4-2024

Implementation and evaluation of AI-based Citizen Question-Answer Recommender (ACQAR) to enhance citizen service delivery in Singapore public sector: A case study

Hui Shan LEE
Singapore Management University, hslee.2020@engd.smu.edu.sg

IMPLEMENTATION AND EVALUATION OF AI-BASED CITIZEN QUESTION-ANSWER RECOMMENDER (ACQAR) TO ENHANCE CITIZEN SERVICE DELIVERY IN SINGAPORE PUBLIC SECTOR: A CASE STUDY

ALVINA LEE HUI SHAN

SINGAPORE MANAGEMENT UNIVERSITY

2024

Implementation and Evaluation of AI-based Citizen Question-Answer Recommender (ACQAR) to Enhance Citizen Service Delivery in Singapore Public Sector: A Case Study

by
Alvina Lee Hui Shan

Submitted to School of Computing and Information Systems in partial fulfilment of the requirements for the Doctoral Degree of Doctor of Engineering

Dissertation Committee:

Venky SHANKARARAM (Supervisor / Chair)
Professor of Information Systems (Education)
Singapore Management University

Eng Lieh OUH (Co-Supervisor)
Associate Professor of Computer Science (Education)
Singapore Management University

Hady Wirawan LAUW
Associate Professor of Computer Science
Singapore Management University

Singapore Management University
2024

I hereby declare that this EngD dissertation is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in this dissertation.

This EngD dissertation has also not been submitted for any degree in any university previously.

Alvina Lee

02 Jan 2024

Implementation and Evaluation of AI-based Citizen Question-Answer Recommender (ACQAR) to Enhance Citizen Service Delivery in Singapore Public Sector: A Case Study

Alvina LEE

Abstract

Government agencies prioritize citizen service delivery to foster trust with the public. Technological advancements, particularly in Artificial Intelligence (AI), hold promise for improving service provision and aligning government operations with citizens' needs. Yet the inherent inflexibility of Service Level Agreements (SLAs) often overlooks the nuances of human emotions and the varied nature of citizen inquiries, exacerbated by a lack of tools to guide appropriate responses. This dissertation aims to address the gaps of overlook of human emotions and non-support for appropriate responses, by exploring the following questions: (1) Can a predictive model incorporating both numeric and textual data effectively forecast SLAs? (2) How does emotion analysis impact the predictive model's efficacy? (3) Does integrating a question-answer recommender, augmented with ChatGPT, improve citizen satisfaction and the efficiency of customer service officers?

To investigate these questions, a final pilot system known as AI Based Citizen Question-Answer Recommender (ACQAR) was developed, employing techniques such as Latent Dirichlet Allocation (LDA), Logistic Regression, use of Empath Library, and ChatGPT.

This dissertation further dive into the AI Based Citizen Question-Answer Recommender (ACQAR) system's implementation within a Singaporean government agency to enhance service delivery. ACQAR attempts to generate contextually aware responses for

customer service officers. The study aims to optimize government-citizen interactions in the digital age, where citizens expect efficient, personalized, and empathetic services. The findings of this pilot system 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.

The study's contributions lie in its novel approach to bridging the gap between SLAs and human emotions in citizen inquiries, shedding light on the potential of AI integration in government service delivery to deliver not only prompt responses, but also appropriate replies. It further offers insights into the practical challenges and implications of AI adoption, proposing strategies for smoother integration and risk mitigation within government agencies.

It is to note that the content of this dissertation is organised with the identification of gaps via the literature review in Chapter 2. This chapter examines AI's evolution in service delivery, emphasizing its potential to transform government services. Research gaps such as traditional use of SLA was unable to detect human emotions in citizen inquiries and guide appropriate responses, lack of AI readiness framework and AI Explainability (XAI) for government agencies were identified.

Chapter 3 presents the case study background. Chapter 4 depicts the first version of the pilot system built which was known as Empath X SLA predictor. Findings showed that the inclusion of human-centric indicators to represent human emotions in the prediction of SLAs does not affect accuracy, in fact it introduces texture to such traditional indicator of citizen delivery standards. Chapter 5 shares about the second version of the pilot system built which was known as Citizen Question Answer System (CQAS). While the accuracy of this second pilot system was not optimal, it sheds light on areas of improvement for such a system, leading to the eventual successful build of ACQAR in Chapter 6.

Chapter 6 outlines ACQAR's design and implementation, including integration details and pilot implementation within the Singaporean government agency. Findings revealed that the pilot system does shorten the time taken by a customer service officer to respond to the citizens' inquiries, while improving citizen satisfaction rates. However, the implementation also revealed that a framework to improve explainable AI (XAI) is required.

Considering the challenges in AI adoption highlighted previously, Chapter 7 assesses a government agency's readiness for AI adoption and proposes a framework for smoother integration and risk mitigation for government agencies. This framework was implemented within the case study and with that, suggested countermeasures were shared.

The dissertation concludes with a summary of findings, contributions, limitations, and recommendations for future research and practice. The research done in this dissertation will contribute to understanding AI's role in public administration, offering insights into practical implementation and challenges associated with AI adoption in the public sector.

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

The landscape of citizen service is undergoing a paradigm shift, driven by the relentless march of technological advancements [2]. Amidst all these, Citizen service delivery must still be prompt and appropriate to ensure efficient governance, meet citizens' needs effectively, and uphold public trust [199]. Further, unprecedented data collection capabilities have opened new avenues for research, prompting a re-evaluation of traditional boundaries in this critical domain to overcome the problem of prompt and appropriate citizen service delivery [103]. One traditional boundary that had always been of consideration will be the use of Service Level Agreements (SLAs). What exactly is SLAs? A Service Level Agreement (SLA) is a formal contract or agreement between a service provider and a customer (in the context of the government agencies, this will refer to citizen) that outlines the specific level of service expected, including quality, availability, responsibilities, and metrics for measuring performance. As shared by Comuzzi et al., SLAs are commonly used in various industries, including Information Technology, telecommunications, and customer service, to ensure that services meet agreed-upon standards and to establish accountability between the parties involved [22]. For example, in Singapore, if a citizen sent an email to inquire about a particular government initiative, the SLA would be within 3 working days from the date of receipt of the inquiry [145].

While Service Level Agreements (SLAs) have long served as the cornerstone for measuring service quality, their limitations are becoming increasingly evident [22]. Citizen service no longer focuses on promptness of service delivery, but as well as appropriateness of the service provided by the government agencies. Hence there is a need to explore alternative approaches that leverage the power of Artificial Intelligence (AI) to address these limitations and usher in a new era of citizen-centric service delivery.

1.1. The Urgency of Rethinking Citizen Service

While Service Level Agreements (SLAs) play a crucial role in shaping citizen-government interactions, they often fall short of expectations for various reasons such as predominantly focus on end-to-end duration and frequency of failed service requests without considering the nature of the request or inquiry [1][2][108]. These shortcomings not only lead to dissatisfaction and eroded trust but also hinder government agencies' ability to effectively serve their constituents. Moreover, literature review in Chapter 2 revealed that the inherent inflexibility of SLAs fails to account for the complexities of human emotions and the diverse nature of citizen inquiries. The first research gap of failure to account for human emotions is meant to be addressed by our first pilot system, the Empath X SLA predictor. A key innovation of this system lies in its departure from traditional approaches to SLA computation, which rely solely on structured, numerical data. Instead, it incorporates the dimension of human sentiment, thereby ensuring that citizen delivery is not only timely but also sensitive to emotional nuances. This will be further discussed in Chapter 4.

1.2. The Rise of AI in Citizen Service

As defined by Stone (2018), AI aims to imbue machines with intelligence, mimicking human responses with minimal intervention [3]. This rapid progression, fuelled by the proliferation of data and advancements in computing power, has led to groundbreaking applications across diverse sectors. From AlphaGo's strategic acumen to IBM Watson's cognitive prowess, AI has convincingly rivalled or surpassed human capabilities in areas previously deemed exclusive to human cognition [4][5][6]. Consequently, there has been a widespread adoption of AI technologies, evident in the recommendation engines of Amazon and Netflix, Samsung's inventory management systems, and DBS's queue management solutions [7][8]. These applications not only enhance customer experiences but also yield cost savings. Governments worldwide are recognizing the transformative potential of AI, with the Singapore Government's digital blueprint serving as a notable example [9].

Amidst the burgeoning influence of AI in citizen service, the focus has extended beyond Service Level Agreements (SLAs) to encompass question-answer systems that aim to expedite information retrieval based on citizens' attributes. Chapter 2 highlights the other research gap whereby SLA can be used to nudge promptness of replies by the service provider, it does not guarantee the appropriateness of replies. The development of the pilot system, Citizen Question-Answer System (CQAS) in Chapter 5 is directed towards closing this research gap by targeting the appropriateness of responses provided by Customer Service Officers (CSOs) to citizens. Serving as an AI enabler, CQAS assists in recommending answers tailored to citizens' inquiries for CSOs' reference. This aims to ensure that citizen service delivery is executed proficiently, thereby enhancing the government agency's public image. Further elaboration on this aspect will be provided in Chapter 5.

1.3. ACQAR: A Pilot System for Enhanced Citizen Service

Incorporating insights from prior pilot systems that integrated lexicon libraries like Empath in SLA prediction [10][11][12] or utilize Question-Answer models, the final pilot system, AI-Based Citizen Question-Answer Recommender System (ACQAR) merges the Empath X SLA predictor with a refined Citizen Question Answer System (CQAS) into a cohesive interface. This system is considered AI-based because of the inclusion of advanced AI technology like ChatGPT. It harnesses the capabilities of advanced large language models, specifically ChatGPT, to craft personalized and precise responses tailored to each citizen's query, considering predicted human sentiment categories and expected service timelines. Because the Question-Answer recommender model within the system is built using the citizens' case data from the case study, the term "Citizen Question Answer Recommender" was used.

Notably, ACQAR adopts a human-in-the-loop approach, ensuring that customer service officers (CSOs) retain decision-making authority in response delivery, thereby upholding transparency and accountability while harnessing the potential of AI. The purpose of ACQAR is to support customer service officers to reply promptly and appropriately to citizens when inquiries are sent in. This final pilot system represents the culmination of these efforts and will undergo trials at the selected Singapore government agency's customer service centre. Chapter 6 will delve into the methodology and results of these trials, while Chapter 3 will provide background on the selected case study.

1.4. Challenges and Opportunities of AI in Government

The adoption of AI in government operations presents both exciting possibilities and significant challenges. Possibilities included the improvement of operational efficiency using AI chatbot and using data to deliver personalised services or responses [146][147]. Yet these are not without challenges that generally branches into the domain of accountability and transparency, in the midst of citizens' concerns about their data privacy and security [148][149][150]. To navigate these challenges and leverage AI effectively, agencies require a robust tool to assess their AI readiness before implementation. Critical questions as follows must be asked:

1. How can we assess a government agency's readiness to adopt AI-enabled citizen service technologies?
2. What corrective measures can be implemented if an agency is not yet ready for AI adoption?

These questions hold immense value for governments worldwide as they grapple with the budgetary and return-on-investment challenges associated with AI implementation. Chapter 7 provides answers to the above questions and proposes a framework for AI readiness assessment while outlining corrective measures. Such a framework aims to empower governments to make informed decisions about leveraging AI to deliver citizen-centric services that meet the expectations of their constituents.

1.5. Need for Explainable AI (XAI)

The adoption of AI in government operations necessitates careful consideration of potential challenges such as data opacity, misinformation, and occasional errors. Several researchers such as Hutson, Lipton etc., have highlighted the challenges associated with data opacity and misinformation in AI systems [151][152][153]. Data opacity refers to the lack of transparency regarding the source, quality, and biases present in the data used to train AI models [151]. Misinformation can arise when AI algorithms make decisions based on incomplete or biased data, leading to inaccurate predictions or recommendations [153]. Due to these challenges, the concept of explainable AI (XAI) appears. XAI refers to the capability of artificial intelligence systems to provide explanations for their decisions or outputs, particularly in situations where transparency and understanding are crucial [86].

To mitigate these concerns as per the definition of XAI and align with core public administration values, this dissertation proposes strategies like prompt engineering and interpretability tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to enhance the explainability of AI models within the unique context of government operations [13][14]. SHAP and LIME provide post-hoc explanations for AI predictions, allowing users to understand the factors driving a particular decision or outcome [154][155]. By ensuring transparency and understanding, we can harness the power of AI responsibly and ethically to deliver exceptional citizen service experiences.

To illustrate the flow of the entire dissertation content, Figure 1 below depicts the progression of ideas and discussions across the chapters.

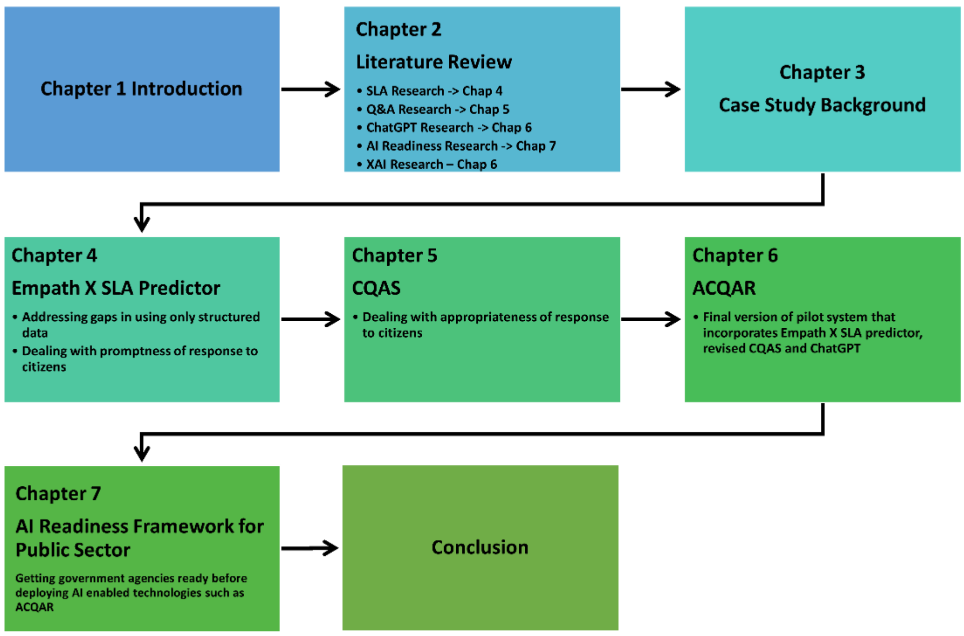

Figure 1 Dissertation Flow

This dissertation delves deeper into these themes via the adaptation of design research guidelines that Hevner et al. (2004) and Bloom et. Al (2004) had advocated in 2004

[198][200]. As per Table 1, we started off with problem relevance in Chapter 1 and with design as an artefact and research rigor, Chapter 4 to Chapter 6 explore the potential of various AI-powered citizen service solutions like Empath X SLA, CQAR and ACQAR while addressing the critical questions of readiness and responsible implementation. As part of design evaluation, we further evaluate the ACQAR by actual implementation in a real-life government agency that is discussed in Chapter 3. Ultimately, we aim to contribute to a future where AI serves as a valuable tool for enhancing citizen engagement, trust, and satisfaction with government services. The research outcomes of these chapters were respectively presented and published at various research conferences under the themes of applied AI and digital government as depicted in Publications Section of this dissertation.

Table 1 Design Science Guidelines

Guidelines	Description [200]	Used in this study
Design as an artifact	Design science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation	The design artefact in this dissertation is the creation of a final pilot system, ACQAR (Chapter 6) and proposed AI readiness framework (Chapter 7) to support the research problem and questions stated.
Problem relevance	Design science research focuses on developing technology-based solutions to important and relevant business problems.	We focused on creating pilot systems to resolve the research problem of the challenges faced by government agencies to deliver citizen services in a prompt and appropriate manner as indicated in Chapter 1.
Design evaluation	One must rigorously demonstrate the utility, quality, and efficacy of a	We evaluated the final pilot system, ACQAR and proposed AI readiness framework by deploying it to trials

	design artifact via a well-executed evaluation method.	in the selected case study depicted in (Chapter 3).
Research contributions	Effective design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.	We applied the well established TOE frameworks, as well as prior predictive algorithms into our own research outputs to form a final product, ACQAR that had been used by the selected case study and the agency is moving forward with such AI-enabled system.
Research rigor	Design science research relies on applying rigorous methods in both constructing and evaluating the design artifact.	We evaluated a total of 4 predictive algorithms and 2 natural language process models respectively in Chapter 4 and 5.
Communication of research	One must present design science research effectively both to technology-oriented and management-oriented audiences.	We published at 6 different applied AI and digital government related conferences to discuss the pilot systems built and frameworks proposed.

Chapter 2

2. Literature Review

In the pursuit of enhancing citizen service delivery, it is imperative to comprehensively review and understand various components shaping the landscape of modern governance. From SLA research to the implementation of advanced AI technologies and finally to AI readiness and explainability, each facet contributes to the intricate tapestry of citizen-government interactions. Therefore, in the following sections, we embark on a journey through literature, exploring past research that underpins each facet. By examining the research that has been done thus far, we aim to glean insights that not only inform how we build the various pilot systems, but also inform strategic decision-making and foster innovation in how government agencies could improve their citizen service delivery.

2.1. Peering into the SLA research

In citizen service delivery, integrating data and analytics into a singular unit to drive service enhancement is a critical area that governments are looking into. This is especially true when better service delivery is key to establishing trust between citizens and government [11][16][17]. SLA monitoring is a sub-set of this field, where research is done to enhance the SLA process, in turn improving the relationship between government and citizens. SLA is the most common mechanism used to establish agreements or expectations on the quality of service between the government who is now like the service provider and the citizens who are customers. Hence its importance cannot be underestimated [18].

A key component in SLA research is the SLA prediction process, which uses available data to forecast the possible end-to-end SLA duration that a particular service will take to complete. SLA prediction started mainly considering only statistical computation or system tracking [19][20]. With these as a basis, SLA design frameworks have been proposed over the years [21]. Concrete variables such as time, frequency, and number of cases resolved were the key factors used for establishing SLA [22]. An example of such an SLA design framework was the one developed by Comuzzi et al. [22]. This highly cited framework states that four roles, namely service customer, software provider, service provider and infrastructure provider and three layers, namely business, software, and infrastructure management, are to be examined. However, despite the fact that the roles and layers are broadly human-centric, the conception of relevant basic data entities covers only design-time and run-time data. These data entities remained the only inputs to the SLA predictive model.

However, with advancements in data analytics, machine learning techniques provide opportunities for further enhancing SLA prediction. This allows the prediction to go beyond the current practice of statistical computation or system tracking, or simply put, forecasting [24]. Slowly more research has surfaced on using analytics to make a prediction, whereby unlabelled data points could be parsed through a predictive model to derive its SLA duration [25][26][27][28][29]. However, the limitation is that such works still focus only on the numerical variables of the dataset, such as end-to-end duration and frequency of failed service requests, which in turn is narrow and limits the insights that an SLA predictive model can provide.

Since SLA is an essential factor in determining service quality that will impact citizens purely relying on the numeric approach to predict SLA has several weaknesses [30][31].

Firstly, based on the time factor and the number of cases resolved, we will not be able to understand why the case ticket fails its SLA [30]. For example, if a citizen sent in an email that uses angry words, one would expect the citizen is not going to be patient to wait for days for his issue to be resolved. Hence, purely using numerical variables, it will not be possible to analyse such human-centric indicators in an SLA prediction model. Secondly, considering that an IT system is built for humans to use, be it the government as the service provider or the citizens, key quality service indicators must go beyond numerical measurements and include textual data as a variable to introduce the human element into the prediction. This is why tools like Empath are developed, so that researchers can leverage upon Empath to elicit the human element from the textual data [11]. Such tools also make incorporating human of human element easier into SLA prediction.

By considering the more human-centred variables such as feelings that include attitude, emotions, moods, and other affectual states by analysing the text contained in the service ticket or case details submitted by the customers, it is possible to gain better insights into the human element of the collected data [31]. Empath library is such a tool that can be used in this study. It can generate and validate new lexical categories on demand from a small set of seed terms. Hence it can be used to help assist in text processing and then in introducing human-centred variables into the SLA prediction model.

In the past, SLA predictive models had not considered such human behaviour aspects, probably due to limitations of available text analytics algorithms and technology limitations in the collection of such textual data. However, with the capability made available by technological advancement, such textual data that could impact the resolution duration can now be included as variables in the prediction process.

Addressing this gap and proposing an SLA predictive model that considers text data is the goal of Chapter 4 – Pilot System 1 – Empath X SLA Predictor.

2.2. Unravelling Question Answer Systems

Question Answering (QA) systems allow users to ask questions and retrieve answers for them automatically, rather than browse through documents or FAQs to find answers. Because of the potential value that it could bring to society, the QA system has evolved into a fast-growing research area that brings together research techniques from Information Retrieval, Natural Language Processing, and Information Extraction [32]. Research work started by Simmon in 1965[33] and Waltz in 1978[34] whereby they both tried to build a question answering system using an English language relational database, to current research works that not only propose novel approaches to the creation of QA system [35], but also the attempt to build QA systems based on different languages [36]. While it has huge potential, the complexity is that these research techniques cut across multiple research domains, making it hard for researchers to apply these techniques when building a QA system [37]. This further explains why the real-world QA systems such as virtual assistants or knowledge management libraries that are embedded within a Customer Relationship Management System (CRM) would not be able to deploy all the techniques. The more pragmatic approach must be to consider which technique would be more suitable based on the context within which the QA system is being deployed and the structure of the documents within the chosen domain [38][39][40].

2.2.1. Structure and Types of QA Systems

A QA system structure comprises three components [41]: question classification, information retrieval and answer extraction. Question classification focuses on ensuring the questions are first processed and classified according to the different topics of relevance [157]. Information retrieval focuses on finding the most relevant documents within the collections of documents that could be mapped to the topic that the question falls under [156]. Finally, answer extraction focuses on assembling the answer to the question asked [166].

To better understand the types of QA systems that are built using the above structure so that we can decide on the QA type for our pilot system, we draw insights from existing QA research based on the four main types of QA systems suggested by Soares and Parreiras (2020)[42], i.e. (i) Information Retrieval QA, (ii) Natural Language Processing QA, (iii) Knowledge based QA and (iv) Hybrid QA.

2.2.2. Information Retrieval QA

This type of QA system uses search engines to retrieve answers, apply filters, rank the answers, and recommend the answer to the user [43]. The strength of this type of QA system is to process enormous amount of information content [41][158]. However, the downside of relying only on information retrieval to build QA systems is that using only cosine similarity between the documents and inquiries may lead to documents being retrieved even when not all keywords in the question are present in the document. This is often referred to as the term mismatch issue [44][159]. Hence the subsequent ranking of the answers might not be accurate [45][46]. For example, when we consider the following question posted to a QA system:

Question: "How long is the AABB initiative"?

The information retrieval could have only focused on the cosine similarity for the keyword "AABB" and searched for all documents containing "AABB" on the Internet.

Hence it might return a ranked answer unrelated to the duration.

2.2.3. Natural Language Processing QA

This type of QA system uses linguistic intuitions and machine learning methods to extract answers. The questions presented to such QA systems are usually analysed using NLP techniques, which are then used to construct a standard database query. The strength of this type of QA system is that it leverages sophisticated algorithms to go down to the level of syntax and semantics of questions [47]. That said, there are still limitations to this approach, whereby the knowledge stored in the structured database can only answer questions asked within the restricted domain [160][161]. On the flip side, such a QA system will be good for handling specific contexts [48]. Using the same example question:

Question: "How **long** is the AABB initiative"?

Repeated training of questions with the term "AABB" in an NLP QA system will allow the QA system to build a database catered solely for the AABB initiative. This approach can increase the accuracy of the answers extracted to reply to the questions related to this topic. Hence an alternative question such as the following will likely return the same correct answer as the previous question:

Question: "How **many months** will the AABB initiative last"?

2.2.4. Knowledge-Based QA

This type of QA system retrieves answers based on a structured data source and standard database queries used to replace word-based searches. It is meant to leverage the use of ontologies, on top of NLP QA approaches [49]. The strength of such QA system is that it would be able to handle complex reasoning and retrieve required information if the metamodel is already in place [50]. However, this also means the natural incompleteness of a knowledge-based QA system limits the question scope that it can answer, hence the high reliance on the person who develops it to update its data source and database of queries [51][162][163]. However, this limitation also guarantees the quality of the answer so long the question can be found in the database. Taking the same question as an example:

Question: “How long is the AABB initiative”?

If the structured data source has various differently phrased questions but with the same meaning, the answers provided to all these questions will be correct.

2.2.5. Hybrid QA

This type of QA combines the techniques from the three types of QA systems namely Information Retrieval, Natural Language Processing and Knowledge Based to enrich the answers generated for the inquiry. Such hybrid QA helps overcome some of the limitations of the previously discussed QA systems [164][165]. An example of such a QA system is IBM Watson [52], whereby deepQA techniques are used and experimented with “Jeopardy”. The system has proven to have high accuracy in answering the questions asked. Another example is using the same question:

Question: "How long is the AABB initiative"?

Within a hybrid QA system, the question will be processed with natural language processing to derive the context that the question is set in, while knowledge-based processing and information retrieval techniques will be further combined to retrieve possible answers from various data sources. The outcome will be a list of ranked recommended answers to the same question, allowing one to consider various answers to one question.

However, a basic requirement for building such Hybrid QA is the availability of multiple sources of data in different formats.

Based on the analysis of the literature review, Hybrid QA was chosen as the preferred approach for building the Citizen Question Answer System (CQAS) described in Chapter 5 due to the following reasons:

- The hybrid QA provides higher accuracy when answering questions to user queries.
- The data requirements for building a hybrid QA are satisfied through the availability of a collection of policy documents in different formats along with case details of past citizen inquiries that were recorded by the customer service officers verbatim.

2.3. Exploring the Landscape of ChatGPT

In the landscape of public administration, the utilization of artificial intelligence (AI) and natural language processing (NLP) technologies has gained prominence as a potential catalyst for enhancing the provision of citizen services within government agencies. Among these, ChatGPT which debuted in late 2022, stands out as a versatile AI-powered conversational agent with the capacity to reshape the dynamics of government-citizen interactions [52][53]. This section examines the role of ChatGPT in citizen service delivery, emphasizing its prospective advantages and delineating the associated challenges that necessitate comprehensive consideration for its effective implementation.

The current research on ChatGPT within the public sector is still lean, yet rapidly evolving. Since OpenAI first launched ChatGPT in Nov 2022, till Sept 2023, in 11 months, a quick search on various research search engines, using key statements such as “ChatGPT and Government”, “ChatGPT in Public administration”, “ChatGPT and citizen service delivery” and “ChatGPT and policies” yielded the following results:

Table 2 Outcomes from Research Search Engines (as of Sept 2023)

Search Terms	**Google Scholar**	**Science Direct**	**Scopus**	**JSTOR**
ChatGPT and Government	10,300	206	10	39
ChatGPT in Public Administration	11,200	158	10	7
ChatGPT and citizen service delivery	1,450	22	1	1
ChatGPT and policies	22,500	418	283	22

As we examine the search results, the research relating to ChatGPT and its use in the public sector can be grouped as follows:

1. Using ChatGPT as a replacement for existing chatbot/information retrieval engine [55][56]
2. Discussions about the use of government data and the implications [57][58][59]
3. Ethical implications and impact on policies [60][61][62][63]

Based on the above and Table 3 below, it is clear that the role of ChatGPT in citizen service delivery has yet to be fully explored beyond being an enhanced chatbot or search engine. The potential use of ChatGPT to enhance citizen service delivery could be extrapolated from current research done in the private sector for customer service [64][65]. These two papers concluded from twenty-one research papers that micro, small, and medium enterprises could benefit from the implementation of ChatGPT for better customer service.

Table 3 Use Cases derived from current research papers

Use Cases	Countries	Research
Using of ChatGPT as a replacement for existing chatbot/information retrieval engine	USA	[66]
Government using ChatGPT to tackle transport challenges	USA & Canada	[67]
Optimise EFL tool/ Materials	UK	[68]
Using ChatGPT with government open data	NA	[57],[58]
Paving way for Medical AI	China	[69]
Military	Iran etc	[59], [70]
ChatGPT in Education Strategies	UAE	[71]
AI Act- Internet Policy Review	Netherlands	[72]
ChatGPT for e-Tourism	Italy	[73]
Balancing ChatGPT and Data Protection	Germany	[63]
ChatGPT for Finance research	Ireland	[74]
Marketing with ChatGPT	USA	[75]

While ChatGPT's rapid response capacity in addressing frequently asked questions streamlines information dissemination, potentially reducing wait times for citizens and improving the efficiency of government responses, there is the chief challenge of it hallucinating [76]. This is of exigent concern as such responses from ChatGPT could impact citizens' trust in governments.

To address this concern, in Chapter 7, we introduce human-in-the-loop by developing the pilot system, known as ACQAR, for CSOs to leverage upon ChatGPT's capability to reply to citizens and not go with the approach of using ChatGPT as a replacement for the existing chatbot for citizens to interact with directly. CSOs as the human-in-the-loop will help to re-craft the response from ChatGPT and ensure the chief challenge of hallucination does not affect citizen delivery. This will establish a win-win situation in that we can incorporate the plus points of ChatGPT while mitigating the downside.

2.4 Existing Frameworks in Assessment of AI Readiness

In information systems (IS) research, readiness has always been a topic of discussion [77]. Chatterjee, Ghosh, and Nguyen (2019) define digital readiness as 'the degree to which an organisation is ready to transform the current organisation digitally'[77]. As AI is classified as a digital technology, we can learn from the digital readiness frameworks developed for other earlier digital technologies.

The Technology Acceptance Model (TAM) was first introduced as a framework by Davis (1985) to leverage the user acceptance process to assess the successful implementation of IS [79]. However, the limitation of TAM is that it lacks consideration of the business context and environment dimensions, hence is more suitable for individual use rather than organisational use [80]. The Technology-Organisations-Environment (TOE) framework was later proposed by Tornatzky et al. (1990) to cover the limitations of TAM, as it analyses a firm from three different dimensions: technology, organisation, and environment[81]. The technological dimension includes all the relevant technologies available within and outside the firm. The organisational dimension describes business characteristics and resources that might influence the adoption process, such as firm size, managerial structure, decision-making, and communication. The environmental dimension refers to the industry's structure, including the firm's competitors, suppliers, customers, and regulatory environment. The versatile nature of the TOE framework was created due to the use of three different dimensions, allowing the flexibility for researchers to apply the framework in their respective domains while studying the possible components that can fall under each dimension that is applicable to the business context or industry.

Although the TOE framework is widely used in research due to its versatile nature, the framework has its limitations as it could be affected by the element of bias due to companies' self-assessments. AlSheibani (2018) proposed an enhanced version of the framework by including the components of relative advantage and compatibility of the AI technology with the organisation [82]. Nortje & Grobbelaar (2020) had seven other components, such as employee culture, strategy, security, etc., on top of the TOE framework to assess AI readiness [83]. This is like Jöhnk et al. (2021), who proposed an AI readiness framework that incorporated the components of culture and strategy [84]. Subsequently, there were even researchers that proposed a TAM-TOE framework to overcome such limitations [167]. But the hybrid framework was not widely adopted by companies or researchers that wanted to look into AI-enabled systems and hence most returned to the use of TOE framework alone and further modified upon it [168].

In the area of AI-enabled CRM systems, Chatterjee et al. (2019) proposed a conceptual readiness framework upon TOE framework that considered integration, auditing, analysis, and regularisation that were not considered in the frameworks proposed by Nortje & Grobbelaar (2020) and Johnk et al. (2021) [77][83][84]. These newly added components are crucial in the current context, just as mentioned in section 2.2, the explainable element, XAI and trusted tech component are essential in AI adoption by organisations, especially government agencies, and will require such components to ensure transparency and accountability.

Further, the component of integration proposed by Chatterjee et al. (2019) is key to an

AI readiness framework used by government agencies when citizens are looking for a one-stop portal to complete all their transactions with the government [77].

Currently, there are existing AI readiness frameworks in place that leveraged upon TOE framework. The most notable framework will be AI Readiness Assessment Framework (ARAF) developed by European Commission Joint Research Centre (2021). It incorporates TOE dimensions to evaluate factors such as technological complexity, organizational readiness, and regulatory environment, providing actionable insights to guide AI adoption strategies [169]. The rest of the AI readiness frameworks are generally discussed by researchers in their respective context but not as well established as ARAF [170][171][172][173][174].

In summary, components such as integration, explainable elements, etc., were also not proposed by Stirling et al. (2017), which indicates that the proposed AI readiness index might require modification to meet the current climate that the public sector is in [85]. Hence, prior to the final implementation of ACQAR, the need to understand the readiness of the organization in the case study in terms of using AI enabled capabilities was necessary. The process of discovering the readiness of the case study gives birth to a proposed AI readiness framework that will be discussed in Chapter 7.

2.5. Understanding Explainable AI (XAI)

The increasing utilization of Artificial Intelligence (AI) across various domains has brought to light the crucial need for explainability and interpretability in these systems. This has led to the emergence of Explainable AI (XAI) as a vital area of research, focusing on methods and techniques that make AI models understandable and transparent [175]. This literature review explores the key aspects of XAI, highlighting its importance, methods, applications, and ongoing challenges.

What exactly is Explainable AI? As defined by Scott et al. (1977), XAI refers to the capability of artificial intelligence systems to provide explanations for their decisions or outputs, particularly in situations where transparency and understanding are crucial [86]. From the early days of AI research, there has been a persistent discourse among scientists advocating for intelligent systems to elucidate their reasoning processes [87]. This emphasis on explanation becomes especially pertinent in contexts involving decision-making.

For instance, consider a scenario where a citizen visits a government agency's website and is presented with recommended content. In such cases, it is imperative for the agency to understand the underlying factors and parameters that influenced the selection of this content for the citizen. By having insight into how the recommended information was derived and the criteria involved in prioritizing certain content, agencies can ensure transparency and accountability in their interactions with citizens. This aligns with the fundamental principle of XAI, which aims to enhance the interpretability and trustworthiness of AI systems by shedding light on their decision-making mechanisms.

Subsequently, as we transitioned from rule-based and feature-based algorithms such as decision trees, linear/logistic regression etc., to contemporary deep learning methods (a subset of machine learning that focuses on learning representations of data through the use of artificial neural networks with multiple layers of abstraction[180]) like neural networks which are complex computational models inspired by the structure and function of the human brain [176][177], Explainable AI (XAI) regained prominence as a central research focus. However, unlike earlier approaches, many of today's deep learning models lack inherent explanatory mechanisms that even the developers of the models cannot explain [178]. This presents a significant challenge, as neither the models themselves nor external components can adequately explain their outputs [179]. Consequently, when deployed, these models often offer limited insights into their decision-making processes, with researchers frequently attributing outcomes to the characteristics of the input data. Neural network is one example of this [88]. Consequently, the emergence of AI "black boxes" became increasingly prevalent, with inference processes remaining opaque to observers and lacking interpretability for humans [89]. Within a black box, there are always opportunities for the element of bias to manifest within yet remain undetected [90]. This challenge was exacerbated with the advent of Large Language Models, including technologies such as ChatGPT and Bard AI, further underscoring the need for transparent and interpretable AI systems [91][92][93].

While Explainable AI (XAI) holds significance, numerous research studies have underscored a trade-off between model explainability and prediction capability [94]. However, this does not necessarily imply a direct impact on human decision-making abilities. Studies have indicated that even with highly explainable models, individuals lacking domain expertise may not experience improved understanding of outcomes, recognition of uncertainty, or calibration of trust in the model [95]. Thus, it becomes evident that the influence of XAI is largely confined to those familiar with the domain and model intricacies. Despite this, researchers continue to prioritize the development of an array of XAI methods. This endeavour aims to assure stakeholders implementing models for widespread use by providing means to explain outputs comprehensively, thereby addressing potential queries or concerns from consumers in the future.

XAI methods primarily can be approached from two perspectives: Interpretability and Model Specific/Agnostic Techniques.

2.5.1. Interpretability

- Local Interpretability: Local interpretability offers a focused examination of individual predictions, offering insights into the rationale behind a particular output. Widely employed techniques in this domain include LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME operates by approximating the model's behaviour in the vicinity of a specific prediction, constructing simpler, more interpretable models to elucidate its reasoning. Conversely, SHAP assigns credit for the prediction across various features, elucidating their distinct contributions to the outcome. These methods contribute to a deeper understanding of model outputs at a granular level, facilitating informed decision-making and fostering trust in AI systems [96].
- Global Interpretability: Global interpretability provides a comprehensive perspective by analysing the overarching behaviour of the model. Techniques such as feature importance analysis shed light on the relative influence of different features on the model's predictions, offering insights into the factors

driving its overall performance [97]. Decision rules, on the other hand, present a structured representation of the decision-making process akin to a flowchart, thereby simplifying intricate models into easily understandable steps for human interpretation. By elucidating the broader patterns and decision-making mechanisms inherent in the model, these methods facilitate a deeper comprehension of its functioning and enhance transparency, thus fostering trust and confidence in AI applications.

- Counterfactual Explanation: Counterfactual explanations delve into alternative scenarios to discern how modifications in input data could impact the model's output, thereby facilitating an understanding of its reasoning and the detection of potential biases. Techniques such as "What-if" explanations prompt the model to predict outcomes for hypothetical situations, enabling insights into how changes in input variables influence predictions [98]. Meanwhile, counterfactual examples generate alternative inputs that would yield divergent outputs, offering a tangible exploration of the model's decision-making process. By probing into counterfactual scenarios, these methods not only enhance interpretability but also aid in uncovering underlying model dynamics and ensuring fairness and robustness in AI systems.

2.5.2. Model-Specific vs Model-Agnostic Techniques

- Model-Specific: Model-specific techniques capitalize on understanding the unique architecture and internal mechanisms of a particular model to offer nuanced insights. For example, within decision trees, it is possible to directly delineate the sequence of decisions leading to a specific prediction [99]. Similarly, in linear models, the coefficients assigned to each feature directly signify their impact on the resulting prediction. While these approaches furnish detailed explanations tailored to the intricacies of specific model types, they are constrained by their applicability solely to those models and may not seamlessly translate to others [100]. Thus, while effective within their scope, these techniques necessitate consideration of their limitations and compatibility with diverse modelling frameworks.

- Model-Agnostic Techniques: Model-agnostic techniques encompass approaches that operate across various models, irrespective of their internal structures, rendering them broadly applicable across diverse contexts [101]. Despite their versatility, these methods may provide explanations that are less granular and tailored to the specific intricacies of individual models. Prominent among these techniques are LIME (Local Interpretable Model-Agnostic Explanations), which constructs simplified, interpretable models surrounding a particular prediction to elucidate its rationale, and SHAP (SHapley Additive exPlanations), which allocates credit for predictions across different features, delineating their respective contributions. Additionally, feature importance analysis ranks feature according to their influence on model predictions, thereby furnishing a global comprehension of the model's behaviour. While model-agnostic techniques offer broad utility and accessibility, their explanations may lack the depth and specificity inherent in model-specific approaches, necessitating a balanced consideration of their trade-offs in interpretability and applicability across various modelling scenarios.

2.6. Conclusion

The literature review spanning from Service Level Agreement (SLA) research to Explainable AI serves as the foundational basis for the development of several frameworks, including the AI Readiness Framework, the 4-Steps Framework for utilizing ChatGPT in government agencies. SLA and QA research provide the groundwork for the creation of Empath X SLA predictor and CQAS (ChatGPT QA System), which are subsequently refined and integrated into the ACQAR framework. Additionally, AI readiness research contributes to a comprehensive examination of this dissertation's scope, encompassing aspects ranging from people and processes to the eventual implementation of pilot systems. Lastly, research on Explainable AI offers valuable insights into ensuring the transparency and comprehensibility of ACQAR, particularly with the inclusion of ChatGPT technology, for the government agencies utilizing it.

Chapter 3

3. Background of the Selected Case Study

All the pilot systems and proposed framework were built based on real-world data and with feedback from a selected government agency in Singapore. This government agency manages training-related programmes and uses an external customer service centre to address individuals' and companies' enquiries about these programmes. The agency also rolled out grants and training allowances to support initiatives so that citizens can upskill, reskill to either perform better in their current job or make a career switch.

3.1. Case Logging and Resolution Process

In the following Figure 2 for anonymity, we use Org-A to represent the Government Agency and ES-B for the External Citizen Service Centre.

Cases are filed when an inquiry from the citizen comes in via a channel such as walk-in, email, letter, telephone, or from web (See Figure 2). The Level 1 customer service officer (L1 CSO) attends to the inquiry and creates it as a case in the customer relationship management system that belongs to Org-A. Based on the nature of the case, the L1 CSO will either resolve it and close the case or escalate it to Level 2 Subject Matter Experts (L2 SME) who will resolve and close the case. The SLA is triggered the moment the case is created in the customer relationship management system when the citizen submits the inquiry.

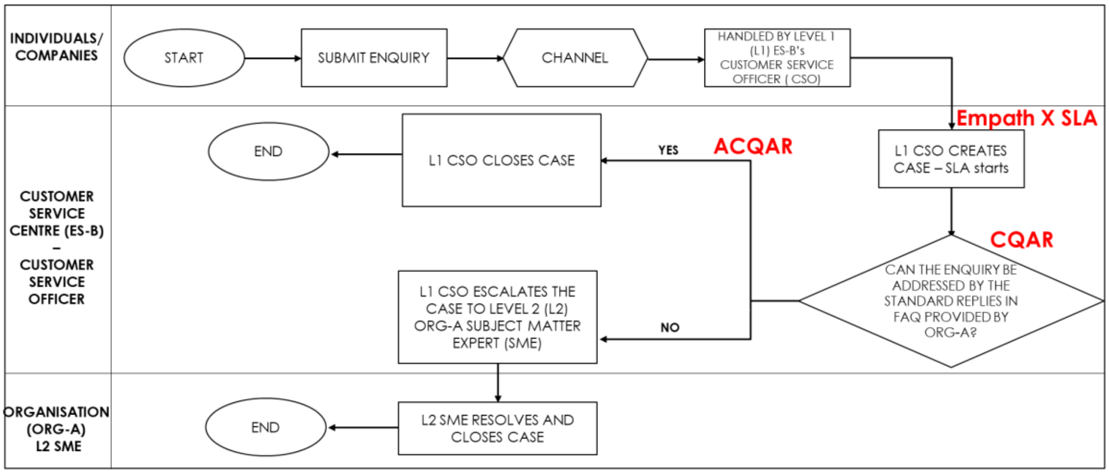

Figure 2 Case Logging and Resolution Process

The customer service centre belongs to a third-party service provider (ES-B), an outsourced vendor of the government agency (ORG-A). Therefore, there is an existing service level agreement (SLA) between the customer service centre and the government agency to specify the terms of service provisioning. The SLA between the customer service centre and the government agency includes the service level objectives (SLO) that state the maximum duration each case category can take to resolve. For example, case records categorised as normal should be resolved within 3 working days, while case records classified as complex should be resolved within 21 working days. The L1 CSO will be the one who categorises the case records for this tiered SLA treatment when the cases are being filed. A case is categorised as normal when a simple inquiry is received that can be resolved via existing information contained in frequently asked questions (FAQs) provided to the CSOs. In contrast, a case will be categorised as complex if it consists of more than two enquiries or more than two L2 SMEs are involved in providing the resolution. Violations of SLAs could result in Org-A imposing penalties on ES-B.

The current situation for Org-A and ES-B is that with the surge in inquiries, the CSOs and sometimes even L2 SMEs are either unable to resolve the inquiries within SLA or unable to resolve the inquiries as per what the citizens might expect. The primary reasons are as follows:

1. Difficulties in prioritization of the cases, other than nature of the case type. For example, a case could be of a normal case type, but because the citizen is impatient and hence within 1-2 working days, decided to escalate his/her inquiry to the senior management of Org A.
2. Challenges in locating information about the inquiry, resulting in more time and resources required.
3. Time and effort required to draft appropriate and helpful responses to citizens and enterprise representatives.

3.2. Dataset used

The dataset used in this research is the case records collected by the Customer Service Centre (ES-B) for the organisation Org-A. It consists of 832,324 records and a total number of 16 variables, as indicated in Table 4.

Table 4 Data Dictionary

Variables	**Description**
Case ID	Unique ID for each case
Customer Type	Indicates whether the case pertains to an individual or organisation
Description	Brief general description of the case
Status	Indicates the status of the case, such as whether it is Open, Closed, Pending etc.
Incoming Channel	Indicates the channel through which the case was reported, such as Customer Walk In, Email, Letter, Telephone Call, Web Channel
Categorisation	Indicates the categories and subcategories of the case, such as IT Technical issue or Programme enquiry etc.

Case Type	Indicates whether the case involves an Appeal, Complaint, Compliment, Enquiry, Feedback or Request etc.
Case Priority	Indicates the importance of the case such as normal, complex etc.
Escalated / Non-Escalated	Indicates whether the case was escalated or not to L2 SME
Created on	The date the case was reported on
Last updated on/at	The date and time the case was last updated
Actual Resolution Date	Indicates the date of the case's resolution (If the case has been resolved or not)
Problem Description	A more detailed description of the case
Clarification/Solution	Contains updates & instructions if the case has not yet been resolved and contains the solution & resolution to the case if it has been resolved
Update from Customer	Updates regarding the case in the form of messages from the customer

A mock-up of the sample case is depicted in Table 5.

Table 5 Mocked-up Data

Variables	**Sample Input**
Case ID	80001234
Customer Type	Individual
Description	Unable to login to the portal
Status	Closed
Incoming Channel	Web Channel
Categorization	IT Technical Issue
Case Type	Complaint
Case Priority	Normal
Escalated / Non-Escalated	Escalated
Created on	09/08/2020
Last updated on/at	15/08/2020
Actual Resolution Date	15/08/2020
Problem Description	Since August, after I received an email to say I can login to the portal. I had tried logging in to the portal. However, despite resetting my password, I am still unable to login. I had also tried to login using different browsers. This is ridiculous that I received such an email, and yet I cannot login. Please advise asap!

Clarification/Solution	A troubleshooting session was arranged with the customer to walk through her login process. She was instructed to clear cache, and she could now login.
Update from Customer	Just want to say thank you for the assistance, as I can login now.

Based on the current SLA adhered by the customer service centre, we computed two other variables from the existing variables, namely "SLA exceeded" and "Resolution Duration", as described in Table 6. These variables will serve as the "ground truth" or actual labels for evaluating how "accurate" the models are and to help with selecting the most suitable model.

Table 6 Derived Data

Derived Variables	**Description**
SLA Exceeded	Binary values, with 1 representing "exceeding SLA" and 0 representing "kept within SLA"
Resolution Duration	Indicates the number of days taken to provide a resolution to the customer. This is computed by taking the difference between the variables: "Created on" and "Actual Resolution Date"

Chapter 4

4. Introduction

Technological advances have enabled the collection of enormous amounts of data during service operations. Given the ubiquitous role of data and analytics in this time and age, citizen service research can be examined beyond mere internal process improvements and allow additional insights via text analytics to inform provisions of new service requirements, thus creating a competitive advantage in enhancing citizen relationships with the government [102][103][104][105]. With the service analysis informed by data analytics, service operations such as Service Level Agreement (SLA) monitoring can be further examined. SLA plays a significant role in the relationship between citizens and the government bodies [1]. It stipulates the quality levels required for the meaningful interaction between two parties. For example, if SLA is adhered, citizens would be satisfied with the service rendered by the government bodies, resulting in the former to be willing to continue to engage the services and enhance trust in the government. However, if SLA was breached, frustration could kick in, resulting in lower trust in whatever the government is promoting or advocating. Therefore, SLA monitoring is a critical process that governments around the world would pay attention to.

A typical data-driven approach in SLA monitoring is the application of machine learning techniques on collected data to determine if SLAs are achieved or breached [106][22]. This is usually done via SLA prediction, which is part of the SLA monitoring process. SLA prediction is of interest to government bodies and citizens because SLA indirectly defines that breach of trust that adversely impacts both the government bodies and the citizens. Therefore, a government would like to know what the citizens' expected SLA would be, so that the officers can attempt to meet expectations and hence not potentially affect the impression of the government, that a citizen could have [107].

Most past efforts to predict SLAs have predominantly focused on the end-to-end duration and frequency of failed service requests; there has been very little research on the analysis of unstructured data such as textual details of service tickets or case records in the SLA prediction process [108]. As a result, rich information is lost during the SLA monitoring process. Part of my research aims to address this gap in SLA monitoring and prediction by including the text contained in the service ticket when predicting SLAs. This is done via the use of text analytics. The outcome of addressing the gap will help government officers to understand and analyse deeper citizen characteristics that could impact the SLA prediction process.

We first incorporated unsupervised learning algorithms in SLA prediction, on top of the use of text analytics [10]. Secondly, we included the use of lexicon libraries with human-validated categories such as Empath. Lexicon libraries such as Empath, have created room for us to analyse human language rich in subtle signs, and help pick up on citizens' feelings that include attitude, emotions, moods, and other affectual states, thus enriching SLA service analysis [11].

In our research, we used text analytics to derive features from the service tickets' textual data. Then we evaluated four different algorithms used in our building of the SLA predictive model that included the features derived using text analytics. From there, we selected the best performing SLA predictive model. After that, we proposed an SLA predictive model using Empath and evaluated it against real-world business process data from the case study government customer service centre based in Singapore.

This chapter is organised as follows: Section 4.1 introduces the research questions, and Section 4.2 depicts the methodology used. The dataset used in the experiment and the outcomes will be discussed in Section 4.3. Section 4.4 concludes and discusses the possible future direction of the work reported in this chapter.

Our work contributes to the body of knowledge in SLA prediction and citizen service delivery, by proposing a predictive model that incorporates text analytics and Empath library to derive human attributes drawn from unstructured data contained in the service tickets or case reports. To the best of the authors' knowledge, there has not been other similar work that uses text analytics and Empath library in SLA predictions. Using the proposed model, government bodies can consider including text analytics in their citizen relationship management systems to enhance their SLA prediction.

4.1. Research Questions

Considering that the current SLA treatment at Org-A is not the best approach for the customer service centre, a new SLA predictive model was considered as an option to support service analysis. Further, as discussed previously, indicators of service delivery quality should go beyond numerical variables to paint a more accurate picture of the customer when making SLA predictions. Hence given the current technological advancement and the existence of rich textual data, text analytics can be scoped into the predictive model, coupled with the use of the Empath library.

To achieve the above, the following research questions are formulated:

4.1.1. Can an effective model be built using both numeric and textual data to help predict the SLA of the case?

4.1.2. Does the use of the emotion analysis impact the effectiveness of the SLA predictive model adversely?

Via the research questions, we aim to achieve the following outcomes:

1. Identify a suitable topic modelling algorithm to derive the features from the textual data.
2. Select a predictive model that can use textual attributes in conjunction with other numeric parameters of the case record, such as resolution duration etc., to help predict the SLA.
3. Incorporate the Empath library so that more human-centric attributes can be included in the SLA predictive model.

4.2. Methodology

This section describes our methodology as shown in Figure 3, to help answer the above research questions The methodology comprises three processes namely Text Analytics

Process, Predictive Modelling Process, and Empath Scoring Process. These three processes are sequential. Text Analytics Process will need to be implemented first so that features from textual data can be derived and included into the Predictive Modelling Process. After the best model is selected from the Predictive Modelling Process, Empath Scoring Process can be implemented so that there can be a comparison between the first predictive model and the predictive model that had incorporated Empath library. The following is a description of each process.

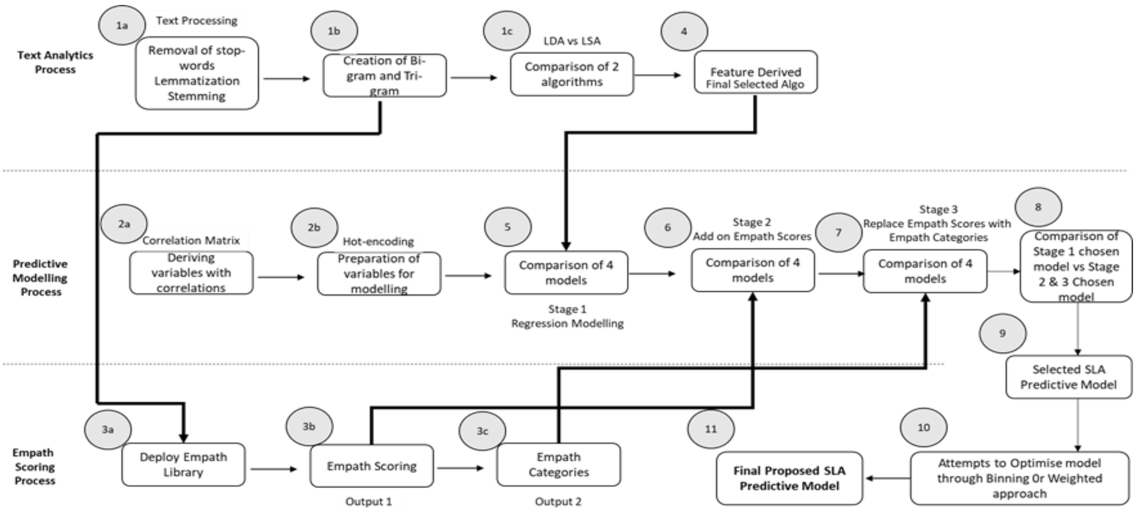

Figure 3 Methodology to derive and build Empath X SLA Predictor

4.2.1. Text Analytics Process

The inclusion of unstructured data (i.e., the variables: Problem Description, Update from Customer) will require additional steps to help prepare the data to enable the different algorithms to derive the features from the processed text. For example, text processing steps are necessary to remove irrelevant words such as "is", "to", "the", etc., (Step 1a) and convert some texts into bi-gram and trigrams (Step 1b).

Two different algorithms (Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSI/LSA)) are applied to the processed text (Step 1c). Subsequently, a comparison is made to decide the final algorithm. The selected algorithm helps to extract the relevant features from the textual description of the case record (Step 4).

4.2.2. Predictive Modelling Process

Within the dataset, sixteen variables were available. After excluding those variables that will be used in text analytics, and other irrelevant variables such as "date" besides "Case ID", a total of six variables remained (see Figure 4). We conduct statistical comparisons between variables before considering them to be included in the predictive model [29][30].

	Case ID	Customer Type	Status	Incoming Channel	Case Type	Escalated/Non-Escalated	Case Priority
Case ID	1.000000	0.045704	0.091245	0.088894	0.010299	-0.021152	-0.035409
Customer Type	0.045704	1.000000	0.001505	0.096304	0.096245	-0.122861	0.062344
Status	0.091245	0.001505	1.000000	0.018409	-0.00107	-0.033089	-0.011921
Incoming Channel	0.088894	0.096304	0.018409	1.000000	0.231490	-0.098793	0.433990
Case Type	0.010299	0.096245	-0.001074	0.231490	1.000000	-0.094290	0.245825
Escalated/Non-Escalated	-0.021152	-0.122861	-0.033089	-0.098793	-0.094290	1.000000	0.162017
Case Priority	-0.035409	0.062344	-0.011921	0.433990	0.245825	0.162017	1.000000

Figure 4 Correlation Matrix

A correlation matrix (see Figure 4) is formed between these six variables and the unique identifier to establish a data table representing the correlations between each variable and the identifier (Step 2a).

Only those that have a positive score when computing correlation are selected. Hence a total of four variables which are "Customer Type", "Case Priority", "Incoming Channel", and "Case Type", are chosen to be included in the SLA predictive model, which is then combined with the features output by the text analytics process and subsequently psychometric outputs from the Empath library.

Regression algorithms are used to build the predictive model for SLA. To facilitate this, the chosen variables undergo hot encoding, where the variables are encoded as binary vectors (Step 2b). A total of four different algorithms namely Logistic Regression, Linear Regression, Multinomial Naive Bayes, and Random Forest are used and defined as per Table 7 below. These models are compared based on their accuracy level which is done using scikit-learn library in Python. The process involves splitting the dataset into training and testing sets, training the model on the training set, making predictions on the testing set, and then evaluating the accuracy of the predictions for each of the algorithms.

Table 7 Summary of Algorithms chosen

Algorithms chosen	**Definition**	**Rationale to the choice**
Logistic Regression	A statistical method used for binary classification tasks. It models the probability that a given input belongs to a particular class using a logistic function [182].	Logistic Regression is a widely used method for binary classification tasks, making it suitable for predicting binary outcomes such as whether a particular case will meet the SLA or not. It produces coefficients that can be interpreted to understand the impact of each feature on the predicted outcome, providing insights into the factors influencing SLA compliance [181].
Linear Regression	A statistical method used for regression tasks. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data points [183].	Linear Regression is a straightforward method for regression tasks, which could be useful if the research involves predicting continuous variables related to SLA performance. It provides coefficients that indicate the strength and direction of the relationship between input features and the predicted outcome, offering

		insights into the factors affecting SLA performance [184].
Multinomial Naive Bayes	A probabilistic classification algorithm based on Bayes' theorem. It is commonly used for text classification tasks, particularly when the features are categorical [185].	Multinomial Naive Bayes is commonly used for text classification tasks, making it suitable if the research involves analysing textual data related to SLA inquiries or documents. While Naive Bayes models are less interpretable compared to linear models, they can still provide insights into the probability distribution of different classes given the input features [186].
Random Forest	An ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification tasks or the mean prediction for regression tasks [187].	Random Forest is an ensemble learning method known for its robustness and flexibility, making it suitable for a wide range of classification tasks, including complex ones with nonlinear relationships. While Random Forest models are less interpretable compared to linear models, techniques such as feature importance can provide insights into the relative importance of different features in predicting SLA compliance [188].

The algorithms are used to build the base model that includes the chosen variables and the features from the text analytics process. In Stage 1, the model with the highest accuracy is selected as the proposed SLA predictive model. The evaluation of the selected model is also done by comparing its accuracy to the accuracy of the majority class in the primary classifier derived from the actual labels, i.e., "Resolution Duration" and "SLA exceeded" (Step 5).

4.2.3. Empath Scoring Process

To include the human element, the Empath library is deployed on the unstructured data (i.e., in the variables: Problem Description, Update from Customer) to derive both the Empath numerical scores and Empath categories (Step 3a – 3c). For example, a case record with the words such as "funding", "allowance", "money" will be sorted into the category of "business". This sheds light that the customer who has initiated this service request is looking into seeking financial assistance.

The four regression models are rerun with these outputs, and the accuracy scores are compared to derive the selected model (Step 6 - 9).

4.3. Results and Discussion

Text pre-processing (Step 1a and 1b) and feature selection are essential steps when dealing with unstructured data such as text. They can help handle the reduction of data dimensionality. Data dimensionality is a constant challenge for models that attempt to include textual data in their data frame. This is because the inclusion of the original textual data generally will increase data dimensionality, in turn increasing the data frame, leading to it being too massive for processing [109].

To manage the data dimensionality issue brought about by textual data, LDA and LSI/LSA algorithms are being used to ensure that features are derived, yet meaningful predictive modelling can still occur later (Step 1c). A coherence score is used to assess the quality of the learned topics by the two algorithms. The algorithm with the higher coherence score can be considered the better approach to topic modelling [110].

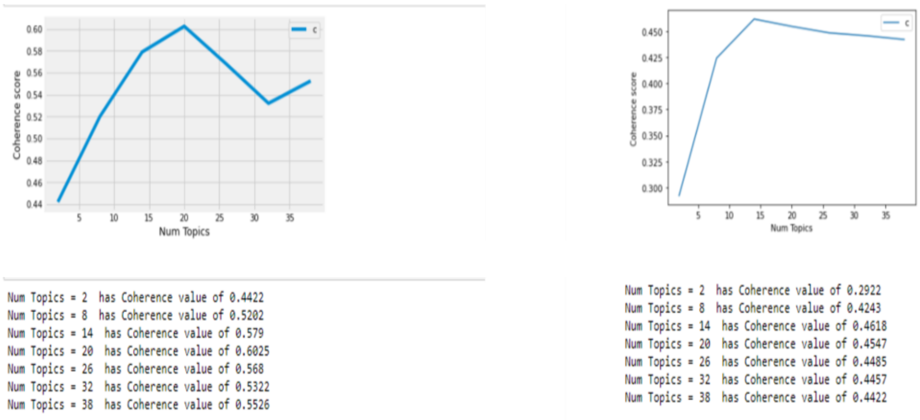

Figure 5 Coherence Outcomes LDA (left) vs LSI/LSA (right)

Using the two variables in the real-world dataset (i.e., Problem Description, Update from Customer), the textual data had been processed and prepared via NLTK Library and Spacy Package in Python (Steps 1a & 1b). After which, the processed data is further processed by two topic modelling algorithms: LDA and LSI/LSA (Step 1c). The LDA model has a higher coherence value than the LSI/LSA model as shown in Figure 5. The higher the coherence score, the more similar the words are within a topic, indicating that the text analytics algorithm works well hence, indicating that the features derived from this algorithm will be more accurate. Hence, based on this dataset, LDA is the better algorithm to be used for text analytics on this data. This answers the first research question, whereby for a dataset that contains textual data on case description LDA is the proposed topic modelling algorithm to derive the features.

As depicted in Figure 3 (Stage 1: Regression Modelling (Step5), four different regression models namely Logistic Regression, Linear Regression, Multinomial Naive Bayes and Random Forest are used to build the predictive model to predict the end-to-end case duration. The input variables include "Customer Type", "Incoming Channel", "Status", "Case Type", and the feature identified from the text analytics process namely

"Total Sentences". This feature represents the weighted list of words derived for each case record. This output can be used to represent the original textual data in the SLA predictive model, while not allowing the issue of data dimensionality to affect the modelling. Hence the use of "Total Sentences", a feature derived from the LDA, a text analytics process, will allow the linguistic aspect of each case record to be included in the prediction of SLA. This will help address the gap in current service analysis research as highlighted in Chapter 2.1.

Since that the predicted SLA duration is a continuous output, hence, for this pilot system, regression models are considered instead of classification models. The business requirement of ES-B and Org-A is to know the predicted duration and not whether the case will pass or fail SLA, hence regression model will be a better fit in this study (Tang and Tang, 2014). Using the above variables as the inputs, the outcomes of the models are as follows: 1. Logistic Regression – Accuracy 75.13%, 2. Linear Regression – Accuracy 74.86%, 3. Multinomial NB – Accuracy 70.24% and 4. Random Forest Classifier – Accuracy 71.68%.

The accuracy score for Logistic Regression is the highest. This means that in comparison against the gold standard (i.e. Resolution Duration stated in Table 5), the SLA predictive model built using logistic regression has an accuracy of 75.13%. Hence it is the best regression model to be used for the SLA predictive model among the 4 models, after the predicted output is compared against the actual real resolution duration. To further validate this, a receiver operating characteristic curve (ROC Curve) is plotted to show the performance of a regression model at all regression thresholds. For the logistic regression model used in this Chapter, the X and Y-axis of the ROC curve (the blue line) are based on two parameters: True Positive Rate and False Positive Rate (See Figure 5). The best-case scenario is depicted by the ROC curve being at a perfect 90-degree angle, whereby our model (shown as the blue line) is very close to being at a 90-degree angle (depicted by the red line), as indicated in Figure 6.

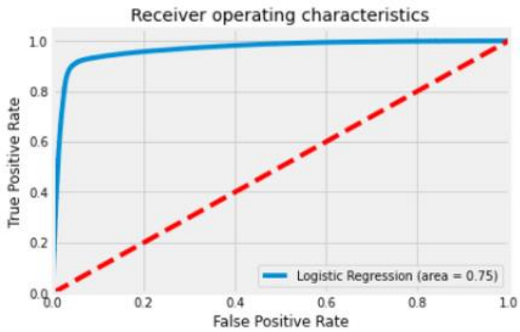

Figure 6 ROC Curve of Stage 1 Logistic Regression.

Therefore, it can be concluded that logistic regression modelling is the better model to build the SLA predictive model for the given data set. This answers our first research question that even if we include textual data into the SLA prediction model, a good accuracy is obtained.

Pre-validated human elements that are made available by tools such as Empath, provide the opportunity for researchers or service providers to link daily words to a broad array of real-world behaviours [111]. For example, researchers had run Empath on truthful and deceptive reviews to identify words that would shed light if the human behind the review is being truthful or deceptive. In the world of today, where citizens can leave reviews on services that they obtained via government bodies, this will help governments to identify the reviews that they should really pay attention to. This will increase operational effectiveness as one need not sift through the massive number of reviews, but just pick up the truthful ones and work from there on enhancing their citizen services. This is like how customers leave product reviews. The human element is introduced to enrich the SLA predictive model by using the Empath library to further enhance the insights gained from the two text-heavy variables (i.e., Problem Description, Update from Customer). By including Empath in the SLA prediction model, we can determine the SLA holistically by taking into consideration the customer's emotions and personal behaviour [112].

Empath library can analyse text across 194 built-in, pre-human validated categories. These categories are highly correlated (r=0.906) with the similar categories in LIWC [11]. As indicated in Figure 3 (Empath Scoring Process), the Empath library uses the processed text as input to derive two outputs, namely Empath Scores (Step 3b) and Empath Categories (Step 3c). Empath Scores are the granular scoring for each topic. At the same time, Empath Categories help to summarise the two variables and place them into the pre-validated categories that are defined in the Empath library using human input. For example, a case record with words such as "irritating", “useless”, “shut up”, "threatening", is placed under the Empath Category of "Aggression". This is indicative that this case is probably associated with an angry customer. Therefore, one would expect case records in this category to have a shorter SLA duration.

The four regression models are rerun with either Empath Scores or Empath Categories, coupled with the other chosen variables (Stage 2 – Step 6, Stage 3 – Step 7). From the outcomes, using only Empath Categories, an accuracy of 0.7512 was attained. For only Empath Scores, an accuracy of 0.7511 was achieved. Thus, inclusion of psychometric elements had not adversely impacted the accuracy of the SLA predictive model, which previously had an accuracy score of 0.7513. See Figure 7 and Table 8 for full comparison.

	model_name	accuracy
0	Logistic regression	0.751172
1	Linear regression	0.747909
2	Multinomial NB	0.7027
3	Random Forest Classifier	0.717132

Stage 2 (Empath Scores only)

	model_name	accuracy
0	Logistic regression	0.75122
1	Linear regression	0.747448
2	Multinomial NB	0.702011
3	Random Forest Classifier	0.714541

Stage 3 (Empath Categories only)

Figure 7 Re-run scores with Empath Scores and Categories

Table 8 Comparison of Accuracy Scores

Algorithms	**With no Empath**	**With Empath Scores**	**With Empath Categories**
Logistic Regression	0.751341	0.751172	0.75122
Linear Regression	0.748573	0.747909	0.747448
Multinomial NB	0.702359	0.7027	0.702011
Random Forest	0.71676	0.717132	0.714541

Given the above outcomes, the second research question is answered, whereby the Empath library in the SLA predictive model does not adversely impact the accuracy. This is likely because, unlike topic modelling which derives only the list of weighted words under each key topic, Empath derives the human-related features, and hence it does not contradict the features from topic modelling. It introduces another dimension that helps the SLA prediction to be more human-centric. As stated in Chapters 2.1 and 3.1, the human element can potentially help inform SLA since it is a critical factor in a service industry.

Additionally, the recommendation would be to incorporate Empath Categories only in the SLA predictive model since it is more accurate than Empath Scores. Furthermore, when the predictive model is used in citizen service platform development, a categorical outcome will be more useful for the customer service officers or government officers to understand rather than providing numerical values. For example, showing an Empath value of 0.8 means nothing to a CSO. However, an Empath category of "Aggression" will alert the CSO that this customer will require the case to be closed as soon as possible before they turn aggressive.

4.3.1. Optimisation of the Chosen Empath X SLA Predictor

To ensure that the model that had been built is the best feasible solution to the business problem that ES-B and Org-A are facing, we attempt to optimise the accuracy of the SLA predictive model from Stage 1 by using 2 other approaches (see Figure 3 – Step 10), i.e., binning, and weighted approaches. The binning approach, also known as discretization, involves dividing continuous numerical features into discrete bins or categories. This can help simplify the model and capture non-linear relationships between the features and the target variable [189]. In this instance, what we do here is to bin the Y, the actual label (i.e., Resolution Duration) into a range such as 4-5 days is labelled as 1, 6-10 days labelled as 2 etc.

Although with this approach, accuracy had improved tremendously to 0.90. However, when the ROC curve (blue line) in Figure 8 is compared to Figure 6, it could be observed that the SLA predictive model that incorporated binning approach is further away from the best-case curve of 90 degrees. Hence this shows that the model is receiving more false-positive outcomes than the original SLA predictive model, even though the latter has an accuracy of 0.75.

Figure 8 Outcome of Binning Approach used on Y

In the real-world dataset, the percentage of case records fulfilling SLA is approximately 70% compared to only 30% not fulfilling SLA. The second approach helps to tackle this imbalance in the real-world dataset. Hence, we rerun the model using a weighted approach (which involves assigning different weights to individual data points or features based on their importance or relevance to the prediction task. This can help improve the model's performance by giving more emphasis to informative or critical data points [190]), whereby in this instance, the majority class "0" (i.e., "fulfilling SLA") was given the weightage of 0.3 to balance up with the minority class "1" (i.e., "not fulfilling SLA") The outcome of an accuracy of 0.748 is derived. Comparing this accuracy score to the accuracy score of the original model (0.75), the original model is still the better option (Step 11). This addresses the final research question and proves that the binning or weighted approach does not optimise the selected SLA predictive model.

4.3.2. Summary of Results

At Stage 1 (see Figure 4), we computed the accuracy scores for different regression models based on the variables with positive correlation and features derived from the topic modelling process. The outcome is that logistic regression has the highest accuracy. A further comparison of the accuracy of this SLA predictive model is made against the accuracy of the primary classifier, which is the frequency of the majority class based on the actual label, i.e., "Resolution Duration". This will help to assess if the accuracy of SLA predictive model is good enough for deployment.

Accuracy of Basic Classifier = (Total number of Case Records that falls within majority class)/ (Total number of Case Records)

It was observed that the first derived SLA predictive model that was built using logistic regression (accuracy of 75.13%, see Figure 8) had an improvement of 7%, as the accuracy of the basic classifier is only 68%.

For the model that incorporated Empath Categories, the accuracy was 75.12%. However, the accuracy of the primary classifier for the majority class is 45%. Therefore, the final SLA predictive model proposed in this chapter has an improvement of 30% accuracy.

4.4. Conclusion

Chapter 4 adds to the body of research work done in SLA prediction by proposing an SLA predictive model that incorporates lexical features from textual data contained in the case records. The experimental evaluation confirmed that an SLA predictive model built using logistic regression provides the best accuracy. Coupled with the use of Empath, the SLA prediction process can be made more human-centric, which is at the heart of service level agreement arrangements [113].

The key contributions from this research are as follows:

1. The study outcomes demonstrate that SLA prediction can go beyond the current practice of statistical computation or system tracking to include textual data by using text analytical approaches such as LDA. This allows the consideration of the nature of inquiry to be considered during prediction.
2. Empath allows the SLA prediction model to consider the human context and emotions when making predictions without impacting the accuracy of the model [114]. In the long run, governments can pick up cases with certain predicted SLA duration in relation with the Empath scoring and categories, to understand the citizens who are sending in the inquiries better. In turn, governments can potentially refine SLA requirements in accordance with citizen archetypes derived from the predicted outcomes.

The work presented can be extended in several ways. Firstly, citizen segmentation can be done in conjunction with the proposed SLA predictive model. This will allow government bodies to understand their citizens better and decide if there is a need to have segmented SLA for different citizen clusters. Secondly, researchers can consider the inclusion of such predictive models in SLA renegotiation frameworks.

Chapter 5

5. Introduction

The Digital Government Blueprint (2020) that the Singapore Government has advocated since 2018 strongly emphasized the need to enhance citizen satisfaction rate by leveraging data and technology to respond to citizens' needs promptly and efficiently [9]. COVID-19 further propelled various government agencies in Singapore to accelerate the use of data and technology aligned with the message of "Digital to the Core and Serve with Heart" [115][116]. The roll-out of various COVID-19 initiatives by the Singapore Government resulted in more incoming inquiries from citizens to respective agencies through telephone, in-person, and email. Even with frequently asked questions (FAQs) published on government websites, the number of inquiries still surged. This surge overwhelmed the Customer Service Officers (CSOs), and they were unable to meet the demands of the citizens. Creating a government chatbot such as AskJamie that is developed and used widely by Singapore Government is one of the solutions to increase citizen satisfaction rate [115][116]. However, such chatbot solutions were not always very successful due to incorrect and inappropriate answers given to citizen inquiries [117][118]. Therefore, a more viable solution is through the introduction of a self-service element for the CSOs, where instead of having to approach subject matter experts to retrieve answers, a software system can help the CSOs respond to citizen inquiries [119][120]. Thus, the respective government agencies began to consider the use of question answering (QA) systems to assist the customer service officers (CSOs) in handling the citizens' surge in inquiries.

With the greater push for citizen engagement and satisfaction, there is a pressing need for governments worldwide to leverage QA systems as part of their digital transformation [121][122][123]. QA systems involve the analysis of a question phrased in natural human speech and then locating a recommended answer to that question within a database of documents [124]. With such systems in place, there is potential operational effectiveness that can be achieved within a customer or citizen service setting. CSOs using the QA system would likely reply more promptly and appropriately to the customers or citizens. This improvement in response speed, in turn, will increase customer or citizen satisfaction rates.

Although the QA system is a widely researched area, the system's effectiveness depends a lot on what data is being input into the system. In the current context of the government agency, data is available in several forms namely frequently asked questions (FAQ) articles, government policies, support documents and citizen inquiries in the form of case records. The challenge is that these are not crafted in the same structure, yet nevertheless, all are needed for developing the knowledge base of a QA system. For example, citizens tend to send in inquiries in the form of long writing, while government policies or support documents tend to be lengthy to ensure details are not missed out. In such situations, it is prudent to use an incremental trial and test approach when developing the QA system in a government context so that citizen service can be delivered promptly and appropriately.

The overarching goal of this research is to elicit the potential issues and possible solutions when implementing a hybrid question answering system (CQAS) that combines techniques of Information Retrieval QA, Natural Language Processing QA and Knowledge Based QA. The data is based on a collection of documents provided to the customer service officers in a government agency. This collection includes website information, FAQs, and policy documents in word, pdf or PowerPoint format. The outcomes of this research will serve as a pre-empt to the government agencies before they embark on QA systems.

5.1. Methodology

The CQAS is built using the Hybrid QA technique. Figure 9 depicts the different techniques used. A brief explanation of how the techniques is applied follows and an example of the complete process flow using sample data is presented in Figure 10.

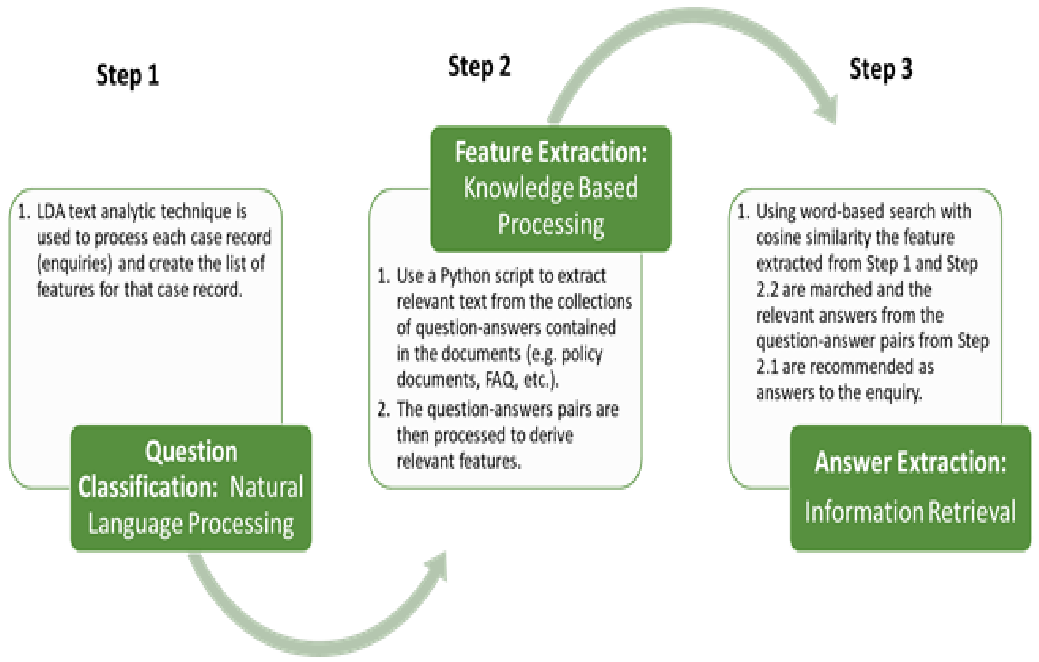

Figure 9 Overview of the Hybrid QA Technique used in CQAS.

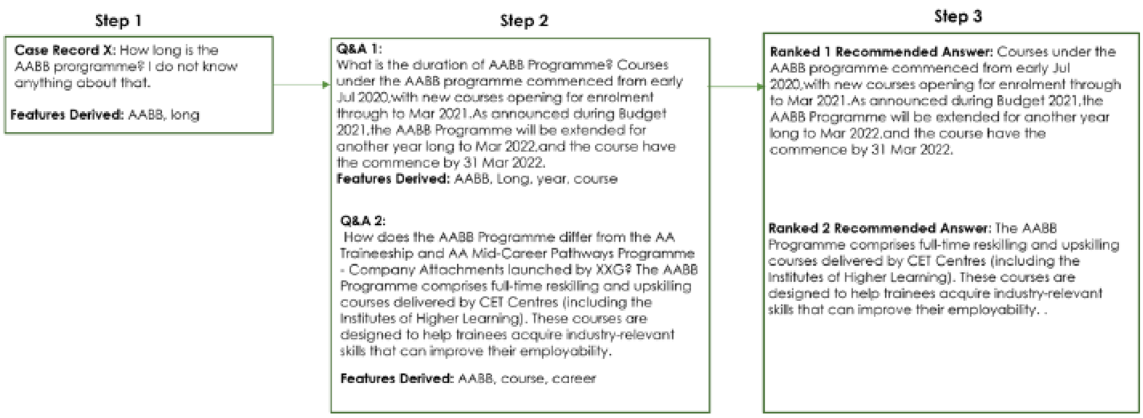

Figure 10 Example showing how the Hybrid QA Technique is applied.

5.1.1. Question Classification (Natural Language Processing Technique): Step 1

CSOs record the citizen inquiries as case records verbatim in text format. This unstructured data, comprising the "Description and Problem Description", as shown in Tables 3 and 4 is processed using Natural Language Processing algorithms to prepare the data and derive the features from the processed text. For example, text pre-processing is necessary to remove irrelevant words such as "is", "to", "the", etc., and convert some texts into bi-gram and trigrams. The reason for using n-grams and tokenization is because some words might not have meaning on their own, however, when combined with other words, the meaning is revealed. For example, the word “stood” is just stand, but the 2-gram phrase of “stood up” has a totally different meaning. It is after preprocessing, then Latent Dirichlet Allocation (LDA) is applied to the processed text to extract the relevant features from the textual description of the case records.

5.1.2. Feature Extraction (Knowledge Based Technique): Step 2

Given that the collections of documents are in several formats including word documents, PDF and PowerPoint, there is a need to restructure these documents into a structured data source. Hence Python script is used for extraction of the relevant textual information from the documents and then structure them as Question-Answer pairs. The question-answers pairs are then processed to derive relevant features that can be used during matching at the answer extraction step.

5.1.3. Answer Extraction (Information Retrieval Technique): Step 3

The features derived from the question classification step are matched to the features derived from the feature extraction step using a word-based and cosine similarity approach. After a match is detected, the corresponding ranked recommended answers are then extracted from the structured data source and shown to the CSO. The CSO uses this ranked list of answers to reply to the customer and subsequently updates the case record.

5.2. Results and Discussion

Before trying out CQAS on the full data set, an initial test was conducted with a total of 50 case records and the complete set of 168 questions with recommended answers was used to test the CQAS. Out of 50, 40 case records had the correct recommended answers matched, giving a very high accuracy of 80%. Subsequently, a total of 261,811 case records submitted by the citizens in 2020 were used for the Question Classification Stage. The same complete set of 168 questions with recommended answers created from the collections of documents was processed at the Feature Extraction and Answer Extraction stages.

5.2.1. Accuracy of CQAS

Despite the original accuracy of 80% for the test model, the final pilot model built using the full set of case records and structured data source only had an accuracy of 33%. To determine the cause for the low model accuracy, a preliminary manual screening was done to see if the recommended answer matched the case records after being processed by CQAS. The outcomes were split into three categories: Accurate, Relevant, and Irrelevant. Examples of the categories are shown in Table 8. This also helps to ascertain if the use of knowledge-based QA approaches had really helped to handle the complexity in the case details database as stated in Section 2.1.3 as key strength of this type of QA system.

The establishment of the "Relevant" category is important to future assessment of CQAS as it will reveal how many of the case records had been matched as "close but not accurate" answers. This category could indicate that while feature matching had done its job, there might not have been ample questions and recommended answer pairs to provide an accurate answer. The preliminary manual evaluation indicated that there were many data points that did fall under the "relevant" category. This also implied that while the use of knowledge-based QA approach had helped to handle this to a certain level, there is need to further review this aspect. Hence, future work will be directed at using the CSOs to assess the recommended answers that fall under the "Relevant" category from the CQAS and then apply human judgement and prepare an appropriate reply to the citizen's inquiry. Such an approach will work in the context of the research undertaken in this chapter since we are not using a chatbot to directly interact with the citizen, but rather the CQAS supports the human, namely CSO, to help answer the inquiry posted by the citizen.

Table 9 Examples of outcomes to ascertain accuracy of CQAS

Problem Description in Case Records	Recommended Answer from CQAS
Accurate – Recommended Answer will be able to answer the inquiry by the citizen	
Customer enquired on the credit claim that was rejected for the course that he had submitted.	If you have already made an initial self-payment for a course in full, you will not be able to use the Credit for this course. If you had been rejected, there is no appeal. You may use your Credit for subsequent courses.
Relevant – Recommended Answer belongs to the same collection of documents that the inquiry is based on	
Customer enquired on the full certificate in his digital X cert Passport. Customer mentioned that the Training Provider has submitted the assessment to the agency on 8 January 2020.	Information provided in the company approved training organization/non-X cert training organization's profile affects the funding your company/ training organization is eligible to receive
Irrelevant – Recommended Answer does not address the inquiry, neither belongs to the same collection of documents	
Customer enquired on how her client can make XXX payment online. Customer mentioned that the client has foreign employees.	Disbursement records from ABC Support Grant payment details are interfaced on-demand.

5.2.2. Accuracy at Question Classification stage is not sufficient to ensure Accuracy of CQAS

It was observed that the coherence score for the question classification stage was 71%, which means that the number of topics extracted to represent each case record has a 71%

accuracy. The topics are likely to reflect the content of the inquiry from the citizens. Despite this, the accuracy at the answer extraction stage was still not ideal since only 33% matched correctly to the case records. A possible reason for this is that there could be more inquiries beyond the collections of documents, i.e., 168 questions with recommended answers. These collections of documents generally only get updated within the first three months of implementing the new initiatives rolled out by the agency. Subsequent inquiries from citizens that are not found within these documents are generally not included as there is currently no systematic approach to capture this source of answers. Further as discussed in Chapter 2.2.2, the virtue of information retrieval QA system, it ideally should have an enormous database to do term matching, hence with the inquiries going beyond the question-answer pairs, accuracy is bound to go down.

To better handle this issue, we propose having a loopback component in the QA system, which allows CSOs to key in their final resolution into the structured data source at the Feature Extraction stage so that matches can be enhanced with new recommended answers. Further mechanisms to tag the case records to the recommended answers can be implemented at the Answer Extraction stage to help train the QA system.

5.2.3. Presence of feedback-related statements increased complexity of CQAS.

Upon reviewing the case records that had fallen within irrelevant or relevant categories (refer to Table 9), we observed that the case records were generally not a clear-cut textual value of only questions but also included feedback. An example is depicted below:

Case Record: "We have received your claim, however you have submitted the claim for the wrong course reference number. As the course has started, I will be grateful if the error can be amended, and my claim can be successful".

As we can see in the example, feedback-related statements such as "I will be grateful", etc., are found. When included in the question classification stage, such a statement is likely to lead to the questions being wrongly identified [125]. A proposed approach to handling this issue will be through introducing a task for CSOs to manually identify the question being asked or allow the CSOs to read the case and then key in the question instead. This approach might facilitate better accuracy during answer extraction later if the question was elicited correctly [126][127].

5.2.4. Review of Typology for Questions that had failed.

Additionally, to better understand how the low accuracy of CQAS comes about, we also studied the typology of the questions that had failed to get an accurate recommended answer. Generally, questions can be defined by the type of answers expected. Based on this school of thought, there are four key question types: factoid, list, definition, and complex question [45]. Going by the definition of a complex question, which is about information in a context and the answer required, will likely be a merge of retrieved passages [128]. Many of the case records that had failed to get an accurate answer fall into this category of complex questions.

We further investigated this by classifying these case records in accordance with the typology proposed by Shah et al. (2012) so that it can help to elicit insights on proposed solutions to deal with the different question types [129]. Examples of the case records and proposed solutions are presented in Table 10 below.

Table 10 Question Typology for Failed Case Records and Proposed Solutions

Category	Definition [	Example of a Case Record	Proposed solution
Ambiguity	Question is too vague or too broad, and for this reason, is misunderstood or causes multiple interpretations.	"Appeal for course"	Create a minimal word count or maximum word count on the input field for the citizen so that overly broad or vague inquiry can be minimized.
Lack of information	Not enough information exists to identify the asker's intended information-seeking goal.	"I would like to re-appeal for my previous case"	Implement an additional feature in the inquiry portal so that when the citizen submits the inquiry, the system can prompt the citizen to indicate his or her case number.
Poor syntax	Question syntax is ill formed, has typos, or has Internet slang that hampers understanding.	"I have een rying to caim credit for a cours but just cannot because system not working. Can you pls assist?"	Introduce autocorrect feature on the portal to help the citizen amend the typos in the inquiry.
Too complex and/or overly broad	Question is too complicated, and a few people have the ability and/or the resources necessary to provide answers, even though enough details are provided to identify the asker's intended information-seeking goal.	"I was refunded in full, but I have not yet received my certificate and two weeks have already passed. Has my appeal been approved? Also why is my school not paid yet?"	Introduce an additional dialogue management feature in CQAS to store and split the sentences until individual questions are being elicit.
Relatedness	Title and/or content poses more than one question (although they are related), so the respondent may be confused in interpreting the asker's intended information-seeking goal.	"Can I find out how to appeal for AB765 course? Can I also know how I can use my credit for it?"	Introduce a local and general label to the collections of documents so that inter-connected questions with its corresponding answers can be surfaced when such question types are sent by citizens.

5.3. Conclusion

In this chapter, using real-world data sets, we presented our experience in building and evaluating a pilot QA system, CQAS for a government agency's customer service centre. This system uses a collection of documents and case records from the domain of training-related government initiatives. CQAS is a hybrid QA system that combines techniques from Natural Language Process QA, Knowledge based QA and Information Retrieval QA.

Our initial assessment showed that despite the high coherence value at the question classification stage, the overall accuracy of CQAS was significantly lower. We did a deep dive into both the case records and the collections of documents and presented the following four key learnings which can be considered for future research work when further enhancing CQAS QA systems:

1. Rather than relying purely on the category "Accurate", it is important to establish an additional category named "Relevant" for the case records when assessing the accuracy of a QA system, as it can help to draw insights if the questions and recommended answers pairs are sufficient though they are not accurate. This is particularly significant in QA systems that have a human in the middle before responding to the customer. The decision to adopt the "Relevant" answers is left to the human, in this instance, the CSO.
2. Inclusion of a feedback loop system so that CSOs can indicate their resolution to inquiries and tag them as recommended answers. This will help to build a more robust QA system, rather than building one based on a legacy dataset. This is specifically significant since the questions and recommended answers pairs are derived from FAQ documents which are meant for human consumption and not machine consumption. Therefore, inclusion of such feedback loop system will help to restructure answers in a format that is for machine consumption.
3. Inclusion of manual question classification mechanism so that CSOs can indicate within the system which part of the case record is a question and thus improve the question classification capability of the QA system.
4. Establishing a question typology for the failed questions that did not have an accurate answer could be recommended as a reply to citizens. This will help the people developing the QA system to further enhance it to specifically cater to the different question types (e.g., ambiguity, poor syntax, etc.) that the CSOs might encounter.

The research done in this chapter contributes to the body of AI applied research in digital government and more specifically, to QA systems to support CSOs responding to citizen inquiries.

Chapter 6

6. 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 [102][103][104][105]. SLAs play a pivotal role in shaping interactions between citizens and government entities, influencing satisfaction and trust in governmental operations [1]. Despite their significance, SLAs can fail due to various factors, hindering government agencies' ability to meet citizens' expectations [2].

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 Chapter 4 and 5, which involved the integration of lexicon libraries like Empath in SLA prediction [10][11][12] and Cosine similarity in CQAS, 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 for replying to citizens.
With consideration that ChatGPT might run into misinformation or hallucinations, human-in-the-loop (HITL) methodology was adopted by having the CSO reviewing the output before sending out to the citizens. What is HITL? The Human-in-the-Loop

(HITL) approach is a methodology that integrates human intelligence and oversight into automated systems or algorithms [191][1920. In HITL systems, humans are actively involved in the decision-making process, providing feedback, guidance, and supervision to improve the performance and reliability of AI-driven processes [193][194]. This approach leverages human expertise, in our instance, the CSO, to handle complex or ambiguous situations, ensure ethical and responsible decision-making, and enhance the overall quality of outcomes [195].

By involving CSOs in the loop, ACQAR ensures that human judgment and expertise are applied to augment and refine AI-generated responses, fostering trust, empathy, and efficiency in citizen service delivery, while considering citizens' 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 [4][5][6]. 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 chapter 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].

Overall, this research attempts to answer the following research question:

- Does integrating a question-answer recommender, augmented with ChatGPT, improve citizen satisfaction and the efficiency of customer service officers?

The answer to the question 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.

6.1. Refinements to CQAS

Our prior research in Chapter 5 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 [136]. 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 [9].

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 complex 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 Chapter 5, the revised CQAR used in the setup of ACQAR has been refined via the following methods:

1. FAQ dataset had been rewritten to avoid the 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%.

6.2. Incorporating ChatGPT

The design of ACQAR was completed with Empath X SLA predictor, enhanced 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 [137].

The consideration of incorporating generative AI technologies is due to its potential to transform unstructured data into intelligently crafted replies. With AI, faster decision making can be fostered [138] and ChatGPT can 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 [53][54]. 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, streamline information dissemination, potentially decrease wait times for citizens and enhance the efficiency of government responses, there is still a significant challenge that arises in the form of hallucinations [55]. 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 11 below:

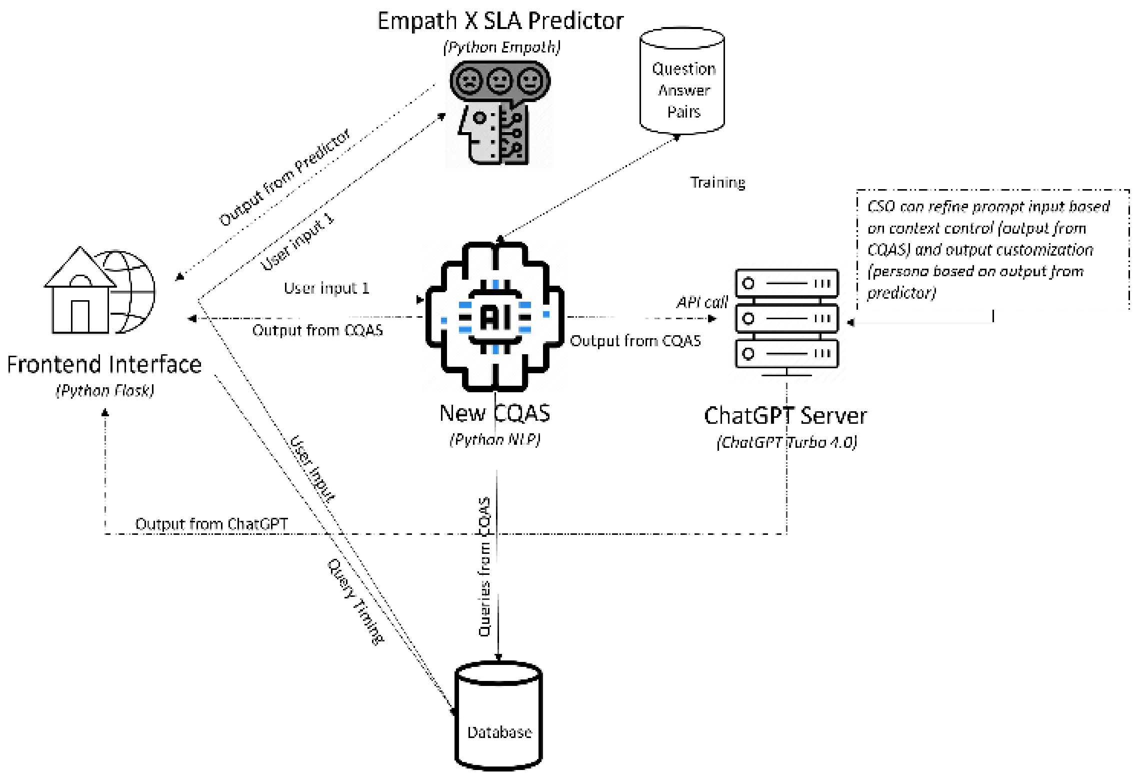

Figure 11 ACQAR System Design

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 frontend user interface is depicted Figure 12 – 14 below. Figure 12 shows the landing page that the CSOs will see the moment they login ACQAR. It was designed to look similar to the Salesforce CRM that they are using, with the Category 1-3 that they will always be referring to or allocating to each citizen's inquiry whenever they received them.

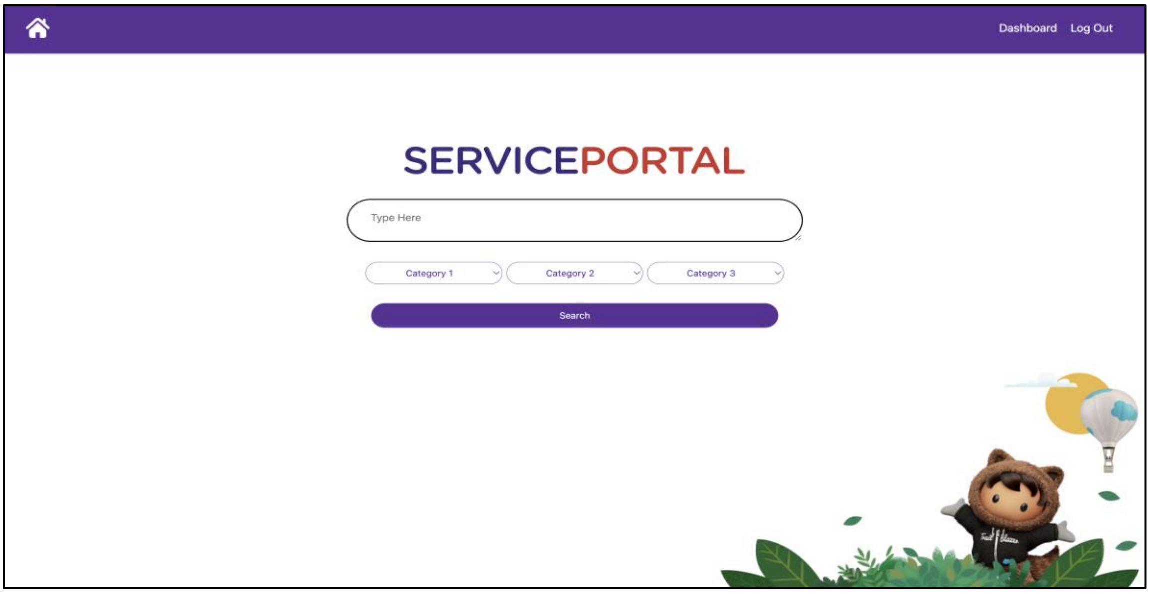

Figure 12 Landing Page

Figure 13 below shows the page whereby after the CSOs had keyed in the citizen's inquiry at the landing page and click "Search", the Question-Answer recommender will breaks the inquiry down into features and based on the similarity scores and recommends the top 10 Frequently Asked Questions (FAQ) to the CSOs. Only if the answers recommended does not fit the inquiry, the CSOs will click escalate button to pass on the case to the next level SME to handle.

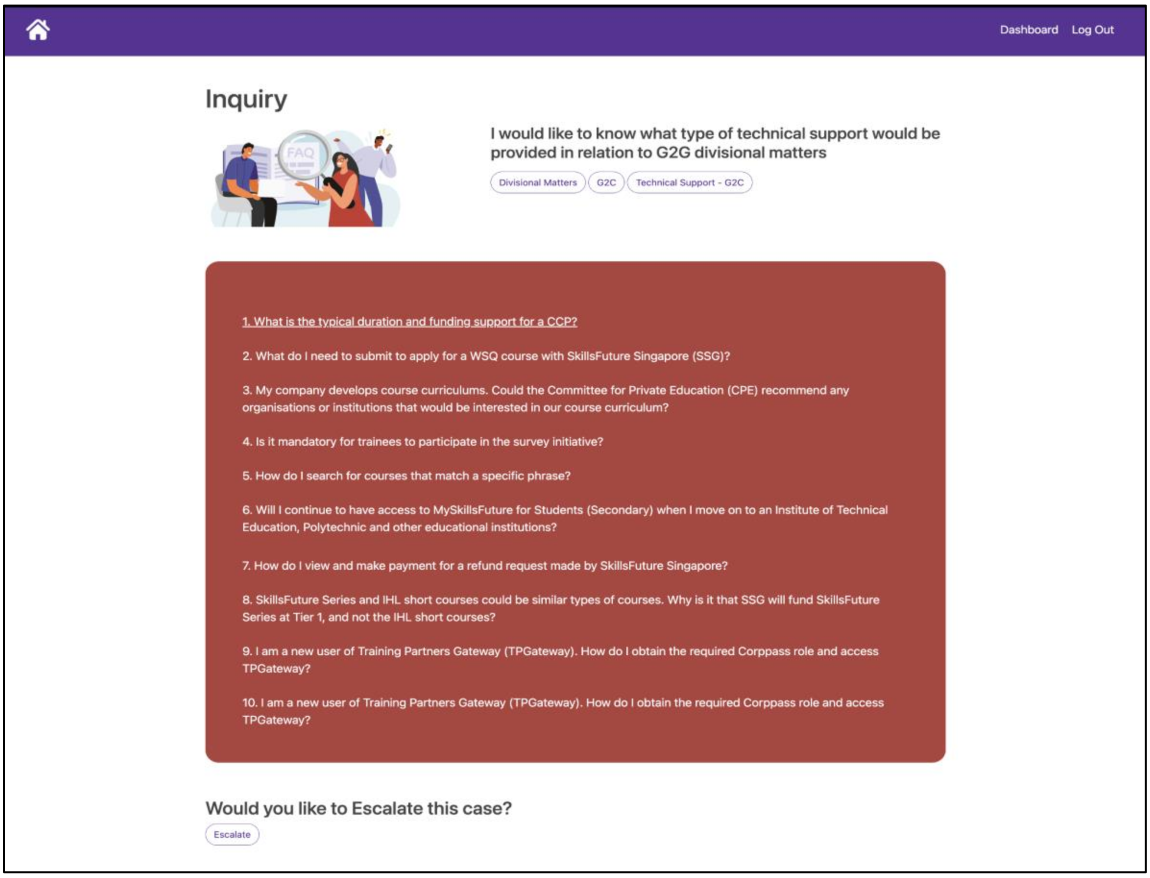

Figure 13 QA Page as an example

Figure 14 below shows the generated answer after the chosen recommended answer and the category of citizen had been input. From here, the CSO will copy and paste the draft reply into Salesforce CRM to further revise before sending to the citizen.

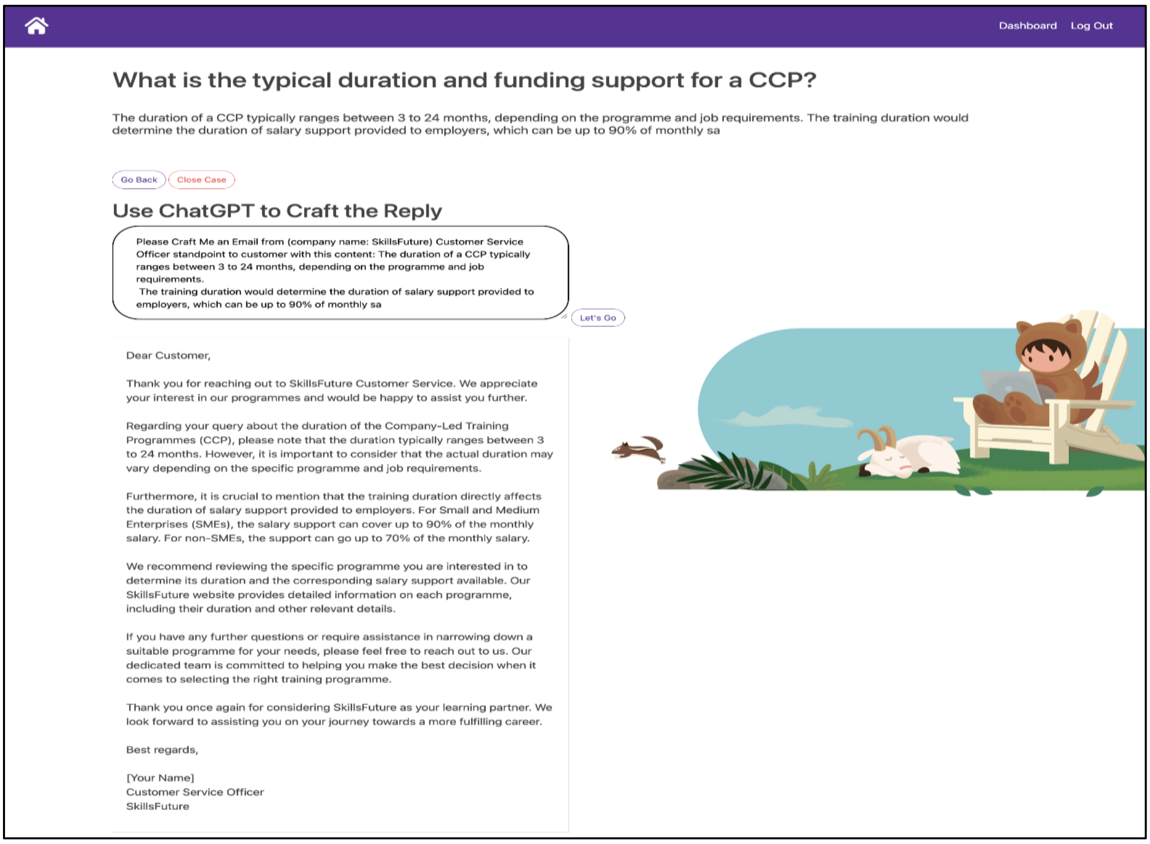

Figure 14 ChatGPT page with prompt template that has recommended answer and empath category

The pilot system enables a CSO to input a citizen's inquiry and select relevant inquiry categories. As discussed previously, 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 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 the use of 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 11 below:

Table 11 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 the standpoint of a customer service officer from XX agency to an *agitated citizen* 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

6.3. Results and Discussion

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 participated 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. Institutional Review Board's approval had been sought for the participation of the 2 CSOs, coupled with the clearance of the survey questions and questions used for focus group discussion (FGD).

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 assistance. 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 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" for these 3 questions. The outcomes are as indicated in Table 12 below:

Table 12 Outcomes of Post-Service Survey

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

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 a 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 observed 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 facilitated superior information retrieval, enhancing the agents' capacity 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. It 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 [196]. Each group of questions are based on the following [197]:

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.
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 13.

Table 13 FGD Questions

	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 data used to train, 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 benefit 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 using 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 accurate and helpful outputs. Quotations from Agent A and Agent B that depicts the concerns are indicated in Table 14:

Table 14 Quotations from Agents during 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."

In summary, 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.

6.4. 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 feedback from the two Citizen Service Officers (CSOs) who 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 the 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 issues not present in the training data, leading to occasional misjudgements.

Furthermore, the agency observed instances of misinformation during the trial experiment. This issue was exacerbated 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.

6.5. Use Cases based on Case Study

The adoption of ChatGPT in government agencies brings immense potential, yet challenges related to transparency and accountability must be navigated for responsible AI integration. Similarly for the case study in Chapter 3, the government agency is exploring various use cases where ChatGPT can be leveraged.

6.5.1. Use Case 1 – Drafting replies to Citizens as per ACQAR

Since the agency works closely with the citizens and organizations on the ground by promoting numerous training initiatives and programs, there is always a lot of inquiries incoming from citizens. This surge in inquiries coupled with the complexity of the information that must be shared with citizens, often resulted in the agency unable to attain a high citizen satisfaction rate. As such, the agency planned to leverage upon ChatGPT to assist in faster crafting of responses to citizens.

6.5.2. Use Case 2 – Faster Creation of Documentation

As a government agency, the number of documents created for the purpose of documentation is alarmingly large. That is why many agencies had put in place document repository systems to facilitate the storage of such documents [139]. The agency in the case study is no stranger to this. The government officers of the agency must create documents ranging from notes of meetings to budget papers. Yet the writing of such documents can be tedious and time consuming. As such, the agency currently is training an internal ChatGPT system for its officers to expedite the drafting of respective documents faster by leveraging past documents as training data for this internal ChatGPT system.

6.5.3. Use Case 3 – Summarisation capability

As the agency leading adult learning and training ecosystem in Singapore, it interacts and governs all the training providers. This also means that the training programmes and courses under this agency's purview are at a record high. Therefore, it is tedious for the agency officers to be screening through numerous programmes or course curriculum to ensure that the training courses offered to citizens are indeed aligned to what the training providers are marketing. ChatGPT proved to be useful when it can easily summarise the lengthy course or programme curriculum and allow the officers to screen and compare with the training providers' marketing content in a more efficient manner. This will minimize the chance that misinformation is provided by the training providers as compared to the actual curriculum that the citizens will undertake.

6.5.4. Learnings from the Use Cases

During the implementation of the respective use cases, 3 key issues were highlighted. First, was the issue of data opacity. When the agency wanted to use Open AI's ChatGPT directly, there was no way to find out where the data would be stored, and which data would exactly be used by the Open AI product. This also means that the agency will face challenges in discerning the origin of the data that the ChatGPT output is based on, or how it has been processed. To overcome this, the agency worked with central government's technology agency to come up with an internal ChatGPT product so that they will know where the data is stored, and which exact dataset was used to train the ChatGPT.

Second, the agency encountered hallucinations from ChatGPT. As an unsupervised learning model, ChatGPT is capable of self-generating new data. While this enhances ChatGPT's human-like responses, the challenge comes about when the model extrapolates or guesses information that it hasn't seen in its training data. This poses a problem for the agency, especially for use case 3, when a large amount of curriculum was fed into the model, and itself generates at times information that was not found within the actual curriculum. This led to occasional misjudgement.

Lastly, the agency during implementation of the use cases, observed that sometimes misinformation can occur. This is especially so when outdated data was ingested by the model already and there is no way to reverse. Hence the output turned out to be factually inaccurate. This could potentially undermine the reputation of the agency.

All the 3 issues pointed to the need for explainable AI (XAI) during the implementation of ChatGPT. In Section 6.6, we will be proposing a 4-Steps framework that incorporates strategies that could enhance this aspect.

6.6. Proposed 4-Steps Framework

The adoption of ChatGPT in government agencies holds great promise but also presents challenges related to transparency and accountability. To ensure responsible AI integration, this framework proposes strategies to enhance AI explainability within the unique context of government operations. The overview of the framework is depicted in Figure 15 below:

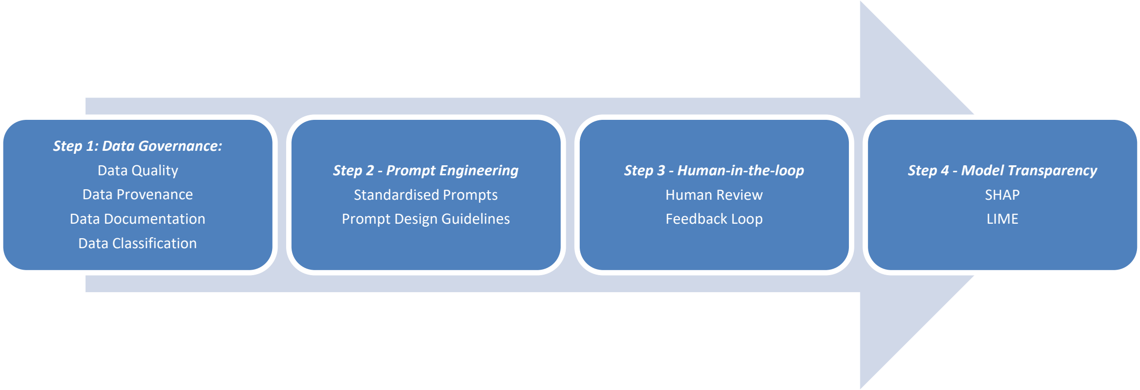

Figure 15 Proposed 4-Steps Framework

6.6.1. Step 1: Data Governance

All AI models relied heavily on the data. Hence the fundamental to enhance AI explainability should start with the focus on data. Considering that government data tend to be of confidential and sensitive nature, there is a need to examine the governing policy of the use of data for ChatGPT [140]. It is proposed that the data governance framework for the use of ChatGPT should include at least 3 key components:

1. Data Quality - Implement rigorous data quality checks and validation procedures to ensure that the data used to train and fine-tune ChatGPT is accurate, up-to-date, and reliable.

2. Data Provenance – We should maintain clear records of data sources, data transformations, and preprocessing steps, allowing government agencies to trace the data legacy and lineage used by ChatGPT.

3. Data Documentation – Creation of comprehensive metadata, data entity relationship diagram and dictionary for datasets should be in place.

4. Data Classification - Streamlining the various datasets into different classification from Official Open, Official Closed, Restricted, Confidential and Secret, will ensure that only certain data is being used within the models based on where the data residency is located.

6.6.2. Step 2: Prompt Engineering

Since ChatGPT's output responses can be varied by prompt inputs, this presents a window of opportunity that government agencies could align the structure of prompt inputs based on the use cases, so that further investigation can take place should any issues occur [141].

Two key components can be of consideration in this aspect:

1. Standardised Prompts – The agency can put in place a standard structure for prompt inputs for common government tasks or inquiries to ensure that ChatGPT provides consistent and reliable responses.

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

6.6.3. Step 3: Human-in-the-loop

Since ChatGPT can potentially hallucinate, introducing human-in-the-loop can help to provide the expert oversight the responses from ChatGPT before the government officers use them [142].

Two key components can be incorporated:

1. Human Review: Implement a human-in-the-loop review system where government experts periodically 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.

6.6.4. Step 4: Model Transparency

Since ChatGPT can potentially provide misinformation, an attempt to make the model more transparent via the use of SHAP or Lime, can assist to know which are the features that contributed to the output response. This will enhance AI explainability of ChatGPT [143][144].

6.7. Conclusion

In conclusion, the findings of this final pilot system 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. However, with that, it means that the case study should consider implementing a framework to assess the AI readiness of her officers before the mass roll-out of ACQAR. 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.

Chapter 7

7. Introduction

As mentioned in Chapter 1.2 and 1.4, AI has demonstrated its capabilities to match up or, in some situations, surpass humans in what was thought to be uniquely human qualities – strategic thinking and decision making. This has led to the mass adoption of AI technologies, coupled with predictions by many researchers to eventually outperform rule-based recommendation or human activities in the upcoming years [130]. This is especially so when AI is founded upon the basis that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it [131]. Hence many are of the belief that AI has the potential to become superior to current rule-based recommendations.

This school of thought is also true of the public sector, whereby citizen service is vital. In utilising government budgets, prudence must be monitored to ensure the 'best value for money'. So, what do citizens look out for in government services? According to Chen (2010), online transactions with the government and feedback channels are the common activities that citizens will participate in [132]. An example of an online transaction with the government agencies will be the application of government relief, while an example of feedback channels will be like Reach-Singapore, whereby citizens are polled regularly on national initiatives to elicit feedback from the ground. One or more government agencies may be involved in handling these citizen-related activities. Yet, Kubicek and Hagen (2000) showed via research that citizens prefer to have a one-stop portal or channel for all the services across government agencies [133]. Further, citizens are looking out for more one-stop citizen-centric services whereby they can interact with any government agency. The latter will be able to cater with a good understanding of the citizens' interaction with the government. Therefore, with such expectations by citizens regarding the provision of services by the government, the adoption of AI will be on the radar of each government to investigate leveraging AI to meet citizens' expectations in the aspect of citizen-centric services.

Hence to leverage AI as part of an agency's digital blueprint, to meet citizens' expectations, agencies will be pressed to have a tool to assess their own AI readiness before they take a leap of faith to implement AI as one of their IT enablers. Coupled with the learning lessons taken from the implementation of ACQAR as indicated in Chapter 6, such tool will be useful prior to the mass implementation of ACQAR to assess the readiness of the case study's officers.

7.1. Proposed Readiness Framework for Public Sector

From the literature review done in Chapter 2.4, it is established that the TOE framework, because of its versatile nature, continues to be applicable in today's settings. Hence, the framework that we propose for implementation in the public sector will be based on TOE. However, the uniqueness of the public sector, such as the presence of highly integrated intranets, robust data governance framework, stringent requirements for security protocols, etc., differentiates the agencies in the public sector from many organisations in the private sector. Hence, the framework applied in the public sector context will have to consider additional sub-criteria under the TOE framework on top of those used in the private sector.

To help identify the sub-criteria that need to be added to the TOE framework in assessing AI readiness in the public sector, a survey protocol was prepared to gather inputs from ICT and operations officers from the case study as indicated in Chapter 3. This is how we derive the proposed AI readiness framework that will later be discussed under Chapter 7.3 and subsequently trial on an actual use case implemented in the case study. From there, this same framework will be deployed prior to the mass implementation of ACQAR.

7.2. Methodology

Data collection occurred between May 2020 and October 2021, whereby officers belonging to the case study that is currently working on the implementation of two projects, i.e., an AI-enabled customer data platform and an AI-enabled customer relationship management system, were surveyed on their future business requirements, experience with current systems and business outcomes for AI-enabled technologies. The inputs were gathered and organised by significant business functions and manually differentiated using themes to form the sub-criteria for the TOE framework. This framework can be further extended to other agencies as it is based on a generic context of citizen service delivery that is a common denominator across the board.

7.2.1. Target Participants

A total of 15 officers from 7 different divisions were surveyed based on their knowledge and experience using the agency's existing marketing platform and CRM system. Their years of service with the agency range from 1 year to 20 years. The average number of years of service is five. Considering their varying work scope, they were able to provide valuable insights into the agency's business landscape and technology landscape from multiple contexts. The work scope of the seven divisions is indicated in Table 15 below:

Table 15 Participants' Profile and Relevance

Division Name	Specific Work scope	Relevance
Division 1	Engagement with citizens on their learning and driving life-long learning initiatives	The participants are part of delivering citizen services and hence will often have to handle citizens' inquiries. They must be aware of what could support their operations better on behalf of the organisation.
Division 2	Engagement with citizens on their upskilling and individual-related initiatives	
Division 3	Engagement with Enterprises on the training of their employees	
Division 4	Branding and marketing of the agency at the national level	The participants market the agency's citizen services and will be the responsible party to draft formal reply to citizens for escalated inquiries. They must be acutely aware of the environmental landscape that the agency is in.
Division 5	Supporting the citizen service delivery by catering to all the inquiries from citizens	The participants are the first point of contact whenever citizens' inquiries come in. They must be aware of what could support their operations better on behalf of the organisation.
Division 6	Implementing and maintaining the IT systems for the agency	The participants are generally the implementation team for the AI capability. Hence would be better to provide insights into the technological context.
Division 7	Supporting the agency in terms of IT infrastructure and aligning the agency's infrastructure with national IT governance	

7.2.2. Survey Design

A total of 12 open-ended questions were used to survey the participants. The survey questions were designed and grouped according to the three contexts of the TOE framework as indicated in Table 16. The description of each context is respectively defined in Chapter 7.3.1, 7.3.3 and 7.3.5.

Table 16 List of survey questions under 3 contexts

Technological Context	Organizational Context	Environmental Context
Have you taken data analytics courses? Please share details of the course, if so.	Can you list down in order of priority what will be the support required to implement AI-enabled CDP and CRM?	What are some of the AI use cases that you know in the market and can be applicable to our agency?
Have you done a data project before to understand the citizens better? Please share details of the project, if so.	What are the current business processes that are not supported by the existing systems and why?	What do you think are the obstacles to implementing AI in the agency?
Have you been involved in projects that deploy analytical models in IT systems?	What AI capability do you think would be required to help you understand the citizen better?	Do you think the agency is comparable to the private sector in terms of technology and why?
How can the current infrastructure or technological systems be adapted to help you do your job better?	Do you think the agency is ready for the implementation of AI and why?	What are the other ways that we could engage citizens more effectively?

7.3. Discussion

The survey results were collated in the form of qualitative feedback. Each participant had provided their responses based on the open-ended survey questions. From there, stop words such as 'is, are' etc are removed. The remaining words were then analysed based on term frequency to identify the possible terms or related concepts when reviewed in accordance with that term. For example, 'data' is a term that has appeared many times. Hence it is listed as one of the sub-criteria shared in Chapter 6.3.2. When going through the detailed feedback that contains the term 'data,' we observed that ten out of fifteen officers related it to data sources to support analytics. Five out of fifteen officers mentioned the term 'data' in relation to data analytics training. With this input, the term 'data' is translated to a sub-criterion under technological context under the TOE framework. Next, we will describe the different contexts used in the framework.

7.3.1. Technological Context

The technological context covers both the internal and external technology context in the TOE framework. Internal context refers to the existing technology stack that the agency is already using. In contrast, external technology context refers to the technology made available in the market not yet adopted by the agency. The government agency's internal context is an important consideration for readiness assessment for the same reason as a company in the private sector. The reason is that internal context estimates how much an agency, or a company can undertake in terms of technology scope and change. Further, the existing technology stack also influences how readily the agency or company will adopt AI. Given that sunk costs are in place and being mindful of budget and cost, both a government agency and its technology partner company will need to justify the monetary benefits that adoption of AI can bring about to offset the sunk costs of the existing technology stack if the latter is to be replaced or enhanced.

AI-enabled systems will likely be found in the external context; however, such new technologies are generally provisioned in tiers within this setting. For example, Amazon web services provide different subscription tiers to cater to other organisations in different stages of analytics readiness. Another example is Salesforce CRM, which also offers tiered CRM solutions based on whether the organisation is at the initial phase of dabbling in predictive insights or at the advanced stage with AI built in to provide recommendations in the form of a subsequent call to action. The presence of such tiered technologies available in the external context is an indication to a government agency or a company that the assessment of readiness is dependent on whether many others in the same sector are adopting the corresponding tier of the new technologies.

7.3.2. Proposed sub-criteria under the Technological Context

From the qualitative data gathered and comparison to existing literature, the results are depicted in Table 17 below. In summary, skillset, infrastructure, data, integration, and security are the key terms that had been constantly used by the participants as they responded to questions under this context.

Table 17 sub-criteria under Technological Context

Sub-Criteria under Technological Context	Definition	Assessment if meet criteria
ICT Expertise Terms: AI, training, analytics	ICT expertise is defined as the agency level of specialized ICT knowledge and skills to provide reliable support in using AI-enabled technologies. The agency is likely to have higher readiness if such expertise is available. [Lin and Lee, 2005]	The agency can assess the percentage of its ICT officers that have taken up AI and data analytics-related training. Further, the agency can also assess the percentage of its officers in terms of request for retraining in AI and data analytics-related training. The higher the percentage, the more ready the agency is [Wang et al., 2019]. This is because it reflects the willingness of the officers to be trained in this aspect.
ICT Infrastructure Terms: cloud, database	ICT infrastructure in a government agency is defined as the physical technology resources, including server rooms, shared government commercial cloud, and shared private government cloud, which provide the foundation to deploy the AI-enabled technologies. Agency will likely deploy such technology if its infrastructure is sophisticated enough [Kowtha and Choon, 2001]. Currently, taking the Singapore government	The agency can assess its readiness by obtaining the number of analytical models deployed within her current ICT infrastructure. If there is none or a low number of such deployments, there are likely challenges to deploying AI-enabled technologies within the existing infrastructure. The agency can also assess its readiness by obtaining the number of officers who are keen to maintain the models deployed

Sub-Criteria under Technological Context	Definition	Assessment if meet criteria
	agency as an example, the shared government commercial cloud will require python or R libraries to be re-packaged and deployed into the environment before deploying AI algorithms. Hence, resources are required for such effort, which translates into the challenges to the agency's readiness.	within her current ICT infrastructure. The higher the number, the likely it is able to be maintained in the long run (Düdder et al., 2021).
Data Terms: Data	Data in the government agency is defined as the data of the citizens and enterprises that had interactions with the agency, such as transactional data, digital web footprints captured from the front-end sites belonging to the agency, etc. Agency with high volume and high variety of data sources are more likely to adopt AI-enabled technologies [Sharman et al., 2022]	The agency can assess its readiness by doing an environment scan of the current data sources and the volume of data available to the agency.
Security & Privacy Risks Terms: Security, Privacy	Security and Privacy risks in the government agency are defined as risks associated with data hosting, firewalls, virtualized and shared resources, and data transfer over the internet and intranet to ensure that data privacy cannot be compromised by hacking. Agency with low security and privacy risks are more likely to be ready to adopt AI-enabled technologies [Subashini and Kavitha, 2011]	The agency can assess its readiness to adopt AI-enabled systems by reviewing its current digital environment regarding security and privacy support.
Integration Terms: cross-agency, system integration, single-sign-on	Integration in the government agency is defined as integrating systems within an agency and across agencies, be it via JSON or API. The more integrated the systems are for an agency, the more likely the agency will adopt AI-enabled technologies [Themistocleous & Irani, 2001]	The agency can assess its readiness by reviewing its enterprise architecture blueprint to determine how integrated its systems are within the agency and with the systems of other agencies.
Audit requirements (IT) Terms: Audit, tracking	IT Audit requirements in the government agency are defined as the log trails captured for every change or adjustment made within the system. The more stringent this is, the longer it takes to deploy AI-enabled technologies [Brundage et al., 2020]	The agency can assess its readiness by reviewing the number of required IT audits per year.

7.3.3. Organisation Context

The organisational context refers to the characteristics and resources of the organisation that impact the implementation decision of the new technology. Examples will be product champions and the presence of a cross-functional task force. Suppose the organisation has champions that are well versed or familiar with AI-enabled technologies in the market. In that case, the organisation will likely be made aware of the possibilities of such technologies. Further, an organisation that has many silo workstreams is unlikely to be able to foster AI innovation, as shared by Najdawi and Shaheed (2021) [134]. This is probably why companies like Netflix, Facebook, etc., encourage the formation of cross-functional teams to drive innovation, in turn driving the faster adoption of AI.

7.3.4. Proposed sub-criteria under the Organisational Context

Under the organisational context, participants tend to use terms such as 'management, support, processes, audit' etc. This highlights that under this context, operations to leverage AI capabilities in the view of the participants that are employees of a government agency, tend to believe that implementation of such capabilities would be smooth with the endorsement of the management, coupled with their business processes being taken care of. The results for this context are presented in Table 18.

Table 18 sub-criteria under Organisational Context

Sub-Criteria under Organisational Context	Definition	Assessment if meet criteria
Senior Management Support Terms: management, forums, approval	Senior management support is defined as the extent to which the management of the government agency will actively support the implementation and management of AI-enabled technologies. This usually comes in the form of approval of budgets, updates required at agency-level forums, and articulation of goals and vision for such technology implementation. The higher the support level, the more likely the agency will implement AI-enabled technologies.	The agency can assess her readiness by knowing the number of approval forums required before procurement of AI-enabled technologies can be achieved. Further, if there is an agency-level forum where the project updates will be reported, the agency is likely ready to embrace this new technology.
Business processes and explainable nature Terms: workflows, operations, processes	Business processes are defined as the operational workflows that a government agency has in place to deliver services to its citizens. The more complex such processes are, the more challenges that the agency will face in translating such processes into the AI-enabled technologies [Clarke, 2019]	The agency can assess its readiness by reviewing the current business processes and derive the percentage of processes that could be translated using AI-enabled technologies with transparency and explainable nature maintained.
Extent of coordination Terms: teamwork, the taskforce	Extend of coordination is defined as the use of different coordination mechanisms while using AI-enabled technologies, such as forming a steering committee to monitor the implementation of such technologies [Pudjianto et al., 2011]. The higher the extent of coordination, the more likely the agency will adopt such technologies [Chatterjee et al., 2019]	The agency can assess its readiness by reviewing the number of cross-functional teams in the agency.
Audit requirements (Business)	Business audit requirements are defined as the audit process of the soundness of the business processes and if measures had been	The agency can assess its readiness by reviewing the steps required in business processes as

Sub-Criteria under Organisational Context	Definition	Assessment if meet criteria
Terms: Audit, tracking	taken to avoid fraud etc. The more stringent this is, the longer it takes to deploy AI-enabled technologies [Brundage et al., 2020]	part of audit requirements. The more steps there are, the more likely the agency is not ready to adopt AI-enabled technologies.

7.3.5. Environmental Context

The environmental context refers to how the organisation conducts its business, such as industry nature, competitors, regulations, etc. For a private sector, it is likely to be impacted by concept of the industry life cycle, as mentioned by Baker (2012) [135]. An example is the textile industry, which is at a maturing stage; innovation might not be as clear-cut as a technology industry. However, the industry life cycle concept does not apply to the public sector. It is more susceptible to innovations in other sectors and the dynamics of global issues.

7.3.6. Proposed sub-criteria under the Environmental Context

Terms such as 'IM8, blueprint, social media' etc are being used frequently in the responses of the participants in relation to the survey questions in this context. It reflects that as government officers, there is acute awareness of the existing regulatory frameworks governing the use of AI. Results for this context are depicted in Table 19 below.

Table 19 sub-criteria for Environmental Context

Sub-Criteria under Environmental Context	Definition	Assessment if meet criteria
Regulatory Environment Terms: IM8, data classification	Governance refers to the regulatory environment that the government agency has to comply with. With proper and supportive governance, the agency will likely be more ready to implement AI-enabled technologies [Pudjianto et al., 2011].	The agency can assess its readiness by reviewing the AI-enabled technologies that the agency is keen to adopt that are compliant with regulations.
Nation Mandate Terms: Digital blueprint, IMDA, Govtech	Nation Mandate refers to the overall direction that the nation or country is moving towards. If the whole government is moving towards digitalization and adoption of AI-enabled technologies, the agency is likely to implement such technologies.	The agency can assess its readiness by reviewing if there is a structured national mandate in place.
Competitive environment	Competitive environment is defined as the landscape that the agency is in. If the private sector implements more AI-enabled	The agency can assess its readiness by comparing its current technology stack against

Sub-Criteria under Environmental Context	Definition	Assessment if meet criteria
Terms: private, new technology, blockchain	technologies, this will accelerate the government agency to consider adopting such innovations [Zhu et al., 2003]	the technology products available in the market.
Social Approach Terms: social media	The social approach is defined as the social media platforms present in the current environment. The more such platforms exist, the more likely the government agency will consider implementing AI-enabled technologies to reach out to the citizens as an effective way to increase the target base of the organization [[Chatterjee et al., 2019]	The agency can assess its readiness by reviewing the number of social platforms that the government agency is engaging its citizens. The higher the number of the social platforms, the higher the probability that her officers are savvy enough to consider AI-enabled technologies.

7.4. Application of the Proposed Framework

In this section, we describe how the TOE Framework with the various sub-criteria was applied to assess the AI readiness of the case study stated in Chapter 3, for the government agency's project titled 'AI-Enabled Customer Data Platform'. The AI-enabled customer data platform is a system that unifies different data sources, such as cookie data, social media data, etc., for a single individual. With that data, the system can further leverage AI capabilities to recommend content to that individual based on past browsing history or social chatter when the individual landed on the agency's website. This technology will empower the agency to provide personalised and targeted information to the citizens through automated analysis of legacy data to find out possible preference trends and to then mass disseminate that information. However, the agency would like to know if their officers are ready for such implementation of an AI-enabled system. Hence, this proposed framework was used to assess the AI readiness of the agency.

7.4.1. AI Readiness for Case Study

Prior to project implementation, using the above framework, the project team is looking to implement an AI-enabled Customer Data Platform and review the agency's readiness for AI by checking against the sub-criteria as per Table 20 below:

Table 20 Review under 3 Contexts

Technological Context		
Sub-Criteria	**What was observed**	**Assessment**
ICT Expertise	80% of the agency's ICT officers had taken and completed the mandate data analytics training, which touches on the application of AI in the public sector.	The higher percentage indicated that the agency is likely to be ready to look into AI-enabled systems.
ICT Infrastructure	Currently, the agency has 60% of its IT applications deployed on the government commercial cloud. Out of the 60%, 100% of the applications involved deploying analytical models, from blockchain technology to fraud detection.	The high percentage indicates that the ICT officers are familiar with analytics-related deployment and will likely face fewer challenges when looking into the deployment of machine learning models in AI-enabled systems.
Data	Currently, the agency has a data lake in place, which helps to consolidate all its data sources in one place to support analytics. Further, considering that the agency manages individual-related initiatives, it has a rich data source with a substantial volume of at least 3 million individuals in Singapore.	The presence of different data sources and high volume is likely to support the data requirements of AI-enabled systems, which generally require a wide variety and volume of data.
Security & Risks	The agency hosted her infrastructure on either government commercial cloud (a form of national private cloud with substantial penetration tests done regularly) or on-premises. The servers are under lock and key.	Due to the sensitive nature of the agency's data, there are potential risks involved should a data leak incident occur. Considering that an effective AI-enabled system will be best deployed on the public cloud, as mentioned in Section 3.2.1, this poses an issue for the agency when assessing AI readiness.
Integration	While the systems within the agency are fully integrated, these systems are only integrated with two other agencies and no external systems.	It is assessed that integration between systems is a norm for the agency. The presence of the AI-enabled customer data platform will allow the agency to integrate with a centralized web analytics system, i.e., WOGAA (Whole-of-Government Application Analytics - wogaa. sg), to obtain more data. Further, the integration of the platform with other social media platforms will eventually enrich the agency's data sources to better support AI initiatives.
Audit Requirements (IT)	There are strict audit requirements for the agency, and this is conducted twice a year.	It is assessed that given the regular IT audits, the agency is familiar and equipped to know how to take steps to ensure that the AI-enabled customer data platform has logs activated to support such audit activities.
Organizational Context		
Senior Management Support	The leadership group of the agency supported the project as they wanted to streamline the engagement approach and ensure personalized and targeted content was recommended to citizens. Further, a specific forum is set up to track the project closely.	This is a strong indication that there will be support for such AI implementation by the agency.

Business Processes and Explainable Nature	Current business processes to track web data and social media activities are in silos. As a government agency, there might be a need to explain if citizens enquired about the personalized content that surfaced when they landed on the agency's website.	It is assessed that the agency will have to think of alternatives to handle the streamlining of business processes and explain the nature of the personalised outreach.
Extent of Coordination	A cross-functional task force for this project was set up.	Since the task force is cross-functional, it will be easier to implement such projects.
Audit Requirements (Business)	The agency's business processes are audited every five years. There is an existing audit framework within the agency to review its business processes internally.	The agency can ensure that the existing framework reviews the business processes impacted by the project before implementation.
Environmental Context		
Regulatory Environment	There is transparent governance set up by Govtech, the centralized technology arm of the Singapore government, about the implementation of cloud solutions and AI.	The agency is assessed to be compliant with the solutioning of the AI-enabled customer data platform, based on the current Singapore regulatory governance, such as the Instructional Manual (IM8) and Risk Assessment of Software-as-a-service (SaaS) products.
Nation Mandate	There is the overall direction for Singapore to move towards the ideal state of a smart nation.	Implementing an AI-enabled customer data project will propel the agency in the same direction as the smart nation mandate.
Competitive Environment	In recent years, AI-enabled technologies have been implemented in the private sector.	Considering that the agency has partnerships with many companies, including Microsoft and salesforce in the private sector, to roll out its initiatives, it is aware of the latest developments. It compares its current technology stack against the technology available in the market.
Social Approach	The agency currently owns social media accounts across five different social media platforms. Citizens are active across different platforms.	The multiple numbers of social media platforms that the agency needs to manage to show that it will have a higher AI readiness. It would be familiar with such platforms and require AI-enabled technologies to support citizen engagement across all social media platforms.

7.4.2. Discussion of the outcomes from the Case Study

Out of the 14 sub-criteria under the three contexts, the case study had a fairly high score of 12 for AI readiness, as indicated in Table 21 below:

Table 21 Overall AI Readiness Score

	Technological Context	**Organizational Context**	**Environmental Context**
Meeting Criteria	5/6	3/4	4/4
Not Meeting Criteria	1/6	1/4	0/4

Although there are three sub-criteria that the project team had assessed that could lead to the agency not being ready for AI-enabled technologies, corrective measures could be taken to address these sub-criteria. Hence the project eventually obtained the budget for the AI-enabled system and awarded the vendor to implement the solution in April 2022. The corrective measures that have been taken are indicated in Table 22 below.

The rationale to include corrective measures based on the proposed framework conducted on an actual case study, is to allow other government agencies to consider such measures and not be deterred by the outcomes of the assessment on their own AI readiness.

Table 22 Corrective Measures that could be taken

Sub-criteria	**Corrective Measures**
Security & Risks	If the agency would like to leverage AI, they will have to take steps to review the data and only allow datasets that have a lower risk to be stored in AI-enabled systems. For this project, the agency had gone with only storing IP addresses, email addresses, and public social media data in the AI-enabled systems.
Business Processes and Explainable Nature	To make sure that the citizens realize that their web data is being used to provide personalized content as they land on the agency's website, a cookie collection clause will prompt for acceptance to allow personalized, targeted content recommendations when the citizens land on the page. If the citizens do not accept it, the personalized content will be removed. In this way, the transparency between the agency and the citizens is maintained. The explainable nature of the AI-enabled customer data platform is also made clear as the citizens will realize that content is recommended based on their agreement to allow the agency to collect their cookie data that helps in providing personalized content as they land on the website of the agency.

7.5. Conclusion

With the pressing need to better serve the citizens by meeting their expectations of service delivery from government agencies, leveraging on AI is one of the key strategies on the radar of each government. In this chapter, we introduced the TOE framework and a survey approach to propose the sub-criteria under each context in this framework applicable to the public sector. This framework aims to assist government agencies in self-assessing AI readiness before adopting AI-enabled technologies.

We used a case study in Singapore that leveraged this framework to demonstrate how this framework can be implemented to assess AI readiness. We also, via the case study, discussed the corrective measures that the agency could take if it is not ready.

The limitation of this work is that it is still considered as exploratory due to insufficient empirical work done in this area for Singapore government. However, we are of the belief that it can be further tried and tested in any other government agency that focuses on citizen service delivery to further ascertain this proposed framework. Future work will extend the approach to AI-based projects in contexts other than citizen service delivery across the Singapore government agencies.

Chapter 8

8. Conclusion

8.1. Key Findings and Contributions

Technological progress has ushered in an era of unprecedented data collection capabilities [24]. The ubiquity of data and analytics in our current landscape calls for a re-evaluation of the traditional boundaries within which citizen service research operates [18]. We started off with problem relevance in Chapter 1 by highlighting the challenges to provide prompt and appropriate citizen service delivery. From there, using real-world data sets, we presented our experience in building three pilot systems and proposed two frameworks as the "design as an artefact" step in Table 1.

This dissertation then attempts to answer the three research questions as stated below:

- Can a predictive model incorporating both numeric and textual data effectively forecast SLAs?
- How does emotion analysis impact the predictive model's efficacy?
- Does integrating a question-answer recommender, augmented with ChatGPT, improve citizen satisfaction and the efficiency of customer service officers?

The research questions prompt the need to do in-depth literature review of research done in the following domains and highlighted the respective research gaps in Chapter 2. A summary of the review is depicted in Table 23.

Table 23 Summary of Research Gaps and Insights

Domain of Literature Review	**Research Gaps/Insights**
SLA Research (Chapter 2.1)	1. Current research in SLA prediction does not consider inclusion of human-centric indicators for human emotions. 2. While SLA may nudge for promptness of replies to citizens, it does not cater for the apprioriateness of the replies.
Question Answer Systems (Chapter 2.2)	1. Every QA system has its limitations, but a hybrid QA model is in better position to navigate these limitations, hence CQAR and ACQAR is built using a hybrid QA model.
ChatGPT (Chapter 2.3)	1. Current ChatGPT research while developing rapidly, the component of citizen delivery related research is still yet to be fully explored beyond an enhanced chatbot or search engine.
Existing Frameworks in Assessment of AI Readiness (Chapter 2.4)	1. TAM framework is not sufficient to be leveraged for modification into an AI readiness framework, yet there are more research that showed that TOE framework is more suitable. There is an established AI readiness framework implemented that is based off TOE framework.
Explainable AI (Chapter 2.5)	1. The current research offers insights to what ACQAR can consider incorporating, i.e. current use of Human-in-the-loop methodology, and use of SHAP and LIME in the future wide scale implementation of ACQAR.

The introduction of the first pilot system built, i.e. Empath X SLA predictor attempts to answer the first two research questions and caters to the promptness of citizen service delivery as indicated as part of the research problem. The outcomes contribute to the body of SLA research by proving that inclusion of human-centric indicators as part of SLA prediction is possible and useful. With such tools, customer service officers (CSOs) cannot only relate to the predicated SLA, but also prioritise the citizens' inquiries based on their predicted human-centric categories. This results in potential promptness of citizen service delivery which will be proved by the final pilot system, ACQAR, which incorporates Empath X SLA predictor.

While Chapter 4 dealt with the promptness of replies to citizens, the second pilot system, i.e. CQAR, in Chapter 5, is built to handle the other requirement – appropriateness of replies. This system served as the prelude to the birth of the final pilot system, i.e. ACQAR in Chapter 6, as it provided the learning lessons on how the hybrid QA model could be refined.

ACQAR integrated the predicted SLA outcomes, Empath categories and recommended output from the QA model, and have these outputs processed through ChatGPT. CSOs utilize a designated prompt template for crafting efficient replies to citizens. Considering the need for design evaluation, it was implemented in an actual customer service centre for a government agency based in Singapore. The key findings of this final implementation and evaluation of ACQAR shows that CSOs after using the system had a notable decrease in their average resolution time. The case study's in-house post-service survey data also indicated an improvement in citizen satisfaction, suggesting that using an AI-based tool like ACQAR does help in responding to citizens' inquiries promptly and appropriately.

All three pilot systems had undergone research rigor during the construction by testing it against various algorithms from LSA to LDA and from Logistic Regression to Random Forest etc. Such rigor allows future researchers to have a basis if they want to construct similar systems.

Finally, the advancement in artificial intelligence has presented government agencies with unprecedented opportunities to enhance citizen service delivery. However, amidst the promises of AI, there exist potential pitfalls such as misinformation and bias in recommendations. ACQAR as a human-in-the-loop system, endeavours to address these challenges. Additionally, due to the few instances of hallucination and misinformation from ACQAR, a proposed 4-Steps framework was introduced. Finally, Chapter 7 examines the end-to-end AI preparedness of the case study, leveraging the proposed AI readiness framework to ensure organizational, technological, and environmental readiness for the future implementation of ACQAR. The pilot trial of ACQAR yielded largely positive results, with valuable feedback from participants informing potential improvements for future widespread implementation.

8.2 Limitations and Future Work

Nonetheless, it is important to acknowledge the limitations of this research, notably its reliance on a single case study and a small number of participants, albeit in a real-world setting within a government agency in Singapore. Future endeavours hold significant potential for advancing this research. One avenue involves expanding similar systems across the entire agency and conducting trials in other domains to provide a more comprehensive understanding of outcomes. Additionally, refining the ACQAR model through the incorporation of techniques such as LIME or SHAP (SHapley Additive exPlanations) holds promise for further enhancing its effectiveness and applicability. These efforts will not only strengthen the current framework but also pave the way for broader implementation and impact.

Acknowledgements

My deepest gratitude goes to my esteemed supervisory committee members, Vice Provost (Education) Venky Shankararaman, Associate Professor (Education) Ouh Eng Lieh, and Associate Professor Hady Wirawan Lauw. Their boundless patience, unwavering guidance, and steadfast support have been the guiding light of my EngD journey. Their presence, always ready to lend an ear and offer sage advice, has been a source of comfort and strength through every challenge and triumph.

To Professor Michelle Chong and the EngD administration office, I extend my heartfelt appreciation for their endless patience and assistance. Their unwavering support in navigating logistical hurdles and coordinating conference funding arrangements has lifted a weight off my shoulders, allowing me to immerse myself fully in my research endeavours.

To my student intern, Hui Min, who worked through weeks to string together my model and work with me to bring ACQAR to life. I truly appreciate her commitment and zealous in delivery of her tasks.

To my cherished colleagues, including Poh Cheng, Richard Lee, Francis Yap, Angelina Soh, Iris Tan, Peggy Lim, Catherine Goh, Kheng Han, Pearlyn Yap, Liu Wei, Kee Weil, Keat Jau, and Dave Cheong, I owe a debt of gratitude for their unwavering encouragement and camaraderie. Their warm smiles and words of encouragement have turned even the toughest days into moments of shared joy and accomplishment.

Lastly, to my beloved family and cherished loved one, “A”, words cannot express the depth of my gratitude for their unwavering love and support. Their boundless patience, understanding, and unwavering belief in me have been my guiding stars through the darkest nights and brightest days. Without their love and encouragement, this journey would have been unimaginable. Their love continues to fuel my passion and dedication, inspiring me to reach greater heights in both my academic and personal endeavours.

Publications included in this Dissertation

1. Chapter 2.1 and Chapter 4: Lee, Alvina Hui Shan, Venky Shankararaman, and Eng Lieh Ouh. "Implementation of Empath X SLA predictive tool for a Government Agency in Singapore." 2022 IEEE International Conference on Big Data (Big Data). IEEE, 2022.

2. Chapter 2.2 and Chapter 5: 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 Hawaii International Conference on System Sciences (pp. 1910-1919).

3. Chapter 2.3 and Chapter 6.1, 7.4: Lee, Alvina Hui Shan, Venky Shankararaman, and Eng Lieh 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). IEEE, 2023.

4. Chapter 2.4 and Chapter 7: Lee, Alvina, Venky Shankararaman, and Ouh Eng Lieh. "Enhancing citizen service management through AI-enabled systems-a proposed AI

readiness framework for the public sector." Research Handbook on Public Management and Artificial Intelligence. Edward Elgar Publishing, 2024. 79-96.

5. Chapter 6.5 and 6.6: Lee, Alvina Hui Shan, Venky Shankararaman, and Eng Lieh 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). IEEE, 2023.

6. Chapter 6.2 and 6.3: [Accepted 0n 30th March 2024] Lee Hui Shan, Alvina, Venky Shankararaman, and Eng Lieh Ouh. " Enhancing Government Service Delivery: A Case Study of ACQAR Implementation and Lessons Learned from ChatGPT Integration in a Singapore Government Agency" DG. O 2024: The 25th Annual International Conference on Digital Government Research. 2024.

Bibliography

[1] 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.

[2] Wanna, J., Butcher, J., & Freyens, B. (2010). Policy in action: The challenge of service delivery. UNSW Press.

[3] Grosz, B. J., & Stone, P. (2018). A century-long commitment to assessing artificial intelligence and its impact on society. Communications of the ACM, 61(12), 68-73.

[4] 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

[5] 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

[6] 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.

[7] Palle, S. (2019). Artificial Intelligence Using DBS-QOS In Banking Organizations. Journal of scientific research & engineering trends, 5(1), 2395-566X.

[8] Prabhu, Sangeetha, & Bhat, Subrahmanya,(2020). Application of Artificial Intelligence in Samsung–A Case Study. International Journal of Case Studies in Business, IT, and Education (IJCSBE), 4(2), 280-292.

[9] GovTech, Singapore. (2020). Digital Government nic. Government Technology Agency. Retrieved April 30, 2022, from https://www.tech.gov.sg/digital-government-blueprint/

[10] Aggarwal, C.C. and Zhai, C., 2012. A survey of text clustering algorithms. In Mining text data (pp. 77- 128). Springer, Boston, MA

[11] 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).

[12] 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.

[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] Pantelis, K., & Aija, L. (2013, October). Understanding the value of (big) data. In 2013 IEEE International Conference on Big Data (pp. 38-42). IEEE.

[16] 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.

[17] 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.

[18] Ameller, D. and Franch, X., 2008, February. Service level agreement monitor (SALMon). In Seventh International Conference on Composition-Based Software Systems (ICCBSS 2008) (pp. 224-227). IEEE.

[19] Patel, P., Ranabahu, A.H. and Sheth, A.P., 2009. Service level agreement in cloud computing.

[20] Adelmeyer, M., Walterbusch, M., Seifert, K. and Teuteberg, F., 2016. Does the Augmentation of Service Level Agreements Affect User Decisions in Cloud Adoption Scenarios? –An Experimental Approach. European Conference on

Information Systems.

[21] Fehrenbacher, D.D., 2016. The Dual Role of Penalty: Service Level Agreement Design and its Effects on Intra-Firm Knowledge Sharing and Commitment of Service Providers. European Conference on Information Systems.

[22] Comuzzi, M., Marquez-Chamorro, A.E. and Resinas, M., 2019, April. A hybrid reliability metric for SLA predictive monitoring. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (pp. 32-39).

[23] Bosse, S., Schulz, C. and Turowski, K., 2014. Predicting availability and response times of IT services. European Conference on Information Systems.

[24] Iordache, G.V. and Pop, F., 2019, September. Predicting service level agreement violations in cloud using machine learning techniques. In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP) (pp. 71-78). IEEE.

[25] Efazati, S., Amerimehr, M.H., Montazeri, A.M., Sabagh, M.R., Alibeigi, M. and Ayazi, F., 2020. Novel Methods For Determining QoS Parameters and Thresholds in End User's Service Level Agreement. AUT Journal of Electrical Engineering, 52(2), pp.205-216.

[26] Pandita, A., Upadhyay, P.K. and Joshi, N., 2020, February. Prediction of service-level agreement violation in Cloud computing using bayesian regularisation. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 231-242). Springer, Singapore.

[27] Nagini, RVSSS and Pratyusha, T., 2021. Machine Learning Approaches for Improving Service Level Agreement in Cloud Computing. Annals of the Romanian Society for Cell Biology, pp.2532-2543.

[28] Vahidov, R., & Neumann, D. (2008, January). Situated decision support for managing service level agreement negotiations. In Proceedings of the 41st Annual

Hawaii International Conference on System Sciences (HICSS 2008) (pp. 52-52). IEEE.

[29] Stamou, K., Kantere, V., & Morin, J. H. (2013, October). SLA data management criteria. In 2013 IEEE International Conference on Big Data (pp. 34-42). IEEE.

[30] Gouryraj, S., Kataria, S., & Swvigaradoss, J. (2021, September). Service Level Agreement Breach Prediction in ServiceNow. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 689-698). IEEE.

[31] Hashmi, S.I., Haque, R., Schmieders, E. and Richardson, I., 2011, July. Negotiation towards service level agreements: A life cycle-based approach. In 2011 IEEE World Congress on Services (pp. 1-8). IEEE

[32] Ravichandran, D., & Hovy, E. (2002, July). Learning surface text patterns for a question answering system. In Proceedings of the 40th Annual meeting of the association for Computational Linguistics (pp. 41-47).

[33] Simmons, R. F. (1965). Answering English questions by computer: a survey. Communications of the ACM, 8(1), 53-70.

[34] Waltz, D. L. (1978). An English language question answering system for a large relational database. Communications of the ACM, 21(7), 526-539.

[35] He, G., Lan, Y., Jiang, J., Zhao, W. X., & Wen, J. R. (2021, March). Improving multi-hop knowledge base question answering by learning intermediate supervision signals. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 553-561).

[36] Alwaneen, T. H., Azmi, A. M., Aboalsamh, H. A., Cambria, E., & Hussain, A. (2022). Arabic question answering system: a survey. Artificial Intelligence Review, 55(1), 207-253.

[37] Wang, M. (2006). A survey of answer extraction techniques in factoid question

answer-ing. Computational Linguistics, 1(1), 1-14.

[38] Voorhees, E. M. (2001). Overview of the TREC 2001 question answering track. In In Proceedings of the Tenth Text REtrieval Conference (TREC).

[39] Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., ... & Sutcliffe, R. (2006, September). Overview of the CLEF 2006 multilingual question answering track. In Workshop of the cross-language evaluation forum for European Languages (pp. 223-256). Springer, Berlin, Heidelberg.

[40] Sasaki, Y., Chen, H. H., Chen, K. H., & Lin, C. J. (2005, December). Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1). In NTCIR.

[41] Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3).

[42] Soares, M. A. C., & Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. Journal of King Saud University-Computer and In-formation Sciences, 32(6), 635-646.

[43] Aberg, J., & Shahmehri, N. (2001, January). Collection and exploitation of expert knowledge in Web assistant system. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences (pp. 10-pp). IEEE.

[44] Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR), 44(1), 1-50.

[45] Kolomiyets, O., & Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24), 5412-5434.

[46] Lampert, A. (2004). A quick introduction to question answering. Dated December 2004.

[47] Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: the

view from here. natural language engineering, 7(4), 275-300.

[48] Clark, P., Thompson, J., & Porter, B. (1999, March). A knowledge-based approach to question-answering. In Proc. AAAI (Vol. 99, pp. 43-51).

[49] Lopez, V., Pasin, M., & Motta, E. (2005, May). Aqualog: An ontology-portable question answering system for the semantic web. In European Semantic Web Conference (pp. 546-562). Springer, Berlin, Heidelberg.

[50] Fertier, A., Barthe-Delanoë, A. M., Montarnal, A., Truptil, S., & Bénaben, F. (2020). A new emergency decision support system: the automatic interpretation and contextualisation of events to model a crisis situation in real-time. Decision Support Systems, 133, 113260.

[51] Xu, K., Feng, Y., Huang, S., & Zhao, D. (2016, December). Hybrid question answering over knowledge base and free text. In Proceedings of COLING 2016, the 26th Interna-tional Conference on Computational Linguistics: Technical Papers (pp. 2397-2407).

[52] Ferrucci, D. A. (2012). Introduction to "this is watson". IBM Journal of Research and Development, 56(3.4), 1-1.

[53] 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.

[54] 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.

[55] 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).

[56] Yang, L., & Wang, J. (2023). Factors influencing initial public acceptance of integrating the ChatGPT-type model with government services. Kybernetes.

[57] Alexopoulos, C., Saxena, S., Loukis, E. N., Rizun, N., Ioanna Maratsi, M., & Ali, M. (2023). ChatGPT Application Vis-A-Vis Open Government Data (OGD): a Research Agenda. Available at SSRN 4378771.

[58] Siebenlist, T. (2023, July). Approaches towards using ChatGPT as an open data companion. In Proceedings of the 24th Annual International Conference on Digital Government Research (pp. 674-675).

[59] Biswas, S. (2023). Prospective Role of Chat GPT in the Military: According to ChatGPT. Qeios.

[60] Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. Internet Policy Review, 12(1).

[61] Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research, 1-28.

[62] Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1112-1123).

[63] Leboukh, F., Aduku, E. B., & Ali, O. (2023). Balancing ChatGPT and Data Protection in Germany: Challenges and Opportunities for Policy Makers. Journal of Politics and Ethics in New Technologies and AI, 2(1), e35166-e35166.

[64] Subagja, A. D., Ausat, A. M. A., Sari, A. R., Wanof, M. I., & Suherlan, S. (2023). Improving Customer Service Quality in MSMEs through the Use of ChatGPT. Jurnal Minfo Polgan, 12(2), 380-386.

[65] Ezenkwu, C. P. (2023, May). Towards expert systems for improved customer services using ChatGPT as an inference engine. Institute of Electrical and

Electronics Engineers.

[66] Biswas, S. S. (2023). Role of ChatGPT in public health. Annals of Biomedical Engineering, 51(5), 868–869.

[67] Kim, J., & Lee, J. (2023). How does ChatGPT Introduce Transport Problems and Solutions in North America?. Findings.

[68] Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. Language Education and Technology, 3(1).

[69] Wang, H., Wu, W., Dou, Z., He, L., & Yang, L. (2023). Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. International Journal of Medical Informatics, 105173.

[70] Esmailzadeh, Y. (2023). Potential Risks of ChatGPT: Implications for Counterterrorism and International Security. International Journal of Multicultural and Multireligious Understanding (IJMMU) Vol, 10.

[71] Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation.

[72] Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. Internet Policy Review, 12(1).

[73] Mich, L., & Garigliano, R. (2023). ChatGPT for e-Tourism: a technological perspective. Information Technology & Tourism, 1-12.

[74] Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. Finance Research Letters, 53, 103662.

[75] Rivas, P., & Zhao, L. (2023). Marketing with chatgpt: Navigating the ethical terrain of gpt-based chatbot technology. AI, 4(2), 375-384.

[76] Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023, April). ChatGPT: Applications, opportunities, and threats. In 2023 Systems and Information Engineering Design

Symposium (SIEDS) (pp. 274-279). IEEE.

[77] Chatterjee, S., Ghosh, S. K., Chaudhuri, R., & Nguyen, B. (2019). Are CRM systems ready for AI integration? A conceptual framework of organizational readiness for effective AI-CRM integration. The Bottom Line.

[78] Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. European Journal of Information Systems, 29(1), 1-8.

[79] Davis, F. D. (1985). A technology acceptance model for empirically testing new end-user information systems: Theory and results (Doctoral dissertation, Massachusetts Institute of Technology). 12407479.

[80] Ajibade, P. (2018). Technology acceptance model limitations and criticisms: Exploring the practical applications and use in technology-related studies, mixed-method, and qualitative researches. Library Philosophy & Practice, 1941.

[81] Tornatzky, L. G., Fleischer, M., and Chakrabarti, A. K. 1990. Processes of Technological Innovation. Lexington books.

[82] AlSheibani, S., Cheung, Y., & Messom, C. (2018). Artificial Intelligence Adoption: AI-readiness at Firm-Level. In PACIS (p. 37).

[83] Nortje, M. A., & Grobbelaar, S. S. (2020, June). A framework for the implementation of artificial intelligence in business enterprises: a readiness model. In 2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-10). IEEE.

[84] Jöhnk, J., Weißert, M., & Wyrtki, K. (2021). Ready or not, AI comes—an interview study of organizational AI readiness factors. Business & Information Systems Engineering, 63(1), 5-20.

[85] Stirling, R., Miller, H., & Martinho-Truswell, E. (2017). Government AI Readiness Index. KOREA, 4, 7-8.

[86] Scott, A. Carlisle, et al. "Explanation capabilities of production-based consultation

systems." American Journal of Computational Linguistics (1977): 1-50.

[87] Neches, Robert, William R. Swartout, and Johanna D. Moore. "Explainable (and Maintainable) Expert Systems." IJCAI. Vol. 85. 1985.

[88] Angelov, Plamen, and Eduardo Soares. "Towards explainable deep neural networks (xDNN)." Neural Networks 130 (2020): 185-194.

[89] Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society Series B: Statistical Methodology 82.4 (2020): 1059-1086.

[90] Wang, Danding, et al. "Designing theory-driven user-centric explainable AI." Proceedings of the 2019 CHI conference on human factors in computing systems. 2019.

[91] Susnjak, Teo. "Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT." International Journal of Artificial Intelligence in Education (2023): 1-31.

[92] Zylowski, Thorsten, and Matthias Wölfel. "An NLP Analysis of ChatGPT's Personality Simulation Capabilities and Implications for Human-centric Explainable AI Interfaces." Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023). 2023.

[93] Ali, Tarek, and Panos Kostakos. "HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)." arXiv preprint arXiv:2309.16021 (2023).

[94] Gunning, David, and David Aha. "DARPA's explainable artificial intelligence (XAI) program." AI magazine 40.2 (2019): 44-58.

[95] Wang, Xinru, and Ming Yin. "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making." 26th international conference on intelligent user interfaces. 2021.

[96] Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

[97] Du, Mengnan, Ninghao Liu, and Xia Hu. "Techniques for interpretable machine learning." Communications of the ACM 63.1 (2019): 68-77.

[98] Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

[99] Huysmans, Johan, et al. "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models." Decision Support Systems 51.1 (2011): 141-154.

[100] Speith, Timo. "A review of taxonomies of explainable artificial intelligence (XAI) methods." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

[101] Dwivedi, Rudresh, et al. "Explainable AI (XAI): Core ideas, techniques, and solutions." ACM Computing Surveys 55.9 (2023): 1-33.

[102] 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.

[103] Davenport, T. H., 2013. "Analytics 3.0." Harvard Business Review, 91(12), 64–72.

[104] 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.

[105] 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.

[106] Becker, M., Borrisov, N., Deora, V., Rana, O. F., & Neumann, D. (2008, January).

Using k-pricing for penalty calculation in grid market. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008) (pp. 97-97). IEEE.

[107] Reta, M. Y. G. G., Hidayanto, A. N., Sandhyaduhita, P. I., & Nazief, B. A. A. (2016, October). Inhibiting factors of Service Level Agreement (SLA) adoption in Government to Citizen (G2C) services: A case of Indonesia. In 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 277-282). IEEE.

[108] Leitner, P., Michlmayr, A., Rosenberg, F. and Dustdar, S., 2010, July. Monitoring, prediction and prevention of sla violations in composite services. In 2010 IEEE International Conference on Web Services (pp. 369-376). IEEE.

[109] El-Amir, H. and Hamdy, M., 2020. Feature selection and feature engineering. In Deep Learning Pipeline (pp. 233-276). Apress, Berkeley, CA

[110] He, J., Larson, M. and De Rijke, M., 2008, March. Using coherence-based measures to predict query difficulty. In European conference on information retrieval (pp. 689-694). Springer, Berlin, Heidelberg.

[111] Tausczik, Y.R. and Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerised text analysis methods. Journal of language and social psychology, 29(1), pp.24-54.

[112] Bodoff, D. and Ho, SY, 2014. THE EFFECTS OF TIME AND NUMBER OF PERSONALISED ITEMS ON USERS' AMOUNT OF SAMPLING. European Conference on Information Systems.

[113] Lusch, R.F. and Nambisan, S., 2015. Service innovation. MIS quarterly, 39(1), pp.155-176

[114] Avery, L., Rotondi, N., McKnight, C., Firestone, M., Smylie, J. and Rotondi, M., 2019. Unweighted regression models perform better than weighted regression

techniques for respondent-driven sampling data: results from a simulation study. BMC medical research methodology, 19(1), pp.1-13.

[115] Niculescu, Andreea I., Ivan Kukanov, and Bimlesh Wadhwa. "DigiMo-towards develop-ing an emotional intelligent chatbot in Singapore." Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures. 2020.

[116] SBram Klievink and Marijn Janssen. 2010. Coordinating e-government service delivery. In Proceedings of the 11th Annual International Digital Government Research Confer-ence on Public Administration Online: Challenges and Opportunities (dg.o '10). Digital Government Society of North America, 209–216.

[117] MOH takes down 'Ask Jamie' chatbot that gave 'misaligned' Covid-19 advice. https://www.todayonline.com/singapore/moh-takes-down-ask-jamie-chatbot-that-gave-misaligned-covid-19-advice

[118] Meyer von Wolff, R., Hobert, S., & Schumann, M. (2019, January). How may i help you?–state of the art and open research questions for chatbots at the digital workplace. In Proceedings of the 52nd Hawaii international conference on system sciences.

[119] Jeffares, Stephen. (2021). The Non-Public Encounter: Self-Service and the Ephemoralisation of Public Service. In The Virtual Public Servant (pp. 125-150). Palgrave Macmillan, Cham.

[120] María Isabel Arias and Antônio Carlos Gastaud Maçada. 2018. Digital Government for E-Government Service Quality: a Literature Review. In Proceedings of the 11th Interna-tional Conference on Theory and Practice of Electronic Governance (ICEGOV' 18). As-sociation for Computing Machinery, New York, NY, USA, 7–17

[121] Lindgren, I., & van Veenstra, A. F. (2018, May). Digital government

transformation: a case illustrating public e-service development as part of public sector transformation. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (pp. 1-6).

[122] Luis F. Luna-Reyes, J. Ramon Gil-Garcia, and José A. Celorio Mansi. 2011. Citizen-centric approaches to e-government and the back-office transformation. In Proceedings of the 12th Annual International Digital Government Research Conference: Digital Gov-ernment Innovation in Challenging Times (dg.o '11). Association for Computing Machinery, New York, NY, USA, 213–218.

[123] Akkaya, C., & Krcmar, H. (2019, June). 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 (pp. 81-90).

[124] Matthew W. Bilotti and Eric Nyberg. 2008. Improving text retrieval precision and answer accuracy in question answering systems. In Coling 2008: Proceedings of the 2nd work-shop on Information Retrieval for Question Answering (IRQA' 08). Association for Computational Linguistics, USA, 1–8.

[125] See A, Liu P J, Manning C D. Get to the Point: Summarisation with Pointer-Generator Networks[J]. Proceedings of the 55th Annual Meeting of the Association for Computa-tional Linguistics, 2017:1073-1083.

[126] Xia Gong, Xiangyi Kong, Zhujun Zhang, Lin Tan, Zixiong Zhang, and Bing Shao. 2019. Customer Service Automatic Answering System Based on Natural Language Processing. In Proceedings of the 2019 International Symposium on Signal Processing Systems (SSPS 2019). Association for Computing Machinery, New York, NY, USA, 115–120.

[127] Ulf Hermjakob. 2001. Parsing and question classification for question answering. In Pro-ceedings of the workshop on Open-domain question answering - Volume 12 (ODQA '01). Association for Computational Linguistics, USA, 1–6.

[128] García-Cumbreras, M. Á., Martínez-Santiago, F., & Ureña-López, L. A. (2012). Architecture and evaluation of BRUJA, a multilingual question answering system. Information retrieval, 15(5), 413-432.

[129] Shah, C., Radford, M. L., Connaway, L. S., Choi, E., & Kitzie, V. (2012). "How much change do you get from 40$?"–Analyzing and addressing failed questions on social Q&A. Proceedings of the American Society for Information Science and Technology, 49(1), 1-10.

[130] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754.

[131] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, 27(4), 12-12.

[132] Chen, Y. C. (2010). Citizen-centric e-government services: Understanding integrated citizen service information systems. Social Science Computer Review, 28(4), 427-442.

[133] Kubicek, H.(Eds. 2000). One Stop Government in Europe. Results from, 11, 1-36.

[134] Najdawi, A., & Shaheen, A. (2021, February). Which Project Management Methodology is better for AI-Transformation and Innovation Projects?. In 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM) (pp. 205-210). IEEE.

[135] Baker, J. (2012). The technology–organization–environment framework. Information systems theory, 231-245.

[136] 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).

[137] Lee, Alvina Hui Shan, Venky Shankararaman, and Eng Lieh 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). IEEE, 2023.

[138] 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.

[139] Abdulkadhim, H. A. I. D. E. R., Bahari, M., Bakri, A., & Ismail, W. (2015). A research framework of electronic document management systems (EDMS) implementation process in government. Journal of Theoretical and Applied Information Technology.

[140] 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).

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

[142] Thrun, M. C. (2022). Identification of explainable structures in data with a human-in-the-loop. KI-Künstliche Intelligenz, 36(3-4), 297-301.

[143] 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.

[144] Abdulkadhim, H. A. I. D. E. R., Bahari, M., Bakri, A., & Ismail, W. (2015). A research framework of electronic document management systems (EDMS) implementation process in government. Journal of Theoretical and Applied

Information Technology.

[145] Public Service Division. (2024, April 8). Service commitment. https://www.psd.gov.sg/about-psd/service/

[146] Panagiotopoulos, Panos, Bram Klievink, and Antonio Cordella. "Public value creation in digital government." Government Information Quarterly 36.4 (2019): 101421.

[147] Valle-Cruz, David, et al. "A review of artificial intelligence in government and its potential from a public policy perspective." Proceedings of the 20th annual international conference on digital government research. 2019.

[148] Poullet, Yves. "Artificial Intelligence and Public Services: the Role of Public Authorities in the Service of the “Third Way” Drawn up by the European Commission." European review of digital administration & law 2.2 (2021): 129-148.

[149] Saldanha, Douglas Morgan Fullin, Cleidson Nogueira Dias, and Siegrid Guillaumon. "Transparency and accountability in digital public services: Learning from the Brazilian cases." Government Information Quarterly 39.2 (2022): 101680.

[150] Bignami, Francesca. "Artificial intelligence accountability of public administration." The American Journal of Comparative Law 70.Supplement_1 (2022): i312-i346.

[151] Hutson, Matthew. "The opacity of artificial intelligence makes it hard to tell when decision-making is biased." IEEE Spectrum 58.2 (2021): 40-45.

[152] Mei, Alex, et al. "Users are the north star for AI transparency." arXiv preprint arXiv:2303.05500 (2023).

[153] Thelisson, Eva, Kirtan Padh, and L. Elisa Celis. "Regulatory mechanisms and algorithms towards trust in AI/ML." Proceedings of the IJCAI 2017 workshop on explainable artificial intelligence (XAI), Melbourne, Australia. 2017.

[154] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model

predictions." Advances in neural information processing systems 30 (2017).

[155] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[156] Formica, Anna, Ida Mele, and Francesco Taglino. "A template-based approach for question answering over knowledge bases." Knowledge and Information Systems 66.1 (2024): 453-479.

[157] Y NIRANJAN, P. R. A. S. H. A. N. T., V. I. J. A. Y. S RAJPUROHIT, and Sanjeev S Sannakki. "CLASSIFICATION OF QUESTIONS USING MACHINE LEARNING TECHNIQUES." International Journal of Computing and Digital Systems 15.1 (2024): 1-8.

[158] Roy, Pradeep Kumar, et al. "Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review." CAAI Transactions on Intelligence Technology 8.1 (2023): 95-117.

[159] Alzubi, Jafar A., et al. "COBERT: COVID-19 question answering system using BERT." Arabian journal for science and engineering 48.8 (2023): 11003-11013.

[160] Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension." ACM Computing Surveys 55.10 (2023): 1-45.

[161] Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension." ACM Computing Surveys 55.10 (2023): 1-45.

[162] Jiang, Lei, and Zuqiang Meng. "Knowledge-based visual question answering using multi-modal semantic graph." Electronics 12.6 (2023): 1390.

[163] Yerragunta, Sudarshan, Subu Kandaswamy, and Rajendra Prasath. "Dynamic Conditional Gating Mechanisms for Question Answering Over Incomplete

Knowledge Bases." Available at SSRN 4377382.

[164] Zhang, Yuteng, et al. "Chinese medical question answer selection via hybrid models based on CNN and GRU." Multimedia tools and applications 79 (2020): 14751-14776.

[165] Esteva, Andre, et al. "COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization." NPJ digital medicine 4.1 (2021): 68.

[166] Lee, Seongyun, Hyunjae Kim, and Jaewoo Kang. "LIQUID: a framework for list question answering dataset generation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 11. 2023.

[167] Polisetty, Aruna, et al. "What determines AI adoption in companies? Mixed-method evidence." Journal of Computer Information Systems (2023): 1-18.

[168] Polisetty, Aruna, et al. "What determines AI adoption in companies? Mixed-method evidence." Journal of Computer Information Systems (2023): 1-18.

[169] European Commission Joint Research Centre. (2021). Artificial Intelligence and Public Administration: Opportunities and Challenges. Luxembourg: Publications Office of the European Union.

[170] Bin Naeem, Salman, et al. "Leveraging the TOE Framework: Examining the Potential of Mobile Health (mHealth) to Mitigate Health Inequalities." Information 15.4 (2024): 176.

[171] Seethamraju, Ravi, and Angela Hecimovic. "Adoption of artificial intelligence in auditing: An exploratory study." Australian Journal of Management 48.4 (2023): 780-800.

[172] Molopa, Tshwarelo. Factors affecting the adoption of artificial intelligence (AI) in the supply chain and logistics lndustry. Diss. University of the Western Cape, 2023.

[173] Smit, Danie, et al. "Exploring the suitability of the TOE framework and DOI

theory towards understanding AI adoption as part of sociotechnical systems." Annual Conference of South African Institute of Computer Scientists and Information Technologists. Cham: Springer Nature Switzerland, 2023.

[174] Abdurrahman, Abdurrahman, Aurik Gustomo, and Eko Agus Prasetio. "Impact of dynamic capabilities on digital transformation and innovation to improve banking performance: A TOE framework study." Journal of Open Innovation: Technology, Market, and Complexity 10.1 (2024): 100215.

[175] Hassija, Vikas, et al. "Interpreting black-box models: a review on explainable artificial intelligence." Cognitive Computation 16.1 (2024): 45-74.

[176] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[177] Lukas, Ondrej, and Sebastian Garcia. "Bridging the Explanation Gap in AI Security: A Task-Driven Approach to XAI Methods Evaluation."

[178] Torres-Martínez, Sergio. "Embodied human language models vs. Large Language Models, or why Artificial Intelligence cannot explain the modal be able to." Biosemiotics (2024): 1-25.

[179] Barbierato, Enrico, and Alice Gatti. "The Challenges of Machine Learning: A Critical Review." Electronics 13.2 (2024): 416.

[180] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[181] Zeng, Xuezhi, et al. "Detection of SLA violation for big data analytics applications in cloud." IEEE Transactions on Computers 70.5 (2020): 746-758.

[182] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. John Wiley & Sons, 2013.

[183] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[184] Biswas, Nirmal Kr, et al. "An approach towards development of new linear regression prediction model for reduced energy consumption and SLA violation in the domain of green cloud computing." Sustainable Energy Technologies and Assessments 45 (2021): 101087.

[185] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." AAAI-98 workshop on learning for text categorization. Vol. 752. No. 1. 1998.

[186] Benila, S., and N. Usha Bhanu. "Service Level Agreement Based Secured Data Analytics Framework for Healthcare Systems." Intelligent Automation & Soft Computing 32.2 (2022): 1277-1291.

[187] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[188] Ansori, Muhammad Isa, Ririen Kusumawati, and M. Amin Hariyadi. "Prediction of service level agreement time of delivery of goods and documents at PT Pos Indonesia using the random forest method." International Journal of Advances in Data and Information Systems 4.2 (2023).

[189] Kumar, Rajeev, and Abhaya Indrayan. "Receiver operating characteristic (ROC) curve for medical researchers." Indian pediatrics 48 (2011): 277-287.

[190] Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

[191] Amershi, Saleema, et al. "Power to the people: The role of humans in interactive machine learning." AI magazine 35.4 (2014): 105-120.

[192] Holzinger, Andreas, et al. "A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop." arXiv preprint arXiv:1708.01104 (2017).

[193] Zanzotto, Fabio Massimo. "Human-in-the-loop artificial intelligence." Journal of Artificial Intelligence Research 64 (2019): 243-252.

[194] Pandey, Rahul. Jointly Improving Performance of Human Annotators and Models in Human-In-The-Loop Machine Learning Systems. Diss. George Mason University, 2023.

[195] Rodríguez, David M., Manuel P. Cuéllar, and Diego P. Morales. "Concept logic trees: enabling user interaction for transparent image classification and human-in-the-loop learning." Applied Intelligence (2024): 1-13.

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

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

[198] John, Blooma Mohan, et al. "Graph-based cluster analysis to identify similar questions: A design science approach." Journal of the Association for Information Systems 17.9 (2016): 2.

[199] Bhatnagar, Subhash. E-government: From vision to implementation-A practical guide with case studies. Sage, 2004.

[200] Hevner, Alan R., et al. "Design science in information systems research." MIS quarterly (2004): 75-105.