IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 December 2023, accepted 24 December 2023, date of publication 4 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3349969

RESEARCH ARTICLE

Intelligent Practices of Large Language Models in Digital Government Services

JIAWEI HAN[1,2], JIANKANG LU[1], YING XU[3], JIN YOU[1], AND BINGXIN WU[1]
[1]College of Cyber Security, Changchun University, Changchun 130022, China
[2]Digital Identity and Blockchain Joint Laboratory, Peking University, Beijing 100871, China
[3]School of Administration, Changchun University, Changchun 130022, China

Corresponding author: Jiawei Han (hanjw78@ccu.edu.cn)

This work was supported in part by the Jilin Provincial Development and Reform Commission Planning Project under Grant 2020C020-2, in part by the Science and Technology Development Center of the Ministry of Education under Grant 2018A04002, in part by the Growth and Climbing Plan Fund Project under Grant ZKP202125, and in part by the Growth Research Fund Project under Grant ZKQD201914.

ABSTRACT Large language models have been widely used in open-domain tasks with significant results, as well as being able to perform zero-sample closed-ended questions based on internal knowledge stored in the parameters during pre-training to answer the task. However, this internalized knowledge may be insufficient or the knowledge may be outdated in responding to government service consultation scenarios, which may result in the inability of the large language models to perform accurate and rigorous answers and provide effective assistance. This issue has attracted widespread attention, but there is a lack of datasets for relevant research. Therefore, in this paper, we take Beijing as an example to collect all kinds of common service counseling questions from its government website and the corresponding official answers as a dataset, which contains the daily counseling questions encountered by the citizens, including common questions about medical insurance, social insurance, provident fund flexible employment, and government-private interaction. Therefore, this paper designs a domain-specific language model (GCALLM) for government service consultation based on this scenario. By fine-tuning the large language models for knowledge injection, the fine-tuned model helps the large language models improve their performance in governmental service consulting scenarios by providing contextual information. And it solves the problem of not being able to answer precisely, allowing for more rigor and accuracy of the answers. In addition, the response information is answered in seven major national languages to improve the construction of digital government consulting services. A large number of experiments have proved that the model can produce accurate responses in this scenario in the field of governmental counseling.

INDEX TERMS Digital government, governmental counseling, GCALLM, large language models, effective assistance.

I. INTRODUCTION

GPT [1], Google's PaLM [2], and other benchmark models [3], [4], [5], [6], [7] are prominent in the field of artificial intelligence. The pretraining on large amounts of data enables these models to have exceptional language understanding and generation capabilities [8]. When it comes to domain-specific problems, large language models exhibit limited performance due to their insufficient pre-training on domain knowledge, and the overwhelming presence of domain-generalized data causes them to prioritize public knowledge, leading to potential oversight of critical domain-specific information [8], [9], [10]. However, the scale and cost of training large language models from scratch are very costly for most companies and researchers. Fine-tuning existing models based on domain-specific data is another option, and several studies have shown efficient strategies for achieving this step, among them P-tuning v2 [11] which rivals full incremental fine-tuning at any scale for question-answering tasks, not only does it reduce the number of parameters that need to be fine-tuned to 0.1% of the original, but by using methods such as model quantization and Gradient Checkpoint, it is possible to improve training efficiency while reducing the amount of

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

memory and computational resources consumed by the model when in use. These methods make the model more lightweight by optimizing and compressing it to reduce the demand for resources. At the same time, these methods have good performance preservation, i.e., they reduce resource consumption while keeping the performance of the model unaffected significantly. These methods make the model more lightweight by optimizing and compressing it to reduce the demand for resources. These methods are effective in maintaining performance, reducing resource consumption, and maintaining the model's unaffected performance.

Along with the rise of the large language models and the effective application of various fine-tuning technologies, the government consulting service should also usher in new development opportunities. Consulting service is one of the important works of the government office, and the excellent dialogue ability of the large language models can accelerate the development of intelligent consulting services. At present, the intelligence level of China's consulting service robots is insufficient, it can only output the original text matching the preset answers, and it can't make personalized answers to the user's questions. The text-to-text (T2TT) translation function with the help of the large language models and SeamlessM4T [12] can solve the demand for intelligent counseling to help the public more easily access, understand, and master the government's processes and work, avoiding problems such as no door to do business, no one to consult, and ineffective communication. Thus, the service is continuously provided around the clock. On the other hand, with the development of smart government, more and more government departments will begin to gradually utilize the large language models to improve the quality of consulting services, which has some limitations in providing accurate answers due to the relatively insufficient knowledge base of the large language models in the field of governmental consulting at present. This means that the model may not be able to give complete and accurate answers to questions involving the domain of governmental consulting.Therefore, this paper proposes a scenario-specific domain model to provide accurate policy information services for people, addressing the long-standing problems of difficulty in use and convergence without wisdom. This is bound to have a positive impact on socio-economic development.

Existing work, on the one hand, focuses on extracting domain-specific knowledge through retrieval-based approaches to enhance the performance of large language models in specific domains [13], [14] or external modules [15], [16]. On the other hand, text embeddings are employed to retrieve potentially relevant text summaries from a corpus of domain-specific documents. The text is transformed into real-valued vectors encoding meanings. Identification of the text involves vectorizing the query question with all text blocks stored in the vector repository, pre-computed using cosine similarity between the query vectors and each block. A small portion of the most relevant data blocks is then appended to the user query, constructing effective prompts that are sent to the significant language model, resulting in a response [17]. However, these methods have some limitations. Therefore, to address these writing problems, this paper proposes to fine-tune a large language model using P-tuning v2 [11], i.e., ChatGLM [6], which fine-tunes a large language model using the domain documents of a government consulting service to construct prompts. This fine-tuned large language model provides knowledge specific to government service consulting services at runtime. It makes it easier to maintain and protect privacy within the domain of government service consulting.

The contribution of this article is to collect common consultation Q&A information from government services in Beijing as a dataset and as a data source for fine-tuning the large language models. The fine-tuned model based on the P-tuning v2 [11] technique of injecting domain knowledge is used to provide good contextual information for the large language models, which is combined with the large language models to generate knowledge in the domain of government service counseling, and a large number of experiments have proved that the method of this paper is superior to the embedded text-based method in the domain of government service counseling. To better build the government service consulting service of digital government, the responsive answers express precise information using seven major languages in the world, which are Chinese, French, Spanish, Portuguese, German, and Japanese. New ideas and methods are provided for digital government reference counseling services to help digital governments better adapt to the information age changes and provide more innovative and valuable services.

II. RELATED WORK

The training of large language models is usually divided into two phases: pre-training and fine-tuning. In the pre-training phase, the model is trained using a large amount of unlabeled text data to learn the basic laws and patterns of language pairs, and the large language models thus acquire basic language comprehension and generation capabilities. In the fine-tuning phase, the model's broad language comprehension is made more adaptable to specific tasks and better responds to domain-specific nuances, thus improving the model's ability to generalize to unknown tasks [18], [19]. However domain-specific tasks often involve complex concepts, technical terminology, and complex relationships between entities [20]. In the absence of targeted guidance, large language models can be seriously illusory.

Recently, many relevant large language models have emerged in the medical [21], [22], [23], financial [24], [25], and legal domains [26], [27]. Efforts have been made to improve the context-generation capabilities of large language models in specific domains using a combination of external knowledge [28]. Among them are external modules to improve the context generation capability of large language models, such as TaskMatrix [7] and AutoGPT [16]. These

approaches are highly dependent on the prompt management of the large language models as well as the availability of external tools. However, for domain-specific scenarios, these external modules are not always effective. Alternatively, the use of chatbots [17], enables the use of large language models for specialized domains without updating the parameters. It is also possible to inject domain knowledge into the model using updating parameters using fine-tuning, and there have been some notable advances in fine-tuning techniques [11], [29], [30], [31], [32]. Among them, P-tuning v2 [11] is an optimization and adaptation implementation of Deep Prompt Tuning [17], [33]. One of its final improvements is to apply continuous prompts to each layer of the pre-trained model, not just the input layer. By increasing the capacity of continuous prompts, and for various settings, P-tuning v2 [11] improves performance comparable to full fine-tuning, and only needs to fine-tune 0.1% ˜ 3 % of the parameters. This paper mainly uses the latter in the field of government service consulting.

At different training stages, the training methods for large language models can be broadly categorized as follows: one is to train from scratch using domain data, Bloomberggpt [24] relies on a large amount of domain data, and the training cost is relatively expensive; one is to fine-tune based on the domain instruction data, e.g., ChatLaw [26]; and another method is to train the domain on the basic large language models based on the domain data, and then fine-tune with the instructions [27].

III. METHOD

A. LM PROMOPTING

Language models are pre-trained by predicting the next tokens based on the previous tokens, known as autoregressive language modeling [34], [35], and training by this method enables zero-sample instruction learning. Specifically, a question and an instruction message are provided to the language model, and the model generates a corresponding answer based on the input text message. More specifically, each input message is first modified into a text string X called a prompt, using a template T for a particular instruction. As follows $T : x \rightarrow X$. There is a question $x =$'' The term of the collective contract ?'', and command templates T: '' Please answer this question:$\{x\}$ '', then the resulting hint will be $T(x) =$ ''Please answer this question: the term of the collective contract ? ''. The prompt is then forwarded to the large language models, through which the answer is generated. There are several challenges with this approach, the large language models rely too much on the knowledge of the parameters, and the lack of knowledge in the governmental domain tends to produce factually incorrect information.

B. GCALLM

To address the limitations of the current prompting scheme with language models, this paper combines the relevant domain knowledge in the fine-tuned model with the question and then forwards it to the large language models to generate the answer through the large language models. While giving full play to the advantages of the large language models, and standardizing its answer generation to ensure that it is sufficiently complete and accurate, this paper adopts a fine-tuned large language model combined with another large language model. The model interaction (shown in FIGURE 1) involves two key steps: obtaining a fine-tuned domain-specific model of the large language models and providing the generated domain-specific knowledge to the large language models.

In the first step, this paper uses the Q&A knowledge captured in the government service consultation documents to form QA Q&A pairs for domain fine-tuning of the large language models. The government service consultation documents are used as a knowledge base that contains a variety of common consultation questions. By utilizing this knowledge base, domain-specific knowledge is infused into the large language models through fine-tuning using P-tuning v2 [11]. To facilitate model fine-tuning, prompts are constructed using the collected dataset of common government service inquiries and each prompt consists of a question and an answer. For example:

{''content'': ''Can 40-year-old female comrades handle flexible employment ?'', ''summary'':'' Can handle.''}

{''content'':'' Flexible Employee Social Security Contributions for the month of the increase in what time to deduct the fee ?'', ''summary'':'' The next monthly deduction for the procedure of adding members in the same month.''}

In the second step, the fine-tuned model, which has been injected with domain knowledge, provides domain-specific Q&A knowledge of the large language models when the user accesses the fine-tuned model D. To follow specifically, first, using a command-specific template T', each input is modified into a text string called a prompt X', as shown in the following $T' : (x, D) \rightarrow X'$. For example, there is a question $x =$ ''Duration of the collective contract ?''. Fine-tuning model tip messages D='' Collective contracts generally have a duration of 3 years, and instruction templates definite information {D}. Based on the above-known information, use the known information for output, the question is:{x}''.Then the hint X' is obtained as ='' Known information: The term of a collective contract is usually 3 years''. Based on the above-known information, use the known information to produce an output that cannot be modified or added to. The output should be in Chinese. The question is: What is the duration of a collective contract ? ''. The prompt is then forwarded to the large language models, through which the answer is generated. This knowledge enrichment enhances the understanding of the task context by the large language models, enabling them to generate more accurate and contextually relevant responses. The obtained response message is passed to the translation module. The module is a large-scale multilingual T2TT model (SeamlessM4T-NLLB [12]), which can understand

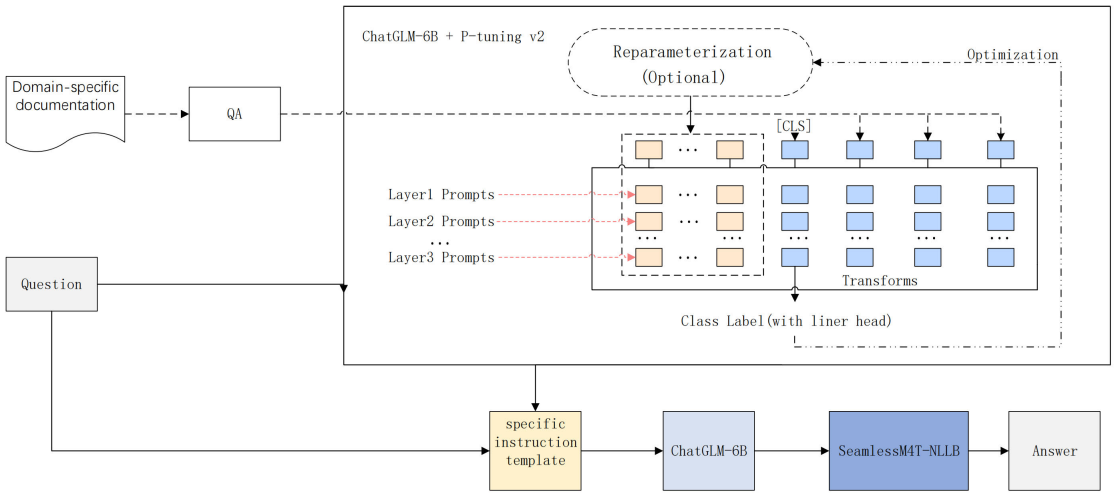

FIGURE 1. GCALLM model framework.

texts in nearly 100 languages and generate corresponding translated texts, as well as consists of a Transformer Text Encoder and a Transformer Text Decoder, which encode the response message through the text encoder, and decode the encoded message through the decoder to translate the response message into other major national languages.

IV. EXPERIMENTS

A. DATASET

In order to solve the use of a governmental consulting scenario dataset, this paper takes Beijing as an example, and collects the common questions of its governmental service consulting and the corresponding official standard answer as a dataset, which includes the frequently asked questions of medical insurance and the content of convenient services and other information.

1) DATA COLLECTION

Medical insurance, social insurance, housing provident fund, and flexible employment are common and popular keywords in government consulting services. The corresponding questions are common consultation questions from citizens, as well as some help and suggestions. Citizens can find relevant questions or provide suggestions on the official website of Beijing City, and the relevant departments will respond promptly. Below are three brief examples of Q&A responses.

Q: How long do you need to be in flexible employment to be eligible for a flexible employment subsidy again?

A: Complete 90 days.

Q: Can you switch to a CPF personal housing loan if you have already taken out a commercial loan?

A: Currently, you cannot transfer a commercial loan to a Housing Provident Fund (HPF) personal housing loan in Beijing.

Q: Can insured persons use their personal accounts at designated retail pharmacies?

TABLE 1. Government data analysis.

Question	Category	Number
Medical insurance	Issues related to medical insurance.	207
Humanities Beijing	Various literary and artistic activities in Beijing.	100
Various consultations	Various problems of government-citizen interaction.	197
Social insurance	Human resources and social security issues.	257
Surplus	Housing provident fund issues.	61
Flexible employment	Employment subsidy issues.	56
Bus opinions	Opinions to the bus company.	122

A: It can be used by insured persons to pay for their personal expenses on drugs, medical devices, and medical consumables at designated retail pharmacies.

In this paper, we will focus on answering questions with short word counts and extracting the answers corresponding to the official departmental responses.

2) DATA FILTERING AND POST-PROCESSING

As the data comes from online websites, the content is complex and contains a large amount of irrelevant content, which makes it difficult to conduct research. To address this problem, this paper deeply samples the collected data, summarizes the problems, determines the patterns, and designs the rules for data filtering. (1) Deletes the name of the matter, the subject of implementation, which is about the direction of the content of the question, as well as the official unit that answers the question. (2) Normalize all connections in the data. (3) Remove some answer contents that are irrelevant to the question, such as:

"Thank you for your interest in the transit industry and we welcome you to continue to monitor our work."

(4) Detecting the length of questions that are too long and replacing them with questions that are displayed briefly on the website, without using detailed question information. (5) Delete some Q&A information whose answers are too long. (6) Delete the name information encrypted by the sender and the time of the consultation. By implementing these rules for data filtering, the quality and usability of the dataset are improved.

3) STATISTICS

The dataset contains 1k Q&A pairs, which are common government service inquiry questions, in which the types of questions are categorized into 7 categories, and the statistical information about the specific dataset is shown in TABLE 1. In the dataset, each question corresponds to a different category, and good categorization information makes it easier for users to understand the specific question information.

B. EXPERIMENTS SETUPS

The experiment takes Beijing as an example, collects a dataset of common questions and corresponding answers for government service consulting. This paper follows the evaluation process of the NLG task and uses a set of metrics to comprehensively evaluate the quality of responses generated by large language models given a question, the Chinese responses of the evaluated responses are compared with the standard Chinese responses. In this experiment, this paper evaluates large language models a fine-tuned model (DLLM), a knowledge base combined with a fine-tuned model (KBDLLM), and a GCALLM, it uses seven metrics to comprehensively evaluate the response quality of the large language models. We evaluated a large language model, ChatGLM [6]. It generates answers from input questions without introducing external domain knowledge.

1) GCALLM

GCALLM uses the P-tuning v2 [11] method to fine-tune ChatGLM [6] on the dataset collected in this paper to obtain a fine-tuned model of government service domain knowledge. Given a question, the domain information is first collected through the fine-tuned model, and then this domain information is assembled in conjunction with the question using a specific prompt template and then sent to the large language model to answer the question.

2) KBDLLM

First, the sentence vector of consulted government affairs information is obtained. This paper uses the text2vec-largechinese model from the Sentence Transformers database to obtain the sentence vector.After acquiring the sentence vector in the knowledge base, FAISS [36] is employed to conduct a similarity search, FAISS [36] uses the inverted indexing approach to accelerate the approximation vector search, the top_k answers obtained, using the appropriate chunk_size to complement the contextual information of these answers, the answers sorted and assembled into contextual information and user input questions using langchain's prompt_template function to generate the large language models input required prompt prompts to guide the fine-tuning of the large language models from the answer. The large language models for fine-tuning are guided to understand and analyze the answers from the references and generate accurate and complete answers, and at the same time, it is emphasized that it is forbidden to generate answers by itself when it cannot come up with an answer and reasonable hints are given.

TABLE 2. Result analysis.

	LLM	DLLM	KBDLLM	GCALLM
BLEU	0.1154	0.1794	0.1361	**0.305**
ROUGE-1	0.1954	0.3536	0.3013	**0.4862**
ROUGE-2	0.0460	0.1730	0.1428	**0.3407**
ROUGE-L	0.1421	0.2921	0.2459	**0.4157**
BERTScore-P	0.5834	0.6975	0.6825	**0.7688**
BERTScore-R	0.6564	0.7143	0.6639	**0.7576**
BERTScore-F1	0.6145	0.7029	0.6688	**0.7564**

ChatGLM results on government data, bolded numbers represent the best results, and the DLLM column shows the evaluation results obtained by directly querying the fine-tuned model. The LLM shows the evaluation results obtained by querying the LLM without applying any enhancement, and the rest of the two columns are a combination of the evaluation results under different enhancement methods.

3) DLLM

On the dataset collected in this article, the P-tuning v2 [11] method was used to fine-tune ChatGLM [6] to obtain a fine-tuning model for knowledge in the field of government services. Directly use the question to access the fine-tuning model answers.

C. METRICS

As the large language models generate content with intelligent and random characteristics, this paper adopts the evaluation index Bert Score [37] based on the language model for evaluation, which uses contextual embedding to calculate the similarity of the markers, and extracts the features through the Bert model by transforming the generating text and the reference text into tokens, respectively, and then calculates the inner product corresponding to each word in the two texts, to construct a similarity matrix, based on which the maximum similarity scores of the two texts are calculated and normalized, and finally the Precision, Recall and F1 values are obtained. BLEU [38], ROUGE-1, ROUGE-2, and ROUGE-L [39] are also used to measure the degree of agreement between the responses generated by the large language models and the standard answers.

D. EXPERIMENTAL ENVIRONMENTS SETTINGS

Our experimental environment uses the deep learning framework of Pytorch version 2.0.1 The NVIDIA graphics card

TABLE 3. Examples of government questions and answers.

Subject: Composed model	
Query:	Can I switch to a CPF personal housing loan if I have already taken out a commercial loan?
Standard Answer:	Currently, you cannot transfer a commercial loan to a Housing Provident Fund (HPF) personal housing loan in Beijing.
DLLM:	you cannot transfer a commercial loan to a Housing Provident Fund (HPF) personal housing loan in Beijing.
LLM:	In general, commercial loans and housing fund personal residence...
GCALLM(Chinese):	Currently, you cannot transfer a commercial loan to a CPF personal housing loan in Beijing.
GCALLM(French):	Actuellement impossible d'effectuer prêts commerciaux transfert logement fonds commun de réserve opérations de prêts individuels au logement.
GCALLM(German):	Derzeit nicht in Peking abzuwickeln kommerzielle Kredite um Wohnungs Kompensationsfonds persönliche Wohnkredite Geschäft .
GCALLM(Japenese):	在北京で手きできない商融住宅公积金人住宅融事 .
GCALLM(Portuguese):	Actualmente não se pode fazer em Pequim empréstimos comerciais transfer habitacional fundos comuns de empréstimos individuais habitacionais .
GCALLM(Spanish):	Actualmente no se puede tramitar en Beijing préstamos comerciales trans vivienda común de depósito personal negocio de préstamos vivienda .
KBDLLM:	The result is not permissible.

driver version is Cuda 12.0, the graphics card model used is RTX 3090 * 1, and the operating system version is Ubuntu 20.04.

V. RESULT

TABLE 2 gives the evaluation results of ChatGLM [6] on seven metrics of the governmental dataset, in general, after fine-tuning the model's enhancement. Some metrics such as BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore-P, BERTScore-R, BERTScore-F1.From Table 2, using only large language models for answering government consultation questions yields suboptimal results. The BERTScore-F1, BERTScore-P, and BERTScore-R values are 0.6145, 0.5834, and 0.6564, respectively. The BLEU score is 0.1154, indicating a low precision match between generated answers and standard answers. The ROUGE-1 value is 0.1954, suggesting a low overlap at the individual word level. The ROUGE-2 value is 0.0460, indicating a low overlap at the two consecutive word levels, and the ROUGE-L value is 0.1412, implying a low long-sequence matching and semantic relevance. Comparatively, DLLM shows improvements over LLM in BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore-F1, BERTScore-P, and BERTScore-R. It enhances precision matching, single-word overlap, consecutive two-word overlap, and semantic relevance in answering government consultation questions. KBDLLM, compared to LLM, enhances precision matching, single-word overlap, consecutive two-word overlap, and semantic relevance. However, compared to DLLM, it exhibits a decrease in precision matching, single-word overlap, consecutive two-word overlap, and semantic relevance. Despite attempts to mitigate over-interpretation through prompts, KBDLLM's performance is inferior to DLLM. GCALLM surpasses LLM, DLLM, and KBDLLM in BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore-F1, BERTScore-P, and BERTScore-R, achieving the highest numerical values. This results in optimal precision matching, single-word overlap, consecutive two-word overlap, and semantic relevance in generating answers compared to standard answers. Its outstanding performance in answering consultation questions is attributed to specific instruction templates and the incorporation of domain knowledge. GCALLM ensures the rigor and accuracy of answers, enhancing consultation services' intelligence, interactivity, and user-friendliness. The SeamlessM4T-NLLB module generates answers in seven different languages, accelerating the development of government service consultations.

VI. CASE STUDY

The responses with ChatGLM [6] before and after the approach combining the three models are given in TABLE 3, due to too much information in the word count, only some of the key information is given and ellipses are used to replace the rest of the content, it can be observed that the large language models causes a wrong understanding of the question and fails to provide the correct answer without any knowledge prompts. Injecting domain knowledge into a large language model to create a domain fine-tuned model with contextual knowledge has yielded promising results in generating responses without any explicit knowledge prompts. However, the completeness of the generated answers falls short when compared to the standard responses. The knowledge-based approach combined with fine-tuned modeling causes the fine-tuned model to generate brief answers due to too much information prompted by the knowledge base. Compared to the standard answers, this method exhibits a relative lack of completeness in information response. After undergoing GCALLM, accurate responses matching standard answers indicate the effectiveness of GCALLM in the context of government service consultations. The provided information is accurately presented in the seven major national languages, accelerating the development of digital government consulting services.

VII. CONCLUSION AND FUTURE WORK

This paper explores the use of the fine-tuning model combined with the large language models and knowledge base combined with the fine-tuning model, which can enhance the user query in the large language models in the government service consultation scenario. Due to the lack of relevant datasets, this paper proposes to use the common government-citizen interaction information on the website of the Beijing Municipal People's Government to construct a Q&A dataset, and then subsequently use a fine-tuning model combined with large language models for the large language models to provide domain-specific knowledge. To evaluate the effectiveness of the model, seven evaluation methods are used to verify the effectiveness of the model, and a large number of experiments prove that the domain-specific language model designed based on the scenarios of Beijing's governmental consulting service is effective.

However, the GCALLM model still suffers from phantom flaws, and fine-tuning the model based on the sequence of data, starting from low quality and progressing to high quality, short samples preceding long samples, and easy tasks preceding difficult ones, remains an area for improvement. In the future, the following steps of experimental exploration will be planned, including the implementation of other fine-tuning techniques and reinforcement learning to enhance the performance of GCALLM.

REFERENCES

[1] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[2] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.

[3] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpaca: A strong, replicable instruction-following model," *Stanford Center Res. Found. Models*, vol. 3, no. 6, p. 7, 2023. [Online]. Available: https://crfm.stanford.edu/2023/03/13/alpaca.html

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[5] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[6] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 320–335, doi: 10.18653/v1/2022.acl-long.26.

[7] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face," 2023, *arXiv:2303.17580*.

[8] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," *CoRR*, vol. abs/2303.12712, 2023, doi: 10.48550/ARXIV.2303.12712.

[9] A. Lecler, L. Duron, and P. Soyer, "Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT," *Diagnostic Interventional Imag.*, vol. 104, no. 6, pp. 269–274, Jun. 2023.

[10] C. M. C. Nascimento and A. S. Pimentel, "Do large language models understand chemistry? A conversation with ChatGPT," *J. Chem. Inf. Model.*, vol. 63, no. 6, pp. 1649–1655, 2023, doi: 10.1021/ACS.JCIM.3C00285.

[11] X. Liu, K. Ji, Y. Fu, W. Lam Tam, Z. Du, Z. Yang, and J. Tang, "P-Tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," 2021, *arXiv:2110.07602*.

[12] S. Communication et al., "SeamlessM4T: Massively multilingual & multimodal machine translation," 2023, *arXiv:2308.11596*.

[13] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W.-T. Yih, "REPLUG: Retrieval-augmented black-box language models," 2023, *arXiv:2301.12652*.

[14] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," 2023, *arXiv:2302.12813*.

[15] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual ChatGPT: Talking, drawing and editing with visual foundation models," 2023, *arXiv:2303.04671*.

[16] S. Gravitas. (2023). *Auto-GPT: An Autonomous GPT-4 Experiment*. [Online]. Available: https://github.com/Significant-Gravitas/Auto-GPT

[17] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, Jun. 2021, pp. 5203–5212, doi: 10.18653/V1/2021.NAACL-MAIN.410.

[18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[19] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[20] C. Ling et al., "Domain specialization as the key to make large language models disruptive: A comprehensive survey," 2023, *arXiv:2305.18703*.

[21] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen, "DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task," 2023, *arXiv:2304.01097*.

[22] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, and T. Liu, "HuaTuo: Tuning LLaMA model with Chinese medical knowledge," 2023, *arXiv:2304.06975*.

[23] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," 2023, *arXiv:2303.14070*.

[24] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.

[25] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-source financial large language models," 2023, *arXiv:2306.06031*.

[26] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "ChatLaw: Open-source legal large language model with integrated external knowledge bases," 2023, *arXiv:2306.16092*.

[27] Q. Huang, M. Tao, C. Zhang, Z. An, C. Jiang, Z. Chen, Z. Wu, and Y. Feng, "Lawyer LLaMA technical report," 2023, *arXiv:2305.15062*.

[28] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, and T. Scialom, "Augmented language models: A survey," 2023, *arXiv:2302.07842*.

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–26. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[30] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," 2023, *arXiv:2305.14314*.

[31] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Punta Cana, Dominican Republic. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059, doi: 10.18653/v1/2021.emnlp-main.243.

[32] Z. Liu, G. Wang, S. Zhong, Z. Xu, D. Zha, R. Tang, Z. Jiang, K. Zhou, V. Chaudhary, S. Xu, and X. Hu, "Winner-take-all column row sampling for memory efficient adaptation of language model," 2023, *arXiv:2305.15265*.

[33] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, vol. 1, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597, doi: 10.18653/v1/2021.acl-long.353.

[34] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
[35] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.
[36] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021, doi: 10.1109/TBDATA.2019.2921572.
[37] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–43. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr
[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.
[39] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

JIAWEI HAN received the M.S. degree in software engineering and the Ph.D. degree in computer science and technology from Jilin University, in 2005 and 2018, respectively. He was a Visiting Researcher with the School of Electronics Engineering and Computer Science, Peking University. He is currently an Associate Professor with the College of Cyber Security, Changchun University, and the Digital Identity and Blockchain Joint Laboratory, Peking University. His research interests include cybersecurity technology, the IoT networks, and data mining.

JIANKANG LU was born in Xinxiang, Henan, China, in July 1999. He received the bachelor's degree from Pingdingshan University, in 2021. He is currently pursuing the master's degree with Changchun University.

YING XU received the M.S. degree in law from Jilin University, in 2004. She is currently with the School of Administration, Changchun University. Her research interests include protection of intellectual property rights and legal protection of network privacy.

JIN YOU was born in Lvliang, Shanxi, China, in September 1995. She received the bachelor's degree from Taiyuan University, in 2020. She is currently pursuing the master's degree with Changchun University.

BINGXIN WU was born in Taizhou, Jiangsu, China, in September 2001. He received the bachelor's degree from the Nanjing Normal University Zhongbei College, in 2023. He is currently pursuing the master's degree with Changchun University.