6-2024

Enhancing government service delivery: A case study of ACQAR implementation and lessons learned from ChatGPT integration in a Singapore government agency

Hui Shan LEE
Singapore Management University, huishan.lee.2006@business.smu.edu.sg

SHANKARARAMAN, Venky
Singapore Management University, venks@smu.edu.sg

Eng Lieh OUH
Singapore Management University, elouh@smu.edu.sg

Enhancing Government Service Delivery: A Case Study of ACQAR Implementation and Lessons Learned from ChatGPT Integration in a Singapore Government Agency

Alvina Lee Hui Shan
Singapore Management University
hslee.2020@engd.smu.edu.sg

Venky Shankararamen
Singapore Management University
venks@smu.edu.sg

Ouh Eng Lieh
Singapore Management University
elouh@smu.edu.sg

ABSTRACT

This paper presents the pilot implementation of AI Based Citizen Question-Answer Recommender (ACQAR) as an attempt to enhance citizen service delivery within a Singaporean government agency. Drawing insights from previous studies on the Empath library's use in Service Level Agreement (SLA) prediction and the implementation of the Citizen Question-Answer system (CQAS), we redesigned the pilot system, ACQAR. ACQAR integrates the outputs from Empath X SLA predictor and CQAS as essential inputs to the ChatGPT engine, creating contextually aware responses for customer service officers to use as responses to the citizens.

Empath X SLA predictor anticipates the expected service response time based on citizens' emotional states, while CQAS recommends answers for faster and more efficient officer responses. This paper provides a comprehensive blueprint for governments aiming to enhance citizen service delivery by fusing sentiment analysis, SLA prediction, question-answer models, and ChatGPT. The proposed system design aims to revolutionize government-citizen interactions, delivering empathetic, efficient, and tailored responses without violating SLAs.

Although the full-scale deployment of ACQAR is pending, this paper outlines a foundational step towards the practical development and implementation of an intelligent system by sharing the trial outcomes of ACQAR. By leveraging ChatGPT, this system holds the potential to significantly enhance citizen satisfaction, foster trust in government services, and strengthen overall government-citizen relationships.

Additionally, the paper addresses inherent challenges associated with ChatGPT, including data opacity, potential misinformation, and occasional errors, especially critical in government decision-making. Upholding public administration's core values of transparency and accountability, the paper emphasizes the importance of AI explainability in ChatGPT's adoption within government agencies. Strategies proposed include prompt engineering, data governance, and the adoption of interpretability tools such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to enhance understanding and align ChatGPT's decision-making processes with these principles.

DGO 2024, June 11–14, 2024, Taipei, Taiwan

ACM ISBN 979-8-4007-0988-3/24/06
https://doi.org/10.1145/3657054.3657130

CCS CONCEPTS

• **Applied computing** → Computers in other domains; Computing in government; E-government; • **Applied Computing** → Document Management and Text Processing; • **Generative AI**;

KEYWORDS

Question Answering, Service Innovation, Citizen Services, Information Retrieval, Text Analytics

ACM Reference Format:
Alvina Lee Hui Shan, Venky Shankararamen, and Ouh Eng Lieh. 2024. Enhancing Government Service Delivery: A Case Study of ACQAR Implementation and Lessons Learned from ChatGPT Integration in a Singapore Government Agency. In *25th Annual International Conference on Digital Government Research (DGO 2024), June 11–14, 2024, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3657054.3657130

1 INTRODUCTION

Technological advancements have ushered in an era characterized by unprecedented data collection capabilities, prompting a re-evaluation of traditional boundaries in citizen service research. Instead of solely focusing on internal process improvements for Service Level Agreement (SLA) compliance, there is an opportunity to expand research into uncharted territories by leveraging lexicon libraries, such as Empath, Question-Answer models, and large language models like ChatGPT [1] [2] [3] [4]. SLAs play a pivotal role in shaping interactions between citizens and government entities, influencing satisfaction and trust in governmental operations [5]. Despite their significance, SLAs can fail due to various factors, hindering government agencies' ability to meet citizens' expectations [6].

In this context, we introduce the blueprint of a pilot system, AI-based Citizen Question-Answer Recommender (ACQAR), designed to address SLA deficiencies within a Singaporean government agency's customer service centre. Leveraging insights from previous research, which involved the integration of lexicon libraries like Empath in SLA prediction [7] [8] [9], ACQAR incorporates Empath X SLA predictor and a refined Citizen Question Answer System (CQAS) within a unified interface. From there, the outputs are passed to ChatGPT to draft the proposed response for the customer service officers to use to reply to citizens. This novel approach which incorporates human-in-the-loop (i.e. customer service officers), aims to provide efficient and tailored responses to citizens while considering their sentiments, expected service timelines, and recommended answers from official government documents.

Simultaneously, the adoption of advanced language models like ChatGPT in government operations has become increasingly prevalent, promising more personalized and accurate responses [10] [11] [12]. However, challenges such as data opacity, potential misinformation, and occasional errors must be addressed to align with core public administration values of transparency and accountability [13]. This paper proposes strategies, including prompt engineering and the use of interpretability tools like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), to enhance ChatGPT's explainability within the unique context of government operations [14].

Section 2 of this paper delves into the preceding works related to lexicon libraries, SLA prediction, and Question-Answer models, and ChatGPT, providing a foundation for ACQAR's development. Section 3 shares the analysis from the implementation of ACQAR after the pilot trial was completed. Section 4 discusses ChatGPT and its applications in government, laying the groundwork for the challenges addressed in Section 5. The latter proposes a comprehensive 4-step framework with strategies to enhance ChatGPT's explainability within the public administration context. The paper concludes in Section 6 with a summary of insights and future plans for ACQAR. Overall, this research contributes a holistic approach to citizen service delivery, introducing the innovative ACQAR system design and addressing the challenges associated with the adoption of generative AI capabilities such as ChatGPT in government operations.

2 ACQAR BUILT UPON PREVIOUS WORKS

Within the realm of citizen service delivery, governments are progressively acknowledging the paramount importance of incorporating data and analytics as a pivotal measure to augment the quality of services provided to constituents. This strategic emphasis on elevating service quality is substantiated by a multitude of studies [15] [16] [17]. Notably, Service Level Agreements (SLAs) have risen to prominence as the primary mechanism for defining the expectations of service quality between governments, assuming the role of service providers, and their citizens. Consequently, the significance of SLAs in this context cannot be overstated [18].

2.1 Empath X SLA Predictor

SLAs for citizen service delivery can fail due to various factors. Common reasons include insufficient resources, bureaucratic red tape, inefficient processes, inadequate training, technological challenges, increased demand during crises, and complex service ecosystems. Our exploration into addressing Service Level Agreement (SLA) failures in government customer service centres, pinpointed three key reasons: a surge in citizens' inquiries, inefficient processes in information retrieval, and inadequate training in citizen delivery. The primary objective of our endeavours is to empower Customer Service Officers (CSOs) to respond efficiently and appropriately, thereby mitigating these SLA failure factors.

The initial phase of our previous research focused on understanding citizens' inquiries and the corresponding SLAs. We proposed the application of text analytics to extract features from textual data in citizens' service tickets. Evaluating four different algorithms, including logistic regression, we identified the best-performing SLA predictive model. Importantly, our study introduced an innovative SLA predictive model utilizing Empath and tested it with real-world data from a government customer service centre in Singapore. The experimental results confirmed that incorporating text analytics and lexicon libraries like Empath enhances the analysis of emotional and attitudinal aspects in citizen interactions, improving the predictive accuracy of SLAs. This, in turn, facilitates government officers in comprehending citizen characteristics influencing the SLA prediction process [9].
Our findings also emphasized the efficacy of incorporating Empath Scores and Empath Categories in the SLA predictive model, maintaining a high accuracy score of 0.7513. The inclusion of Empath introduces a human-centric dimension to SLA prediction, recognizing the pivotal role of the human element in the service industry. The recommendation to integrate Empath Categories in the SLA predictive model (Empath X SLA predictor) stems from their utility in providing a more meaningful categorical outcome. This approach assists Customer Service Officers and government officials in understanding and responding appropriately to citizen interactions, directly contributing to overcoming the SLA failure reason associated with inadequate training of CSOs in citizen service delivery.

While the Empath X SLA predictor addresses the sentiments of citizens, our subsequent exploration delved into Question-Answering models. The aim is to assist CSOs in enhancing their turnaround time in information retrieval, thereby enabling more efficient responses to citizens.

2.2 Citizen Question-Answer System – CQAS & refinements

Our prior research delves into the practical implementation and insights derived from a hybrid Citizen Question Answering System (CQAS) in the context of government service delivery. Combining Information Retrieval QA, Natural Language Processing QA, and Knowledge-Based QA techniques, the CQAS aims to improve citizen engagement and satisfaction in digital government services [19]. This initiative is driven by the evolving landscape of government-citizen interactions, exemplified by the Singaporean government's Digital Government Blueprint (2020), emphasizing the importance of leveraging data and technology to promptly address citizen needs [20].

Utilizing real-world data from a government agency's customer service centre, the research incorporates a diverse range of document types, including Frequently Asked Questions (FAQs) in the form of government policies, support documents, and case records. Key insights from the pilot implementation include the proposal to redefine accuracy assessment by introducing a "Relevant" category for case records. This nuanced approach recognizes the importance of responses that, while not entirely accurate, remain relevant and informative. Furthermore, the study emphasizes the involvement of Customer Service Officers (CSOs) in system improvement through feedback loops, such as indicating the categories under which citizens' inquiries may fall—an invaluable step in adapting QA systems to the dynamic nature of citizen queries. Additionally, manual question classification mechanisms, like restructuring the FAQ dataset, are proposed to enhance the system's capabilities. Lastly, the establishment of question typologies is suggested to address various query types, tackling issues such as ambiguity and poor syntax commonly encountered by CSOs.

With the insights derived from the previous work, the revised CQAR used in this paper has been refined via the following methods:

1. FAQ dataset had been rewritten to avoid issues such as ambiguity and poor syntax.

2. The dataset was then further restructured in a standard format of question-answer pairs and consolidated using the agency's new Customer Relationship Management (CRM)System.

3. Categories were tagged to all question-answer pairs and a filtering mechanism was incorporated into the new CQAS to increase accuracy from 33% to 76%.

2.3 Incorporating ChatGPT

The design of ACQAR was completed with Empath X SLA predictor, new CQAS, and ChatGPT, and trained using a real-world dataset from a Singapore government agency's customer service centre. This government agency manages training-related programmes and uses an external customer service centre to address individuals' and companies' inquiries about these programmes [21].

The consideration of incorporating generative AI technologies was because of the potential that it could transform how unstructured data into intelligently crafted replies. With AI, faster decision making can be fostered [30] and ChatGPT can actually act as an efficient digital assistant. Further, within the realm of public administration, the adoption of artificial intelligence (AI) and natural language processing (NLP) technologies has emerged as a significant driver with the potential to elevate the delivery of citizen services within government agencies. Notably, ChatGPT, introduced in late 2022, has garnered attention as a versatile AI-powered conversational agent capable of transforming the dynamics of government-citizen interactions [22] [23]. This section critically analyses the role of ChatGPT in citizen service delivery, highlighting its anticipated benefits and outlining the challenges that demand thorough consideration for its successful implementation.

While ChatGPT's ability to swiftly address frequently asked questions streamlines information dissemination, potentially decreasing wait times for citizens and enhancing the efficiency of government responses, a significant challenge arises in the form of hallucinations [24]. This poses a pressing concern, as responses generated by ChatGPT may impact citizens' trust in government due to the potential inaccuracies or misrepresentations.

In response to this concern, we implement a human-in-the-loop approach through the development of ACQAR, so that Customer Service Officers (CSOs) can utilize ChatGPT's capabilities in a more measured approach while responding to citizens. This approach deviates from using ChatGPT as a direct replacement for the existing chatbot in citizen interactions. By having CSOs act as the human-in-the-loop, they play a crucial role in refining ChatGPT's responses, ensuring that the prominent challenge of hallucination does not adversely impact citizen service delivery. This strategy aims to create a mutually beneficial scenario, enabling the incorporation of ChatGPT's strengths while effectively mitigating its potential drawbacks. The overall design of ACQAR is indicated in Figure 1 below:

ACQAR consists of four main components: 1. A backend Question-Answer model that recommends answers based on input (citizens' inquiries) and outputs from Empath X SLA predictor, providing predicted categories such as "Agitated" and a 3-day SLA prediction; 2. A database capturing input, output, and the duration taken for the Customer Service Officer (CSO) to close or escalate a case (citizens' inquiry); 3. Integration with ChatGPT Turbo 4.0; and 4. A frontend user interface for CSO input and output.

The pilot system enables a CSO to input a citizen's inquiry and select relevant inquiry categories. As discussed in Section II, ACQAR is built upon the new CQAS, whose accuracy is enhanced by incorporating a category filter. Consequently, the pilot system allows the CSO to input categories to activate the CQAS through the interface. After the CSO submits the inquiry, the backend CQAS returns the top 10 recommended Frequently Asked Questions (FAQs) related to the current citizen's inquiry, while the Empath X SLA predictor provides the corresponding Empath category and predicted SLA.

As the human-in-the-loop for ACQAR, the CSO acts as a second layer to filter the recommended FAQ list, choosing the answer most relevant to the citizen's inquiry. After selecting the FAQ, the CSO proceeds to the next stage, where the recommended answer is pre-populated as prompt input into ChatGPT Turbo 4.0. The CSO can further refine the inputs in the input box with the Empath category before using ChatGPT to generate the answer.

A prompt framework is implemented to ensure CSOs optimize ChatGPT's capabilities. In the case of citizen service delivery, a combination of context control and output customization is considered. In this context, the recommended answer contributes to context control, while the Empath category serves as part of output customization, resembling a Persona Pattern. An example of a prompt input for a citizen inquiry is illustrated in Table 1.

3 RESULTS FROM IMPLEMENTATION OF ACQAR

This pilot trial aimed to assess the impact of the human-in-the-loop AI system, ACQAR, on the efficiency and satisfaction of citizen interactions within a government agency call centre. Two Customer Service Officers (CSOs) with 1-2 years of experience took part in a trial in November 2023, and the trial tested two primary hypotheses: 1) ACQAR would lead to a decrease in average resolution time, and 2) citizen satisfaction would improve compared to relying solely on operational CRM systems.

The trial involved two phases: Baseline (Phase 1) and Intervention (Phase 2). During Phase 1, CSOs responded to citizen queries using the existing operational CRM system, without ACQAR aid. In Phase 2, CSOs used ACQAR alongside the CRM system to handle citizen queries. The average resolution time data revealed noteworthy improvements:

1. Agent A's average resolution time decreased from 14 hours 28 minutes to 11 hours 15 minutes.

2. Agent B's average resolution time decreased from 11 hours 29 minutes to 11 hours 13 minutes.

These findings suggest a considerable decrease in average resolution time for both CSOs after the introduction of ACQAR.

Further, post-service survey was implemented for the cases that both agents had resolved in October 2023 and during the trial in

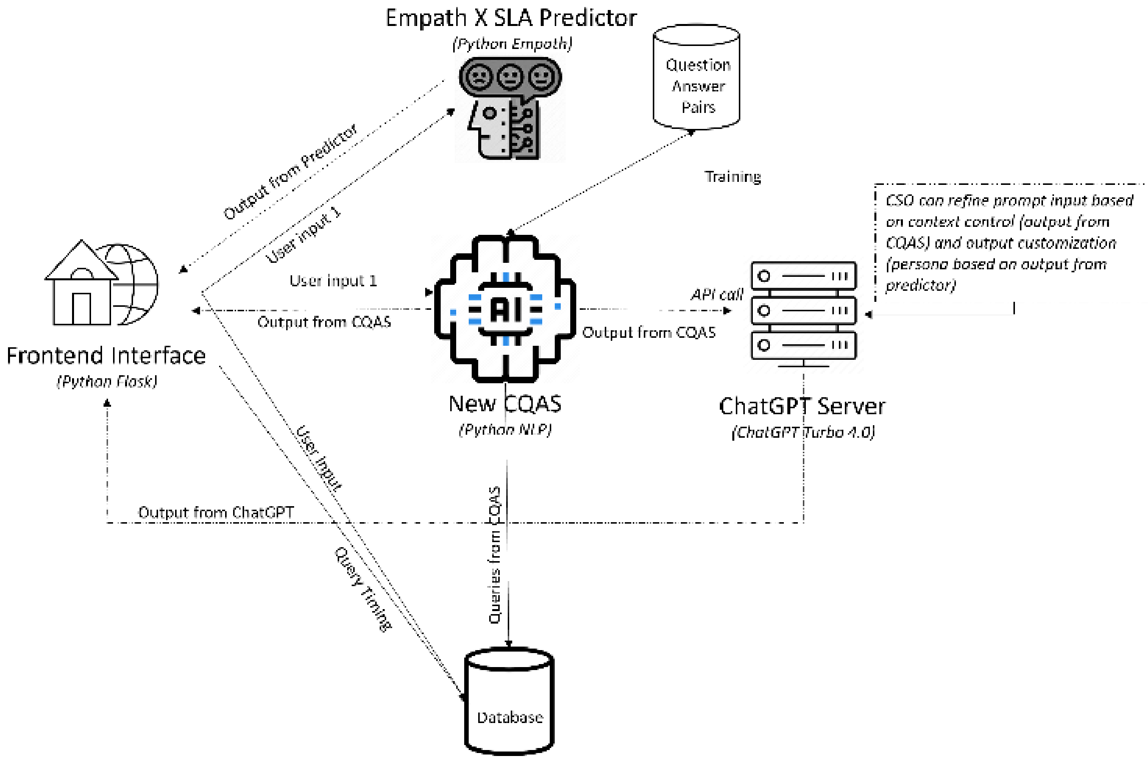

Figure 1: ACQAR System Design

November 2023. The post-service survey by the agency had a total of 3 questions stated below:

1. How well did we understand your concern?
2. How well did we address your issue?
3. How well was your overall experience with our service?

Citizens were to rate “poor”, “fair” and “good”. The outcomes are as indicated in Table 2 below:

Upon analysing the post-service survey data for Agents A and B, several notable trends emerge. Overall, both agents experienced an enhancement in citizen satisfaction across all three questions during the trial period in November 2023 compared to October 2023. The November data revealed minimal ”poor” ratings, indicating general satisfaction with the performance of both agents. Notably, ”fair” ratings decreased, while ”good” ratings exhibited a substantial increase. Agent B received marginally higher ”good” ratings than Agent A across all questions and time periods.

Examining specific categories, both agents demonstrated a consistent decrease in ”fair” ratings and an increase in ”good” ratings for understanding citizen concerns. The most substantial improvement was seen in the ”Addressing Issue” category, with a significant decline in ”fair” ratings and a corresponding rise in ”good” ratings. Similarly, both agents displayed notable progress in garnering ”good” ratings for the overall citizen experience. All in all, this answers the research question that integrating a question-answer recommender, augmented with ChatGPT, can improve citizen satisfaction and the efficiency of customer service officers.

Potential explanations for this improvement include the use of ACQAR, which may have helped superior information retrieval, enhancing the agents’ ability to comprehend and address citizen concerns effectively. The increased efficiency gained through ACQAR could have contributed to shorter resolution times, thereby fostering a more positive overall experience for citizens. However, it is essential to acknowledge the limitations of the survey results, as the observed improvements may not be solely attributable to ACQAR, and external factors could have influenced citizen satisfaction during the trial period.

Finally, a Focus Group Discussion (FGD) with the participating CSOs was conducted after the trial. t covers the following questions that was designed in accordance with the TOE-TAM framework that other researchers when conducting interviews for adoption of technological tools had used [32]. Each group of questions are based on the following [31]:

1. Perceived Usefulness - refers to the degree to which an individual believes that using a particular technology would enhance their job performance or productivity. It assesses the user’s subjective perception of the benefits and advantages associated with adopting the technology.

2. Perceived ease of use - Perceived ease of use refers to the extent to which an individual believes that using a particular technology would be free from effort or difficulty. It assesses the user’s perception of the simplicity, intuitiveness, and user-friendliness of the technology.

Table 1: Example of how ACQAR works with context control and output customized framework for prompt input structure.

Citizen's inquiry	Can I know right now whether I am entitled to extra training funds if I am 40 years old and above? Please get back to me immediately as you all always take so long, and it is irritating!
Empath X SLA Predictor Output	Empath Category: Agitated Predicted SLA: 3 days
New CQAS output	A one-off training credit of $500 will be provided to every Singapore Citizen aged 40 to 60 (inclusive) as of 31 December 2020. This will be on top of the training credit top-up for Singapore Citizens aged 25 and above.
Prompt Input Note: Underlined: context control Italic and underlined: output customization - Persona.	Please craft me an email reply from <u>the standpoint of a customer service officer from XX agency</u> to an <u>*agitated citizen*</u> with this content: A one-off training credit of $500 will be provided to every Singapore Citizen aged 40 to 60 (inclusive) as of 31 December 2020. This will be on top of the training credit top-up for Singapore Citizens aged 25 and above.
ChatGPT Turbo 4.0 Output	Subject: Re: Inquiry Regarding Training Credit for Singapore Citizens Dear [Citizen's Name], I hope this email finds you well. Thank you for reaching out to XX Agency with your inquiry regarding the one-off training credit of $500 for Singapore Citizens aged 40 to 60 as of 31 December 2020. I understand your concerns, and I'm here to provide you with the necessary information. The one-off training credit of $500 for Singapore Citizens aged 40 to 60 (inclusive) as of 31 December 2020 is indeed a valuable initiative. It is designed to support lifelong learning and skills development for citizens in this age group. This credit is provided in addition to the training credit top-up available for Singapore Citizens aged 25 and above. We understand the importance of this initiative and are committed to assisting citizens like you in making the most of this opportunity for personal and professional development. If you have any further questions or need clarification on any aspect of the training credit, please do not hesitate to contact us at [Agency Contact Information]. Our dedicated team is ready to assist you and ensure that you can access and utilize your training credit effectively. We appreciate your interest in this program and look forward to supporting your educational and skills development goals. Thank you for being an engaged and informed Singapore Citizen. Warm regards, [Your Name] Customer Service Officer XX Agency

Table 2: Outcomes of Post-Service Survey

	Oct 2023			**Nov 2023**		
Questions	**1**	**2**	**3**	**1**	**2**	**3**
Agent A	Poor - 3	Poor - 5	Poor - 5	Poor - 3	Poor - 4	Poor - 4
	Fair – 15	Fair – 18	Fair – 19	Fair – 12	Fair – 10	Fair – 12
	Good - 17	Good - 12	Good - 11	Good - 20	Good - 21	Good - 19
Agent B	Poor - 2	Poor - 3	Poor - 4	Poor - 2	Poor - 3	Poor - 3
	Fair – 16	Fair – 15	Fair – 18	Fair – 12	Fair – 11	Fair – 12
	Good - 16	Good - 17	Good - 13	Good - 21	Good – 21	Good - 20

3. Relative advantage - Relative advantage refers to the degree to which a new technology is perceived as superior to existing alternatives or practices.

4. Compatibility - Compatibility refers to the extent to which a new technology is perceived to be consistent with existing organizational practices, values, and norms. It assesses the alignment between the technology and the organizational context, including technical infrastructure, workflows, and cultural factors.

5. Complicatedness - Complicatedness refers to the perceived complexity or difficulty associated with understanding and using a new technology. It evaluates the user's perception of the learning curve, training requirements, and potential challenges in mastering the technology.

There is no scoring matrix and the analysis of the recorded outputs in the form of transcript is based on manual identification of themes in the form of area of concerns as depicted in Table 3.

Table 3: Questions for FGD

TOE-TAM Categories	*Introduction*	1. Can you share your overall experience with the pilot system, including how well you think you've been using it and your general impressions?
Perceived Usefulness	*Technology Integration and Use*	2. Describe how you incorporated the pilot system's auto-recommendation of FAQs and ChatGPT into your daily interactions with citizens. What were the main benefits you observed?
Perceived Usefulness	*Impact on Workflow*	3. How has the pilot system influenced the efficiency of your work and the effectiveness of your responses to citizen inquiries?
Complicatedness	*Challenges and Limitations*	4. Were there any challenges or limitations you encountered when using the pilot system? Can you provide examples of situations where the technology fell short or presented difficulties?
Compatibility	*User Feedback and Improvement*	5. Did you have opportunities to provide feedback on the pilot system and its features during the experiment? Were there any suggestions or recommendations you shared with the team?
Complicatedness	*Training and Adaptation*	6. Reflect on the training and support you received during the experiment. How well-prepared did you feel when starting to use the new system?
Perceived Usefulness	*Long-Term Adoption*	8. Do you see the pilot system, including auto-recommendations and ChatGPT, becoming a permanent part of your workflow for assisting citizens? Why or why not?
Perceived ease of use	*Overall Satisfaction*	9. On a scale from 1 to 10, how satisfied are you with the pilot system and its features for citizen service? Please explain your rating.
Relative advantage	*Future Technology Considerations*	10. What lessons have we learned from this experiment, and how can similar technology be improved for your work in the future?
Compatibility	*Conclusion*	11. Is there anything else you'd like to share about your experience with the pilot system and the impact it had on your role in serving citizens? Any final thoughts or insights you'd like to provide?

The key highlights of the FGD revealed that design and prior training are key considerations when it comes to the use of AI capabilities. The participants are of the view that while the AI capabilities will help their work but may not necessarily boost their efficiency if not design properly. Furthermore, if there is no proper training, even with the presence of generative AI capabilities like ChatGPT, mistakes can still happen or the output from the system would not be ideal.

In terms of the system design of ACQAR, selecting FAQs before input into the AI system (ChatGPT) might be less effective than training a large language model (LLM) and using it directly. This is coupled with the feedback that the prompt template used by the system may need to be revised to provide more exact and helpful outputs.

Quotations from Agent A and Agent B that depicts the concerns are indicated in Table 4 below:

In summary, the data shows that ACQAR has the potential to enhance the efficiency of CSOs in resolving citizen inquiries and enhance citizen satisfaction rate. However, it's crucial to note that the sample size was small, and the study duration was limited. Therefore, these findings should be interpreted with caution. The FGD highlights areas for improvement in the design and implementation of ACQAR, including the potential effectiveness of training a large language model directly and refining the prompt template for better guidance to the AI system. Further research with a larger

Table 4: Extracted Quotations from FGD

Areas of concern	Quotations from Agent A and B during their Focus Group Discussion
Not design properly	Agent A: "I have to cut and paste the answers into CRM after generated. This one slows me down."
	Agent B: "The system should be inside CRM then we can directly have it show inside the email box to reply to citizens. So, need to design this better."
Data inputs to ACQAR	Agent A: "The drafting is good, but there are times that answers recommended not answering inquiries."
	Agent B: "Ya, if there is no such answer, then draft also no use. End up also we had to google."
	Agent A: "why the answer recommender cannot be directly inside ChatGPT? Then save one step?"
Training related.	Agent A: "not everyone know how to ask the ChatGPT to draft properly. So, the template given is good."
	Agent B: "true that. If the template is made available like in this pilot system, then people might use it better. Else I think hor many of us not trained enough to ask the ChatGPT the right questions to get good answers."

sample size and an extended duration is recommended to validate and build upon these initial findings.

All in all, the data indicates that ACQAR has the potential to enhance the efficiency of CSOs in resolving citizen inquiries and enhance citizen satisfaction rate. However, it's crucial to note that the sample size was small, and the study duration was limited. Therefore, these findings should be interpreted with caution. The FGD highlights areas for improvement in the design and implementation of ACQAR, including the potential effectiveness of training a large language model directly and refining the prompt template for better guidance to the AI system. Further research with a larger sample size and an extended duration is recommended to validate and build upon these initial findings.

4 IMPLICATIONS OF USING CHATGPT

In the context of the agency's case study, a series of challenges emerged during the trial of ACQAR, shedding light on significant issues related to the use of ChatGPT. Three key challenges were particularly pronounced. Firstly, the issue of data opacity surfaced, encompassing concerns about how data is stored and potentially accessed. This is followed by during the trial experiment, feedback from the two Citizen Service Officers (CSOs) raised concerns about potential hallucinations, prompting a request for an improved prompt template. Lastly, the incorporation of internet resources in ChatGPT's training led to the issue of misinformation.

In the agency's attempt to utilize OpenAI's ChatGPT directly, the challenge of data opacity became apparent. There was uncertainty about where the data would be stored, and the agency faced difficulties in discerning the origin and processing of the ChatGPT output. To address this, the agency collaborated with the central government's technology agency to develop an internal ChatGPT product, ensuring clarity on data storage and the exact dataset used for training.

The agency also grappled with hallucinations from ChatGPT, particularly concerning in drafting of responses to citizens, whereby recommended FAQ and the Empath category of the inquiry was served as inputs to ChatGPT. The unsupervised learning nature of ChatGPT allowed it to self-generate data, sometimes resulting in information extrapolation or guessing not present in the training data, leading to occasional misjudgments.

Furthermore, the agency seen instances of misinformation during the trial experiment. This issue was worsened when outdated data was ingested by the model, with no mechanism for reversal, resulting in factually inaccurate outputs that could potentially undermine the agency's reputation, if not for the human-in-the-loop nature of ACQAR.

These three challenges underscore the crucial need for AI explainability in ChatGPT's implementation. In the subsequent section, we will propose a 4-Steps framework integrating strategies aimed at enhancing AI explainability, addressing the intricacies posed by data opacity, hallucinations, and misinformation.

5 PROPOSED 4-STEPS FRAMEWORK

The adoption of ChatGPT in government agencies brings immense potential, yet challenges related to transparency and accountability must be navigated for responsible AI integration. We presented the proposed 4-Steps framework (See Figure 2) at the IEEE Big Data conference in 2023 [25], addresses these challenges through strategies aimed at enhancing AI explainability within the unique context of government operations.

Step 1 focuses on Data Governance, recognizing that AI models heavily rely on data. In the government context, where data is often confidential and sensitive, the framework suggests a comprehensive approach [26]:

1. Data Quality: Implement rigorous checks and validation procedures to ensure accurate, up-to-date, and reliable data for training and fine-tuning ChatGPT.

2. Data Provenance: Maintain clear records of data sources, transformations, and preprocessing steps to trace the data legacy used by ChatGPT.

3. Data Documentation: Create comprehensive metadata, data entity relationship diagrams, and dictionaries for datasets.

4. Data Classification: Streamline datasets into different classifications to control data usage within the models based on residency of the data and separated into Official Open, Official Closed, Restricted, Confidential and Secret.

Step 2 emphasizes Prompt Engineering, acknowledging the impact of prompt inputs on ChatGPT's output variability [27].:

1. Standardised Prompts: Implement a standardized structure for common government tasks to ensure consistent and reliable responses.

2. Prompt Design Guidelines: Develop guidelines for designing effective prompts that yield informative and unbiased answers.

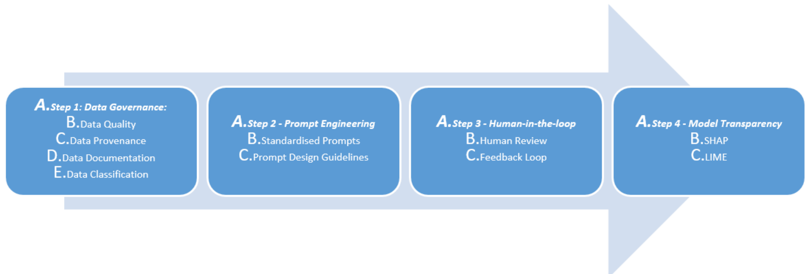

Figure 2: 4-Steps Framework

Step 3 introduces a Human-in-the-loop approach to mitigate potential hallucinations:

1. Human Review: Implement a periodic human-in-the-loop review system where government experts assess ChatGPT's responses for accuracy and ethical considerations.

2. Feedback Loop: Establish a mechanism for government administrators to provide feedback on model performance and address issues promptly.

Step 4 Model Transparency, aims to address potential misinformation by enhancing ChatGPT's explainability:

1. SHAP or LIME: Use interpretability tools like SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) to make the model more transparent and understand the features contributing to output responses [28] [29].

6 CONCLUSION

In conclusion, the findings of this paper shed light on the potential of the AI-based Citizen Question-Answer Recommender (ACQAR) in improving the efficiency of Citizen Service Officers (CSOs) in government agencies. The pilot trial revealed a notable decrease in average resolution time for CSOs after the implementation of ACQAR, suggesting enhanced responsiveness in addressing citizen inquiries. Additionally, the post-service survey data indicated an improvement in citizen satisfaction, particularly in the understanding of concerns and the overall experience.

However, it is crucial to approach these findings with caution due to the small sample size and limited study duration. The insights gleaned from the Focus Group Discussion (FGD) underscored areas for improvement in the design and implementation of ACQAR, including the potential effectiveness of training a large language model directly and refining the prompt template for more accurate outputs.

As with any innovative system, ACQAR has its limitations. The issues of potential hallucinations and misinformation, inherent to ChatGPT's capabilities, were observed during the study. Moreover, the reliance on data from the internet for ChatGPT training and where the data is stored, raises concerns about data opacity and privacy, necessitating careful data governance measures.

The proposed 4-Steps framework, while offering valuable strategies for enhancing AI explainability in government operations, also has limitations. The framework suggests the use of SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) for model transparency. However, the effectiveness of these tools needs empirical validation, and future work should include building SHAP into ACQAR and comparing it against another implementation of LIME into ACQAR.

In future research, there is intention to roll out this to a larger sample size. Refining the prompt templates and the model of ACQAR will also be a priority, with a specific focus on incorporating SHAP or LIME for increased model transparency. This iterative approach aims to enhance the system's accuracy, reliability, and ethical considerations, paving the way for more robust and responsible AI integration in government agencies.

REFERENCES

[1] Chakravarti, B., & Varma, V. (2008, November). An enterprise architecture framework for building service-oriented e-governance portal. In TENCON 2008-2008 IEEE Region 10 Conference (pp. 1-6). IEEE.
[2] Davenport, T. H., 2013. "Analytics 3.0." Harvard Business Review, 91(12), 64–72.
[3] Smith, G., Ofe, H. A., & Sandberg, J. (2016, January). Digital service innovation from open data: exploring the value proposition of an open data marketplace. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 1277-1286). IEEE.
[4] Engel, C.T. and Ebel, P., 2019. Data-driven service innovation: a systematic literature review and development of a research agenda. European Conference on Information Systems.
[5] Nugraha, Y., & Martin, A. (2017, April). Towards the classification of confidentiality capabilities in trustworthy service level agreements. In 2017 IEEE International Conference on Cloud Engineering (IC2E) (pp. 304-310). IEEE.
[6] Wanna, J., Butcher, J., & Freyens, B. (2010). Policy in action: The challenge of service delivery. UNSW Press.
[7] Aggarwal, C.C. and Zhai, C., 2012. A survey of text clustering algorithms. In Mining text data (pp. 77- 128). Springer, Boston, MA.
[8] Fast, E., Chen, B. and Bernstein, M.S., 2016, May. Empath: Understanding topic signals in large-scale text. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 4647-4657).
[9] Lee, A. H. S., Shankararaman, V., & Ouh, E. L. (2022, December). Implementation of Empath X SLA predictive tool for a Government Agency in Singapore. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 2297-2304). IEEE.
[10] Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). Chatgpt: A meta-analysis after 2.5 months. arXiv preprint arXiv:2302.13795.
[11] Kurian, N., Cherian, J. M., Sudharson, N. A., Varghese, K. G., & Wadhwa, S. (2023). AI is now everywhere. British Dental Journal, 234(2), 72-72.
[12] Qu, Y., Bai, B., & Zhang, Z. (2023, April). The New Generation of Artificial Intelligence Technology ChatGPT Causes: Potential Legal Risks and Regulatory Countermeasures. In 2023 8th International Conference on Computer and Communication Systems (ICCCS) (pp. 1205-1211). IEEE.
[13] Fayyad, U. M. (2023). From Stochastic Parrots to Intelligent Assistants—The Secrets of Data and Human Interventions. IEEE Intelligent Systems, 38(3), 63-67.
[14] Carvalko, J. (2023). GPT–A Paradigm Shift for the Twenty-First Century.
[15] Fast, E., Chen, B. and Bernstein, M.S., 2016, May. Empath: Understanding topic signals in large-scale text. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 4647-4657).
[16] Pantelis, K., & Aija, L. (2013, October). Understanding the value of (big) data. In 2013 IEEE International Conference on Big Data (pp. 38-42). IEEE.

[17] Su, K. K., & Sun, W. (2012, June). Assessing citizen service platform system success in government 2.0. In 2012 IEEE Symposium on Robotics and Applications (ISRA) (pp. 396-399). IEEE.

[18] Di Modica, G., Tomarchio, O. and Vita, L., 2009. Dynamic SLAs management in service-oriented environments. Journal of Systems and Software, 82(5), pp.759-771.

[19] Shan, A. L. H., Shankararaman, V., & Lieh, O. E. (2023). Learnings from Implementing a Pilot Hybrid Question Answering System for a Government Agency in Singapore. In HICSS (pp. 1910-1919).

[20] Digital Government Blueprint (DGB) 2020. GovTech, Singapore. Retrieved Nov 22, 2021, from https://www.tech.gov.sg/files/media/corporate-publications/dgb-public-document_30dec20.pdf.

[21] A. H. S. Lee, V. Shankararaman and E. L. Ouh, "Extending the Horizon by Empowering Government Customer Service Officers with ACQAR for Enhanced Citizen Service Delivery," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 1952-1958.

[22] Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. The Lancet Digital Health, 5(3), e105-e106.

[23] Abdullah, M., Madain, A., & Jararweh, Y. (2022, November). ChatGPT: Fundamentals, applications and social impacts. In 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-8). IEEE.

[24] Cardoso, A. G. (2023). Do we need a Chat-GPT-Gov? The importance of technology for effective access to public information. The importance of technology for effective access to public information. (January 7, 2023).

[25] A. H. S. Lee, V. Shankararaman and E. L. Ouh, "Vision Paper: Advancing of AI Explainability for the Use of ChatGPT in Government Agencies – Proposal of A 4-Step Framework," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 5852-5856.

[26] Pradhan, R., Lahiri, A., Galhotra, S., & Salimi, B. (2022, June). Explainable AI: Foundations, Applications, Opportunities for Data Management Research. In Proceedings of the 2022 International Conference on Management of Data (pp. 2452-2457).

[27] Deldjoo, Y. (2023). Fairness of ChatGPT and the Role Of Explainable-Guided Prompts. arXiv preprint arXiv:2307.11761.

[28] Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. arXiv preprint arXiv:2301.13852.

[29] Chai, A., Vezvaei, A., Golab, L., Kargar, M., Srivastava, D., Szlichta, J., & Zihayat, M. (2023, June). EAGER: Explainable Question Answering Using Knowledge Graphs. In Proceedings of the 6th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (pp. 1-5).

[30] Cigdem Akkaya and Helmut Krcmar. 2019. Potential Use of Digital Assistants by Governments for Citizen Services: The Case of Germany. In Proceedings of the 20th Annual International Conference on Digital Government Research (dg.o 2019). Association for Computing Machinery, New York, NY, USA, 81–90.

[31] Rogers, E. M. (2003). Diffusion of Innovations (5th ed.). Free Press.

[32] Abdekhoda, M., Dehnad, A., & Zarei, J. (2019). Determinant factors in applying electronic medical records in healthcare. East Mediterranean Health Journal, 25(1), 24-33.