3355.2025.1.zzxyj202501010]

[24] Nizon-Deladoeuille M, Stefánsson B, Neukirchen H, Welsh T. Towards supporting penetration testing education with large language models: An evaluation and comparison. In: Proc. of the 11th Int'l Conf. on Social Networks Analysis, Management and Security. Gran Canaria: IEEE, 2024. 227–229. [doi: 10.1109/SNAMS64316.2024.10883774]

[25] Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, Leike J, Schulman J, Sutskever I, Cobbe K. Let's verify step by step. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[26] Bai YS, Lv X, Zhang JJ, Lyu HC, Tang JK, Huang ZD, Du ZX, Liu X, Zeng AH, Hou L, Dong YX, Tang J, Li JZ. LongBench: A bilingual, multitask benchmark for long context understanding. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 3119–3137. [doi: 10.18653/v1/2024.acl-long.172]

[27] Guha N, Nyarko J, Ho DE, *et al*. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: NeurIPS, 2023. 44123–44279.

[28] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.

[29] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual

Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[30] Zhang TY, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with BERT. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[31] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[32] Luo K, Liu Z, Xiao ST, Zhou T, Chen YB, Zhao J, Liu K. Landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 3268–3281. [doi: 10.18653/v1/2024.acl-long.180]

[33] Sun Q, Cetin E, Tang YJ. Transformer-squared: Self-adaptive LLMs. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025.

[34] Sarthi P, Abdullah S, Tuli A, Khanna S, Goldie A, Manning CD. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[35] Yu JS, Shi JL, Yang LE, Xiao D, Yang EH. Automatic Transformation of enhanced dependencies in Chinese. Journal of Chinese Information Processing, 2023, 37(10): 26–33 (in Chinese with English abstract).

[36] Mao KL, Dou ZC, Mo FR, Hou JW, Chen HN, Qian HJ. Large language models know your contextual search intent: A prompting framework for conversational search. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 1211–1225. [doi: 10.18653/v1/2023.findings-emnlp.86]

附中文参考文献

[4] 王友奎, 张楠, 赵雪娇. 政务服务中的智能问答机器人: 现状、机理和关键支撑. 电子政务, 2020(2): 34–45. [doi: 10.16582/j.cnki.dzzw.2020.02.004]

[5] 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. 电子技术应用, 2023, 49(4): 128–132. [doi: 10.16157/j.issn.0258-7998.223038]

[6] 范怡帆, 邹博伟, 徐庆婷, 李志峰, 洪宇. 常识问答研究综述. 软件学报, 2024, 35(1): 236–265. http://www.jos.org.cn/1000-9825/6913.htm [doi: 10.13328/j.cnki.jos.006913]

[7] 王芳, 魏中瀚, 连芷萱. 政务智能问答系统评价指标体系构建与测评问题编制. 图书情报知识, 2023, 40(6): 98–111. [doi: 10.13366/j.dik.2023.06.098]

[9] 耿云飞. 面向政务领域的任务型对话系统的研究与实现 [硕士学位论文]. 长沙: 中南民族大学, 2022.

[20] 马家宽, 王亚沙, 李刚, 梅宏. 电子政务需方参与活动对项目绩效的影响分析. 软件学报, 2012, 23(10): 2679–2694. http://www.jos.org.cn/1000-9825/4244.htm [doi: 10.3724/SP.J.1001.2012.04244]

[23] 田先红. 算法嵌入、信息赋能与基层政府治理能力建设. 政治学研究, 2025(1): 119–134, 190. [doi: 10.3969/j.issn.1000-3355.2025.1.zzxyj202501010]

[35] 余婧思, 师佳璐, 杨麟儿, 肖丹, 杨尔弘. 汉语增强依存句法自动转换研究. 中文信息学报, 2023, 37(10): 26–33.

作者简介

王赛玥, 硕士生, 主要研究领域为机器学习, 大模型测试时增强.

胡晋武, 博士生, 主要研究领域为大语言模型, 强化学习, 计算机视觉.

王宇丰, 博士生, 主要研究领域为大模型, 强化学习, 具身智能.

胡宇, 博士, 主要研究领域为机器学习, 多组学分析, 数据挖掘.

高浩然, 工程师, 主要研究领域为智能全栈开发技术.

邱舟强, 教授级高级工程师, 主要研究领域为机器学习.

谭明奎, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 视觉生成, 大数据分析, 大规模凸优化.