An Automated Government Affairs Response Reference Generation System Based on Large Language Models and Multi-Agent Systems

Authors: Fang Keyuan, Xu Kewei, Xu Kewei

Date: 2025-02-13T00:00:00+00:00

Abstract

In the digital governance era, government platforms need to respond to citizens' government inquiries in a timely and effective manner. However, existing government Q&A systems are primarily based on manual responses, with limited assistance from automated processing algorithms, making it difficult to efficiently handle the massive volume of citizen government consultation demands in the big data era. Therefore, government platforms in the digital governance era need to establish more effective and intelligent Q&A systems to respond to citizens' government consultations. Today, large language models (LLMs) are expected to assist government platforms in processing citizens' government consultations in an automated and effective manner. LLMs can enhance the efficiency of interaction between government platforms and citizens, providing natural language responses to various types of citizen consultations. However, existing general LLMs have limited understanding of domain-specific expressions in government affairs and cannot yet produce effective responses like platform staff. This study constructs a specialized reference generation system for government consultation responses (GovLLM) based on LLMs and a vector database of historical government consultation Q&A, utilizing multi-agent technology. After inputting new citizen consultations, the system can generate practical and effective example answers for platform staff to reference when handling citizen consultations. The system demonstrates superior text generation performance compared to baseline models, which is beneficial for improving the efficiency and effectiveness of government platforms when responding to citizen consultations.

Full Text

Preamble

Automatic Government Response Reference Generation System Based on Large Language Models and Multi-Agent

Fang Keyuan[1], *Xu Kewei*[1] †

[1]Innovation, Policy and Entrepreneurship Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 510000, China

Abstract: In the digital governance era, government platforms must respond to citizen inquiries in a timely and effective manner. However, existing government Q&A systems rely primarily on manual responses with limited assistance from automated processing algorithms, making it difficult to efficiently handle the massive volume of citizen consultation demands in the big data era. Therefore, government platforms need to establish more effective and intelligent Q&A systems to address citizen inquiries. Large language models (LLMs) now offer the potential to help government platforms process citizen consultations automatically and effectively, improving interaction efficiency and providing natural language responses to various types of inquiries. Nevertheless, existing general-purpose LLMs have limited understanding of domain-specific expressions in government affairs and cannot yet generate effective responses like platform staff. This study constructs a response reference generation system specifically for government inquiries (GovLLM) based on LLMs, a historical vector database of government consultation Q&A, and multi-agent technology. When new citizen inquiries are input, the system generates practical and effective example answers for platform staff to reference when handling consultations. The system demonstrates better text generation performance than baseline models, enhancing both the efficiency and effectiveness of government platforms in responding to citizen inquiries.

Keywords: digital government; large language models; Q&A system; information system

Funding: This work was supported by the 2024 Youth Project of Guangdong Provincial Philosophy and Social Science Planning (GD24YGL37), the 2024 Basic and Applied Basic Research Special Youth Doctor "Sailing" Project of Guangzhou (No. SL2023A04J01836), the Guangzhou University-Enterprise Joint Funding Project Basic Research Program (2023A03J0165), and the Guangzhou Municipal Education Bureau University Research Project (2024312153).

Author Biographies: Fang Keyuan (1998-), male, from Jieyang, Guangdong, Ph.D. candidate, research interests include natural language processing, e-government, and data mining. Xu Kewei (1989-), male (corresponding author), from Nanyang, Henan, Assistant Professor, Ph.D. supervisor, Ph.D., research interests include citizen participation, smart cities, big data analytics, and public

policy (coreyxu@hkust-gz.edu.cn).

1 Related Research

The digital transformation of government services is a critical component of enhancing national governance capacity. Modern governance requires inter-departmental service and information integration [1], placing data value at the core of government service reform [2]. With the widespread adoption of digital government platforms, the volume of online citizen consultations has grown rapidly, with increasingly diverse demand types, posing greater challenges to government platform responsiveness [3]. In this context, government Q&A systems have emerged as effective tools for responding to citizen inquiries in the digital government era, facilitating interaction and communication between government platforms and citizens [4]. Intelligent and efficient government Q&A systems serve as the primary channel for meeting citizens' diverse consultation needs and as an important avenue for municipal departments to understand citizen demands. Therefore, constructing intelligent and efficient Q&A systems holds significant value for both citizens and municipal departments [5].

China attaches great importance to e-government development and has proposed the overarching goal of building a "responsive government" [6], requiring government departments at all levels to actively respond to citizen demands and provide substantive solutions. Advancing responsive government construction is both a policy requirement and a crucial measure for improving governance effectiveness.

Traditional Q&A systems have been constructed primarily through three approaches [7]. The first is information retrieval-based systems that search for short text fragments within document collections to answer questions [8]. The second is knowledge matching systems that map natural language questions to queries on structured databases [8]. The third approach applies machine learning and deep learning methods to train models that learn mappings between questions and answers [9]. While these three methods can achieve relatively good results in general-purpose Q&A systems, they have limitations when applied to digital government platforms. First, these methods have relatively low understanding of specialized and fixed expressions in the government service domain. Second, their output responses often deviate from users' original intentions. These shortcomings are particularly evident in digital government platforms, where citizen inquiries are numerous and varied, especially given significant policy differences between provinces and cities in China [10]. This requires government platforms to have robust historical data retrieval capabilities and to generate practical responses that account for regional diversity. Current government Q&A systems struggle to meet these requirements, resulting in most systems still relying primarily on manual responses [11], which cannot satisfy citizens' growing demands for government consultations. Therefore, this

study attempts to apply the latest large language models (LLMs) to address these research gaps.

In recent years, LLMs have made significant advances, particularly in question-answering tasks [12]. LLMs possess the ability to generate human-like and professional text across different domains, fundamentally reshaping the AI landscape [13]. Existing research has successfully applied LLMs to solve domain-specific problems in finance [14] and medicine [15]. However, their application and research in the government domain remain relatively limited, indicating significant research potential in this area. While LLMs show great promise for government applications, several issues need resolution. First, LLMs occasionally generate content that does not align with facts [16]. Second, their generated text can be hollow and templated, lacking practical and useful information [17]. Unlike Q&A systems in other domains, government platform responses must strictly conform to objective facts and effectively solve citizens' problems to enhance trust and satisfaction in municipal departments [18]. Therefore, this study seeks to establish an LLM-based government response reference generation system to fill this research gap.

The contributions of this study can be summarized as follows:

a) With the overarching goal of building a responsive government platform, this study constructs a government reference response generation system (GovLLM) based on LLMs and a historical government Q&A vector database. Using multi-agent technology, the system achieves a 5-11% performance improvement over baseline methods, filling the research gap of LLMs in this domain and demonstrating LLMs' powerful capabilities in understanding government text.

b) The generated response references provide practical and effective solutions rather than hollow template answers. Additionally, the system incorporates modules including a historical vector database, agent design, location filtering, question category classification, and ranking based on question and response attributes, which improve response generation efficiency and ensure that response content aligns with objective facts in the consultation' s region of origin, thereby reducing misinformation. This enhances system usability and reliability, promoting citizen trust and satisfaction with government platforms.

c) The system not only achieves significant improvements in efficiency and text quality but also helps reduce manual labor costs. Traditional government consultation platforms rely on manual responses with high associated labor costs. The response references generated by this system reduce the time and effort that government staff must invest in handling consultations, facilitating efficient platform operation.

2 Data Overview

The dataset used in this study comprises Q&A data from the People' s Daily Online government message board in China. This dataset includes text data of Q&A exchanges between citizens and the People' s Daily Online platform, where citizens submit government consultation questions through the message board and staff provide targeted responses on the government platform. The dataset contains a total of 130,258 entries, covering government message board Q&A data from all provinces across China for the full year of 2022. Additionally, it includes field information such as question timestamp, question type (urban construction, environmental protection, transportation, education, finance, employment, tourism, enterprise, agriculture, culture and entertainment, healthcare, public security), province/city of the question, and user likes. This field information reflects the diverse information about citizen questions and facilitates subsequent targeted screening and retrieval of Q&A data for system construction.

The selection of data from this year was primarily based on considerations of data volume and balance across types. This year had the largest volume of government questions in the past five years, providing sufficient sample data for model training and testing. Furthermore, the distribution of government question data across different types and time periods in this year showed high equilibrium, helping to prevent the model from overfitting to specific question types or time periods. Therefore, the system construction in this study is based on training and testing with this dataset.

---

3 System Framework

The GovLLM system constructed in this study, shown in Figure 1 [Figure 1: see original paper], consists of two main components. The first component is LLM fine-tuning. The system uses Qwen-14B, developed by Alibaba Cloud, as the base LLM, which demonstrates superior understanding of Chinese text. The LLM is fine-tuned based on the training set to develop the ability to understand government consultation contexts. When new government consultation questions are input, the fine-tuned LLM accepts the new question and passes it to the second component. The second component is retrieval enhancement based on multi-agent systems. In addition to being used for LLM fine-tuning in the first component, the training set is used in the second component to construct a historical Q&A vector retrieval library generated by BERT. BERT (Bidirectional Encoder Representations from Transformers) plays a core role in constructing the historical Q&A vector retrieval library. Through its deep bidirectional Transformer architecture, BERT can understand and generate high-quality context-related vectors that precisely represent and match historical Q&A data, thereby enabling efficient and accurate retrieval functionality. The government consultation questions input in the first component are processed

through extraction and screening by multiple agents (Summary Agent, Classification Agent) to obtain historically similar questions from the same region that are recent and have high like counts. Finally, a Government Response Agent generates reference answers for new questions by combining the new question with historical similar question responses, thereby serving rapid and effective responses to government consultations.

3.1 LLM Fine-Tuning

Although the base LLM used in this study performs excellently in understanding general Chinese text and has accumulated extensive corpus knowledge, there remains significant room for improvement in specific domains and downstream tasks. Particularly in the government consultation domain, LLMs require deeper contextual understanding and domain-specific expression capabilities. To this end, this study employs fine-tuning techniques to adjust the parameters of the base LLM to adapt to the government consultation Q&A dataset. Fine-tuning enables the base LLM to better understand knowledge in the government consultation domain, thereby effectively meeting practical application requirements.

Given the enormous number of parameters in LLMs, full-parameter fine-tuning would consume substantial computational resources across multiple batches. Therefore, this study adopts an efficient parameter fine-tuning method—Low-Rank Adaptation [27] (LoRA)—to adapt to government consultation tasks while reducing resource consumption. The core of LoRA technology lies in freezing the parameter weights of the pre-trained LLM and introducing two matrices, A and B, to replace the parameters changed during fine-tuning. During fine-tuning, only matrices A and B are updated rather than all model parameters. This method simulates parameter changes through low-rank decomposition, achieving indirect training of LLMs with minimal parameter increments, thereby better adapting to government consultation response tasks under existing resource constraints. After LoRA fine-tuning, the base LLM is used for subsequent agent construction and text generation tasks.

3.2 Multi-Agent Retrieval Enhancement

The second part of the GovLLM system primarily involves enhanced retrieval of historically similar questions based on multi-agent systems. Compared to direct retrieval enhancement, multi-agent-based enhanced retrieval can not only extract main text information and reduce noise but also share key information through collaborative operation, maximizing the retention of contextual information in government consultation questions. This enables more accurate retrieval of historical questions and makes LLM reasoning and generation more effective.

3.2.1 Historical Q&A Vector Database Construction To make historical government Q&A text data more easily retrievable and matchable, this study transforms the historical government question texts in the training set into a

text vector index library. Specifically, the BERT pre-trained model is applied for semantic encoding of texts, storing each historical government consultation question text in the form of sentence vectors. This vector index library can be used for screening and querying, allowing the retrieval of specific sub-vector libraries based on question categories and answer like counts, and enabling rapid retrieval of historically similar questions by calculating similarity between new question vectors and vectors stored in the vector database.

3.2.2 Agent Design The GovLLM agent system consists of three core agents, each performing different tasks to improve the efficiency and accuracy of text reasoning and retrieval. The functions of these three agents are defined through the Role and Content parameters in each conversation with the base LLM.

Summary Agent: The primary function of the Summary Agent is to extract and summarize key information from citizen government consultation questions. By analyzing consultation texts, this agent can identify and extract three core elements: problem background, current situation, and core demands. Similar to data cleaning staff in government platforms, it not only simplifies questions and reduces information bias caused by different citizen expression styles but also provides precise and simplified problem descriptions for subsequent processing steps. The design purpose of the Summary Agent is to lay the foundation for classification and response generation through accurate information extraction.

Classification Agent: The Classification Agent is responsible for categorizing summarized government consultation questions into predefined consultation categories. Through prompt engineering, this agent automatically assigns questions to one of the following categories: urban construction, environmental protection, transportation, education, finance, employment, tourism, enterprise, agriculture, culture and entertainment, healthcare, public security, or livelihood. The Classification Agent resembles staff responsible for question categorization and distribution in government platforms. Its design purpose is to correctly classify each input question and obtain a specific category database of historical questions based on the question category, thereby enhancing retrieval targeting and search efficiency.

Government Response Agent: The Government Response Agent is the key component for generating targeted responses based on summarization and classification retrieval results. This agent assumes the role of government response staff, comprehensively considering the core elements provided by the Summary Agent and the classification results from the Classification Agent, while retrieving historically similar and recent questions from the same region with high like counts to generate targeted responses for new questions. The design purpose of the Government Response Agent is to simulate the process of government staff responding to citizen consultations, leveraging the base LLM' s text understanding and reasoning capabilities to reduce manual intervention, improve response speed, and enhance citizen satisfaction.

The prompt templates for the three agents are shown in Table 1 . These three agents operate collaboratively, establishing a complete information transmission chain that collectively constitutes an efficient and intelligent government consultation processing system aimed at improving response quality and efficiency through automated and intelligent means.

3.2.3 Enhanced Retrieval Based on Vector Similarity and Question Attributes Based on the constructed historical question vector index library and three core agents, the GovLLM system efficiently and accurately retrieves historically similar government consultation questions from the same region that are recent, most similar, and have high answer like counts. Specifically:

First, the Summary Agent analyzes newly input government consultation questions, extracts key information, simplifies the text, and generates question vectors. Simultaneously, the Classification Agent categorizes the input question, obtains the question category, and further screens the same-category sub-library from the historical question index library to narrow the retrieval scope and improve efficiency. Based on this, the system employs a cosine similarity matching algorithm for coarse screening of historical question vectors in the sub-library to identify candidate questions with high similarity to the new question.

Next, the system performs fine screening and ranking of the coarse-screened results based on the new question' s location and temporal information to ensure that selected questions are geographically and temporally close to the new question and have high citizen satisfaction with responses. The screening criteria for this step are: same province/city, same category, recent time, and high like count, thereby obtaining the top-ranked refined results to retrieve the Top K (customizable, default is 5) historically similar questions and their responses after fine screening.

Finally, the Government Response Agent combines the extraction results from the Summary Agent and the responses from the refined historical similar questions to generate reference responses for government staff to reference and modify during their replies. This process not only improves the automation level of question processing but also ensures the effectiveness, relevance, and timeliness of responses, thereby enhancing the quality and efficiency of government consultation responses.

Through this system design, GovLLM can effectively utilize historical data to provide intelligent processing solutions for government consultation questions, achieving the research goal of building a responsive government.

3.3 System Evaluation

This study employs BERTScore as the primary metric for evaluating the quality of system-generated text. Compared to traditional evaluation metrics such as BLEU and METEOR, BERTScore offers significant advantages. It can capture

underlying word relationships and semantic information rather than simply lexical arrangements, making BERTScore more effective for evaluating semantic similarity between texts. BERTScore correlates better with human judgment, with evaluation results closer to human intuitive perception of text similarity. It is less susceptible to common word interference because it relies not merely on n-gram matching mechanisms but considers the entire sentence' s contextual information. Through these combined advantages, BERTScore leverages BERT model' s pre-trained contextual embeddings and measures the matching degree between words in generated and reference texts via cosine similarity [28].

The process first converts generated and reference texts into embedding representations, then uses the BERT model to evaluate similarity scores between these embeddings, thereby reflecting semantic similarity between the two text segments. Specifically, RBERT (Recall BERT) measures the sum of maximum similarity between reference and generated texts divided by the total number of reference texts, representing recall. PBERT (Precision BERT) measures the sum of maximum similarity between generated and reference texts divided by the total number of generated texts, representing precision. FBERT (F1 BERT) is the harmonic mean of precision and recall, providing a balanced similarity evaluation metric. The relevant metrics are shown in formulas (1) to (3):

Where x represents the reference text for evaluating generated text quality, xi represents the i-th reference text, y represents the system-generated text, and yj represents the j-th system-generated text.

---

4 Experimental Results and Analysis

4.1 Test Data

This study selected the top 1,000 Q&A pairs ranked by answer like counts from the total dataset as the test set, which was excluded from the training set. Questions from the test set were used as input to obtain system-generated texts, while answers from the test set served as reference texts for evaluating system-generated text quality. This study measured and quantified the quality of generated texts by calculating BERTScore between system-generated texts and reference texts, thereby assessing the system' s ability to answer government questions.

4.2 Baseline Reference Models

To evaluate GovLLM' s performance, this study selected several common baseline models for comparison, including GPT-4o, Qwen-14B, Kimi, and ERNIE. GPT-4o is the latest large language model developed by OpenAI. Kimi is the latest multilingual dialogue system developed by Moonshot AI, while ERNIE is the latest semantic understanding model launched by Baidu. These models collectively constitute the baseline models for this study, enabling accurate

measurement of GovLLM' s performance.

4.3 Generated Text Quality Testing

Based on the aforementioned test dataset, this study conducted generated text quality tests on GovLLM and other baseline models, with the primary metric being the average BERTScore of these models based on the test set reference texts. The test results include the models'overall performance on the full dataset and their performance on specific categories of government questions, as shown in Table 2 .

The results shown in Table 2 indicate that GovLLM outperforms the four baseline models on the full dataset, with approximately 10% performance improvement. This means GovLLM-generated texts are highly similar to high-like-count response texts, demonstrating high text quality and application value. Additionally, for specific question types, GovLLM maintains equivalent performance to that on the full dataset, while some baseline models exhibit significant performance fluctuations on specific type test sets. Benefiting from the Classification Agent design in GovLLM, the system can targetedly retrieve historical questions and responses of the same type as the test question from historical data and generate type-specific responses accordingly. This design enables GovLLM to maintain stable and excellent generated text quality across different types of government questions.

4.4 Parameter Sensitivity Testing

Although GovLLM' s overall framework is fixed, it contains numerous adjustable parameter settings. Different parameter configurations significantly affect GovLLM' s efficiency and performance. Moreover, practical application scenarios differ substantially from experimental scenarios, often lacking sufficient computational resources. Therefore, to test GovLLM' s sensitivity to these parameters and identify the most suitable parameter adjustment scheme for practical application scenarios, this study conducted controlled variable tests on Top K values (number of historical similar responses input into the Government Response Agent), fine-tuning data volume, and historical vector library data volume, using FBERT as the evaluation metric. The results are shown in the figures.

Figure 2 [Figure 2: see original paper] shows that under full fine-tuning, increasing the Top K value significantly improves FBERT when K is small, thereby enhancing generated text quality. However, when K exceeds 5, FBERT shows a slight downward trend. This effect is more pronounced when using the full historical vector dataset. A possible explanation is that when fewer similar responses are provided to the Government Response Agent, these responses are highly relevant to the new question and of good quality, allowing the Agent to quickly learn patterns from similar responses and generate high-quality answers. However, when K exceeds a certain threshold, the relevance and quality of similar responses decrease, causing the Agent to learn noise information from

low-quality responses, thereby reducing answer quality. Therefore, in practical applications, the optimal K threshold should be determined through sensitivity testing. For the GovLLM system, K can be set to 5.

Figure 3 [Figure 3: see original paper] explores the impact of historical vector library data usage on generated text quality under full fine-tuning. The experiment found a non-linear relationship between vector database usage and FBERT. When vector database usage is small, FBERT increases significantly with usage, but after exceeding a certain threshold, the growth rate slows and eventually stabilizes. Additionally, K value selection also affects FBERT, with K=5 better utilizing the historical vector database than K=3 to improve text generation quality. In practical applications, although using the full historical vector database can fully utilize data, it also increases retrieval time and reduces efficiency. For the GovLLM system, under full fine-tuning and Top K=5 conditions, selecting 70% vector data usage can achieve over 90% of the FBERT value while reducing retrieval time, meeting application scenarios with limited computational resources but certain text quality requirements.

Figure 4 [Figure 4: see original paper] analyzes the impact of fine-tuning data volume on generated text quality. Fine-tuning data volume shows a positive correlation with FBERT, but the improvement effect is limited with slow growth. In contrast, higher historical vector database usage can significantly improve FBERT. In practical applications, LLM fine-tuning helps the model quickly understand the context and expression patterns of the government domain, improving text quality. However, the marginal benefit of large-scale fine-tuning data training is low, requiring reasonable allocation of fine-tuning data and vector library data proportions based on computational resources and efficiency needs to maximize system performance.

4.5 Ablation Experiments

To examine the performance contribution and effectiveness of the fine-tuning module, vector database module, agent module, and retrieval-ranking module based on question attributes in the overall GovLLM system, this study removed some modules from GovLLM to form five new models and conducted generated text quality tests on the full test set. The results are shown in Table 3 . In Table 3, -RFT (remove finetune) refers to removing the fine-tuning module, -RVD (remove vector database) refers to removing the vector database, -RA (remove agent) refers to removing all agent modules (replaced with direct Q&A prompts), and -RR (remove ranking) refers to removing the ranking module based on question and response attributes (only recalling historically similar responses based on similarity matching).

Table 3 shows the ablation experiment results, which test and record the metrics and their changes after progressively adding new modules to the base model. The results indicate that although fine-tuning training enables LLMs to understand corpus in the government domain, its improvement effect on gener-

ated text quality is relatively limited. The introduction of the historical vector database module significantly improves the model' s F1 BERT value, as the historical vector database explicitly guides and supervises LLM text generation through prompt engineering, with more significant improvement effects than fine-tuning training. The agent module and the retrieval-ranking module based on question-response attributes contribute to LLM performance improvement by extracting key information, determining question types, narrowing retrieval scope, and integrating high-quality responses from historically similar questions. The attribute retrieval-ranking module ensures regional applicability and effectiveness of response content by screening out similar high-quality responses from the same category and region, excluding irrelevant data. In summary, GovLLM' s four modules each have distinct functions and utilities, providing different degrees of performance improvement to the base model, enhancing generated text quality, improving retrieval and generation efficiency, and strengthening the effectiveness and practicality of text content in the government domain.

4.6 Results Analysis and Discussion

Table 4 shows the generated response texts from GovLLM and other baseline models for the same input question. The results demonstrate that baseline models do not capture the important location information of Guangzhou, so although their responses appear to answer the question, their content cannot directly serve Guangzhou citizens, reducing the practicality and effectiveness of their responses. This represents an area where current government Q&A systems need improvement. GovLLM, benefiting from the collaborative action of the historical vector database and agent system modules, generates response texts that not only align with the actual situation in Guangzhou but also provide feasible solutions, significantly improving response effectiveness and practicality. Citizens can solve their government problems following its content, and government platforms can edit and modify the generated reference responses to form final replies, improving the efficiency of government platform question responses.

Based on the experiments and results, this study yields several meaningful findings discussed from both algorithmic and organizational management perspectives.

From an algorithmic perspective, this study provides a practical and scalable solution for LLM applications in the government domain. First, the module design (vector database, agents, etc.) fully leverages LLMs' pattern recognition and text generation capabilities, making generated content more controllable and effective for targeted adjustment and modification by relevant personnel, enabling cross-domain applications. Second, parameter sensitivity testing quantifies the impact of different parameter selections on final system performance, helping relevant personnel select appropriate parameters based on test results and actual computational resources for application in real scenarios. Finally, ablation experiments quantify the improvement effects of different modules, helping rele-

vant personnel prioritize optimization of specific modules with limited resources to maximize system performance. In summary, this study not only provides a system framework applicable to government question responses but also offers relevant experimental results to facilitate its practical deployment.

From an organizational management perspective, this study provides an effective tool system for achieving the goal of building a "responsive government." First, the proposed GovLLM system helps improve government platforms' big data processing capabilities, particularly regarding the efficiency of responding to large volumes of government questions. Staff no longer need to spend time and effort on tedious search and retrieval work but only need to verify and modify generated reference responses, improving response efficiency and reducing unnecessary resource investment. Second, based on the agent system and retrieval-ranking modules, GovLLM can integrate government question-answering tasks from all provinces and municipalities within a single system, which helps eliminate information gaps in government services between provinces and municipalities and build a "one-stop" government service platform, providing an example solution for building a nationwide responsive government. For citizens, they no longer need to cumbersomely query different provincial and municipal government Q&A platforms but can achieve government Q&A for different regions through a single platform. For governments, managing and maintaining a unified government platform is more efficient, and policy information gaps between provinces and municipalities are significantly reduced. Finally, this study also promotes the process of building "responsive governments" across government platforms. The system not only helps government platforms respond to large volumes of government questions in a timely and efficient manner but also improves the effectiveness and practicality of responses, fully considering actual regional differences across China, with response content being more specific and aligned with local actual conditions for citizens. This is consistent with the core concept of building a "responsive government" in China.

5 Conclusion

This study successfully constructed the GovLLM system, an LLM-based government reference response generation system that improves model performance by 5-11% compared to baseline models, effectively expanding LLM application capabilities in the government domain. The GovLLM system integrates multiple modules including a historical vector database, agent system, and retrieval-ranking, improving response efficiency and text quality while ensuring content aligns with regional actual conditions, mitigating LLM text generation hallucinations, and enhancing system usability. Additionally, the system helps reduce the labor costs of pure manual responses and improves the operational efficiency of government platforms. Overall, the GovLLM system provides a practical and effective technical solution for building a "responsive government," advancing the intelligent process of responsive government platforms.

6 References

[1] Alvarenga A, Matos F, Godina R, CO Matias J. Digital transformation and knowledge management in the public sector[J]. Sustainability, 2020, 12(14):5824.

[2] Gamper J, Augsten N. The role of web services in digital government[C]//Proc of Conference on Electronic Government. Berlin: Springer, 2003: 74-82.

[3] Bélanger F, Carter L. The impact of the digital divide on e-government use[J]. Communications of the ACM, 2009, 52(4):132-135.

[4] Lee-Geiller S, Lee TD. Using government websites to enhance democratic e-governance: conceptual model evaluation[J]. Government Information Quarterly, 2019, 36(2):208-225.

[5] Bastos D, Fernández-Caballero A, Pereira A, Rocha NP. Smart city applications to promote citizen participation in city management and governance: a systematic review[J]. [5], 2022, 9:89-118.

[6] Chen Wenquan, Yu Yujie. Research on the responsiveness and path of service-oriented government construction in the network environment: taking the 2013 consecutive replies of secretaries and governors of five provinces (cities) to netizens' messages as an example[J]. Chinese Public Administration, 2014, (7): 74-77.

[7] Gupta P, Gupta V. A survey of text question answering techniques[J]. International Journal of Computer Applications, 2012, 53(4):1-8.

[8] Daniel J, James H. Speech and language processing[M]. 2nd ed. Upper Saddle River: Prentice Hall, 2014.

[9] Hao Tianyong, Li Xinxin, He Yulan, Wang Fu Lee, Qu Yingying. Recent progress in leveraging deep learning methods for question answering[J]. Neural Computing and Applications, 2022, 1:1-9.

[10] Dong Hui, Yu Siwei, Jiang Ying. Text mining on semi-structured e-government digital archives of China[C]//Proc of 2009 Second Pacific-Asia Conference on Web Mining and Web-based Application. Piscataway, NJ: IEEE Press, 2009: 11-14.

[11] Gao Shangsheng, Gao Li, Li Qi, Xu Jianjun. Application of large language model in intelligent q&a of digital government[C]//Proc of the 2023 2nd International Conference on Networks, Communications and Information Technology. New York: ACM Press, 2023: 24-27.

[12] Nigam SK, Mishra SK, Mishra AK, Shallum N, Bhattacharya A. Legal question-answering in the Indian context: efficacy, challenges, and potential of modern AI model[OL]. (2023-09-26)[2024-11-06]. https://arxiv.org/abs/2309.14735.

[13] Tang Ruixiang, Yu-Neng Chuang, Xia Hu. The science of detecting llm-generated text[J]. Communications of the ACM, 2024, 67(4):50-59.

[14] Wu Shijie, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G. Bloomberggpt: a large language model for finance[OL]. (2023-03-30)[2024-11-06]. https://arxiv.org/abs/2303.17564.

[15] Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine[J]. Nature Medicine, 2023, 29(8):1930-1940.

[16] Baek J, Jeong S, Kang M, Park JC, Hwang SJ. Knowledge-augmented language model verification[OL]. (2023-10-19)[2024-11-06]. https://arxiv.org/abs/2310.12836.

[17] Tan Bowen, Yang Zichao, AI-Shedivat M, Xing EP, Hu Zhiting. Progressive generation of long text with pretrained language models[OL]. (2020-01-28)[2024-11-06]. https://arxiv.org/abs/2006.15720.

[18] Welch EW, Hinnant CC, Moon MJ. Linking citizen satisfaction with e-government to trust in government[J]. Journal of Public Administration Research and Theory, 2005, 15(3):371-391.

[19] Katsonis M, Botros A. Digital government: a primer and professional perspectives[J]. Australian Journal of Public Administration, 2015, 74(1):42-52.

[20] Zhang Yuhao. Digital Government Construction Perspective: a Study on the path to promote the improvement of government public management capacity[J]. Frontiers in Business, Economics and Management, 2022, 6(3):35-40.

[21] Patsoulis G, Promikyridis R, Tambouris E. Integration of chatbots with Knowledge Graphs in eGovernment: getting a passport[C]//Proc of the 25th Pan-Hellenic Conference on Informatics. New York: ACM Press, 2021:425-429.

[22] Schwarzer M, Düver J, Ploch D, Lommatzsch A. An Interactive e-Government Question Answering System[J]. Labor & Workforce Development Agency, 2016:74-82.

[23] Beltrán A, Ordoñez S, Monroy S, Melo L, Duarte N. Question-answering systems specific domain e-government[OL]. (2016-04-01)[2024-11-06]. http://dx.doi.org/10.13140/RG.2.1.2491.6889.

[24] Rodsawang C, Thongkliang P, Intawong T, Sonong A, Thitiwatthana Y, Chottanapund S. Designing a competent chatbot to counter the COVID-19 pandemic and empower risk communication in an emergency response system[J]. Outbreak, Surveillance, Investigation & Response (OSIR) Journal, 2020, 13(2):71-77.

[25] Hua Bin, Kang Yue, Fan Linhan. Knowledge modeling and association Q&A for policy texts[J]. Data Analysis and Knowledge Discovery, 2022, 6(11):79-92.

[26] Cui Congmin, Shi Yunmei, Yuan Bo. Research on relation extraction method for government documents[J]. Computer Technology and Development, 2021, 31(12):26-32.

[27] Hu EJ, Shen Yelong, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang Lu, Chen Weizhu. Lora: Low-rank adaptation of large language models[OL]. (2021-01-17)[2024-11-06]. https://arxiv.org/abs/2106.09685.

[28] Zhang Tianyi, Kishore V, Wu Felix, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert[OL]. (2019-04-21)[2024-11-06]. https://arxiv.org/abs/1904.09675.

Note: Figure translations are in progress. See original paper for figures.