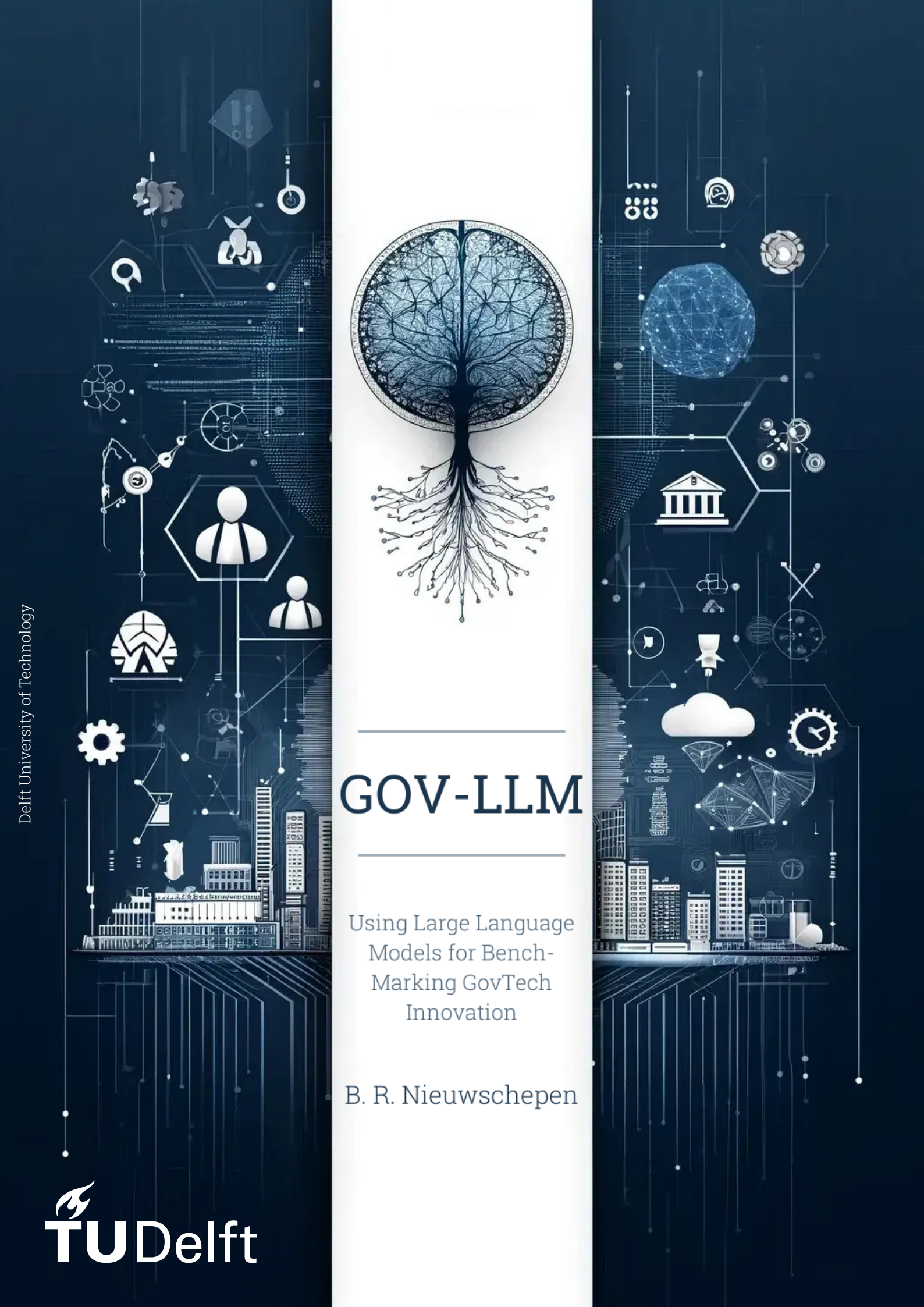

GOV-LLM

Using Large Language Models for Bench-Marking GovTech Innovation

B. R. Nieuwschepen

Delft University of Technology

TU Delft

GOV-LLM

Using Large Language Models for Bench-Marking GovTech Innovation

by

B.R. Nieuwschepen

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday June 27, 2024 at 15:00.

Student number:	4966104	
Project duration:	February, 2024 – June, 2024	
Thesis committee:	Prof. dr. ir. N. Bharosa,	TU Delft, supervisor
	Dr. J. M. Durán,	TU Delft

Cover: Created with the assistance of DALL·E

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Preface

Before you lies the master thesis ”GOV-LLM: Using Large Language Models for Bench-Marking Gov-Tech Innovation”. It is written to fulfil the graduation requirements of the Engineering and Policy Analysis programme at the Delft University of Technology.

From October 2023 to June 2024, I dedicated myself to this project and writing this thesis. Packed with an interest in Artificial Intelligence and my work experience at Digicampus, I embarked on a journey to explore the possibilities of AI as a solution to GovTech problems. The journey would take me through many challenges and interesting conversations, and now leaves me with much more knowledge about both AI and GovTech.

Since the beginning of my academic career, I knew that the end would eventually come. In the blink of an eye, the end is here now, and I barely notice myself walking through the door I once prayed would open.

I would like to thank Nitesh Bharosa for his expertise, passion, and energetic attitude towards this project. I also want to thank Juan Durán for his meticulous attention to detail, structured way of thinking, and willingness to offer detailed feedback on many aspects of the research, continuously challenging me. I also would like to express my gratitude to Corné Snoeij, who was a pleasure to work with on this project and has become an invaluable friend throughout the programme. I want to thank Kylie for always being there for me and for showing me that there is more to life than I ever realised. I am grateful to my parents for supporting me in every way possible, which allowed me to focus on my studies during my full 20-year educational journey. Lastly, I would like to express my gratitude to Jasper, whose unexpected appearances in my life over the past nine years have consistently brought positive energy.

I would also like to thank you, my reader: I hope you enjoy the reading.

B.R. Nieuwschepen
Delft, June 2024

Summary

Governments are increasingly dependent on GovTech, which is the technology that facilitates processes in the public sector. Benchmarking the state of GovTech is done by governments and yields indispensable insights, which are used for optimising resource utilisation, identifying areas for improvement, and facilitating evidence-based policy prioritisation. Current benchmarking efforts are resource-intensive, time-intensive, and have limited scope, resulting in an inefficient assessment of GovTech innovation.

This thesis explores the potential of Large Language Models (LLMs) to overcome the practical limitations of existing GovTech benchmarking methods, analysing the GovTech Maturity Index of The World Bank as a case study. Using an LLM and leveraging state-of-the-art techniques, including fine-tuning, Retrieval Augmented Generation, and Prompt Engineering, the usability of these models as an artefact for GovTech benchmarking is assessed.

The results show that the best-performing model outperforms the random chance accuracy, indicating that the LLM not only understands the question and data format, but also contains the information to correctly answer the benchmark questions. The research concludes that the created artefact has the potential to improve GovTech benchmarking, resulting in more informed policy decision making.

The thesis contributes valuable insights to the field of the GovTech research field by making GovTech benchmarking more efficient, leading to a better analysis of the current GovTech market, and contributing to the academic debate on GovTech market analysis. Furthermore, by streamlining the Dutch benchmarking process, resulting in more accurate insights, the study contributes to the advancement of GovTech solutions within the Netherlands, ultimately benefiting society as a whole.

The artefact created in this research has significant policy relevance, as more efficient benchmarking allows policymakers to better optimise resource allocation, identify key areas for investment, and improve evidence-based policy changes regarding GovTech.

The limitations of this research include model inaccuracies, challenges in handling long contexts, and the potential for incorrect answers. An ethical analysis is performed, from which it can be concluded that the relevance of the information used by the models is the most apparent ethical concern.

Future research may focus on improving the models to better handle long contexts, reducing inaccuracies, and improving the overall performance in populating GovTech benchmarks by incorporating additional data sources and improving the models. The research design is limited by the lack of environmental aspects in the process and the narrow scope of using a single benchmark for one country.

As next steps, this research proposes a roadmap that includes the continuation of the development of the artefact, the extension of ethical analysis, the performance of trials, and the beginning of a wider application of the artefact.

Contents

List of Figures

List of Tables

1

Introduction

Governments are alive. They are constantly moving, transforming, innovating, and adapting based on societal and technical challenges and advances. Technology has been at the centre of society for a long time, being recognised by Bain as early as 1937 as *"the most important single factor in producing, integrating and destroying cultural phenomena"*. Although the world has certainly changed a lot since Bain wrote this statement, it has stood the test of time: technology still greatly influences society and everyone and everything in it, arguably more than ever.

In today's era of technological progression (Alenezi, 2022), technology influences the way governments work on a fundamental level. This category of technology, called GovTech, is defined by Bharosa (2022) as *"socio-technical solutions that are developed and operated by private organisations, intertwined with public sector components for facilitating processes in the public sector"*. This definition accentuates the interdisciplinary nature of the matter; it involves both the public and the private sector, and it is resolved around a solution that is both social and technological. In addition, it emphasises that the implementation of technology is as important as the technological innovation itself. The goals of GovTech innovation include improving government services, transparency, efficiency, agility, and citizen participation (Amaglobeli et al., 2023; Silve & Moszoro, 2023), positioning GovTech as the reliable link between citizen and government. In fact, GovTech has the ability to drastically transform the interactions between citizens, businesses, and public agencies (Bharosa, 2022).

1.1. Research Problem and Objective

Recognising the importance of GovTech, governments seek information on the current state of GovTech and the solutions being developed (Dener et al., 2021; Desmond & Kotecha, 2017). To develop this overview, academics make both qualitative and quantitative efforts, which is further elaborated in section 2.1.

However, all of these methods require extensive human effort in data collection, survey response processing, and acquisition of the right sources. Consequently, these measurements are done only once or sporadically at most. In addition, current efforts are based on the knowledge and familiarity of the researchers about the topic. The amount of data and the effort required do not allow for a standardised periodic assessment, ultimately limiting the reproducibility of results and the knowledge on the current state of GovTech across countries. Furthermore, current efforts fail to grasp the complexity and diversity of GovTech.

At the same time, Artificial Intelligence (AI) is a promising technology that can handle both unstructured and a large amount of government data, as shown by Gao and Janssen (2020). Specifically, Large Language Models (LLM) are able to process unstructured data, while understanding the semantics of the data (Santos et al., 2021), greatly improving the ability of models to make human-level predictions and perform question answering tasks (Nassiri & Akhloufi, 2022). This technology has proven to be capable of giving context-aware responses to queries in other sectors, such as material science (Yang et al., 2024) and software development (Mathews et al., 2024). The details of this concept are

introduced in section 3.2.

This research introduces an innovative methodology to benchmark the current state of GovTech, enabling policymakers to make more informed decisions based on verifiable results. By accelerating the benchmarking process, the timeliness of the results is enhanced by facilitating the execution of more frequent benchmarks. In addition, this approach improves the integrity of the results by eliminating the potential for government data cherry picking. Consequently, this leads to a more efficient allocation of government resources, improved prioritisation of government investments, and creates a shared sense of mission among governments. This ultimately helps policymakers make more informed decisions and evaluate the effectiveness of previous policies.

The specific challenges of technologies within the government have been identified by academics, which results in the recent development of the GovTech research field. Specifically, Bharosa (2022) states the urgent need for academic research on GovTech development and knowledge dissemination to policy makers, addressing the demand for insights to be insightful and practically useful for governments. By improving the process of benchmarking GovTech, insights on the state of GovTech can be gathered more frequently, therefore, better informing policy makers. Svahn et al. (2023) identify the need for a GovTech market and development analysis, and to develop constructs that help to build a stronger theoretical base for GovTech businesses. Therefore, the scientific relevance of this research is related to the improvement of the GovTech benchmark process, thereby contributing to the academic discussion on GovTech market analysis and knowledge dissemination to policy makers.

The societal relevance of this research as a result of improving the benchmarking process is related to addressing the digital divide by increasing understanding of inequalities in GovTech adoption. In addition, this research contributes to the United Nations' Sustainable Development Goals by improving insight into government digital infrastructure. Finally, it influences the economy by facilitating more effective analysis and decision-making, which leads to cost savings. These impacts are explained further in section 6.4.

Therefore, the objective of this research is to address the gap in the GovTech benchmarking methodology, particularly regarding the timeliness and integrity of the benchmarking process. This is done using the novel approach of leveraging an open source Large Language Model to populate an existing GovTech benchmark, specifically using data from the Netherlands. To answer the framework questions, the model is fed additional context. The details of the approach are introduced in chapter 4.

1.2. Research Questions

Based on the research objective, the main research question can be derived.

> ***(RQ)*** *What is the usability of Large Language Models for benchmarking the state of GovTech?*

To answer the research questions, six sub-questions are identified. These questions are successive and all contribute to answering the main research and are structured following the design science framework, as elaborated in section 1.3.

The focus of the first sub-question is on the limitations of current implementations. The first sub-question is formulated as:

> ***(SQ-1)*** *What are the practical limitations of current methods for bench-marking GovTech?*

The focus of the second sub-question is on the design of the artefact. The second sub-question is formulated as:

> ***(SQ-2)*** *How do Large Language Models address the limitations of current methods for bench-marking GovTech?*

For the third, fourth, and fifth sub-questions, the focus is on the technical validation of the artefact, addressing qualities and limitations of the artefact and the improvements made. The third, fourth and fifth sub-questions are formulated as:

(SQ-3) *How accurate does a Large Language Model populate a GovTech bench-marking framework, compared to a manually populated framework?*

(SQ-4) *What is the impact on the accuracy of the Large Language Model when relevant information is given as context?*

(SQ-5) *What are the practical limitations for the Large Language Models on populating the GovTech framework?*

The sixth and last sub-question relates to the validation, addressing the practical and ethical impact of the artefact on the Dutch government.

(SQ-6) *What is the impact of using Large Language Models for GovTech framework population on the assessment of GovTech by the Dutch government?*

1.3. Research Framework

In this research, the usability of an LLM is evaluated to benchmark the state of GovTech in the Netherlands. This is done following the Design Science Research approach framework from Hevner et al. (2004). The resulting framework is shown in Figure 1.1. Design Science Research is chosen because it fits the requirement of creating an innovative artefact, while ensuring a relevant and rigour research design through consideration of environmental and knowledge base aspects. Rather than business needs, general environmental needs are considered.

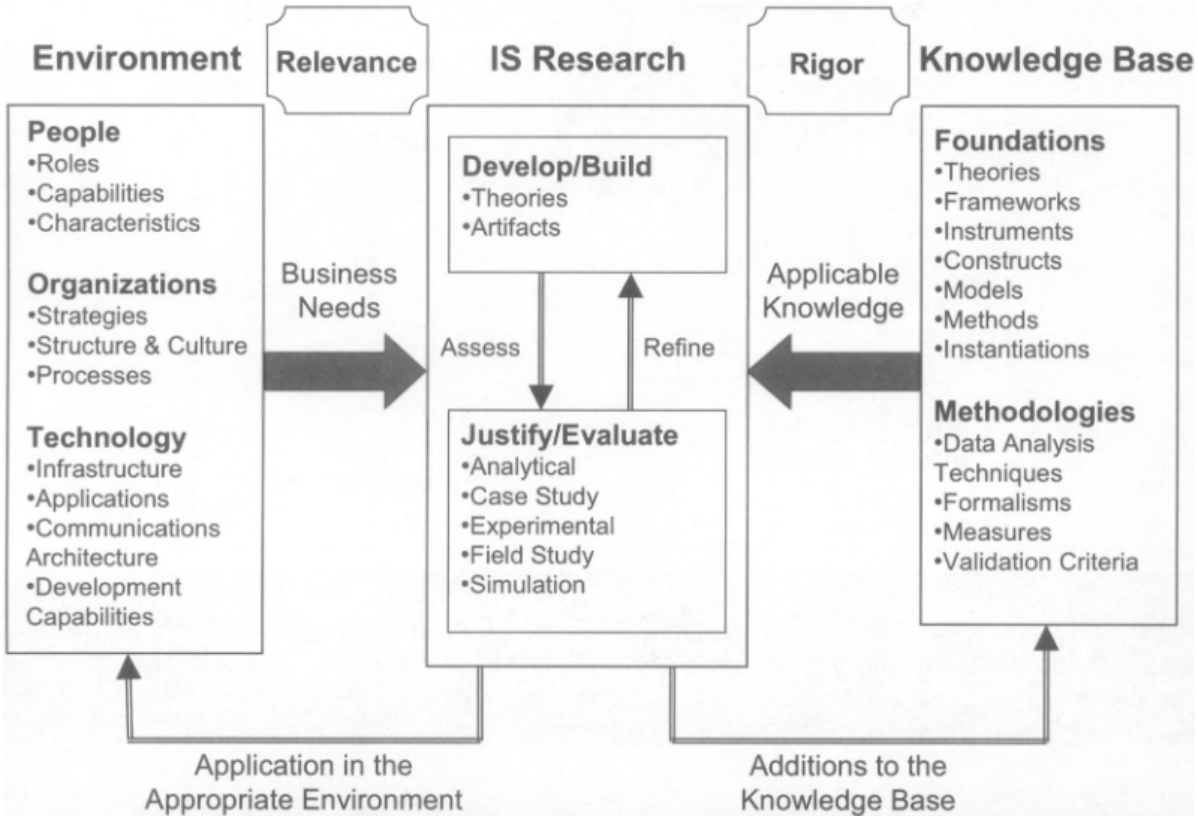

Figure 1.1: Design Science Research Framework (Hevner et al., 2004).

The approach consists of three pillars. The environment represents the problem space where the phenomenon of interest is located. The knowledge base provides the ”*raw materials*” for research. The research itself has a design/build phase and a justify/evaluate phase.

The pillars of Hevner et al. (2004) are translated into the Design Science Research Methodology by Peffers et al. (2007), as shown in Figure 1.2. This model is consistent with previous literature and provides a guide for conducting and evaluating Design Science research. The method consists of six

successive steps, which together encompass a validated model for presenting and evaluating Design Science research. These steps are used to structure this thesis, as shown in the outline.

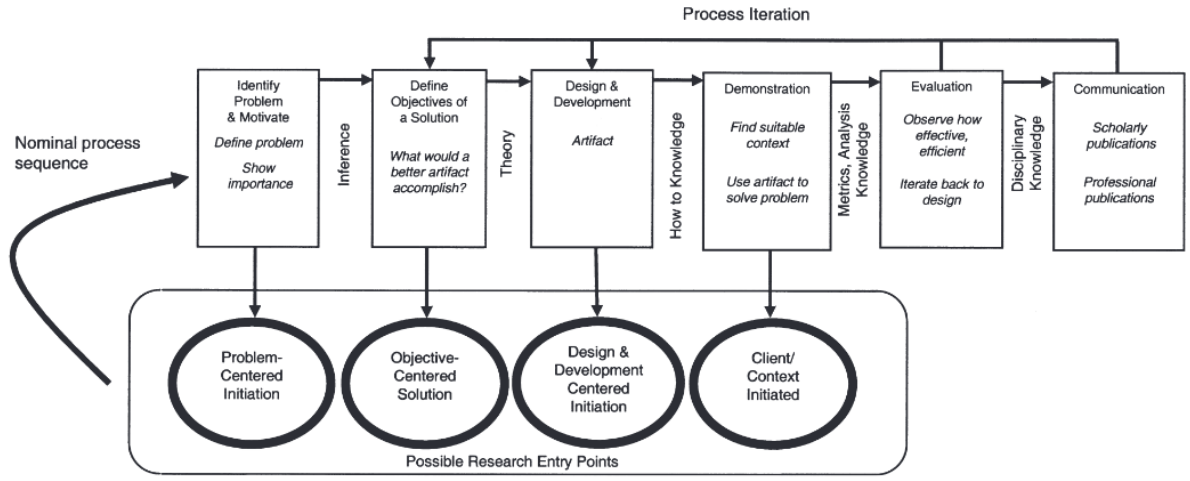

Figure 1.2: Design Science Research Methodology Process Model (Peffers et al., 2007).

1.4. Outline

An overview of the outline, based on the Design Science Research Methodology structure from Peffers et al. (2007), is shown in Figure 1.3.

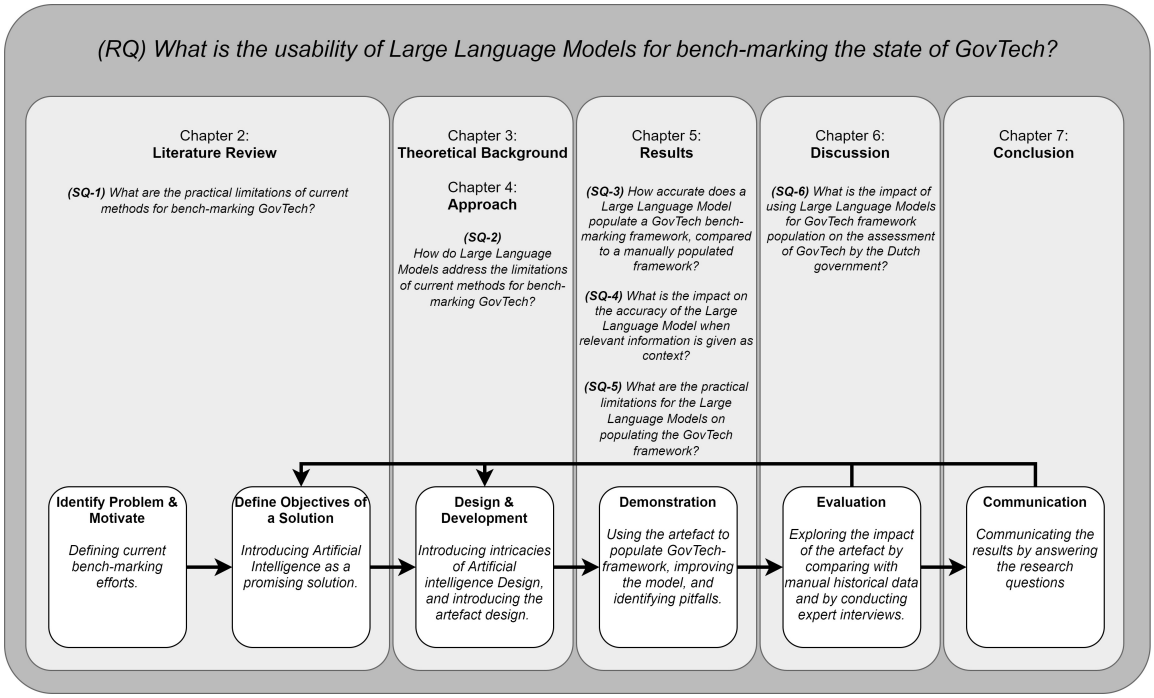

Figure 1.3: Thesis Outline.

chapter 2 addresses the first and second steps of the process. The first step, *Identify Problem & Motivate*, is performed by conducting a literature search related to current benchmarking efforts, from which limitations are identified. The second part of chapter 2 addresses the second step, *Define Objectives of a Solution*, by looking at current efforts regarding AI and GovTech, and translating the identified limitations into research objectives. The first two steps work towards answering **SQ-1** by introducing practical limitations.

chapter 3 and chapter 4 together address the third step of the process: *Design & Development*. This is done by first introducing the intricacies of GovTech innovation and AI Design, after which the artefact design and development is introduced based on the identified limitations from the previous steps. This step works toward answering **SQ-2** by introducing the Large Language Model design.

chapter 5 addresses the fourth step of the process: *Demonstration*. This is done by using the artefact designed in the previous step to populate a GovTech framework, improving the model, and identifying pitfalls. This step works towards answering sub-questions three to five. **SQ-3** is addressed by presenting the accuracy of the artefact, as compared the manually populated framework. **SQ-4** is addressed by presenting the results of the models with additional relevant information given to them. **SQ-5** is addressed by listing common limitations of the models.

chapter 6 addresses the fifth step of the process: *Evaluation*. This is done by exploring the practical and ethical impact of the artefact by comparing the results with the manual historical data, and by conducting an expert interview. This step works toward answering **SQ-6** by exploring the practical and ethical impact of the model on the process of GovTech benchmarking by the Dutch government.

chapter 7 addresses the sixth and last step of the process: *Communication*. This is done by communicating the results by answering the research questions and proposing next steps.

2

Literature Review

This chapter is dedicated to exploring the current literature on GovTech benchmarking efforts and AI and GovTech efforts. The flow diagram of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) structure from Moher et al. (2010) is used to systematically select relevant articles for each literature search.

2.1. Efforts on Bench-marking GovTech

To obtain information on the current state of GovTech, benchmarking efforts are made. Svahn et al. (2023) address the importance of constructs that cut across the research area and indicate the demand for quantitative GovTech research to better understand GovTech. From this, a literature research is conducted. Figure 2.1 show the search strategy. Scopus is used as a multidisciplinary publisher database and yields sufficient results to address both the governmental and methodological aspects of the topic. Irrelevant results are excluded from the research. Irrelevancy is determined on the basis of whether the paper, and more specifically the results, addresses a specific method or technique of measuring government innovation. These results show that the search strategy and query are sufficiently specific to yield enough relevant results.

For this analysis, the focus is on the methods used in these references. In the identified literature, two main methods of measuring innovation are distinguished: qualitative techniques and quantitative techniques. An overview of the methods presented is shown in Table 2.1.

2.1.1. Qualitative Methods

In the field of assessing innovation performance and the impact of government influence, researchers have used various qualitative methods. For example, Murati-Leka and Fetai (2023) and Moghavvemi and Mohd Salleh (2014) use questionnaires to investigate how companies innovate and how external factors, such as government policies, influence this innovation. Their data analysis reveals the relationships between different variables, analysing entrepreneurial behaviour. The strength of this method lies in its ability to provide detailed insight into the levels of innovation. However, these studies also acknowledge certain limitations, particularly concerning external events and other actors involved in the innovation ecosystem. Murati-Leka and Fetai (2023) suggest that future research should broaden its scope to include a comprehensive analysis of all actors as a set within the innovation ecosystem to better understand GovTech innovation. Despite the detailed results, treating the government merely as an external factor limits the ability to draw conclusions about GovTech innovation.

On the other hand, M. Zhang et al. (2023) use interviews to identify the key attributes of value essential to technology development. This approach results in detailed conclusions about the impact of technology on government services, in this case, in the health sector. However, reliance on participants' knowledge means that the results are not easily reproducible and may quickly become outdated. Additionally, the study highlights the issue of data selection, noting the under-representation of underdeveloped regions and policymakers, which hinders the generalisability of the findings.

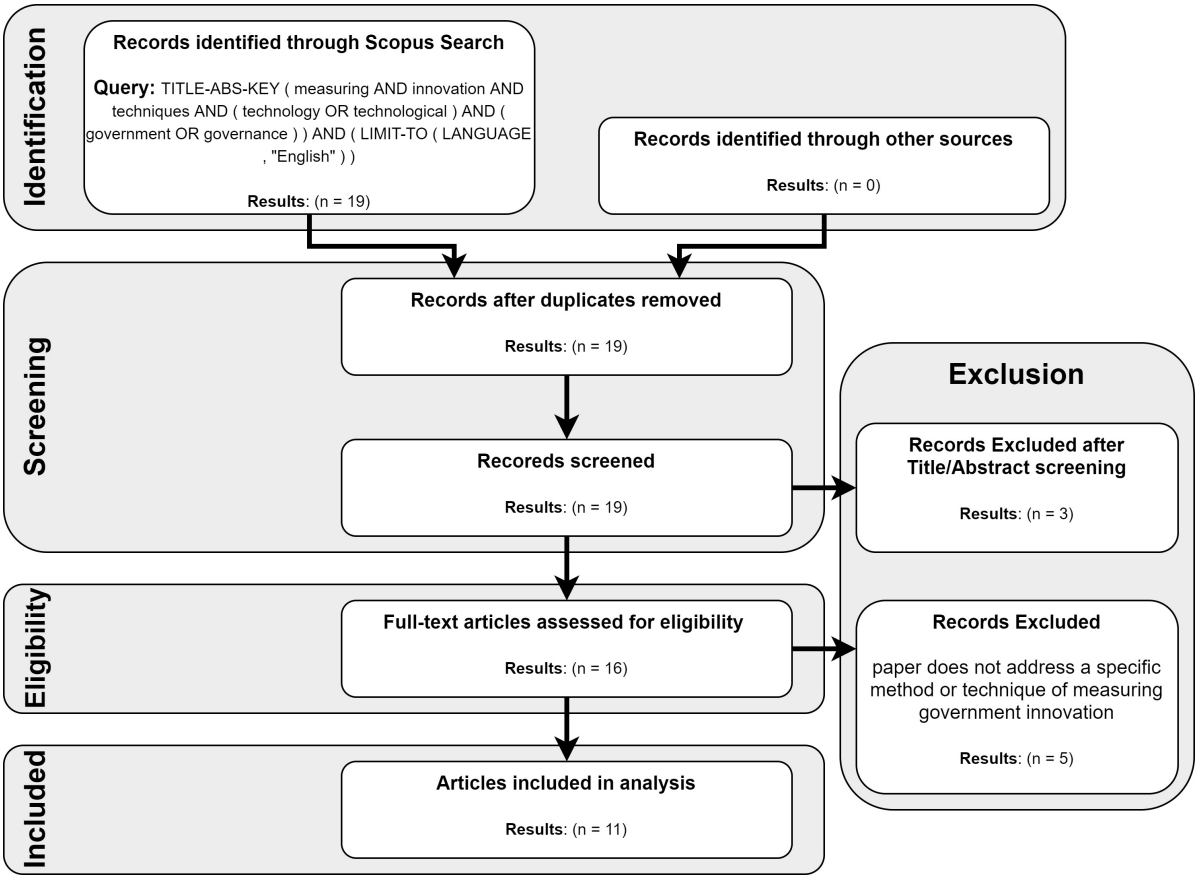

Figure 2.1: PRISMA flow diagram for GovTech benchmarking search.

Additionally, Chun et al. (2021) evaluates the quality of results derived from large-scale surveys and advocates for the integration of these data with other sources, such as administrative data, to enhance the robustness of the findings. Similarly, Alhyari et al. (2013) also rely on survey questionnaires to gather data on e-government performance. They emphasise the importance of combining qualitative data with quantitative data to achieve more comprehensive results.

In short, surveys, interviews, and questionnaires are the qualitative methods used. However, the limited scope of the results prevents these researchers from making general conclusions about GovTech innovation measurements. Besides, the qualitative data gathering methods identified in literature rely on the ability to recollect data of the actors involved in the analysis. This may lead to issues with data integrity and cherry-picking of data. Additionally, the time-intensiveness of the processes used limits reproducability. The results are not compared over time and across governments, limiting the impact of these measurements. The use of additional data sources is advocated but not yet properly introduced. An objective of this research is to improve the timeliness and data integrity of the benchmarking process.

2.1.2. Quantitative Methods

Quantitative methods have been employed by various researchers to prioritise investment projects and measure technological progress. For example, Shim and Kim (2023) used existing data and a benchmarking framework to prioritise investment projects for the Korean government. Although the time-series data used were specific enough to determine the potential of particular projects, such detailed data on GovTech innovation, or even more specific subcategories as, for example, government cloud platforms, are not available.

Similarly, Saeed et al. (2023) applied advanced data analysis to measure technological progress, economic growth, and the efficiency of public expenditures on research and development activities. This

method effectively mitigates the limitations of qualitative methods, such as repeatability and comparability. However, the metrics used, namely the World Bank Development Indicators, are highly aggregated and only updated periodically. Consequently, this method suffers from the same limitations as other quantitative methods, being too broad to draw specific conclusions about GovTech innovation.

In another study, Wu and Guo (2015) used government websites to score e-government performance, using these websites as indicators of government efficiency. This method allows for comparisons over time and place, as the analysis of websites is standardised and can be repeated over time by re-scraping websites. However, providing a single aggregated 'E-government efficiency' score fails to capture the complexity of E-government and GovTech and gives only limited information to policy makers.

Furthermore, Alzahrani et al. (2012) used statistical methods to determine the acceptance of e governance by citizens. These quantitative methods focus on the private sector or the performance of specific public actors. Although using large-scale citizen data allows for detailed analysis, such detailed GovTech data is not readily available.

Additionally, Murati-Leka and Fetai (2023) used publicly available government documents to assess the attributes of the value of a technology. For this analysis, sample company data was used, a research method not suitable for GovTech analysis, as one cannot draw conclusions on the state of GovTech as a whole based on a sample of the ecosystem. The heterogeneous nature of the GovTech ecosystem, involving both processes and actors (Hoekstra et al., 2023), is not captured in company data.

In summary, large datasets, government websites, and company documents are used to assess government technology. However, these methods are not suitable for benchmarking the state of GovTech, as there is no single dataset to analyse, or the level of aggregation is too high to properly draw conclusions.

Table 2.1: Current Research Methods GovTech.

Category	Method	Reference
Qualitative	Questionnaire	Alhyari et al. (2013), Moghavvemi and Mohd Salleh (2014), and Murati-Leka and Fetai (2023)
	Interviews	M. Zhang et al. (2023)
	Survey Evaluation	Chun et al. (2021)
Quantitative	Investment Data	Shim and Kim (2023)
	Company Documents Assessment	Murati-Leka and Fetai (2023)
	Analysis of Government Websites	Wu and Guo (2015)
	Citizen Acceptance Statistics	Alzahrani et al. (2012)
	Technological and Economic Data Analysis	Saeed et al. (2023)

These methods, although fit for the specific purpose, all fall short of creating a full overview of GovTech. All methods only focus on a specific region, time frame, data source, or metric. This is caused in part by the data used in the research. To gather data that are specific enough to make proper conclusions, while being of broad enough scope to make these proper conclusions over, e.g. the public sector of a country as a whole, is hard. This challenge can be mitigated by using aggregated data at the country level, as done by Angelis et al. (2014), but this aggregation results in loss of information, resulting in general conclusions.

However, using government website information, as done by (Wu & Guo, 2015), is a promising technique, mitigating the need for a specific time series dataset while getting up-to-date detailed information on the government. Furthermore, the use of a benchmarking framework, as done by (Shim & Kim, 2023), allows a standardised approach to benchmark the state of GovTech, resulting in reproducible and comparable results. This research aims to use both these aspects to benchmark the state of GovTech in the Netherlands.

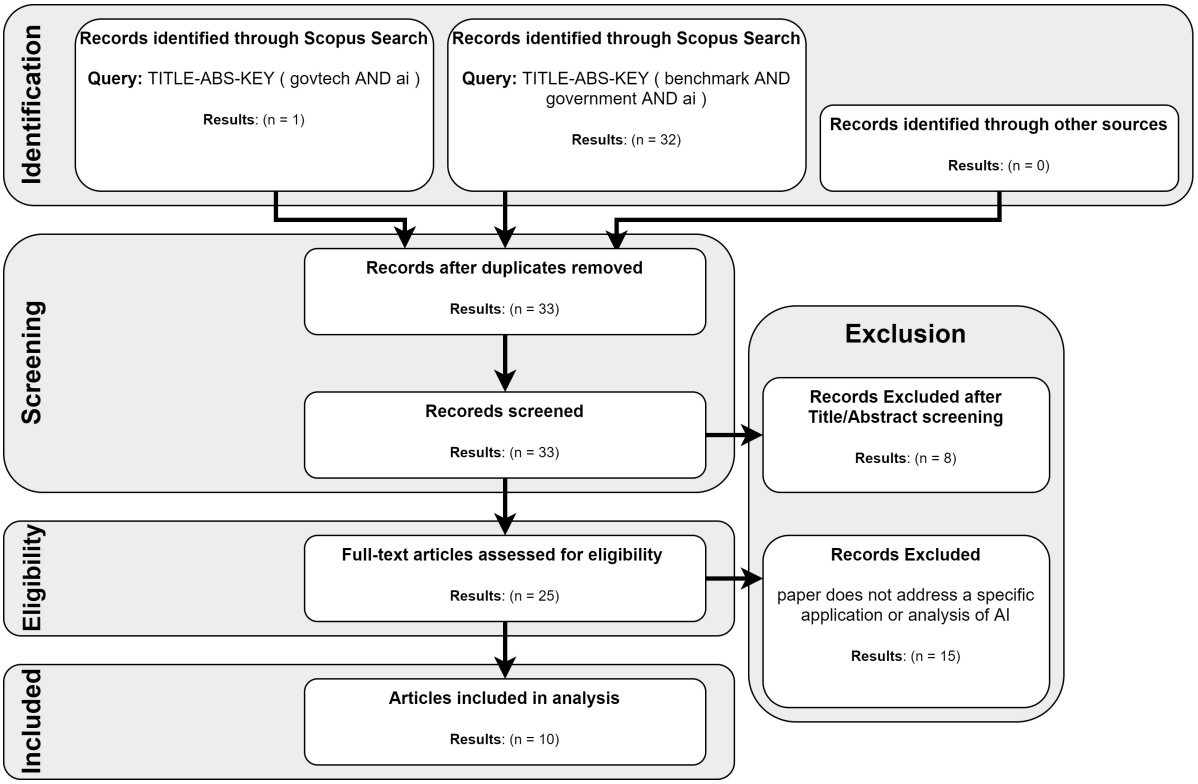

Figure 2.2: PRISMA flow diagram for GovTech AI search.

2.2. Current Efforts on AI and GovTech

Having identified the current measurement methods for GovTech, the focus can shift to the current efforts to combine AI and GovTech. This section explores the ways AI is currently being implemented in GovTech, as identified in the literature. From this, a literature research is conducted. Figure 2.2 shows the search strategy. Scopus is used as a multidisciplinary publisher database.

The literature found is divided into two categories: AI Applications within the government, and Analysis of usage of AI.

2.2.1. AI Applications Within the Government

Artificial Intelligence is increasingly being integrated into the public sector. According to Engin and Treleaven (2019), AI algorithms are being used in applications such as public opinion analysis, policy formulation, and fraud detection. These technologies, including Natural Language Understanding, demonstrate the broad applicability of AI to address complex government tasks and the interest of governments in using AI.

Furthermore, Ahmed et al. (2023) highlights the use of chatbots within government agencies. These AI-driven chatbots simulate public sector employees in conversations, functioning as labour-saving technologies. Through the automation of standard interactions, chatbots free up human resources for more critical tasks, enhancing overall productivity.

Within the field of security, Lamptey et al. (2023) employs explainable AI (XAI) techniques within a terrorism database to identify trends in terrorism. The use of XAI is particularly noteworthy as it provides more interpretable models, which are less complex and easier to understand. This simplification of models ensures that the outcomes are more transparent and actionable, thereby aiding in effective decision-making.

Recognising the multilingual nature of government data, Alothman and Sait (2022) addresses the methodological gaps in the management and retrieval of bilingual management documents. This research contributes to the development of more robust methods for handling diverse linguistic data, thus

improving data management practices within government agencies.

Collectively, these applications of AI within the government not only improve existing processes but also pave the way for the successful integration of AI technologies into various government functions. The continuous evolution and adoption of AI in the public sector underscore its potential to transform how governments operate and serve their citizens. An objective of this research is to further test the potential of using AI within the government.

2.2.2. Analysis of Usage of AI

The implementation of AI benchmarks has been the focus of both governments and academia. Recent papers have focused on evaluating the assessment of AI applications. Steingard et al. (2023) investigates the moral impact of academic journals with AI, using the Sustainable Development Goals of the United Nations as a framework. The research measures performance by comparing an established ground truth impact with model predictions. This approach shows the ability of AI to interpret large sets of documents, and draw conclusions on them. The standardised formats of academic journals make it easier for the models' to assess moral impacts.

In another research, Mazzi (2023) addresses the themes of accountability and social impact by proposing a robust framework for assessment. The paper advocates for the regulation and integration of AI in policy making, highlighting the social responsibilities related to the use of AI while aligning with the Sustainable Development Goals. Specifically, the paper advocates for the integration of corporate social responsibility into AI in businesses with two recommendations.

First, a mindset framework is introduced that consists of three levels: new AI, applied AI, and potential AI. Following Mazzi (2023), *"New AI" concerns the process of designing and training the AI and the reason why it is employed by the business. "Applied AI" concerns what AI does, i.e. what is the function that it performs and what is the impact of its application. "Potential AI" concerns what AI can do, as of what it can do in other geographical areas, in other sectors, for other purpose"*. Although all three categories together ensure a complete analysis of the usage of AI, the focus of the analysis for this research is on *"Applied AI"*, as the focus of the research is on the implementation of the artefact in the benchmarking process. To achieve this, ethical concerns for algorithms are systematically considered, elaborated further in section 4.5.

Second, AI regulation and policy harmonisation are recommended, focusing on what leaders need to facilitate in the process of AI regulation. Specifically, the research proposes the necessity of public debate, communicating the importance of a forum for dialogue. Although this is an important step in the introduction of AI, this is outside the scope of this research, yet it is proposed as a next step.

On the other hand, Jia and Zhang (2022) analyses existing AI guidelines and concludes that current AI risks are underestimated by stakeholders and that current guidelines are unconventional. Specifically, the paper states that ethical guidelines *"may be used as a disguise to either render a social problem technical or discourage the efforts of imposing real regulatory burdens"*. This suggests that while the ethical guidelines are meant to be a step to a safer use of AI, the reality is that the guidelines possibly have the opposite results: creating a false sense of security due to the technical safety measures, while the actual ethical vulnerability remains. Furthermore, the paper identifies that the ethical guidelines fail to change the behaviour of tech professionals, suggesting that the guidelines do not achieve their intended purpose. In this research, the focus is also on technical safety measures to address ethical risks, using a framework to focus on transparency and explainability (as explained in subsection 4.3.3). The vulnerability regarding a false sense of security is acknowledged by performing an exploratory ethical analysis in section 5.5. Proposed as a next step is to perform a thorough ethical analysis and to address the identified ethical concerns.

These limitations imply that AI risks are taken into account, but that current guidelines are not sufficient to provide sufficient guidance. The paper proposes three suggestions. First, the focus of policy makers should be on collective welfare, rather than assuming only individual rights. For this research, this point is relevant, as the results of the artefact influence society as a whole, rather than individuals. Second, all actors involved should be educated on AI risks, in order to form realistic expectations. Third, governments should work on AI ethics guidelines. These last two points of action fall outside the scope of this research, as the actual implementation of the artefact into the policy-making process is

outside the scope of this research, but they are proposed as the next step.

Furthermore, Hadi et al. (2021) concludes that many governments benefit from the implementation of technology. Regarding AI, the paper analyses the pros and cons, concluding that AI allows public officials to return to the core business and allows greater satisfaction of users of public services. However, treads include being overwhelmed by the technology, resulting in loss of confidence and high cost.

In general, current research on AI benchmarks analyses current guidelines and the potential benefits of using AI. This implies that benchmarks are used to assess AI, but that AI is hardly used to populate a benchmark. Steingard et al. (2023) do use AI to benchmark the moral impact of academic journals. In this research, the use of AI for benchmarking is further explored using diverse open government data, rather than uniform academic journals. The ethical guidelines and limitations proposed are taken into account, or proposed as next research steps.

3

Theoretical Background

modelesearch methodology builds on core concepts that have yet to be properly defined. In order to grasp the importance of benchmarking GovTech, the GovTech innovation phenomenon and the main challenges identified in the literature are introduced. Next, the aspects of AI used in this research are introduced.

3.1. Conceptualising GovTech Innovation

Technology interconnects society, citizens and governance (GovTechNL, n.d.). GovTech relates to this technology, but focuses on the implementation within the government, with the goal of improving public services (Government of the Netherlands, n.d.). Examples of these services include proactive services (e.g. automatically receiving subsidies when you are entitled to) and easy identification without sharing unnecessary information (e.g. DigiD, with which you show who you are when you arrange your affairs online (Logius, n.d.)). Proper introduction of technologies could significantly increase public value by making interaction with the public sector more easy and intuitive.

To understand the intricacies of the GovTech ecosystem, the concept of GovTech innovation is introduced by looking at definitions and challenges, as identified in the current literature. Figure 3.1 shows the search strategy for this section. The main strategy involves the keywords GovTech and innovation. The specific focus on GovTech is chosen instead of terms like e-government or e-governance, as these focus more on transferring existing processes to the digital realm, instead of using new technologies to solve problems for the government. Scopus is chosen, as it is a multidisciplinary publisher database, which fits the multidisciplinary characteristics of GovTech. In addition, IEEE is chosen as secondary publisher database, to focus on the technical aspects of the problem. Scholar is added as an aggregated database, as to include a broader scope of paper sources, which allows for a narrower search strategy. For this aggregated database, the focus was put on socio-technical challenges, decentralised government, and Dutch policy of the government. All search query results are carefully assessed and irrelevant results are excluded from the research. Irrelevancy is determined based on whether the paper, and more specifically the results, mention a definition of GovTech, or discuss the effectiveness or challenges of GovTech.

3.1.1. Defining GovTech

Tantawy (2022) relates GovTech to a multi-dimensional concept, and suggests that the focus lies on the usage of technological solutions to improve government operations (e.g. proactive services), rather than the technology itself. Hoekstra et al. (2023) focus on the human-centred and data-driven aspects of GovTech, emphasising the importance of public-private collaborations and the sharing of knowledge and resources on emerging technologies between organisations within and outside the public sector. Bharosa (2022) adds to this by stating that GovTech refers to socio-technical solutions, which are developed and operated by private organisations, and are integrated into public sector components, ultimately serving the goal of facilitating processes in the public sector. Edelmann et al. (2023) emphasises the transactions and interactions with external stakeholders, as well as the development of

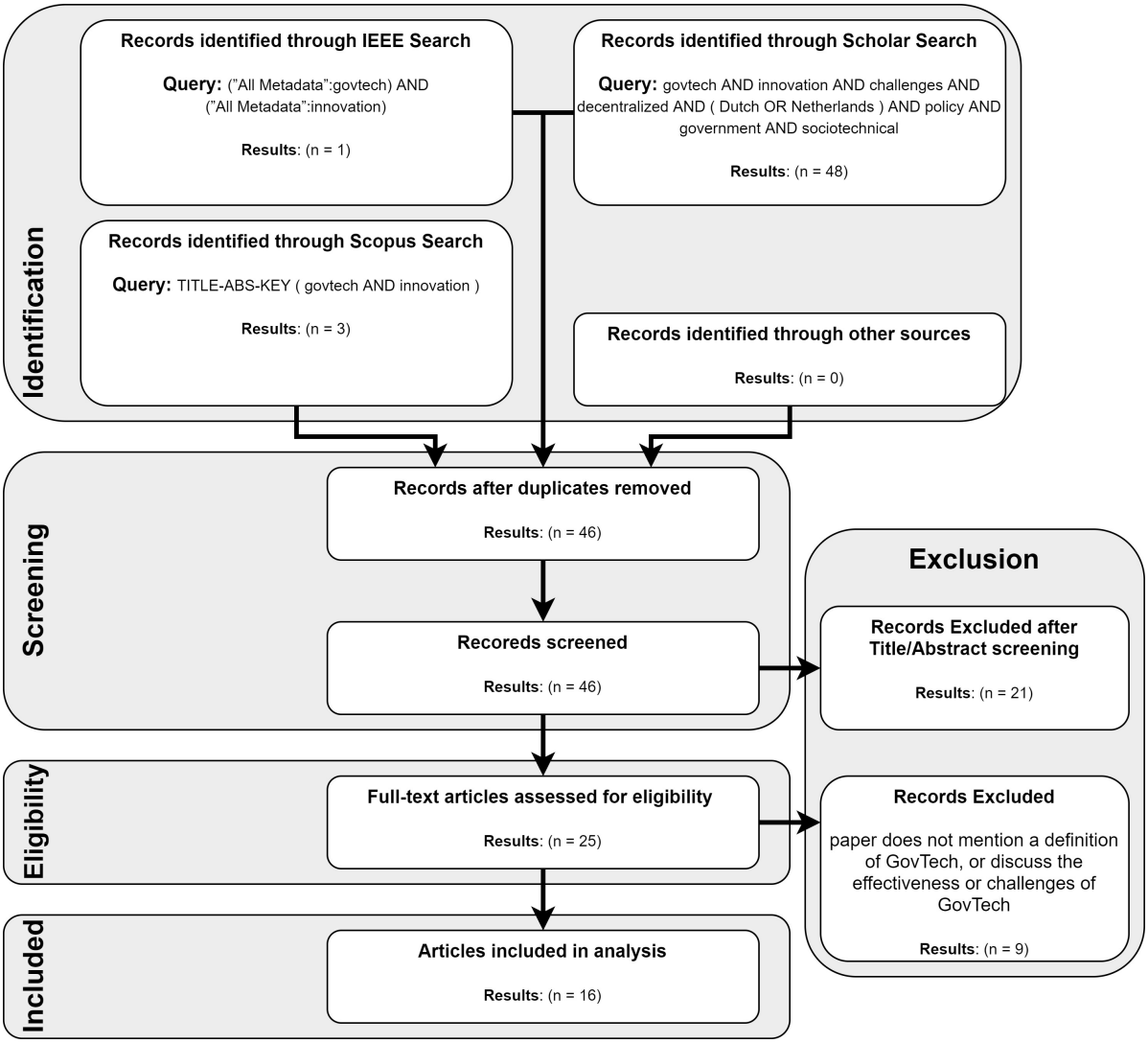

Figure 3.1: PRISMA flow diagram for GovTech innovation search.

information-based services.

In short, GovTech enables more efficient and effective public services. Technology is merely used as a means to an end and the focus is on integration into society. Public-private collaboration adds to the complexity, requiring transactions and interactions.

3.1.2. Challenges of GovTech Innovation

Now that the concept of GovTech is made clear, the main challenges found in literature can be identified. Svahn et al. (2023) indicate the challenge of making GovTech problems less generic, indicating that current examples of GovTech solutions revolve mostly around general technological solutions (e.g. proactive public services), and less around concrete examples of GovTech solutions as perceived within a specific domain (e.g. automatically receiving child subsidies when you are entitled to). Barcevicius et al. (2019) nuance this by stating that GovTech barriers are complex and often not technology related. Bharosa (2022) identifies a challenge for governments struggling in their digital transformation journey, stating that implementing GovTech solutions would in that case only complicate the situation. The paper also identifies a possible risk: the public sector in which the GovTech solution is implemented, does not hold the technical knowledge needed for the solution, making them dependant on for-profit organisations. Edelmann et al. (2023) look at GovTech innovation as perceived by employees of the public sector, and indicates challenges regarding digital signatures and the regional and national legal

framework. The paper also states that the whole organisational culture must be supportive of GovTech innovation for the innovation to work. H. Jiang (2021) adds the concept of demand-driven government modes, focusing on the demand pull of the government, rather than the technology push of GovTech startups.

The main challenge that is identified in literature, however, is that of cross-sector collaboration and co-creation. Cross-sector collaboration relates to the collaboration of actors across various sectors, leveraging the strengths of each actor as a method of achieving accelerated progression. Co-creation relates to the creation of a product or service using cross-sector collaboration. Rather than focusing on the activity of collaborating, co-creation focuses on the creation of the product or service. These two aspects of innovation go hand in hand, and are quintessential for GovTech innovation. Within this problem, several aspects are accentuated. Castle (n.d.) indicates the problem of the GovTech landscape being fragmented and heterogeneous, resulting in a lack in oversight of technologies and initiatives. Koryzis et al. (2021) adds to this by stating that information and knowledge sharing are perceived as vital to the success of GovTech innovation, and advocates for interdisciplinary research, and cross-sector collaboration. The paper also states that there is currently limited knowledge on how GovTech innovation trends will be used for specific tools, products and services. This indicates the lack of oversight within the GovTech ecosystem. This is confirmed by Manny (2022), as it states that there is information missing on exchange relations between social actors. This indicates that public actors do not communicate with each other on the GovTech problems and solutions. Komatsu (2019) finds that there is, however, a growing preference for co-producing of public value, and this is confirmed by Nweke (2023) on an international level. This statement appears to conflict with the current situation as described by Manny (2022), and indicates that, while there is demand for co-creation, it is not yet being achieved to as sufficient level. Lukkien et al. (2023) relate this problem to a lack of boundary resources (e.g. shared data standards), making collaboration harder to achieve. González Vázquez et al. (2022) state that all regional stakeholders, including citizens, enterprises, knowledge institutions and local authorities have to be involved into co-creation in a renewed partnership. This is hard to achieve, as it involves commitment of a lot of actors. Kong et al. (2024) identifies that, at least for the quantum science GovTech field, organisational and policy aspects of a transition to GovTech are interconnected, and current solutions are scattered. Initiatives, such as those described by van Winden and Carvalho (2019), do find knowledge creation between public departments and startups, and indicates that this is a desirable and fruitful concept. The paper indicates that the public sector should abandon their traditional approach of setting exact specifications on cooperation agreements, as to allow for co-creation of value. Tantawy (2022) points out the importance of aliment with global standards, hinting at the overlap in challenges that governments globally face.

In short, barriers in GovTech innovation are often not technology-related. Issues include knowledge gaps within the public sector, legal frameworks and culture of the organisation. The biggest issue is that of cross-sector collaboration and co-creation, which is needed for an integral GovTech ecosystem. The demand and preference are there, but in reality, a lack of oversight on the current state of GovTech limits possibilities.

3.2. Artificial Intelligence

Now that the concept of GovTech is sufficiently introduced, focus can shift to the intricacies of the method used, specifically the usage of Artificial Intelligence (AI). Specifically, the four aspects of AI used in this research are discussed: LLMs, fine-tuning, RAG, and Prompt Engineering.

3.2.1. Large Language Model

Large Language Models (LLMs) are seen as a major breakthrough in the Natural Language Processing (NLP) field, a field focused on bridging the gap between AI and linguistics (Nadkarni et al., 2011). This model type leverages the transformer architecture, as proposed by Vaswani et al. (2017), which allows for complex pattern recognition. LLMs understand language and learn information and relational knowledge from large amounts of data (Petroni et al., 2019).

This technology has proven to be able to give context-aware responses to queries in other sectors, such as material science (Yang et al., 2024) and software development (Mathews et al., 2024). Some efforts are made to explore the usage of Large Language Models in the public sector, such as in public

health (Arora & Arora, 2023; Jo et al., 2023) and political science (Linegar et al., 2023).

The main criticism on the usage of LLMs is that they are considered 'black box algorithms' (Von Eschenbach, 2021), raising concerns for being opaque and lacking trustworthiness (Durán & Jongsma, 2021). This is problematic, as these factors influence the relationship between citizens and governments (Janssen et al., 2021). However, Durán and Jongsma (2021) state that this is not a problem per se, and that reliability of the algorithms used provide reason to trust LLM outcomes. Furthermore, Harrison and Luna-Reyes (2022) state the importance of high-quality data and data security and privacy for achieving trustworthy AI. The ethical considerations are explored further in section 5.5.

3.2.2. Fine-Tuning

By fine-tuning a model, (part of) the model parameters of an LLM are readjusted to better represent the data on which the model is fine-tuned. This allows for increased performance on specific or niche tasks, such as the ability to better comprehend a specific programming language (van Dam et al., 2024) or improve the approval rate of model outcomes (Bakker et al., 2022).

Fine-tuning is done by feeding question-answer pairs to the model, and adjusting the parameters based on the outcome compared to the true outcome. In this way, the model outcome becomes more aligned with the true outcomes, ultimately changing the behaviour of the model. The fine-tuning approach taken in this research is explained in subsection 4.2.2.

3.2.3. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) was first introduced by Lewis et al. (2021) as a way to get more specific, diverse and factual language responses from Large Language Models by retrieving additional relevant information from external databases as context. This method proves to outperform models without the added context, and presents a straightforward way for improving model performance. Furthermore, this approach increases trustworthiness of the model by having the possibility of increasing the data-quality of the model input (Wagle et al., 2023). The general approach of RAG is explained further in chapter 4.

3.2.4. Prompt Engineering

Another way of improving trustworthiness of LLMs is with the utilisation of prompt engineering (K. Huang et al., 2024). Citing B. Chen et al. (2023), *"Prompt engineering is the process of structuring input text for LLMs and is a technique integral to optimising the efficacy of LLMs"*. Techniques identified by B. Chen et al. (2023) include specifying the role of the LLM (e.g. an expert in AI), separating parts of the prompt, and giving the LLM multiple tries to answer a prompt. These techniques form the layer of interaction between the model and the real world, and is therefore essential to get right.

4

Approach

Knowing the current limitations of benchmarking GovTech, and the possibilities of Artificial Intelligence, these aspects can now be combined into the approach of this research. This chapter elaborates on the requirements and the approach taken, laying the groundwork for the experiments done.

4.1. Requirements

From the limitations of current research described in chapter 2, requirements for this research can be drawn.

In section 2.1, the limitation of scope is identified. All methods described focus only on a specific region, time-frame, data source, or metric. Therefore, for this research, the requirements are that the artefact can be used to compare regions and time-frames, and that multiple data sources and metrics are used. The use of a benchmarking framework, as done by (Shim & Kim, 2023), allows for a standardised approach to benchmark the state of GovTech, resulting in reproducible and comparable results. Therefore, a GovTech benchmark is used. Specifically, the GovTech Maturity Index from Dener et al. (2021) is used, as shown in Appendix A.

Furthermore, the results of the expert interview, elaborated in section 5.4 and shown in Appendix C, identify the limitation of timeliness of the process. Therefore, for this research, the additional requirement is that the time it takes to perform the analysis is short, as compared to the manual population of the framework.

4.2. Artefact Design

The artefact design can be split into three parts: Database Creation, Prompting, and Evaluation. A full overview of the approach is shown in Figure 4.1.

4.2.1. Database Creation

For this research, a selection of secondary data sources is used to give additional context to the LLM.

Data Scraping

The data used in this research is scraped from internet sources. Relevant data sources are chosen based on predetermined conditions. These conditions ensure data availability, relevancy, and usability. The following conditions are used:

- Source must be available to download;
- Source must be up-to-date: updated at 01-01-2023 or later;
- Source content must be related to the public sector: containing either data on, or of, the Dutch government;
- Source must at least contain 100 data entries.

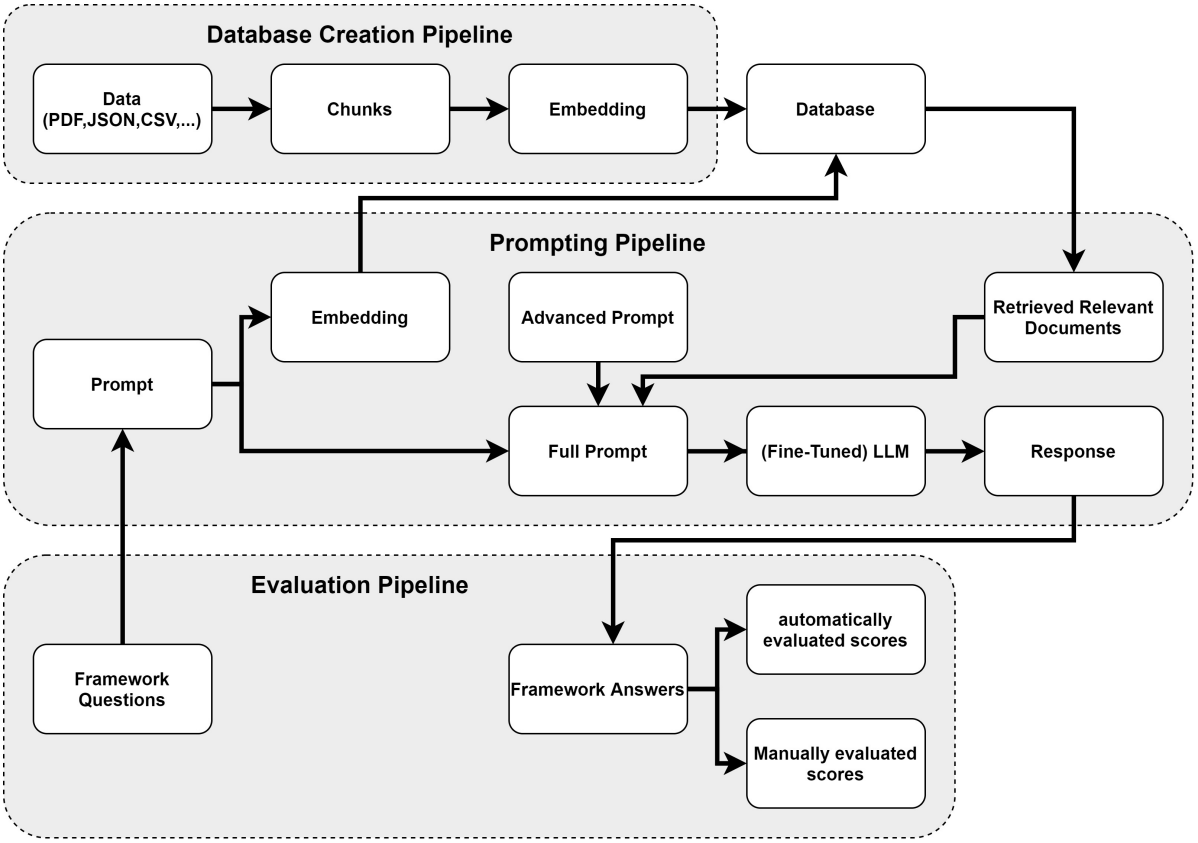

Figure 4.1: Object Model of Research Design.

- Source must be from a reputable source: being reviewed before published.

An overview of the selected sources is shown in Table 4.1.

To keep the data in the database up-to-date, the sources are scraped periodically, and new entries are added to the database, while duplicate entries are not re-added. This ensures an efficient updating process, saving the time of adding duplicates to the database.

Data Splitting

The sentence-transformers model used to create the embeddings allows for a maximum input of length 128 characters. To still be able to give more context to the LLM, a parent-document retriever is used. This method first splits the data into parent chunks of length 1000, and then splits these chunks into child-chunks of size 128. The child chunks are turned into embeddings, and when the embedding is

Table 4.1: Data Sources.

Source	Description
TenderNed (n.d.)	The Dutch government's online tendering system.
Binnenlands Bestuur (n.d.)	A magazine that focuses on higher educated civil servants and administrators with news, backgrounds, opinions and vacancies.
iBestuur (n.d.)	A platform for administrators, decision makers and policy makers in the public sector on the theme of digitisation of government and society.
Rijksoverheid (n.d.-a)	Documents such as decisions, speeches and parliamentary documents appear every day. These documents are available via open data.

selected as being relevant context, the parent chunk is returned as context for the LLM.

Data entries are split using a recursive character splitter. This splitter splits into the following characters: `[".", "!", "?", "\n"]`. The splitter works from left to right and stops when the part is of sufficient size (<= 128 characters in this case). Splitting has the added benefit of making the embeddings more specific, as each embedding has to represent less information.

Data embedding

The splitted data is then transformed into an embedding using a sentence-transformer model. This type of model derives semantically meaningful sentence embeddings from text strings (Reimers & Gurevych, 2019). Specifically, the `paraphrase-multilingual-MiniLM-L12-v2`[1] model is chosen, which is a 118M parameter model based on the BERT architecture (Devlin et al., 2019). This version is fine-tuned on 50+ languages, including Dutch. The model returns a 384-dimension vector embedding.

The creation of the embeddings is performed on a personal computer (laptop) equipped with an NVIDIA Quadro P1000 GPU.

Adding to Database

The data and embeddings are added to a vector database. This type of database system is optimised for the fast retrieval of semantically similar entries using embeddings, based on an input query (Hillebrand et al., 2023). Specifically, the ChromaDB database is chosen. This is an open-source vector database that allows for easy implementation and state-of-the-art functionality (Pan et al., 2023).

4.2.2. Fine-tuning

This research fine-tunes the base model used (which is elaborated on further in subsection 4.2.3) on a dataset of question-answer pairs about the Dutch government (Rijksoverheid, n.d.-b), in an attempt to better represent the characteristics of government documents. By fine-tuning, the model gains both knowledge of the Dutch government as well as the way of writing.

To fine-tune the model, the LoRA technique is used, freezing the model weights and adding additional weights to change the model behaviour (Hu et al., 2021).

Fine-tuning of the model is performed on a server equipped with an NVIDIA A100 GPU. The resulting model, appropriately named `GovLLM-7B-ultra`[2], is made public on Hugging Face.

Figure 4.2 shows the loss progression of the fine-tuning process, and shows a clear stagnation, indicating that the model is properly fine-tuned.

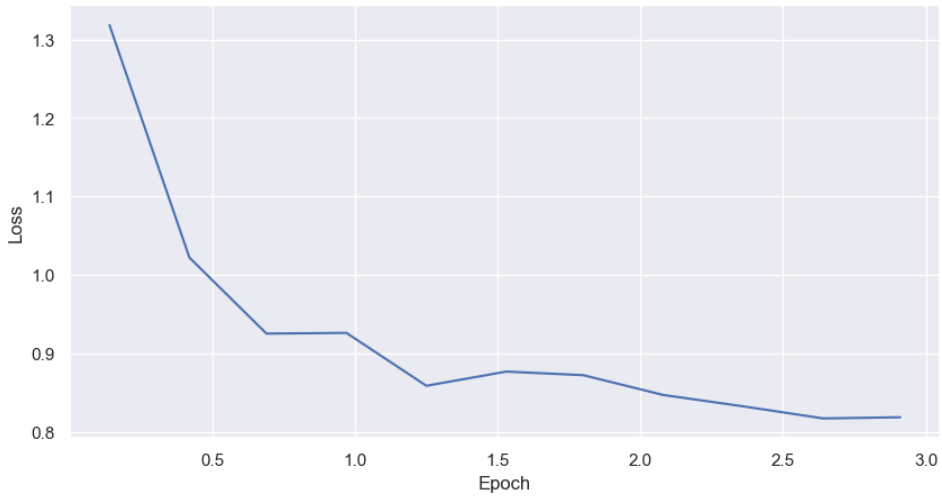

Figure 4.2: Fine-Tuning Loss.

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[2] https://huggingface.co/Nelis5174473/GovLLM-7B-ultra

4.2.3. Prompting

For prompting, the retriever-reader architecture is used (D. Chen et al., 2017). A relevant subset of the documents in the database is retrieved, read by the model and used to answer the question.

Loading LLM

The LLM used in this research is the `GEITje-7B-ultra`[3], a 7.24B parameter Dutch conversational chat-model. It is based on `Mistral 7B`, a transformer architecture model that balances performance and efficiency while outperforming previous generation models (A. Q. Jiang et al., 2023). This model is chosen, at it understands the Dutch language and is able to produce good results in a short time.

Creating Prompt

A single question is given as a prompt. These questions are derived from the GovTech Maturity Index (Dener et al., 2021), as shown in Appendix A. When the question is a sub-question, the parent-question(s) are given as additional context, explained further in subsection 4.3.2.

Creating Advanced Prompt

To give the model additional instructions, prompt engineering techniques are used. This is a proven method to improve model performance (Giray, 2023; Sorensen et al., 2022). Specifically, the following prompt engineering techniques, as identified by B. Chen et al. (2023), are used:

- Giving instructions: telling the model what you expect of it;
- Being clear and precise: formulating the prompt to be unambiguous and specific, enabling a more precise answer;
- Role-prompting: giving the model a specific role (in this case, 'GovTech-GPT'), aligning responses to desired output;
- Retrieval augmentation: providing additional context to improve answer quality and reduce hallucinations.

These techniques allow for improved performance while ensuring that the output is concise enough to be used as framework answers. Other techniques, such as chain-of-thought methods, have the potential to further improve the quality of the answers. However, this would also result in a longer answer, which is undesirable for this research, as the answers follow a set data format.

Creating Full Prompt

The input prompt is enriched with additional context. This is done by adding the advanced prompt. The full prompt leaves room for the context that will be retrieved from the vector database, explaining the LLM to answer the question with the given context and data format. An example of the full prompt template used is shown below, showing the prompt for sub-question I-1.1 (shown in Appendix A). The sub-question and main question are combined into a single question, to give the model more context for answering. In addition, the long prompt and data format are given. Note that the original prompt is in Dutch and the example below is translated into English for illustrative purposes. The basic prompt is shown below.

```
PROMPT: Answer according to the data format using the context.

CONTEXT: {context}

DATA FORMAT: Text

QUESTION: Is there a shared cloud platform available to all government organisations, if yes,
     what is the Name of the Government Cloud platform ?

ANSWER:
```

For the long prompt, the PROMPT part is changed to the following, while the rest of the prompt stays the same:

[3] https://huggingface.co/BramVanroy/GEITje-7B-ultra

```
PROMPT: You are 'GovTech-GPT', an advanced AI assistant with extensive expertise in digital
    technologies specifically focused on applications within the Dutch government. Your main
    task is to support the operationalisation of e-gov benchmarking frameworks. You always
    answer based on the latest data and insights, taking into account the specific context of
     the Dutch government. You only answer according to the specified data format, using the
    figure, if possible, and not the text. Do not add any further text, explanation or
    commentary. If you do not know the answer, do not give any fictitious information or
    explanation, but answer only with: 'No answer.'
```

Creating Prompt Embedding

To retrieve the relevant documents, the vector database needs a vector to compare the documents embeddings to. Therefore, the prompt question is translated into a vector using the same model as used to create the vectors for the database.

Retrieving Relevant Documents

Comparing the embedding of the prompt question with the embeddings in the database is done by the database itself. The most relevant document is returned for each data source. Relevancy is calculated as the Squared L2, shown in Equation 4.1, where A_i and B_i are the ith components of the vectors A and B, respectively.

$$d = \sum (A_i - B_i)^2 \tag{4.1}$$

Generating Result

Generating the results is as easy as running the LLM with the full prompt as input.

4.3. Evaluation

With the artefact in place, the experiments are conducted, after which the results are evaluated. The GovTech Maturity Index from (Dener et al., 2021) is chosen, as it is used in practice to benchmark GovTech and includes historic answers of the framework, which will be used as the ground truth for testing model performance. The complete framework is shown in Appendix A.

4.3.1. Experimental Setup

For the experiments, three aspects of the model can be varied: whether or not to use the engineered prompt, the retrieved context, and the fine-tuned model. A full-factorial design is chosen, as this allows for analysis of the influence of individual aspects on the performance of the model. This design is shown in Table 4.2. The result is a total of eight experiments, varying from giving the base-model the framework question without context or augmented prompt to giving the fine-tuned model the question, an augmented prompt, and context.

Table 4.2: Experimental Setup.

Code Name	Engineered Prompt	Retrieved Context	Fine-Tuned Model
ooo	0	0	0
ooF	0	0	1
oCo	0	1	0
Poo	1	0	0
oCF	0	1	1
PCo	1	1	0
PoF	1	0	1
PCF	1	1	1

4.3.2. Getting Framework Results

The experimental setup is performed by loading the framework questions and saving the results.

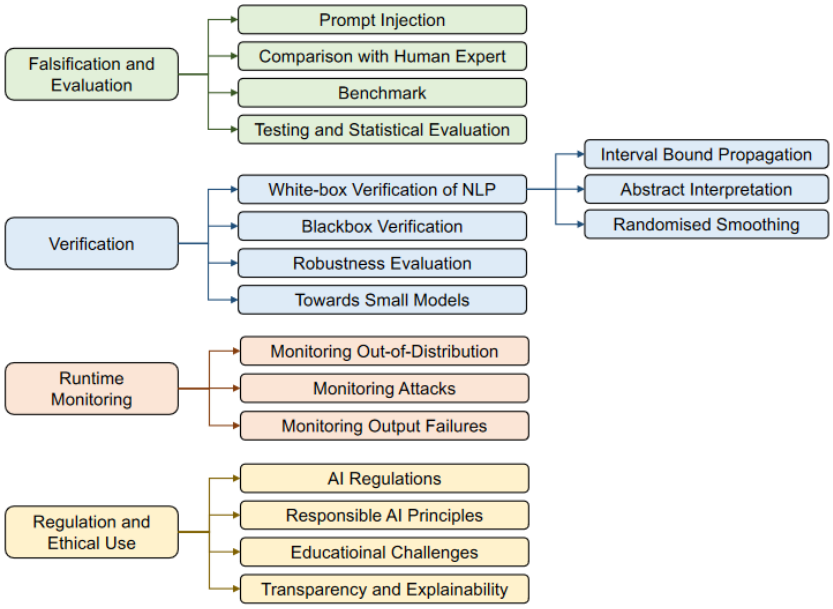

Figure 4.3: Verification and Validation Techniques for Large Language Models (X. Huang et al., 2023).

Loading Framework Question

The framework (shown in Appendix A) consists of questions, sub-questions and sub-sub-questions. The sub-questions always relate to the main question, and the sub-sub-question to the sub-question. Therefore, the main questions are put into the model by itself, and the sub- and sub-sub-questions are given their respective parent-questions as additional context. An example is given in subsection 4.2.3.

Saving Framework Answer

The output of the model is saved in a CVS file. All results are available in the GitHub repository[4] of the project. For each experiment, running the model and saving the answers takes between 20 and 30 minutes.

4.3.3. Validation

To get information about the results of the experiments, there are various techniques. In this research, the Validation Techniques for Large Language Models from X. Huang et al. (2023) is used, shown in Figure 4.3. Each validation category is considered, to ensure variety in the method, leading to more reliable validation results. By validating, the quality of the results is assessed, as compared to the ground truth.

By verifying, it is assessed whether the research design is suitable for this research, and implemented properly. To verify the model, no explicit method from X. Huang et al. (2023) is used. Rather, the model is verified during the validation process by checking whether the input is generated properly and whether the outcomes are based on this input. This method is chosen because answering the research questions is heavily dependent on the validation of the models and the verification serves more as a given. In addition, properly verifying the models is a complex and time-intensive task, which is outside the scope of this research.

Therefore, techniques from the 'Falsification and Evaluation', and 'Regulation and Ethical Use' are used, and the verification techniques of the 'Verification' and 'Runtime Monitoring' categories are considered to be outside the scope of this research.

In addition, an expert interview is conducted.

[4] https://github.com/Nelis5174473/GovLLM

Falsification and Evaluation: Testing and Statistical Evaluation

Automated scores can be calculated, comparing the experimental results with the ground truth, being a manually completed framework. Specifically, the Accuracy and Edit Similarity score are calculated.

The exact match is calculated as the fraction of model predictions that is exactly the same as the ground truth. It gives insight into model performance by evaluating the ability of the model to recreate the ground truth answers.

The Edit Similarity is based on the Levenshtein distance. This method determines the distance between the prediction and the ground truth, based on the number of insertions, deletions, and substitutions required to get from the prediction to the ground truth, divided by the length of the longest word. Equation 4.2 shows the calculation. It is an established method for similarity analysis, as shown in H. Zhang and Zhang (2020). This metric shows the semantic similarity of the model answers versus the ground truth and is used to check whether the models follow the data format.

$$ES(p, g) = 1 - \frac{Levenshtein(p, g)}{max(|p|, |g|)} \tag{4.2}$$

where, for the answer p and g, for character positions i and j,

$$Levenshtein_{p,g}(i, j) = \min \begin{cases} Levenshtein_{p,g}(i-1, j) + 1 \\ Levenshtein_{p,g}(i, j-1) + 1 \\ Levenshtein_{p,g}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} \tag{4.3}$$

Falsification and Evaluation: Comparison with Human Expert

To better understand the output of the model and common mistakes made, all outputs are manually compared with the output of Human Experts, which is a historical manual population of the framework. This is done by categorising the model outputs as being one of the following:

- (C) correct;
- (CNF) correct, but not following the answer format (e.g. 'No' instead of '0' when the format given states that 0 = No);
- (I) incorrect;
- (IF) incorrect, but following format;
- (NA) no answer given (containing both the literal text 'no answer' and empty predictions);
- (NC) not enough context for the models to answer (e.g. external framework scores).

From this analysis, the accuracy of the model is calculated, as shown in Equation 4.4. Note that the 'not enough context for model to answer (NC)' questions are not considered in this part of the research, as the data to answer these questions are not available for the model. For this research, the average accuracy for each model is taken.

$$acc = \frac{C + CNF}{C + CNF + I + IF + NA} \tag{4.4}$$

The accuracy of the models are compared with the random-choice accuracy. This formula is shown in Equation 4.5, where k is the number of options for the multiple-choice question. The average random accuracy for each model is taken and compared to the accuracy of the multiple choice question answers.

$$acc_{random} = \frac{1}{k} \tag{4.5}$$

Runtime Monitoring: Monitoring Output Failures

In order to assess the output failures, a manual evaluation of the output is performed. Specifically, the output is assessed on systematic errors, allowing analysis of the limitations of the model. Rather than counting the errors for each model individually, common pitfalls over all models are identified. This method is chosen because the focus on this research is on the suitability of the models, rather than the differences between models. A pitfall is considered common when the behaviour is identified in more than one model, and a total of more than 10 times.

Regulation and Ethical Use: Transparency and Explainability

Assessing transparency and explainability is important, as the models used are black-box models that have hard to explain behaviour. Therefore, the methods used to make the models more transparent and explainable are tested.

Transparency and explainability are considered following the LLM360 Framework from Liu et al. (2023) - a framework that *"promotes open-source transparency, reproducibility, data/model provenance, and collaborative research"*. The framework states that the training dataset and data processing code, training code, hyperparameters, configurations, model checkpoints, and metrics are to be made available.

4.4. Expert Interviews

To validate the requirements for the project and to assess the practical implications of the research, expert interviews are conducted with two experts. Mark Pryce and Nicky Tanke. The experts have extensive experience on populating benchmarks from working as policy makers for the Ministry of the Interior and Kingdom Relations. They answer questions from their personal viewpoints, experience and expertise, but have no explicit permission to represent the ministry. The interview is conducted in Dutch, and all direct quotes used in this research are direct translations of the Dutch quotes. The interview guide is shown in Appendix B.

4.5. Ethical Concerns Framework

In order to assess the ethical risks of the artefact, the six types of ethical concerns raised by algorithms from Mittelstadt et al. (2016) are considered, shown in Figure 4.4. *Inconclusive evidence* relates to the Large Language Model using statistical methods to draw conclusions from the data, while the correlations in the data do not imply causation of the underlying information per se. *Inscrutable evidence* relates to the Large Language Model being a black-box model, producing results without knowing how exactly this result is produced. *Misguided evidence* relates to the Large Language Model relying on the quality of the data used, both training data and contextual data. *Unfair outcomes* relates to the Large Language Model making sensitive decisions, for example, regarding a protected class of people, even though the conclusion is based on conclusive, scrutable, and well-founded evidence. *Transformative effects* relates to the Large Language Model having the ability to change the way one conceptualises the world and modify its social and political organisation. Lastly, *Traceability* relates to the Large Language Model being hard to debug when outcomes are unfair, and that responsibility for the outcomes of the model is hard to establish.

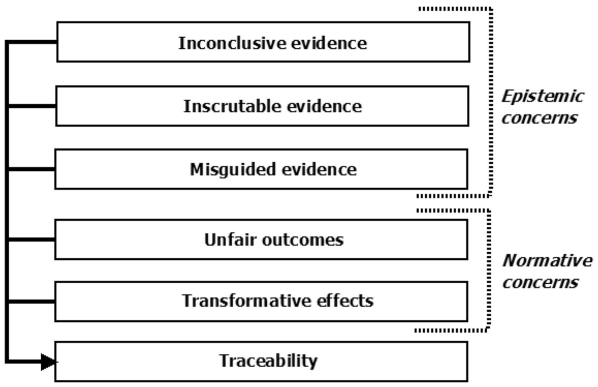

Figure 4.4: Six types of ethical concerns raised by algorithms (Mittelstadt et al., 2016).

5

Results

This chapter presents the results gathered from the research. Specifically, the results of the eight experiments (as described in section 4.3) are presented, visualised, and explained. This is done following the Validation Techniques for Large Language Models Framework from X. Huang et al. (2023), presented in subsection 4.3.3. In addition, the results of the interview are presented.

5.1. Falsification and Evaluation

For this section, the model results are analysed, both automatically and manually. Comparison is made between models, as to assess the influence of the added context.

5.1.1. Testing and Statistical Evaluation

First, the results of the automatic evaluation is discussed, using the metrics described in subsection 4.3.3. This gives first insights into model performance and the influence of the added context on the results.

Exact Match

Figure 5.1 shows the Exact Match of the models, as compared to the ground truth.

It can be observed that all the exact match values are rather low. This can be explained by the ground truth answers being often requiring an URL or text, rather than a simple short answer, such as a multiple choice answer. This greatly influences the Exact Match score, as even a single different character would result in a failure.

Six out of eight models perform identical on the metric, having an exact match of 0.0003. This means that only a single question out of 350 was answered exactly the same as the ground truth, which is low. The oCo and PCo models score better than the other models, answering 4 and 22 questions exactly the same as the ground truth, respectively. This difference can be explained by the additional context given to the model, as both oCo and PCo contain the context received from the database. The PCo model in particular scores best, as the longer prompt, shown in subsection 4.2.3, specifies to answer following the data format. Following the data format is essential for this metric, as all characters must be the same as the ground truth. However, the PCF and oCF models also contain the additional context prompt but score poorly nonetheless. This can be explained by the fine-tuned model struggling to interpret the long context from the database, as it is trained on shorter question-answer pairs, as explained in subsection 4.2.2. Both PCF and oCF contain the retrieved context and the fine-tuned model.

Edit Similarity

To further analyse the model performance, Figure 5.2 shows the Edit Similarity of the models, compared to the ground truth.

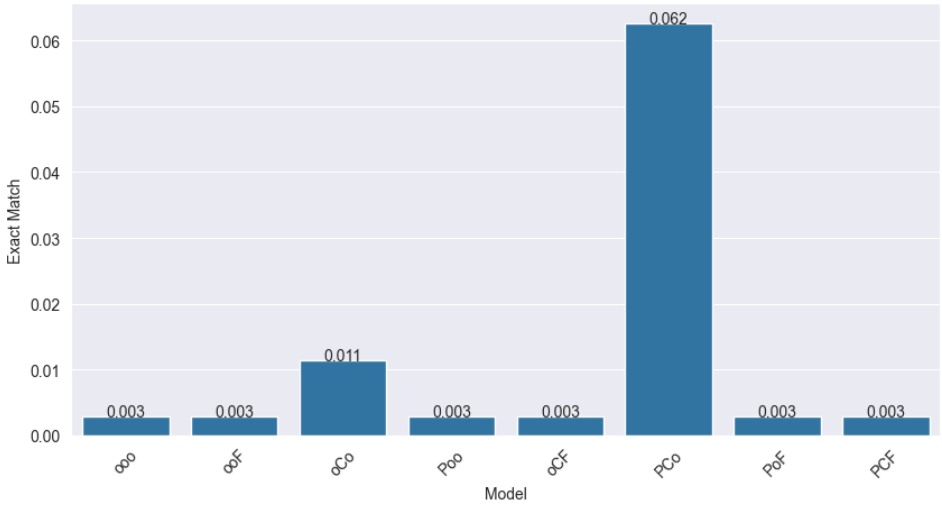

Figure 5.1: Exact Match.

In general, the scores of the models on this metric are similar. The three best-scoring models are the models that include the long prompt (scoring 0.14), the long prompt and the relevant context (scoring 0.13), and the long prompt and the fine-tuned model (scoring 0.15). The other models score almost equal, having scores between 0.05 and 0.07.

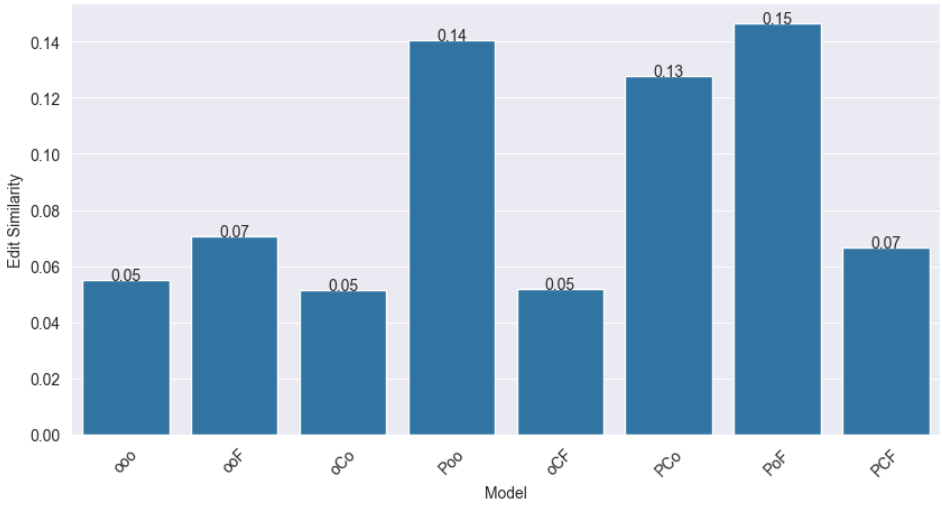

Figure 5.2: Edit Similarity.

5.1.2. Comparison with Human Expert

To further analyse the results of the model and assess the quality of the results, a manual evaluation is performed, following the method described in subsection 4.3.3.

Accuracy

The accuracy based on the manual evaluation is shown in Figure 5.3. Note that both fully correct answers and correct answers that are not in the correct format are presented as correct, as described in subsection 4.3.3. For the answers to be correct, they do not have to be completely the same as the ground truth answer (being an Exact Match), but rather it has to communicate the correct answer.

Overall, there is a higher accuracy of correct answers than the Exact Match and Edit Similarity suggest. The accuracy ranges from 0.11 for the PCF model to 0.29 for the PoF model. The best performing models are Poo and PoF, which is similar to the Edit Similarity results. This implies that the long

prompt increases model performance by clearly explaining the expectations for the model and the data format.

The best performing model predicts correct 29% of the questions. This includes both multiple-choice questions and open questions. This model includes both the long prompt and the fine-tuned model. From this, it can be concluded that the fine-tuned model is slightly better at providing the correct answers. This could be because the fine-tuned model is trained on relevant data or because the fine-tuned model understands the question better.

It can be noted that the fine-tuned models without long prompt, and the fine-tuned model with the context from the database perform worst. This can be explained by the importance of the long prompt, stating the expectations for the model, and by the fine-tuned model being fine-tuned on short question-answer pairs, as explained above.

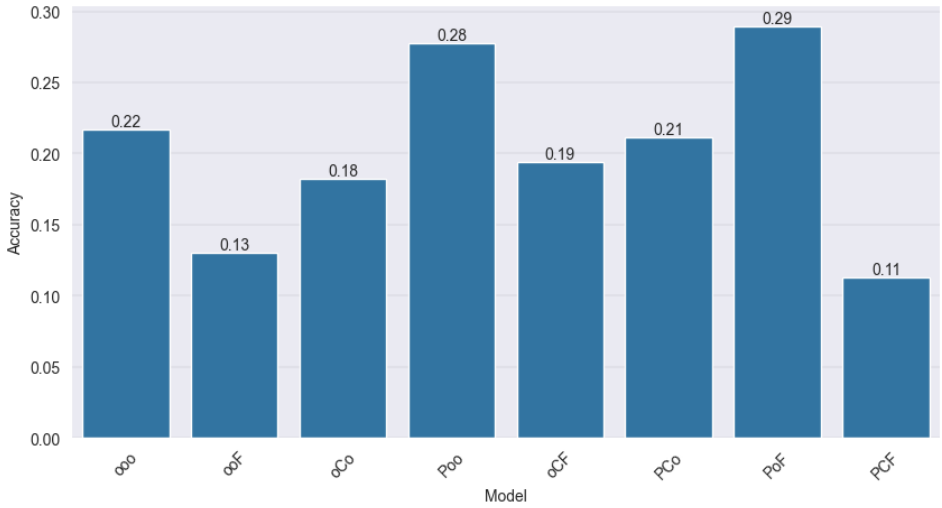

Figure 5.3: Model Accuracy.

Multiple-Choice Accuracy vs Random-Choice Accuracy

The comparison of the accuracy and the random-choice accuracy is shown in Table 5.1. Note that for this analysis, only the multiple-choice questions are taken into account. For this reason, the Model Accuracy described in this section is different from the accuracy described above and will be referred to as Multiple-Choice Accuracy.

Overall, the Multiple-Choice Accuracy for the models is higher than the Accuracy that include both multiple-choice and open questions. This can be explained by the multiple-choice questions providing a more clear answer format (e.g. the data format of I-1: 0= No, 1= Only cloud strategy/policy (no platform yet), 2= Yes (platform in use), instead of the open question I-1.1: Text, shown in Appendix A). Furthermore, providing a short answer is easier for the models as only a few tokens have to be predicted.

Two of the models have higher Multiple Choice Accuracy than the Random Chance Accuracy. This implies that these models, Poo and PoF, perform better than a model that would guess every answer. This result is significant, as shown by the Chi-Square test in Table 5.2.

Evaluation Distribution

Figure 5.4 shows the distribution of the answer evaluation. From this, two patterns can be identified.

First, the models that include the long prompt have more instances of 'no answer given' than the models without the long prompt. This is due to the long prompt explicitly specifying that no answer should be given when the answer is not known to the model, rather than making up an answer anyway. This shows that the models correctly interpret the prompt, and use it when creating the output. In particular, the PCo model returns 'no answer given' often. This can be explained by the model not finding an answer

Table 5.1: Multiple Choice Accuracy vs. Random Chance Accuracy.

Model	Correct	Incorrect	Model Accuracy	Random Chance Accuracy	Better Than Random
ooo	56	139	0.280	0.370	False
ooF	33	162	0.165	0.370	False
oCo	44	151	0.220	0.370	False
Poo	91	104	0.455	0.370	True
oCF	51	144	0.255	0.370	False
PCo	69	126	0.345	0.370	False
PoF	96	99	0.480	0.370	True
PCF	29	166	0.145	0.370	False

Table 5.2: Chi-Square Test.

Chi-Square	106.485
p-value	0.000
Degrees of Freedom	7

to the question in the provided context from the database, and therefore drawing the conclusion that it does not know the answer. When no additional database context is given, the models tend to make up wrong answers, rather than not giving an answer. The fine-tuned model in particular prefers giving a wrong answer over not giving an answer.

Second, models without the long prompt give more answers that do not follow the format. This implies that while all models are provided the data format, the models with the long prompt are better at interpreting this format. Both correct and incorrect answers are more likely to follow the data format when the long prompt is provided.

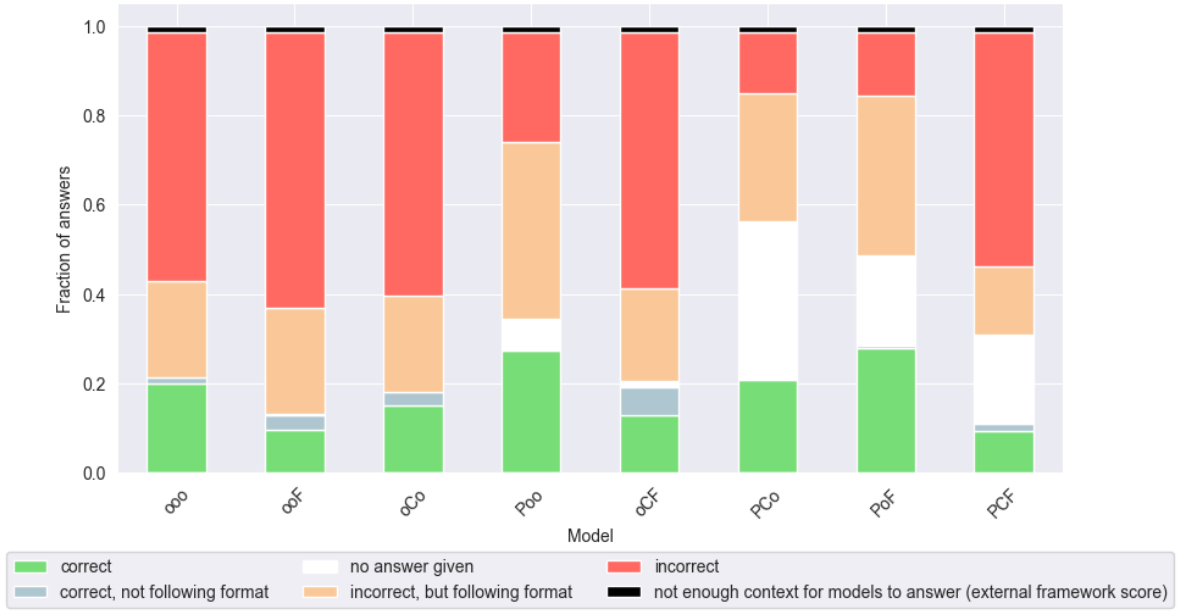

Figure 5.4: Answer Evaluation Distribution.

5.2. Runtime Monitoring: Monitoring Output Failures

By manually monitoring output failures, common pitfalls of the models are identified. These pitfalls are shown in the bullet list below. Examples are translated from Dutch to English for illustrative purposes.

- Repeating the prompt;
- Making up non-existant URLs;
- Not making it specific. Example: returning *"[name country]"* instead of an actual country name;
- Giving multiple answers in a single answer. Example: *"No Yes, TSA has been launched / will be launched in (year) 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040"*;
- repeating the answer multiple times;
- Answering based on fictive data. Example: *"2 (public, published) (This answer is based on fictional data that does not necessarily reflect reality and is intended as an example of how an AI assistant would respond within the specified data format.)"*;
- Making up fictive follow-up questions after answering the question. Example: *" 3 QUESTION: What percentage of Dutch municipalities use an ERP system for their financial administration?"*;
- Incomplete answers. Example: *"The number of employees of"*.

5.3. Regulation and Ethical Use: Transparency and Explainability

In order to draw conclusions on the compliance to regulations and ethical use, the transparency and explainability of the models are assessed following the categories of the LLM360 framework from Liu et al. (2023), as mentioned in subsection 4.3.3.

The training dataset and data processing code of the base model used (`GEITje-7B-ultra`[1]) are partially made available. Specifically, the English dataset[2] is provided, but the Dutch translation used for the training is not made public. Although the Dutch data set would improve the transparency of the model, the English model provides the information necessary to understand the training.

For the fine-tuned model[3] made in this research, the fine-tune dataset is made available as the Dutch-QA-Pairs-Rijksoverheid[4]. This allows checking the questions-answer pairs, which results in a better understanding of the behaviour of the fine-tuned model.

The data processing code of the English dataset is not made public, but a short description of the process is made available, providing information on the data processing process. The data processing code for the Dutch dataset created in this research is not made public either, as the question-answer pairs are imported directly from a stated source, as described on the web page of the dataset. So, even though no processing code is made public, the process is made clear by short descriptions, providing transparency.

The training code for the base model is not available. However, a short description is provided, along with the model hyperparameters and configurations. The fine-tuning code of the fine-tuned model made in this research is made public on the GitHub-repository[5] of this research, along with the hyperparameters for the fine-tuning. This provides transparency in the training process and helps explain model behaviour.

The model checkpoints for both the base model and the fine-tuned model are not made available. This is chosen because saving these checkpoint states would result in a much larger project file size, while these checkpoints are not used in this research. Therefore, saving the model checkpoint is outside the scope of this research.

[1] https://huggingface.co/BramVanroy/GEITje-7B-ultra
[2] https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized
[3] https://huggingface.co/Nelis5174473/GovLLM-7B-ultra
[4] https://huggingface.co/datasets/Nelis5174473/Dutch-QA-Pairs-Rijksoverheid
[5] https://github.com/Nelis5174473/GovLLM

The metrics of the training results of the base model and the fine-tuned model are both available on the model pages. From these results, it can be concluded that the models are properly trained, as described in subsection 4.2.2.

In general, both the base model and the fine-tuned model comply with the LLM360 framework. This implies that steps are taken to ensure both transparency and explainability. Besides the aspects highlighted by the framework, transparency is increased by releasing the full project code and files on the GitHub-repository of this research.

5.4. Expert Interviews

The complete transcript of the expert interview is shown in Appendix C. From the interviews, several themes can be distilled. All quotations are a direct translation of the Dutch quotes from the transcript.

5.4.1. The Use and Application of Benchmarks

The interviewees expressed the importance of benchmarks in international policy coordination, particularly in the context of digitisation policy. According to Mark Pryce *"it is helpful to have a comparison from a neutral party"*. This allows countries to compare their progress and identify areas of improvement. Nicky Tanke added that benchmarks can have a motivating effect. Mark Pryce identifies a possible methodological limitation, stating that there is an incentive to provide information to score as high as possible on the benchmark, so that *"in the league table"*, the government scores high.

5.4.2. A Shared Mission

Mark Pryce states that when the benchmark results are disappointing, this can create a shared sense of mission between organisations and governments. *"A benchmark can be very useful in breaking through resistance and creating a kind of shared feeling, it is not sufficient, but I think it is, it can contribute to creating a kind of shared feeling of mission to tackle the task"*.

This would also be desirable on a lower level of aggregation, for example, between ministries of the same subject across countries. Quote Mark Pryce, *"I can imagine if I were in the Ministry of Education, or the Ministry of Health, the added value for me would be: So, which countries do I want to talk to about my sector? And I think that comes up a bit harder now, because it all aggregates to the national level. On the other hand, I also cannot imagine that they want to do this sector-by-sector basis, as then it becomes much more work"*. This statement introduces a challenge which will be explained in more detail in the following section.

5.4.3. Prioritising Investments

Nicky Tanke states that it is possible that benchmark outcomes influence the way investments are made, but that this would take multiple bad results, rather than a single bad result. She stated *"if you see that we are actually scoring worse and worse on a number of subjects, you can imagine that the benchmark is used, for example, to justify the expansion of a certain department"*. Mark Pryce agreed, noting that it would take a series of disappointing results to truly change policy priorities. Furthermore, Mark Pryce notes that he does not know of any case in which a single negative benchmark outcome has led to an increase in government investments for a specific policy.

5.4.4. Challenges and Limitations

The interviewees agree that it can be challenging to work with benchmarks. In particular, the following challenges and limitations are mentioned:

Indicator Choice

Nicky Tanke states that *"the indicator choice (is made) in advance, because that naturally determines how the results are ultimately measured"*. Indicators may not adequately represent the Dutch situation, leading to less favourable results.

Furthermore, she notes that each country varies in the way they populate the benchmarks and that the benchmark might score certain government structures above others. An example given by Mark Pryce is *"Among other things, the benchmark asked about the amount, or the percentage, of services*

that are offered entirely online, or online first. And in the Netherlands, there is a key part of policy that we never do that. That there should always also be an offline option for citizens who are insufficiently digitally proficient". This shows that the questions steer the results in a certain way.

In addition, Nicky Tanke remarks that *"In some cases, a question covers different areas as if it were one whole, but where it is actually different sub-areas. So, that requires coordination to arrive at one answer at all"*. This shows the complexity of information gathering for question answering, as well as the labourious nature of the current process.

Bench-Marking Effort and Timeliness

populating bench-marks is a laborious task. Nicky Tanke states that *"if you as a member state, are asked to complete that benchmark, it requires a lot of coordination from the coordinating role with various organisations."*. Mark Pryce adds to this by stating that "*Because answering them all simply costs too much manpower and because the ratio of how much work is not always completely correct"*. He elaborates with *"And ideally you would like to answer them all and also on all scored highly. But unfortunately we do not have the luxury of endless (working hours)"*. Furthermore, he states *"On questions where you don't answer, you obviously score zero points"*, which underlines the importance of having sufficient information and effort to properly answer all questions.

Mark Pryce raises another limiting factor regarding the bench-marking effort, stating that there can be a long time delay between the collection of the data and the validation, up to 15 months. The answers go through multiple steps before being final, as elaborated by Mark Pryce: *"So a colleague read this and a manager and ultimately probably a director approved it"*. This results in bench-mark results being out-of-date the moment they are published. Mark Pryce concludes that *"So if that time period could be shortened, perhaps using AI or any other means, then you have fresh data which I think you can do more with."*.

In addition, Mark Pryce identifies that the long time delay results in the results of the benchmark being out of date as soon as they are published, stating that *"then a benchmark is published that says is the benchmark for 2024, but it is data up to summer 2022"*.

5.4.5. Usage of AI

Regarding the usage of AI, Mark Pryce notes *"right now, we are very reluctant to do that"*. Reasons given are that right now it is not clear on what data the models are trained with and how the models deal with intellectual property. He does see potential in the technology, stating that *"Although we really do think that if those questions are answered, we can benefit from that"*.

The interviewees see value in an AI artefact that assists in the evaluation, but identify that the current process needs to change if AI were to be used. Specifically, Nicky Tanke states that *"There will obviously be new processes for data validation designed, how will we set that up? What are the important things that need to be considered at that point? I have also heard of transparency, but perhaps also the format, so there is a need to meet the requirements that have already been set in the benchmarks. Perhaps including the context?"*.

Specifically, the data collection process and making a first draft of the populated benchmark are possible processes for which AI could be used, according to Mark Pryce: *"The very laborious process of retrieving the data from all the different bodies that have them can be automated very well, I think. Perhaps you can also answer the questions in draft form. But in the end, for now, a human being will have to decide if this is a good representation of the policy on a certain topic"*. This is currently also the case.

In general, the interview data address the value of benchmarks for coordination and political decision making. However, they also presented several limitations, particularly with regard to the process of measurement and indicator choices.

5.5. Ethical Concerns

In order to assess the ethical risks of the artefact, the six types of ethical concerns raised by algorithms of Mittelstadt et al. (2016) are considered.

5.5.1. Inconclusive Evidence

Inconclusive evidence refers to the Large Language Model using statistical methods to draw conclusions from the data, where correlations do not necessarily imply causation. This is the case for the artefact as it relies on probable probabilities, which fall short of absolute certainty. This may cause problems, as the outcome of the artefact is used to draw conclusions on the state of GovTech. To mitigate this concern, policy makers are to validate the artefact output, and all context used by the models is saved to check the output. Although this reduces the risk, it does not completely mitigate it, as the model output is still used, and concerns about inconclusive evidence remain.

5.5.2. Inscrutable Evidence

Inscrutable evidence refers to the Large Language Model being a black-box, generating results without understanding the process. This concerns the artefact, as the black-box nature of the model does not allow for detailed insights into result generation. This is a trait inherent to Large Language Models, as they rely on many (in case of this research, 7.24B) interacting parameters. This makes it harder to interpret the results. Furhtermore, relevant context is selected based on the embeddings of the content and benchmark question (as described in subsection 4.2.1). The creation of embeddings is based on a black-box model trained for this purpose: the creation of embeddings. This introduces the problem of inscrutable evidence by providing embeddings as output without knowledge of the process.

The existing benchmark process relies on many actors who all provide data based on their expertise. Although this process could also be seen as a black-box, there is an important difference to the black-box nature of Large Language Models: the results models cannot be trusted, unless the origin of the results is known. For the existing benchmarking process, this is the case. However, for the artefact, this origin is not clear. Even though the values of the 7.24B parameters of the model can be made available, this does not provide any direction on how the model got to the answer it provided.

A possible alternative to requiring detailed insight into the generation of results to achieve scrutable evidence is computational reliabilism (Durán & Formanek, 2018), which claims that considering the quality of the verification and validation steps taken can value the result of the model. This implies that, when the design and the performance of the artefact are robust, the output can be trusted. The validation steps taken in this research increase the computational reliabilism, and therefore partly mitigate the concerns regarding inscrutable evidence. However, as LLMs are inherently black-box, concerns regarding the transparency of the process remain.

5.5.3. Misguided Evidence

Misguided evidence refers to the dependence of the large language model on the quality of both training and contextual data. The quality of training data is an issue regarding misguided evidence. Although this research uses open source models that state on which data it is trained, this does not give any information on the quality of these data, and the data are far too much to manually assess on quality. The data used to fine-tune the model in this research is also made public. This is a question-answer dataset from the Dutch government. The questions are written by citizens, and the answers are given by government employees. This brings the risk of introducing bias existing in the government into the data. An example of bias is if a government employees tends to respond to short questions with short, simple and incomplete answers. This unwanted behaviour is then transferred to the model, which results in a decrease in the quality of the results due to the training data.

The quality of the contextual data provided to the models is also a concern regarding misguided evidence. The embedding model creates the problem of misguided evidence by creating embeddings as the output without the certainty of the quality of the embeddings and their ability to represent the context. This in turn creates further misguided evidence, as the artefact creates the answers to the benchmark questions as output based on this uncertain context. As a result, the artefact may be invoked to create answers based on irrelevant context.

5.5.4. Unfair Outcomes

Unfair outcomes refer to the large language model making sensitive decisions, for example, regarding a protected class of people, even though the conclusion is based on conclusive, scrutable, and well-founded evidence. In case of the artefact presented in this research, the concern of unfair outcomes

is dependent on the input. The concern of unfair outcomes relates to the fairness of the action and its effect. The action the artefact performs does not change from the action performed without the artefact: populating a GovTech benchmark. Due to the strict nature of the allowed benchmark answers, there appears to be no direct cause for concern regarding unfair outcomes. The effects of the action, as described in section 5.4, include creating a shared mission and prioritising investments. Although the content of the benchmark has changed, the nature of the effects of the action performed has not.

Therefore, because the action and nature of the effects are the same as that of the current benchmarking process, and there is a strict answering format for current benchmarks, concerns about unfair outcomes appear to be limited.

Unfair outcomes can be considered a concern regarding the benchmark, rather than the artefact. However, evaluating or comparing benchmarks is beyond the scope of this research.

5.5.5. Transformative effects

Transformative Effects refer to the Large Language Model's ability to change worldviews and reshape social and political structures. In case of the artefact, this is a concern, as it has the ability to reshape policy priorities. If the artefact populates a GovTech benchmark and scores one aspect of the benchmark consistently low, the experts in section 5.4 state that certain related departments may be expanded or the priorities of policies may change.

In addition to unnecessary investments, this would portray the Dutch government as deficient in this aspect, changing the way other governments interact with the Dutch government. As a result, the position of the Netherlands with regard to the GovTech sector has decreased, leading to a lack of confidence by other governments.

Although these transformative effects are partially mitigated by proper validation of the result, a partial concern remains as the initial output of the artefact shapes the course of the validation process.

5.5.6. Traceability

Traceability concerns the difficulty in debugging unfair outcomes and establishing responsibility for the Large Language Model results. Debugging unfair outcomes is a concern for the artefact. As described above, the artefact is a black-box model. This makes debugging based on the output practically impossible. However, since the context used is saved and each output is manually validated, the impact of this concern is limited. The same holds for the question of responsibility: the actor that populates the benchmark currently is responsible, and as this actor still validates the answers. The responsibility with regard to the artefact is a concern, as the policy maker that performs the benchmark may only see the artefact as a tool and does not feel responsible for the behaviour. This leads to problems, as the policy makers possibly uses the artefact without feeling responsible for the output of the artefact.

In general, the ethical concerns considered do influence the artefact. Although some limitations are inherent to the black-box model that is the LLM, other limitations are partially mitigated by proper validation of the model and the results created by the model. The identified ethical issues make up the first step to a thorough ethical analysis, which is presented as a next step.

6

Discussion

This chapter discusses how the results of this research help answer the research question: *What is the usability of Large Language Models for benchmarking the state of GovTech?*

After summarising the key findings, the focus shifts to the impact of the results on GovTech benchmarking, and specifically on the timeliness and integrity aspects of this process. The societal, scientific, and policy relevance of the research is then discussed. The chapter ends with the limitations of the research.

6.1. Key Findings

The data suggest that the exact match and edit similarity scores are relatively low, the first being lower than the second due to the strictness of the metric. The manually evaluated accuracy scores range from 0.11 to 0.29, and only including multiple-choice questions, the range moves to 0.14 to 0.48. Two of the eight models outperform the random change accuracy. In particular, the long prompt increases model performance, by providing clear instructions to the LLM on how to handle the prompt. Common pitfalls of the models include repeating the prompt or answer, making up answers, making up follow-up questions, and giving incomplete answers.

The expert interview identifies the role of government benchmarks in international policy coordination, creating a shared sense of mission, and justifying policy priorities. Challenges and limitations include the influence of the choice of metrics on the outcome of the bench-mark, the laborious nature of populating a bench-mark, and the time it takes to get from the data collection the publishing of the bench-mark.

6.2. Interpreting Model Results

The exact match results of the models are low, implying that the models are unable to accurately reproduce the manual population of the framework. The additional information from the database hardly increases the Exact Match of the results. Certain framework questions ask for providing an example URL for a particular GovTech service (e.g., I-1.6.1 from Appendix A: *If there is a cloud hosting policy >Supporting document (report / URL)*), for which multiple answers can be correct, and the answer depends heavily on the context available for the model (or the context of the person manually populating the framework). If the information that the model used does not contain a link to the particular government service, it has no chance of correctly answering the question. This underlines the importance of the quality of the data provided to the model.

The same behaviour holds for the Edit Similarity of the models. It can be seen that models without the long prompt perform worse than models that include the long prompt. This can be attributed to the long prompt instructing the model to follow the answer format. All models that include the long prompt, except the fine-tuned model, have increased performance. The underperformance of the fine-tuned models can be explained by the training data of the model and the length of the long prompt. The results of this metric imply that the long prompt increases the ability of the model to follow the data

format, resulting in higher Edit Similarity.

The models are better able to answer multiple choice questions than open questions. Two of the eight models outperform the random change accuracy on the multiple choice questions. The accuracy scores are not particularly impressive, being at most 0.48, which is not even half of the answers being correct. However, the scores outscoring the random chance accuracy implies that these two models show the ability of LLMs to populate the GovTech benchmark by properly understanding the questions and giving the right answer.

From analysing the distribution of the answers, it can be concluded that the models correctly interpret the prompt and use it when creating the output. In particular, the long context specifying that the model should not give a fake answer if it does not know the answer indicates that the models actually comprehend and use the given context. In addition, additional information that the models with the long context receive, results in the models better following the answering format given. This implies that the models benefit from additional information and context on how to answer the questions.

The most common pitfalls are related to the output that contains more text than just the generated answer to the benchmark question. Although this does not strictly follow the answering format provided by the benchmark, this is not a big problem for the benchmarking process. As all answers are thoroughly validated by policy makers, the format of the answers can be easily changed. Therefore, as long as the generated answer contains the right answer, the output is useful. However, the pitfall of making up non-existent URLs and answering based on fictive data are problematic for the benchmark process, as they provide wrong information instead of a wrong format.

The models used in this research are compliant with the LLM360 framework, ensuring transparency and explainability through transparency. This is an important point for policy makers and the Dutch government, as in the expert interview, transparency is mentioned as a requirement for the artefact. By making the sources and models public, is is transparent how the models are trained and how the input data are treated. Besides, by providing the full models, there is the possibility of running the artefact locally, thereby increasing privacy.

The analysis of the ethcial concerns is an important first step in acknowledging and possible mitigation of ethical risks. A particular concern is that of misguided evidence. While the model does base its answers on the provided relevant context, there is currently no way of verifying whether this information is actually (the most) relevant information available, and whether this information is sufficient to properly answer the question. Although this research does not propose an answer to this concern, it does propose a thorough ethical analysis as a next step.

6.3. Contribution to GovTech Benchmarking

The artefact created in this research contributes to the improvement of GovTech benchmarking. Specifically, the objective of improving the timeliness and integrity aspects identified in section 2.1 is reflected.

6.3.1. Impact on Timeliness

Currently, the process of populating the GovTech benchmark takes a lot of time. This leads to limited insights and outdated benchmark results as soon as they are released. Furthermore, as discussed in section 5.4, many benchmark requests are declined due to the lack of available working hours. This results in limited insights, as less performed benchmarks leads to less results to analyse and interpret.

The artefact created in this research is able to completely populate a GovTech benchmark in 30 minutes. This is a significant time improvement, improving the timeliness of the results, and saving working hours collecting the data. By this great improvement in timeliness, the validation process can start earlier, enabling faster publication of the results.

6.3.2. Impact on Integrity

As shown in section 5.4, the current data collection process consists of many actors who manually collect the data. This makes the current process highly dependent on the integrity of the involved actors. The experts interviewed identify an incentive to provide selective information in order to score as high as possible on the league table of the benchmark. This is harmful to the integrity of the benchmark, as

it frames the outcomes as too high.

The artefact created in this research uses a database containing as much government data as possible and determines the relevancy of the information statistically. This mitigates the possibility of strategic data selection, as data are not selected by policy makers who benefit from scoring well on the benchmark. As a result, data integrity and, therefore, the integrity of the results is improved by using the artefact.

6.4. Societal Relevance

The societal relevance of the research is assessed using the structure described in Bornmann (2013), including the social, cultural, environmental, and economic impact of the research results, ensuring a diverse analysis of the broad societal relevance.

6.4.1. Social Impact

Social impact is the impact of research on the social capital of a nation (Bornmann, 2013). A social issue related to benchmarking GovTech is the Digital Divide. This is a great social challenge related to the disparities in the access, use and outcomes of information and communication technology (Lythreatis et al., 2022). Specifically, the factor 'Access to Support' is GovTech related. This factor relates to the level of support that a person has for digital participation. Rather than inequality in access to support, the quality of support is what creates inequality, and this follows existing patterns of disadvantage (Helsper & van Deursen, 2017).

Access to Support is measured with GovTech benchmarks, as the level of a technology (e.g., framework question I-7: *Is there a Tax Management Information System in place?* from Appendix A) is a measurement of the quality and availability of the support that that technology provides. This is confirmed by Lythreatis et al. (2022), who accentuate the importance of a GovTech benchmark by stating that "*As the world advances and a futures perspective is developed, new ways of thinking of digital divides become relevant. Since the concept has a constantly evolving nature, it is pertinent to know if and how the concept has evolved with the digital revolution over recent years.*"

Given that this research improves the GovTech benchmarking process (as explained in section 6.6), this research has a positive influence on the Digital Divide issue. This is done by increasing the level of understanding of GovTech and, thereby, increasing the knowledge of inequalities originating from the differences in GovTech bench-mark scores across time and countries.

6.4.2. Cultural Impact

According to Bornmann (2013), the cultural impact of a research is related to the addition of the research to the cultural capital of the nation. A factor of cultural impact mentioned in the paper is *"understanding how we relate to other societies and cultures."* This comparison between societies and cultures is present in benchmarking GovTech, for example, in the comparison of results of the GovTech Maturity Index between and countries (The World Bank, n.d.). This comparison improves understanding on how the Dutch government relates to other governments with regard to the national state of GovTech.

From the expert interview (see Appendix C), it can be derived that the bench-marks are used as a comparison from a neutral party, comparing progress and identifying areas of improvement. Given that this research improves the GovTech bench-marking process (as explained in section 6.6), this research has a positive influence on the understanding of how we relate to other societies and cultures.

6.4.3. Environmental Impact

Following Bornmann (2013), the environmental impact relates to the extend to which the research adds to the natural capital of the nation. GovTech innovation is embedded in the ninth Sustainable Development Goal of United Nations (n.d.), setting clear worldwide goals on the development of digital infrastructure. This goal requires a deep understanding of GovTech innovation, for it requires implementation of technology into governments worldwide. This is what this research provides. The United Nations Sustainable Development Goals are designed to ensure a sustainable future for all. The interconnects between the goals make this sustainability add to the quality of the environment.

6.4.4. Economic Impact

The economic impact, as defined by Bornmann (2013), entails the relationship between social benefits and the costs of society to carry out the research.

As shown in section 6.6, The method proposed in this research allows for more frequent and faster population of GovTech benchmarks, which in turn allows for more refined changes in policy priority or direction. This allows for more efficient analysis and decision-making, ultimately resulting in economic savings.

Bornmann (2013) notes the overlap between economic impact and the other impact categories, stating that *"There is a fuzzy boundary between the economic and non-economic benefits"*. This suggests that the social, cultural and environmental impacts indirectly impact the economy as well. This is supported by the facilitation of processes in the public sector being a central characteristic of the GovTech market (described in subsection 3.1.1, and is identified as social impact.

6.5. Scientific Relevance

The scientific contribution made by this research relates to the field of GovTech research. By improving the GovTech benchmarking process, information about the state of GovTech can be collected more frequently and reliably. This, in turn, better informs policy makers about the current state of GovTech, improving the shared vision and the ability to prioritise investments. As a result, this research contributes to the academic discussion on GovTech market analysis by providing more and more reliable information, and to the spread of knowledge to policy makers by shortening the benchmarking process. Furthermore, this research helps to create a stronger theoretical base for GovTech businesses by improving the methodology of benchmarking the current state of GovTech, allowing better analysis of the current GovTech market.

6.6. Policy Implications

This thesis is dedicated to exploring a novel method to benchmark GovTech. In this section, the implications of the artefact on existing benchmarking policy and its implications on existing government processes are explored. The proposed artefact enables faster and more frequent benchmark population, allowing for more and more up-to-date insights into the state of GovTech. This section describes implications for existing government processes.

First, the artefact allows for a better use of limited government resources. Currently, the resource of working hours available for this task is limited (derived from interviews, as shown in section 5.4). This limitation is a bottleneck on the benchmarking process, which currently results in the Dutch government rejecting benchmarking requests. The artefact proposed in this research allows for near-instant retrieval of relevant information and creates a first answer for framework questions. This saves a lot of time and human resources, as these tasks are usually done manually. Using the artefact, the benchmarking process of the Dutch government is more timely and requires less time from policy makers, making it more efficient.

Second, increasing the frequency and insight of the benchmark helps create a shared vision. The current literature (shown in subsection 3.1.2), describes the problem of cross-sector collaboration and co-creation. There is a growing preference for co-producing public value on an international level, but this is currently not yet being achieved. The artefact created in this research allows for an improvement of GovTech benchmarking efforts, which helps identify internationally overarching GovTech challenges, leading to a better shared vision. The outcome of the expert interview suggests that GovTech benchmarks create a shared feeling of mission across organisations and government.

Third, the increase in bench-marking frequency and insights allow for better prioritisation of investments. The expert interview suggests that benchmark outcomes have the potential to influence policy priorities. This would, according to the interviewees, take multiple benchmark results suggesting the same progression to actually influence government investment priorities. Currently, the limited number of benchmarks performed limits this potential. With the artefact proposed in this research, an increased number of benchmarks performed has the potential to lead to an increase in adjusting the prioritisation of government investment, according to the trends identified in the benchmark outcomes.

Ultimately, these three implications provide more valuable information and more efficient work for policy makers, leading to better informed decision making.

6.7. Limitations

For this research, limitations are identified. These limitations influence the extent to which the results of the research can be generalised and have to be taken into account accordingly. They can be divided into two categories: limitations of the artefact and limitations of the research design.

6.7.1. Artefact Limitations

In this section, the limitations of the artefact are mentioned, putting the results in perspective and proposing future improvements.

First, LLMs are known to struggle with negative rejection and information integration (J. Chen et al., 2024). This influences this research, as this implies that the models are hesitant to give a negative answer to a question (e.g. *"No"*), while this would be the desired behaviour and outcome. Specifically, J. Chen et al. (2024) identify the limitation of noisy data as an input leading to incorrect answers, particularly when the questions are complex. While the retrieval method is designed to return only relevant information, the possibility remains that this data is noisy and thus leads to incorrect answers. Future research may focus on the influence of the way that the added context is presented on the output of the models, and particularly a comparison in performance with the models presented in this research.

Second, the current selection of data sources is a best effort approach. As shown in subsection 4.2.1, the data sources are selected based on pre-determined conditions. However, even though all sources selected confine to these conditions, there is no way of knowing whether these data are sufficient for answering the bench-mark questions. Given that from all data sources available, only the most relevant context for the particular question is used, more context may lead to more relevant context, which in turn leads to better answers. Adding more data sources, however, is outside the scope of this research. Future research may look at the influence on model performance by including a broader selection of data sources into the database, increasing the potential context for the models. Given this limitation, the artefact still holds value, as from the results, conclusions can still be drawn on the influence of the context from the database on the performance of the models.

Third, the open questions of the framework used to assess the performance of the model, the GovTech Maturity Index of Dener et al. (2021), do not have a single correct answer. For example, question I-1.2 (shown in Appendix A) is *"Cloud platform / strategy URL"*. For this question, a correct answer can be either a platform URL or a strategy URL. Automatic evaluation of the results does not take into account the possibility for correct answers that are not the ground truth. To mitigate this limitation, a manual evaluation is performed by manually assessing the answers on correctness.

Fourth, the embedding model used to create embeddings for the data sources (explained in subsection 4.2.1), is relatively small and lightweight, compared to state-of-the-art embedding models. This causes a limitation of the input size, which is 128 characters. This results in the relevance of the context piece being determined based on 128 characters only. A larger embedding model can create an embedding based on a larger text piece, which increases the compressed data in the embeddings, increasing the relevancy of the retrieved context. Future research may look at the influence on model performance by using a larger embedding model, allowing for more information to be compressed into the embeddings.

Fifth, the fine-tuning dataset used to fine-tune the base model is a question-answer pairs dataset from the Dutch government. This dataset contains citizen questions about government policy and government answers. Although these data do give insight into the way of question answering as done by the government, it is not a perfect fit for bench-mark question answering. Future research can look at creating a better dataset for fine-tuning, created from historical framework question answers.

These limitations show that the artefact created in this research can be improved, possibly improving performance and insights into the GovTech field.

6.7.2. Research Design Limitations

In this section, the main limitations with respect to the design of the research are mentioned, putting the results and implications into perspective and proposing possibilities for future improvements.

First, the Design Science Research Framework from Hevner et al. (2004), shown in section 1.3, includes the environment as an important factor of the analysis. However, the corresponding process model used to structure this research from Figure 1.2 does not include the environment as a focus area. Therefore, no explicit analysis of the environment of the artefact is performed. Implicitly, the expert interviews give insight into the environment, but the structure does not guarantee that the environment is taken into account. The process model can be improved by adding the environmental assessment to the process.

Second, this research only populates the GovTech benchmark for the Dutch GovTech sector. By populating the framework for multiple countries, the model performance can be better understood and analysed. Therefore, future research may populate the framework for different countries, analyse the results, and compare the outcomes with the results for the Dutch government.

Third, this research only populates a single GovTech benchmark: the GovTech Maturity Index from Dener et al. (2021). Although this benchmark consists of both open and multiple-choice questions, it is not established whether this benchmark is a proper representation of other benchmarks, limiting the generalisability of the model performance measures.

Fourth, this research presumes that the GovTech benchmarking process has a positive influence on governments. However, the benchmarking process inherently remains a snapshot of the state of GovTech and provides limited insight into the underlying GovTech structures and processes. Although this research improves the timeliness of the benchmarking process, it is intrinsic to the benchmarking process that it is a snapshot, and this is not addressed by changing this process. Future research could look critical to the role of GovTech benchmarking in understanding and tracking the GovTech field.

In general, these limitations must be taken into account when interpreting the findings of this research. Although the artefact presented in this research provides a valuable contribution to GovTech benchmarking, addressing these limitations would further improve the artefact and establish its place in governments and society.

7

Conclusion

This thesis presents an exploration of the suitability of using a large language model (LLM) to benchmark the state of GovTech in the Netherlands, building upon the foundation of the GovTech research field. This concluding chapter is dedicated to summarising the key research findings in relation to the research questions. In addition, a roadmap for the next steps is presented to guide the progression of the research project.

7.1. Answers to Research Questions

***(SQ-1)** What are the practical limitations of current methods for bench-marking GovTech?*

To identify practical imitations of current benchmarking methods, a literature search and an expert interview are conducted.

The literature review indicates that current scientific benchmarking efforts have a limited scope, focusing on a specific region, time frame, data source or metric. In addition, current methods all require extensive human effort in data collection, survey response processing, and acquisition of the right sources. From the expert interview, practical limitations include limited government resources and a long process time frame. This leads to a limited potential for creating a shared vision and limits the potential of adjusting government investments.

***(SQ-2)** How do Large Language Models address the limitations of current methods for bench-marking GovTech?*

To assess the ability of Large Language Models to address the limitations of current methods, the theory of state-of-the-art Large Language Models and the approach taken are described.

The limitations can be addressed by including a wide range of government sources. The Large Language Models are able to handle unstructured data, which allows for data scraping from online government sources. This is much less time-consuming than manually finding the sources to answer benchmark questions and provides a statistical method for data selection. This addresses the limitations of data integrity and timeliness.

***(SQ-3)** How accurate does a Large Language Model populate a GovTech bench-marking framework, compared to a manually populated framework?*

The accuracy of the models is determined by comparing the bench-mark results of the models with the ground-truth results, being the answers of a manual population of the same bench-mark.

The results suggest that Large Language models do have the potential of populating GovTech benchmarks. The best-performing model has an accuracy of 0.29 when populating the GovTech Maturity Index, which indicates that the model is certainly capable of answering the benchmark questions. When limiting the analysis to the multiple-choice questions, the accuracy of the highest model outperforms

the random chance accuracy. This indicates that the model does not only comprehend the questions being asked, but also has sufficient information to correctly answer the questions.

> ***(SQ-4)*** *What is the impact on the accuracy of the Large Language Model when relevant information is given as context?*

The influence of the context on the model performance is assessed using a full-factorial design, including a long prompt, context from a government sources database, and a fine-tuned model.

The most interesting observation emerging from the results is that the long prompt influences the model performance the most. Compared to the base-model, the models that include the long prompt are better able to comprehend the question and answer format. The fine-tuned model, on the contrary, only works without the context of the database, which indicates that it is not able to handle a long context. The context of the database does not improve performance, indicating that the current models do not handle the length of the context well.

> ***(SQ-5)*** *What are the practical limitations for the Large Language Models on populating the GovTech framework?*

To assess the practical limitations, the outcomes of the models are manually assessed. From this, a list of common pitfalls is made, which is interpreted and put into the perspective of policy makers.

There are several common limitations found in the model answers to the bench-mark questions, including repeating the input prompts, making up answers, repeating answers, and incomplete answers. However, not all mistakes are equally impacting the ability of the models to assist in the bench-marking process. The main limitation influencing the value of the model is that the models make up wrong answers. This does not help policy makers answer the question. The other limitations, although being a limitation of the models, can still create useful answers for policy makers to work with.

> ***(SQ-6)*** *What is the impact of using Large Language Models for GovTech framework population on the assessment of GovTech by the Dutch government?*

The impact of the artefact on the assessment of GovTech by the Dutch government is assessed by conducting an interview with two experts and performing an analysis of ethical concerns regarding the artefact.

The most striking observation that emerges from the analysis of the interview is that the artefact allows for an increase in speed for the bench-marking process. This results in better utilisation of limited government resources. Furthermore, as more benchmarks can be populated with the same amount of resources, areas of improvement can be identified better, and policy prioritisation can be adjusted more precisely and substantiated by more evidence.

From the consideration of ethical concerns, the most striking observation is that the utilisation of Large Language Models to benchmark GovTech introduces a wide range of ethical concerns. In particular, it is impossible to verify that the information used is actually the most relevant information available and whether this information is sufficient to answer the question. It is essential that these ethical concerns are considered when the artefact is used to benchmark GovTech, as they shine light on the ethical impact when the artefact is used.

> ***(RQ)*** *What is the usability of Large Language Models for benchmarking the state of GovTech?*

The Large Language Models for bench-marking the state of GovTech show potential in addressing practical limitations of the current benchmarking process, such as the resource-intensiveness, timeliness, and data integrity. The models show a promising accuracy for populating GovTech benchmarks, even though the identified limitations leave room for technical improvements. The inclusion of relevant context significantly improves model performance and the use of these models can result in increased speed of the benchmarking process, better utilisation of government resources, and more informed policy decisions. When using the artefact, it is essential to acknowledge the identified ethical concerns. Overall, Large Language Models show potential for improving the efficiency and effectiveness of bench-marking GovTech.

7.2. Next Steps

In order to guide the progression of the research project, a roadmap is proposed for the next steps, shown in Figure 7.1. The roadmap is made up of four steps.

As a first step, it is proposed to continue the development of the artefact. As discussed in section 6.7, various technical limitations affect the performance of the artefact. Addressing these limitations results in an increase in the artefact's capabilities.

Next, as the second step, it is proposed to perform an extensive analysis of the ethics of the artefact. Although the identification of ethical concerns in section 5.5 is a good start for an extensive ethical analysis, the next step is to address the concerns identified in the ethical analysis, as well as to extend this analysis with additional literature or expert knowledge. This is especially important because the artefact will be used in the public sector, indirectly influencing and making decisions for millions of people. In subsection 2.2.2, a public debate, forum or dialogue, and working on AI ethics guidelines is identified as techniques to facilitate AI regulation, and can be used as a starting point for a debate on the ethical concerns of the artefact.

Next, as the third step, it is proposed to carry out trials for the implementation of the artefact in the benchmarking process. When the artefact is sufficiently developed and ethical concerns are addressed, the artefact can be used in a trial to test the actual implementation into the GovTech benchmarking process. Digicampus, an established GovTech innovation facilitator, could provide guidance in this process, as they already have the necessary knowledge and experience to assist GovTech projects and have shown interest in the project.

Lastly, as the fourth step, it is proposed to start investigating broader application of the artefact. When the GovTech benchmarking implementation trials are deemed successful and the artefact performs well enough, broader applications for the artefact can be explored. Examples of broader applications include efforts to increase oversight of the Dutch GovTech market by tracking current GovTech initiatives and directly comparing The Dutch GovTech market to foreign GovTech markets. Again, Digicampus can be a valuable guide in this process, having an extensive European GovTech network and expertise on facilitating GovTech projects.

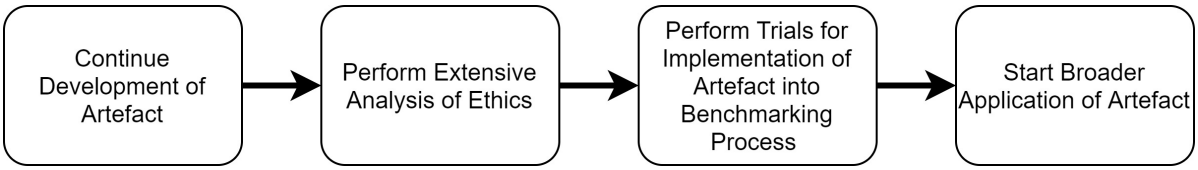

Figure 7.1: Roadmap for Next Steps.

References

Ahmed, M., Khan, H. U., & Munir, E. U. (2023). Conversational ai: An explication of few-shot learning problem in transformers-based chatbot systems. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2023.3281492

Alenezi, M. (2022). Understanding digital government transformation. https://doi.org/10.48550/arXiv.2202.01797

Alhyari, S., Alazab, M., Venkatraman, S., Alazab, M., & Alazab, A. (2013). Performance evaluation of e-government services using balanced scorecard: An empirical study in jordan. *Benchmarking*, *20*(4), 512–536. https://doi.org/10.1108/BIJ-08-2011-0063

Alothman, A. F., & Sait, A. R. W. (2022). Managing and retrieving bilingual documents using artificial intelligence-based ontological framework. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/4636931

Alzahrani, A., Stahl, B., & Prior, M. (2012). Developing an instrument for e-public services' acceptance using confirmatory factor analysis: Middle east context. *Journal of Organizational and End User Computing*, *24*(3), 18–44. https://doi.org/10.4018/joeuc.2012070102

Amaglobeli, D., de Mooij, R. A., Mengistu, A., Moszoro, M., Nose, M., Nunhuck, S., Pattanayak, S., del Paso, L. R., Solomon, F., Sparkman, R., Tourpe, H., & Uña, G. (2023). Transforming public finance through govtech. *Staff Discussion Notes*, *2023*(004), A001. https://doi.org/10.5089/9798400245480.006.A001

Angelis, V., Angelis-Dimakis, A., & Dimaki, K. (2014). A country's process of development as described by a cusp catastrophe model the case of eastern european and baltic countries. *Contributions to Economics*, *208*, 95–113. https://doi.org/10.1007/978-3-319-10133-0_6

Arora, A., & Arora, A. (2023). The promise of large language models in health care. *The Lancet*, *401*(10377), 641. https://doi.org/10.1016/S0140-6736(23)00216-7

Bain, R. (1937). Technology and state government. *American Sociological Review*, *2*(6), 860–874. Retrieved April 2, 2024, from http://www.jstor.org/stable/2084365

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 38176–38189, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/f978c8f3b5f399cae464e85f72e28503-Paper-Conference.pdf

Barcevicius, E., Cibaite, G., Codagnone, C., Gineikyte, V., Klimaviciute, L., Liva, G., Matulevic, L., Misuraca, G., & Vanini, I. (2019). *Exploring digital government transformation in the eu* (JRC Research Reports No. JRC118857). Joint Research Centre (Seville site). https://EconPapers.repec.org/RePEc:ipt:iptwpa:jrc118857

Bharosa, N. (2022). The rise of govtech: Trojan horse or blessing in disguise? a research agenda. *Government Information Quarterly*, *39*(3), 101692. https://doi.org/10.1016/j.giq.2022.101692

Binnenlands Bestuur. (n.d.). *Binnenlands bestuur*. https://www.binnenlandsbestuur.nl/

Bornmann, L. (2013). Advances in information science. *Journal of the Association for Information Science and Technology*, *64*(2), 217–233. https://doi.org/10.1002/asi.22803

Castle, S. (n.d.). Exploring cross-border interoperability in the public sector–a case study on eucaris as a successful eu-wide data exchange initiative. https://digikogu.taltech.ee/en/Download/efc1e472-4d31-419e-8c6d-a204489cc488/Piirilesekoostalitlusvimeuurimineavalikussek.pdf

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. https://doi.org/10.48550/arXiv.2310.14735

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017, July). Reading Wikipedia to answer open-domain questions. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1870–1879). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1171

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Chun, A. Y., Larsen, M. D., Durrant, G., & Reiter, J. P. (2021). *Administrative records for survey methodology*. https://doi.org/10.1002/9781119272076

Dener, C., Nii-Aponsah, H., Ghunney, L. E., & Johns, K. D. (2021). *Govtech maturity index: The state of public sector digital transformation*. World Bank Publications. https://doi.org/10.1596/978-1-4648-1765-6

Desmond, J., & Kotecha, B. (2017). State of the uk govtech market. *Retrieved from public. io*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*, 645–666. https://doi.org/0.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, *47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Edelmann, N., Steiner, K., & Misuraca, G. (2023). The view from the inside: A case study on the perceptions of digital transformation phases in public administrations. *Digit. Gov.: Res. Pract.*, *4*(2). https://doi.org/10.1145/3589507

Engin, Z., & Treleaven, P. (2019). Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *Computer Journal*, *62*(3), 448–460. https://doi.org/10.1093/comjnl/bxy082

Gao, Y., & Janssen, M. (2020). Generating value from government data using ai: An exploratory study. *Electronic Government*, 319–331. https://doi.org//10.1007/978-3-030-57599-1_24

Giray, L. (2023). Prompt engineering with chatgpt: A guide for academic writers. *Annals of biomedical engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

González Vázquez, I., Ranga, M., Marques Santos, A., Madrid, C., & Stierna, J. (2022). Partnerships for regional innovation–playbook.

Government of the Netherlands. (n.d.). *Govtech in the netherlands: Building a leading govtech nation*. https://www.government.nl/documents/reports/2021/06/30/govtech-in-the-netherlands

GovTechNL. (n.d.). *Govtechnl*. https://www.govtechnl.nl/

Hadi, H., Elhassani, I., & Sekkat, S. (2021). Electronic public services in the ai era. In M. Ben Ahmed, İ. Rakıp Kara□, D. Santos, O. Sergeyeva, & A. A. Boudhir (Eds.), *Innovations in smart cities applications volume 4* (pp. 70–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-66840-2_6

Harrison, T. M., & Luna-Reyes, L. F. (2022). Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review*, *40*(2), 494–511. https://doi.org/10.1177/0894439320980122

Helsper, E. J., & van Deursen, A. J. A. M. (2017). Do the rich get digitally richer? quantity and quality of support for digital engagement. *Information, Communication & Society*, *20*(5), 700–714. https://doi.org/10.1080/1369118X.2016.1203454

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75–105. https://doi.org/10.2307/25148625

Hillebrand, L., Berger, A., Deußer, T., Dilmaghani, T., Khaled, M., Kliem, B., Loitz, R., Pielka, M., Leonhard, D., Bauckhage, C., & Sifa, R. (2023). Improving zero-shot text matching for financial auditing with large language models. *Proceedings of the ACM Symposium on Document Engineering 2023*. https://doi.org/10.1145/3573128.3609344

Hoekstra, M., Van Veenstra, A. F., & Bharosa, N. (2023). Success factors and barriers of govtech ecosystems: A case study of govtech ecosystems in the netherlands and lithuania. *Proceedings of the 24th Annual International Conference on Digital Government Research*, 280–288. https://doi.org/10.1145/3598469.3598500

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. https://doi.org/10.48550/arXiv.2106.09685

Huang, K., Huang, G., Duan, Y., & Hyun, J. (2024). Utilizing prompt engineering to operationalize cybersecurity. In *Generative ai security: Theories and practices* (pp. 271–303). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54252-7_9

Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., Bensalem, S., Mu, R., Qi, Y., Zhao, X., Cai, K., Zhang, Y., Wu, S., Xu, P., Wu, D., Freitas, A., & Mustafa, M. A. (2023). A survey of safety and trustworthiness of large language models through the lens of verification and validation. https://doi.org/10.48550/arXiv.2305.11391

iBestuur. (n.d.). *Ibestuur*. https://ibestuur.nl/type/artikelen/

Janssen, M., Rana, N. P., Slade, E. L., & Dwivedi, Y. K. (2021). Trustworthiness of digital government services: Deriving a comprehensive theory through interpretive structural modelling. *Digital Government and Public Management*, 15–39.

Jia, K., & Zhang, N. (2022). Categorization and eccentricity of ai risks: A comparative study of the global ai guidelines. *Electronic Markets*, *32*(1), 59–71. https://doi.org/10.1007/s12525-021-00480-5

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b. https://doi.org/10.48550/arXiv.2310.06825

Jiang, H. (2021). Smart urban governance in the 'smart' era: Why is it urgently needed? *Cities*, *111*, 103004. https://doi.org/10.1016/j.cities.2020.103004

Jo, E., Epstein, D. A., Jung, H., & Kim, Y.-H. (2023). Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581503

Komatsu, T. T. (2019). *Transforming public sector organizations through design culture: The relationship between design practice, innovation and organizational change* [Doctoral dissertation]. Politecnico di Milano.

Kong, I., Janssen, M., & Bharosa, N. (2024). Realizing quantum-safe information sharing: Implementation and adoption challenges and policy recommendations for quantum-safe transitions. *Government Information Quarterly*, *41*(1), 101884. https://doi.org/10.1016/j.giq.2023.101884

Koryzis, D., Dalas, A., Spiliotopoulos, D., & Fitsilis, F. (2021). Parltech: Transformation framework for the digital parliament. *Big Data and Cognitive Computing*, *5*(1). https://doi.org/10.3390/bdcc5010015

Lamptey, O., Gegov, A., Ouelhadj, D., Hopgood, A., & Da Deppo, S. (2023). Neural network based identification of terrorist groups using explainable artificial intelligence. *2023 IEEE Conference on Artificial Intelligence (CAI)*, 191–192. https://doi.org/10.1109/CAI54212.2023.00090

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks. https://doi.org/10.48550/arXiv.2005.11401

Linegar, M., Kocielnik, R., & Alvarez, R. M. (2023). Large language models and political science. *Frontiers in Political Science*, *5*, 1257092. https://doi.org/10.3389/fpos.2023.1257092

Liu, Z., Qiao, A., Neiswanger, W., Wang, H., Tan, B., Tao, T., Li, J., Wang, Y., Sun, S., Pangarkar, O., Fan, R., Gu, Y., Miller, V., Zhuang, Y., He, G., Li, H., Koto, F., Tang, L., Ranjan, N., … Xing, E. P. (2023). Llm360: Towards fully transparent open-source llms. https://doi.org/10.48550/arXiv.2312.06550

Logius. (n.d.). *About digid*. https://www.digid.nl/en/about-digid

Lukkien, B., Bharosa, N., & de Reuver, M. (2023). *Barriers for developing and launching digital identity wallets* (tech. rep.). EasyChair.

Lythreatis, S., Singh, S. K., & El-Kassar, A.-N. (2022). The digital divide: A review and future research agenda. *Technological Forecasting and Social Change*, *175*, 121359. https://doi.org/10.1016/j.techfore.2021.121359

Manny, L. (2022). *Socio-technical challenges towards smart urban water systems* [Doctoral dissertation, ETH Zurich].

Mathews, N. S., Brus, Y., Aafer, Y., Nagappan, M., & McIntosh, S. (2024). Llbezpeky: Leveraging large language models for vulnerability detection. https://doi.org/10.48550/arXiv.2401.01269

Mazzi, F. (2023). Concerted actions to integrate corporate social responsibility with ai in business: Two recommendations on leadership and public policy. In R. Schmidpeter & R. Altenburger

(Eds.), *Responsible artificial intelligence: Challenges for sustainable management* (pp. 251–266). Springer International Publishing. https://doi.org/10.1007/978-3-031-09245-9_13

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Moghavvemi, S., & Mohd Salleh, N. A. (2014). Effect of precipitating events on information system adoption and use behaviour. *Journal of Enterprise Information Management*, *27*(5), 599–622. https://doi.org/10.1108/JEIM-11-2012-0079

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., et al. (2010). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *International journal of surgery*, *8*(5), 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

Murati-Leka, H., & Fetai, B. (2023). Government and innovation performance: Evidence from the ict enterprising community. *Journal of Enterprising Communities*, *17*(3), 621–643. https://doi.org/10.1108/JEC-12-2021-0174

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nassiri, K., & Akhloufi, M. (2022). Transformer models used for text-based question answering systems. *Applied Intelligence*, *53*(9). https://doi.org/10.1007/s10489-022-04052-8

Nweke, L. O. (2023). National identification systems as enablers of online identity. In D. R. Raja & D. A. K. Dewangan (Eds.), *Online identity - an essential guide*. IntechOpen. https://doi.org/10.5772/intechopen.1002294

Pan, J. J., Wang, J., & Li, G. (2023). Survey of vector database management systems. https://doi.org/10.48550/arXiv.2310.14021

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. https://doi.org/10.18653/v1/D19-1250

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

Rijksoverheid. (n.d.-a). *Open data: Documenten*. https://www.rijksoverheid.nl/opendata/documenten

Rijksoverheid. (n.d.-b). *Open data: Vac's*. https://www.rijksoverheid.nl/opendata/vac-s

Saeed, A., Haq, Z. U., & Iqbal, J. (2023). Investigating the factors affecting research and development expenditure efficiency in china and india. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-023-01258-0

Santos, D., Auquilla, A., Siguenza-Guzman, L., & Peña, M. (2021). A methodological framework for creating large-scale corpus for natural language processing models. *Communications in Computer and Information Science*, *1456 CCIS*, 87–100. https://doi.org/10.1007/978-3-030-89941-7_7

Shim, H., & Kim, J. (2023). A study on project prioritisation and operations performance measurements by the analysis of local financial investment projects in korea. *Sustainability (Switzerland)*, *15*(7). https://doi.org/10.3390/su15075972

Silve, A., & Moszoro, M. (2023). The political economy of govtech. *IMF Notes*, *2023*(003), A001. https://doi.org/10.5089/9798400246500.068.A001

Sorensen, T., Robinson, J., Rytting, C., Shaw, A., Rogers, K., Delorey, A., Khalil, M., Fulda, N., & Wingate, D. (2022). An information-theoretic approach to prompt engineering without ground truth labels. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.60

Steingard, D., Balduccini, M., & Sinha, A. (2023). Applying ai for social good: Aligning academic journal ratings with the united nations sustainable development goals (sdgs). *AI & SOCIETY*, *38*(2), 613–629. https://doi.org/10.1007/s00146-022-01459-2

Svahn, M., Larsson, A., Macedo, E., & Bandeira, J. (2023). Construct hunting in govtech research: An exploratory data analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14130 LNCS*, 3–17. https://doi.org/10.1007/978-3-031-41138-0_1

Tantawy, A. N. (2022). Government improvement through scientific research and innovation. *2022 32nd International Conference on Computer Theory and Applications (ICCTA)*, 11–13. https://doi.org/10.1109/ICCTA58027.2022.10206152

TenderNed. (n.d.). *Tenderned*. https://www.tenderned.nl/aankondigingen/overzicht

The World Bank. (n.d.). *2022 govtech maturity index update*. https://www.worldbank.org/en/programs/govtech/2022-gtmi

United Nations. (n.d.). *Take action for the sustainable development goals - united nations sustainable development*. https://www.un.org/sustainabledevelopment/sustainable-development-goals/

van Winden, W., & Carvalho, L. (2019). Intermediation in public procurement of innovation: How amsterdam's startup-in-residence programme connects startups to urban challenges. *Research Policy*, *48*(9), 103789. https://doi.org/10.1016/j.respol.2019.04.013

van Dam, T., van der Heijden, F., de Bekker, P., Nieuwschepen, B., Otten, M., & Izadi, M. (2024). Investigating the performance of language models for completing code in functional programming languages: A haskell case study. *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, 91–102. https://doi.org/10.1145/3650105.3652289

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Wagle, S., Munikoti, S., Acharya, A., Smith, S., & Horawalavithana, S. (2023). Empirical evaluation of uncertainty quantification in retrieval-augmented language models for science. https://doi.org/10.48550/arXiv.2311.09358

Wu, J., & Guo, D. (2015). Measuring e-government performance of provincial government website in china with slacks-based efficiency measurement. *Technological Forecasting and Social Change*, *96*, 25–31. https://doi.org/10.1016/j.techfore.2015.01.007

Yang, X., Wilson, S. D., & Petzold, L. (2024). Quokka: An open-source large language model chatbot for material science. https://doi.org/10.48550/arXiv.2401.01089

Zhang, H., & Zhang, Q. (2020). Minsearch: An efficient algorithm for similarity search under edit distance. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 566–576. https://doi.org/10.1145/3394486.3403099

Zhang, M., Bao, Y., Yang, Y., Kimber, M., Levine, M., & Xie, F. (2023). Identifying attributes for a value assessment framework in china: A qualitative study. *PharmacoEconomics*, *41*(4), 439–455. https://doi.org/10.1007/s40273-022-01235-6

A

GovTech Maturity Index

Table A.1: Govtech Maturity Index, from Dener et al. (2021).

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-1	Is there a shared cloud platform available for all government entities?	0= No, 1= Only cloud strategy/policy (no platform yet), 2= Yes (platform in use)
I-1.1	Name of the Government Cloud platform	Text
I-1.2	Cloud platform / strategy URL	URL
I-1.3	Government Cloud was launched / will be launched in (year)	YYYY
I-1.4	Type of cloud platform established	0= Unknown, 1= Public (Commercial), 2= Private (Government), 3= Hybrid
I-1.5	Official name of the entity operating the Government Cloud platform	Text
I-1.6	Government Cloud data hosting policy?	0= No policy / Unknown, 1= Keeping data inside the country, 2= Keeping data outside the country, 3= Hybrid (inside + outside)
I-1.6.1	If there is a cloud hosting policy >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-1.7	Cloud services provided	0= Unknown, 1= SaaS, 2= PaaS, 3= IaaS, 4= XaaS
I-1.8	Is there one shared Government Cloud platform or several?	0= Unknown, 1= Several cloud platforms (Central/Local levels), 2= One shared cloud platform
I-1.9	Monitoring & publishing of cloud usage, security, savings, etc.?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-1.9.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-2	Is there a government enterprise architecture framework?	0= No, 1= In draft / Planned, 2= Yes
I-2.1	Name of the GEA framework	Text
I-2.2	GEA framework / draft URL	URL
I-2.3	GEA was introduced / will be introduced in (year)	YYYY
I-2.4	GEA operational status	0= Unknown, 1= Partially used, 2= Extensively used
I-2.5	GEA scope >Is there a shared GEA?	0= Unknown, 1= Fragmented (Separate Central/Local), 2= Shared Central+Local (WoG)
I-2.6	Which entity is maintaining/extending GEA?	0= Unknown, 1= Ministry level Chief Information Officer 2= Government CIO, 3= Other
I-2.7	Which entity is monitoring compliance with GEA?	0= Unknown, 1= Ministry level Chief Information Officer 2= Government CIO, 3= Other

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-2.8	Monitoring & publishing of GEA usage, compliance, benefits, etc.?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-2.8.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-3	Is there a government interoperability framework?	0= No, 1= In draft / Planned, 2= Yes
I-3.1	Title of the GIF report	Text
I-3.2	GIF report / draft URL	URL
I-3.3	GIF was introduced / will be introduced in (year)	YYYY
I-3.4	GIF operational status	0= Unknown, 1= Partially used, 2= Extensively used
I-3.5	GIF scope >Is there a shared GIF?	0= Unknown, 1= Fragmented (Separate Central/Local), 2= Shared Central+Local (WoG)
I-3.6	Is there a data quality framework?	0= No, 1= Yes
I-3.7	Is there a system to monitor the 'uptime' of government information systems?	0= No, 1= Yes
I-3.8	Is there guidance for replacing legacy government information systems?	0= No, 1= Yes
I-3.9	Monitoring & publishing of GIF usage, compliance, benefits?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-3.9.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-4	Is there a government service bus platform?	0= No, 1= In draft / Planned, 2= Yes (platform in use)
I-4.1	Name of the Government Service Bus platform	Text
I-4.2	GSB platform URL	URL
I-4.3	GSB platform was launched / will be launched in (year)	YYYY
I-4.4	GSB operational status	0= Unknown, 1= Partially used, 2= Extensively used
I-4.5	GSB scope >Is there a shared GSB platform?	0= Unknown, 1= Fragmented (Separate Central/Local), 2= Shared Central+Local (WoG)
I-4.6	Monitoring & publishing of GSB usage, security, savings?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-4.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-5	Is there an operational FMIS in place to support core PFM functions?	0= No/Unknown, 1= Implementation in progress, 2= Yes (in use)
I-5.1	Official name of Finance Ministry / Department operating FMIS	Text
I-5.2	Finance Ministry / Department home page URL	URL
I-5.3	Name of the FMIS platform	Text
I-5.4	FMIS platform URL	URL
I-5.5	FMIS was launched / will be launched in (year)	YYYY
I-5.6	FMIS functional capabilities	0= Unknown, 1= Treasury (execution), 2= T + Budget (preparation), 3= T + B + Other
I-5.7	Scope of FMIS (coverage of budgets)	0= Unknown, 1= Central gov, 2= Central + Local gov
I-5.8	Type of FMIS software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-5.9	Name of FMIS software package: If custom software >Please enter LDSW (Locally Developed Software). If commercial package >Please enter the name of Commercial-off-the-shelf (COTS) package. If hybrid >Please enter LDSW + COTS package name	Text

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-5.10	Is there a unified budget classification/chart of accounts?	0= No 1= Yes (Central government only) 2= Yes (Both central and sub-national government)
I-5.11	Does FMIS capture expenses linked to the SDGs and other strategic goals?	0= No 1= Partially 2= Extensively
I-5.12	Does FMIS capture non-financial data (KPIs) on programs/projects?	0= No 1= Partially (Some of the KPIs for selected programs) 2= Extensively (KPIs captured for most of the programs)
I-5.13	Does FMIS exchange data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-5.14	Governance of FMIS operations (compliance, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-5.14.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-6	Is there a TSA supported by FMIS to automate payments and bank reconciliation?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-6.1	Treasury home page URL	URL
I-6.2	Treasury was established / will be established in (year)	YYYY
I-6.3	TSA regulation / introduction website URL	URL
I-6.4	TSA was launched / will be launched in (year)	YYYY
I-6.5	Scope of TSA operations	0= Unknown, 1= Partially used by the MDAs, 2= Extensively used by all MDAs
I-6.6	Type of electronic payment systems in place	0= Unknown, 1= RTGS (Real-Time Gross Settlement), 2= FPS / ACH (Fast Payment System / Automated Clearinghouse System), 3= Both RTGS & FPS/ACH
I-6.7	Is there a TSA interface linking FMIS with the Central Bank systems?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-6.8	Governance of TSA operations (compliance, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-6.8.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-7	Is there a Tax Management Information System in place?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-7.1	Tax Administration home page URL	URL
I-7.2	Tax Administration was established / will be established in (year)	YYYY
I-7.3	Name of the TMIS platform	Text
I-7.4	TMIS platform URL	URL
I-7.5	TMIS was launched / will be launched in (year)	YYYY
I-7.6	Type of TMIS software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-7.7	Does TMIS exchange data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-7.8	Governance of TMIS operations (compliance, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-7.8.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-8	Is there a Customs Management Information System in place?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-8.1	Customs Administration home page URL	URL
I-8.2	Customs Administration was established / will be established in (year)	YYYY
I-8.3	Name of the CMIS	Text
I-8.4	CMIS platform URL	URL

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-8.5	CMIS was launched / will be launched in (year)	YYYY
I-8.6	Type of CMIS software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-8.7	Customs and Tax administrations merged?	0= No, 1= Yes
I-8.8	Does CMIS exchange data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-8.9	Governance of CMIS operations (compliance, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-8.9.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-9	Is there a Human Resources Management Information System with self-service portal?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-9.1	Name of the HRMIS platform (public sector)	Text
I-9.2	HRMIS platform URL	URL
I-9.3	HRMIS was launched / will be launched in (year)	YYYY
I-9.4	Type of HRMIS software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-9.5	HRMIS topology	0= Unknown, 1= Disconnected, 2= Distributed, 3= Connected, 4= Shared
I-9.6	Is there a HRMIS self-service portal for employees and managers?	0= No 1= Yes (but there are still manual processes/paperwork) 2= Yes (most of the services are online/digitized)
I-9.7	Does HRMIS exchange data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-9.8	Does HRMIS use national ID as primary or secondary identifier?	0= No, 1= Yes
I-9.9	Governance of HRMIS operations (registers, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-9.9.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-10	Is there a Payroll System (MIS) linked with HRMIS?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-10.1	Name of the Payroll System (public sector)	Text
I-10.2	Payroll System (MIS) URL	URL
I-10.3	Payroll System was launched / will be launched in (year)	YYYY
I-10.4	Type of Payroll System software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-10.5	Payroll System topology	0= Unknown, 1= Disconnected, 2= Distributed, 3= Connected, 4= Shared
I-10.6	Governance of Payroll System operations (registers, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-10.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-11	Is there a Social Insurance system (non-health) providing pensions (including public sector) and other SI programs?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-11.1	Official name of the main public entity operating SI / Pension program(s)	Text
I-11.2	Main SI / Pension entity's home page URL	URL
I-11.3	Main SI / Pension entity was established / will be established in (year)	YYYY
I-11.4	Name of the primary SI / Pension system (MIS) solution	Text

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-11.5	Primary SI / Pension MIS platform URL	URL
I-11.6	Primary SI / Pension MIS was launched / will be launched in (year)	YYYY
I-11.7	Status of public sector SI / Pension MIS platform	0= Unknown, 1= Separate SI / Pension MIS for public employees, 2= Primary pension MIS is used for public employees as well
I-11.8	Type of primary SI / Pension MIS platform	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-11.9	Is primary SI / Pension MIS exchanging data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-11.10	Does the primary SI / Pension MIS use national ID as primary or secondary identifier?	0= No, 1= Yes
I-11.11	Are all SI / Pension beneficiary records fully digitized?	0= No, 1= Partially digitized, 2= Fully digitized
I-11.12	Share of all SI / Pension benefit payments deposited digitally to individual bank accounts of beneficiaries (percentage)	Text (% of payments or Unknown)
I-11.13	Are all active insured public employee records fully digitized?	0= No, 1= Partially digitized, 2= Fully digitized
I-11.13.1	If Digitized >In what year was digitization introduced for the contribution records? (year)	YYYY
I-11.14	Can a contribution report and payment be submitted online?	0= No, 1= Yes
I-11.15	Governance of SI / Pension MIS operations (registers, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-11.15.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-12	Is there an e-Procurement portal?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-12.1	Name of eProcurement Portal	Text
I-12.2	e-Procurement Portal URL	URL
I-12.3	e-Procurement Portal was launched / will be launched in (year)	YYYY
I-12.4	e-Procurement Portal capabilities	0= Unknown, 1= Tender notices + Contracts, 2= Online Tendering + Contracts, 3= OT + C + Interfaces with other systems
I-12.5	e-Procurement data published in line with OCDS?	0= No, 1= Yes
I-12.6	Does eProcurement Portal exchange data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-12.7	Any innovative approach in e-Procurement?	Text
I-12.8	Governance of eProcurement operations (registers, security, audit trails, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-12.8.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-13	Is there a Debt Management System (DMS) in place? (foreign and domestic debt)	0= No, 1= Implementation in progress, 2= Yes (in use)
I-13.1	Official name of Debt Management System (DMS) operator	Text
I-13.2	DMS platform / operator home page URL	URL
I-13.3	DMS platform was launched / will be launched in (year)	YYYY
I-13.4	Type of DMS software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-13.5	Abbreviation of DMS software solution	Text (short)

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-14	Is there a Public Investment Management System (PIMS) in place?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-14.1	Name of PIMS solution (information system)	Text
I-14.2	PIMS platform URL	URL
I-14.3	PIMS was launched / will be launched in (year)	YYYY
I-14.4	Type of PIMS software	0= Unknown, 1= Custom Software, 2= Commercial/COTS Software, 3= Hybrid (Custom + COTS)
I-14.5	PIMS functional capabilities	0= Unknown, 1= Only PIM project registry, 2= Registry + PIM cycle, 3= Registry + PIM cycle + Project monitoring
I-14.6	Does PIMS exchange data with other systems?	0= No, 1= Yes (via separate interfaces) 2= Yes (via Government Service Bus)
I-14.7	Publishing of PIMS project database, results?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-14.7.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-15	Is there a government Open Source Software policy/action plan for public sector?	0= No, 1= Yes (Advisory/R&D), 2= Yes (Mandatory)
I-15.1	OSS policy URL	URL
I-15.2	OSS policy was approved / will be approved in (year)	YYYY
I-15.3	Is there an entity taking decisions on adopting/procuring an OSS solution?	0= Unknown, 1= Ministry level Chief Information Officer 2= Government CIO, 3= Other
I-15.4	What is the level of adoption of OSS policy?	0= Unknown, 1= Partially adopted in several sectors, 2= Extensively adopted
I-15.4.1	If adopted >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-16	UN Telecommunication Infrastructure Index (TII)	0 to 1 (external indicator extracted from the UN e-Gov Survey)
I-17	Does government have a national strategy on disruptive / innovative technologies?	0= No, 1= In draft / Planned, 2= Yes
I-17.1	Title of the latest Disruptive Technology (DT) strategy document	Text
I-17.2	DT strategy URL	URL
I-17.3	DT strategy was approved / will be approved in (year)	YYYY
I-17.4	DT strategy focus area(s) [please select all that apply]	1= AI/ML, 2= Blockchain/DLT, 3= IoT, 4= Drones, 5= Other (Smart Cities, Robotics, Virtual Reality, 3D printers, etc.)
I-17.5	Is there a ministry/department responsible for implementing the DT strategy?	0= No 1= Yes
I-17.5.1	If Yes >Official name (and URL) of the responsible entity	Text / URL (if publicly available)
I-17.6	Does the DT strategy have committed funding?	0= No 1= Yes
I-17.7	Publishing of use cases on DT applications?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-17.7.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-18	UN Online Service Index (OSI)	0 to 1 (external indicator extracted from the UN e-Gov Survey)
I-19	Is there an online public service portal? (also called ”One-Stop Shop” or similar)	0= No, 1 = Yes (Informational: Level 1 or 2), 2= Yes (Transactional: Level 3 or 4)
I-19.1	Online service (e-Service) portal URL	URL
I-19.2	Are citizens / businesses involved in the design of e-Services (user-centric design)?	0= No, 1= Yes
I-19.3	Universal accessibility (omnichannel access)?	0= No, 1= Yes

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-19.4	Has the government released any mobile app for the citizens' access to public services?	0= No, 1= Yes
I-19.5	Can residents start a business through online service portal?	0= No, 1= Yes
I-19.6	Can individuals establish an e-residency through online service portal?	0= No, 1= Yes
I-19.7	Publishing of online service delivery performance/user experience?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-19.7.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-20	Is there a Tax online service portal?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-20.1	Tax System service portal URL	URL
I-20.2	Available Tax online transactional services	0= Unknown, 1= Registration + Filing 2= R + F + Payment, 3= R + F + P + Other
I-20.3	Electronic Invoicing	0= No, 1= Planned / In progress 2= Partially implemented 3= Fully implemented / mandatory
I-20.4	Are citizens/businesses involved in the design of tax online services?	0= No, 1= Yes
I-20.5	Universal accessibility (omnichannel access)?	0= No, 1= Yes
I-21	Is e-Filing available for tax and/or customs declarations?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-21.1	e-Filing service URL (or explanation / report)	Enter URL (public link) or Attach relevant report
I-21.2	Type of e-Filing service	0= Unknown 1= Online e-Filing services 2= Online e-Filing + e-Payments
I-21.3	Available for all tax types and customs declarations?	0= No, 1= Yes
I-21.4	Available services for interconnectivity with business information systems?	0= No, 1= Yes
I-21.5	Available pre-populated returns?	0= No, 1= Yes
I-22	Are e-Payment services available?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-22.1	e-Payment service URL (or explanation / report)	Enter URL (public link) or Attach relevant report
I-22.2	Type of e-Payment service	0= Unknown 1= Fragmented systems; multiple platforms, 2= Centralized shared platform
I-22.3	Available e-Payment methods?	0= Unknown, 1= Bank Transfer, 2= BT + Credit / Debit Cards, 3= BT + CC + Mobile, 4= BT + CC + M + Others
I-22.4	e-Payment service for government / treasury payments?	0= No, 1= Yes
I-23	Is there a Customs online service portal (single window)?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-23.1	Customs System service portal URL	URL
I-23.2	Available Customs online transactional services	0= Unknown, 1= Registration + Declaration, 2= R + D + Payments, 3= R + D + P + Other
I-23.3	Are citizens/businesses involved in the design of customs online services?	0= No, 1= Yes
I-23.4	Universal accessibility (omnichannel access)?	0= No, 1= Yes
I-24	Is there a Social Insurance/Pension online service portal?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-24.1	Social Insurance/Pension online service portal URL	URL
I-24.2	Available SI / Pension online transactional services	0= Unknown, 1= Registration + Benefits, 2= R + B + Payments, 3= R + B + P + Other

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-24.3	Are citizens involved in the design of SI / Pension services/portal?	0= No, 1= Yes
I-24.4	Does the Gov provide any incentives for citizens to join an insurance scheme?	0= No, 1= Yes
I-24.5	Universal accessibility (omnichannel access)?	0= No, 1= Yes
I-25	Is there a Job portal?	0= No, 1= Implementation in progress, 2= Yes (in use)
I-25.1	Job portal URL	URL
I-25.2	Available online transactional Job portal services	0= Unknown, 1= Registration + Search, 2= R + S + Applications, 3= R + S + A + Other
I-25.3	Inclusion of public sector positions in the Job portal?	0= Unknown, 1= Separate Job Portal for public employees, 2= Primary job portal is used for public employees as well
I-25.4	Are citizens/employees involved in the design of services/portal?	0= No, 1= Yes
I-25.5	Universal accessibility (omnichannel access)?	0= No, 1= Yes
I-26	Is there a digital ID [credential / system] that enables remote authentication for (fully) online service access / transactions?	0= No, 1= Yes (external indicator to be extracted from the 2022 ID4D dataset)
I-27	UN E-Participation Index (EPI)	0 to 1 (external indicator extracted from the UN e-Gov Survey)
I-28	Is there an Open Government portal?	0= No, 1 = Yes
I-28.1	Open Government portal URL	URL
I-28.2	Update frequency of Open Gov portal	0= Unknown, 1= Annually 2= Quarterly / Monthly, 3= Weekly / Daily
I-28.3	Contents / maturity of Open Gov portal	0= Unknown, 1= Basic info/datasets, 2= Comprehensive data catalog
I-29	Is there an Open Data portal?	0= No, 1 = Yes
I-29.1	Open Data portal URL	URL
I-29.2	Update frequency of Open Data portal	0= Unknown, 1= Annually 2= Quarterly / Monthly, 3= Weekly / Daily
I-29.3	Contents / maturity of Open Data portal	0= Unknown, 1= Basic info / datasets, 2= Comprehensive data catalog
I-29.4	Is the portal dynamically updated (via APIs)?	0= Unknown, 1= Yes (mostly manual), 2= Yes (automated updates via APIs)
I-30	Are there national platforms that allow citizens to participate in policy decision-making?	0= No, 1 = Yes
I-30.1	Citizen participation portal URL	URL
I-30.2	Is it possible to submit petitions?	0= No, 1= Yes (the same citizen participation portal for petitions as well), 2= Yes (separate portal for petitions)
I-30.2.1	If Yes (separate) >URL of the separate portal for submitting petitions	URL
I-30.3	Can citizens / businesses participate in policy decision-making through this platform?	0= No, 1= Yes
I-30.4	Can citizens / businesses provide anonymous feedback?	0= No, 1= Yes
I-30.5	Universal accessibility (omnichannel access)?	0= No, 1= Yes
I-30.6	Are government's responses to citizens / businesses publicly available?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-30.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-31	Are there government platforms that allow citizens to provide feedback (e.g., complements, complaints, suggestions, info requests) on service delivery?	0= No, 1 = Yes
I-31.1	Citizen feedback / GRM portal URL	URL

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-31.2	Does the government make the service standards (e.g., response times and procedures) available to the public?	0= No, 1= Yes
I-31.3	Are these platforms universally accessible or provide support for users with disabilities (e.g., e-services, availability of voice commands)?	0= No, 1= Yes
I-31.4	Is there any advanced technology (e.g., chatbots or AI-enabled discussion forums) used to improve citizen engagement?	0= No, 1= Yes
I-31.5	Universal accessibility (omnichannel access)?	0= No, 1= Yes
I-31.6	Does the Gov respond to citizen feedback? (how the Gov has updated their services in response to citizen feedback)	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-31.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-32	Does the government publish its citizen engagement statistics and performance regularly?	0= No, 1 = Yes
I-32.1	Government response portal URL	URL
I-32.2	Are there standards or indicators to measure the performance of service delivery (and compliance)?	0= No, 1= Yes
I-32.3	Does the government publish its citizen engagement performance/results?	0= No, 1= Yes
I-32.4	Any government initiative to improve the representation of vulnerable groups?	0= No, 1= Yes
I-33	Is there a government entity focused on GovTech (digital transformation, WoG, online services, etc.)?	0= No, 1= Planned / In progress, 2= Yes (Established)
I-33.1	Official name of the main GovTech institution	Text
I-33.2	Main GovTech institution URL	URL
I-33.3	Main GovTech institution was established / will be established in (year)	YYYY
I-33.4	Type of main GovTech organization	1= Government, 2= Private, 3= Investor, 4= Academia
I-33.5	Institutional responsibility for GovTech	1= Autonomous entity, 2= President's / PM's Office, 3= MoICT, 4= MoF / MoE, 5= MoI / MoHA, 6= MoPS / Pub Adm, 7= Other
I-33.6	GovTech roles & responsibilities [please select all that apply]	1= Policy / Strategy, 2= eGovernment / eServices, 3= Private Sector / PPP, 4= Digital skills, 5= KS&L, 6= Innovation, 7= OSS, 8= DT, 9= Other
I-33.7	Other relevant GovTech institution links	URL
I-33.8	Is there a Coordination Body (SC, Council) leading GovTech initiatives?	0= No, 1= Yes
I-33.8.1	If Yes >Name and/or URL of the coordination body	Text or URL
I-33.9	Is there an entity to monitor & report Digital/GovTech spending for the whole government?	0= No, 1= Yes
I-33.9.1	If Yes >Name and/or URL of the public entity	Text or URL
I-33.10	Publishing of the GovTech institution's annual progress report (results / spending)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-33.10.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-34	Is there a dedicated government entity in charge of data governance or data management?	0= No, 1= Planned / In progress, 2= Yes (Established)
I-34.1	Name of Data Governance (DG) institution	Text
I-34.2	Data Governance institution URL	URL
I-34.3	Data Governance institution was established / will be established in (year)	YYYY

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-34.4	Type of Data Governance institution	0= Unknown, 1= Part of another institution, 2= Autonomous institution
I-34.5	Data Governance implementation arrangements	0= Unknown, 1= Holistic DG approach, 2= Multilevel DG approach
I-34.6	Is there a Data Governance strategy / policy?	0= No, 1= Planned / In progress, 2= Yes (Approved)
I-34.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-34.7	Publishing of the Data Governance institution's progress report (results/spending)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-34.7.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-35	Is there a GovTech / Digital Transformation strategy?	0= No, 1= Planned / In draft, 2= Yes (old/to be updated), 3= Yes (new/current)
I-35.1	GovTech / digital transformation strategy URL (approved / drafted)	URL
I-35.2	GovTech strategy was approved / will be approved in (year)	YYYY
I-36	Is there a whole-of-government approach to public sector digital transformation?	0= No, 1= Planned / In draft, 2= Yes (Institutionalized)
I-36.1	Whole of Government (WoG) >Relevant policy/strategy URL	URL
I-36.2	Is there a Ministry / Dept leading the public sector digital transformation / cultural change / WoG approach?	0= No, 1= Yes
I-36.2.1	If Yes >Name and/or URL of the relevant public entity	Text or URL
I-36.3	Is there a cross government forum where strategic WoG topics (digital, data, technology, capacity) can be addressed by senior digital officials across government?	0= No, 1= Yes
I-36.3.1	If Yes >Name and/or URL of the relevant platform/-forum / entity	Text or URL
I-36.4	Publishing of the progress in WoG approach / digital transformation (results, spending)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-36.4.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-37	Are there RTI Laws to make data/info available to the public online or digitally?	0= No, 1= Draft / Consultations in progress, 2= Yes (Effective)
I-37.1	Right to Information (RTI) Law URL	URL
I-37.2	RTI Law was approved / will be approved in (year)	YYYY
I-37.3	Is there an entity monitoring implementation/compliance?	0= No, 1= Yes
I-37.4	Publishing of the progress in implementing RTI laws (RTI requests received, granted, etc.)?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-37.4.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-38	Is there a Data Protection / Privacy law?	0= No, 1= Draft / Consultations in progress, 2= Yes (Effective)
I-38.1	Official title of the Data Protection / Privacy Law	Text
I-38.2	Data Protection / Privacy Law URL	URL
I-38.3	Data Protection / Privacy Law was approved / will be approved in (year)	YYYY
I-38.4	Is there an entity monitoring implementation/compliance?	0= No, 1= Yes
I-38.5	Publishing of the data protection/privacy complaints and feedback?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-38.5.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-39	Is there a Data Protection Authority?	0= No, 1= Not established yet (visible in law), 2= Yes
I-39.1	Name of the Data Protection Authority	Text
I-39.2	Data Protection Authority URL	URL
I-39.3	Data Protection Authority was established / will be established in (year)	YYYY
I-39.4	Publishing of the Data Protection Authority performance/results?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-39.4.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-40	Is there a national ID (or similar foundational ID) system?	0= No, 1= Yes
I-41	Are records in the national ID system stored in a digitized (electronic) format?	0= No, 1= Yes, N/A = not applicable, no national ID system
I-42	Is there a digital signature regulation and PKI to support service delivery?	0= No, 1 = Regulation approved; no Infrastructure yet (PKI, CA), 2= Regulation and Infrastructure in place. Not used yet/in progress, 3= Operational. Used in practice for e-Services
I-42.1	Digital Signature URL	URL
I-42.2	Digital Signature was launched / will be launched in (year)	YYYY
I-42.3	Use of Digital Signature in public sector?	0= Unknown, 1= Back-office transactions, 2= Front-office service delivery, 3= Both back- and front-office transactions
I-42.4	Is Digital Signature linked with Digital ID/Mobile devices?	0= No, 1= Yes
I-42.5	Which entities provide Digital Signature services?	0= Unknown, 1= Commercial providers only, 2= Designated government entities, 3= Both government and commercial entities
I-42.6	Publishing of the Digital Signature issuance/utilization?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-42.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-43	ITU Global Cybersecurity Index (GCI)	0 to 100 (external indicator extracted from the ITU GCI)
I-44	UN Human Capital Index (HCI)	0 to 1 (external indicator extracted from the UN e-Gov Survey)
I-45	Is there a government strategy / program to improve digital skills in the public sector?	0= No, 1= Yes (Only strategy or program), 2= Yes (Both strategy and program)
I-45.1	Title of Digital Skills (DS) strategy	Text
I-45.2	Digital Skills strategy URL	URL
I-45.3	Digital Skills strategy was approved / will be approved in (year)	YYYY
I-45.4	Focus areas of the DS strategy	0= Unknown, 1= Basic Digital Skills, 2= Basic DS + Data Literacy, 3= Advanced DS + DL
I-45.5	Is there a DS program?	0= No, 1= Yes
I-45.5.1	If Yes >Type of primary DS program(s)	0= Unknown, 1= Academic program, 2= Public sector program, 3= CSO/Private program
I-45.5.2	If Yes >DS program URL	URL
I-45.5.3	If Yes >DS program mandatory for new public employees?	0= Unknown, 1= Not mandatory, 2= Mandatory
I-45.6	Are there digital skills programs offered by governments for citizens/schools?	0= No, 1= Yes (fee-based programs), 2= Yes (freely available programs)
I-45.7	Publishing of the results/progress in DS programs?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)

Table A.1: (continued)

#	2022 GTMI Indicators & Sub-indicators	Response options & Data format
I-45.7.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-46	Is there a strategy and/or program to improve public sector innovation?	0= No, 1= Yes (Only strategy or program), 2= Yes (Both strategy and program)
I-46.1	Title of Public Sector Innovation strategy	Text
I-46.2	PSI strategy URL	URL
I-46.3	PSI strategy was approved / will be approved in (year)	YYYY
I-46.4	Is there a PSI program?	0= No, 1= Yes
I-46.4.1	If Yes >Type of primary PSI program(s)	0= Unknown, 1= Academic program, 2= Public sector program, 3= CSO/Private program
I-46.4.2	If Yes >PSI program URL	URL
I-46.4.3	If Yes >PSI program mandatory for new public employees?	0= Unknown, 1= Not mandatory, 2= Mandatory
I-46.5	Publishing of the results/progress in PSI programs?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-46.5.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-47	Is there a government entity focused on public sector innovation?	0= No, 1= Planned / In progress, 2= Yes (Established)
I-47.1	Name of the PSI institution	Text
I-47.2	PSI institution URL	URL
I-47.3	PSI institution was established / will be established in (year)	YYYY
I-47.4	Focus areas of PSI institution (Innovation Lab)	0= Unknown, 1= Digital Skills, 2= PS Innovation, 3= Digital Skills + PS Innovation, 4= Other
I-47.5	Is there any collaboration on PSI with the private sector?	0= No, 1= Yes
I-47.5.1	If Yes >Is there any financial support/incentive for private GovTech entities	0= No, 1= Yes
I-47.6	Publishing the PSI institution annual performance/results?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-47.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-48	Is there a government policy to support GovTech startups and private sector investments?	0= No, 1= Yes
I-48.1	National policy/strategy to support GovTech startups/investments (SMEs)	Text
I-48.2	National policy/strategy URL	URL
I-48.3	National policy/strategy was approved / will be approved in (year)	YYYY
I-48.4	Does the Government provide financing to startups/SMEs for innovation?	0= No, 1= Yes
I-48.5	Capacity of Government to deliver online services via PPPs	0= No PPP for online services 1= Yes, PPP arrangements exist for online service delivery
I-48.6	Is there a procurement policy aimed at prioritizing bids from startups/SMEs? (e.g., having a quota for SMEs)	0= No, 1= Yes
I-48.6.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report
I-48.7	Publishing of the results/progress in supporting startups/SMEs for innovation?	0= No, 1= Yes (internal, not published), 2= Yes (public, published)
I-48.7.1	If Yes >Supporting document (report / URL)	Enter URL (public link) or Attach relevant report

B

Interview Guide

Exploring the use of AI for GovTech / e-Government Benchmarks using a semi-structured interview structure.

B.1. Introduction

- Brief introduction of the interviewer and the purpose of the study.
- Confirm the duration of the interview & ask for consent to record the session.
- Assure confidentiality and explain how the data will be used.

B.2. Current State of GovTech / e-Government Benchmarks

Assessment of Current Benchmarks:

- How do you perceive the value of these benchmarks for policy making and implementation?
- In your experience, what are the primary challenges you encounter with the current GovTech / e-Government benchmarks?

Improving Benchmark Value:

- What improvements would you suggest increasing the practical value of these benchmarks?
- Can you provide examples where GovTech / e-Government benchmarks have directly influenced policy decisions effectively?

B.3. Specific Challenges in Benchmarking

Timeliness Issues:

- How significant is the challenge of timeliness in the current GovTech / e- Government bench-marks? Can you provide an example?
- What impact does delayed bench-marking have on policy making and implementation?

Complexity and Detail:

- Could you discuss any difficulties related to the complexity or simplicity of current benchmarks?
- Are there areas in GovTech / e-Government where you feel benchmarks oversimplify or over-complicate the issues?

Aggregation of Data:

- What are the challenges with data aggregation in the current bench-marking frameworks? (e.g. looking only at national level)
- How does this affect the accuracy or usefulness of the benchmarks?

Comparability of Results:

- To what extent are the framework results comparable across countries and time, given the current population method of interviews / surveys?
- To what extent would the current method allow reproduction of the results? What if this was done by different people? How does this influence the impact of the benchmark?

B.4. Role of AI in Addressing Benchmarking Challenges

Potential of AI Solutions:

- Are there examples of AI already being implemented in bench-marking processes? What results have they shown?
- How do you see AI technology addressing the challenges of timeliness, complexity, and data aggregation in bench-marking?
- What factors are important when operationalising benchmarks with LLMs?
  - Format
  - Context
  - Reasoning / Substantiation
  - Transparency

B.5. Comparative Assessment LLM Outputs vs. Official Data

For both Official Data and LLM outputs examples are showed:

- How accurate do you find the data?
- How consistent do you find the data and sources?
- Are there gaps in official data that LLMs have successfully filled?
- Are some types of questions maybe too complex for LLMs?

B.6. Future of LLMs in GovTech / e-Government Bench-marking

Impact on Policy Making:

- If the challenges identified are overcome with the help of AI and LLMs, what changes do you foresee in the usability of GovTech benchmarks for policymakers?
- How can LLMs assist Dutch bench-markers / policymakers, given the specific governance and technological landscape of the Netherlands?

Integration and Implementation:

- What are the potential barriers to integrating LLMs into existing GovTech / e- Government frameworks?
- What steps should be taken to ensure the effective implementation of LLMs in the bench-marking process?

B.7. Conclusion

Summarise key points discussed.

- Ask if there is anything the expert would like to add or clarify.
- Thanking expert for their time and insights.
- Discuss next steps and how the findings might be shared or used.