C. Snoeij

AI for GovTech

Exploring the use of LLMs for GovTech Benchmark Operationalization

AI for GovTech

Exploring the use of LLMs for GovTech Benchmark Operationalization

By

Corné Snoeij

Student number: 5174473

In partial fulfillment of the requirement for the degree of

Master of Science
In Engineering and Policy Analysis

At the Delft University of Technology,
To be defended publicly on June 28, 2024 at 10:00 AM.

First supervisor:	Prof. dr. ir. N. Bharosa	TU Delft
Second supervisor:	Dr. J. Durán	TU Delft

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

Preface

You are about to read a thesis report that explores the use of Artificial Intelligence (AI), specifically Large Language Models (LLMs), for operationalizing Government Technology (GovTech) benchmarks. This thesis was written as part of the course EPA2942 Master Thesis EPA for the Master of Engineering Policy Analysis at the Technical University Delft. From October 2023 to June 2024, I dedicated myself to this project and the writing of this thesis.

With AI ushering in a new era, and the world watching in awe, my curiosity got the best of me to dive into this field. Inspired by the work at Digicampus, I envisioned an LLM that could answer all questions about GovTech, analyze it thoroughly, and tackling seemingly impossible ideas. As I progressed, I learned to focus on a specific application: using LLMs to operationalize GovTech benchmarks. With this thesis now finished, numerous new possibilities for future applications of these models have emerged.

I want to express my gratitude to Nitesh Bharosa, whose enthusiasm and contagious energy made working on this project a joy each week. A special thanks to Juan Durán, whose critical eye has shaped this thesis into what it is today. I also want to thank Berend Nieuwschepen; collaborating with you on this project was a pleasure. Lastly, my deepest gratitude goes to Anne, my constant support at home.

I hope this thesis contributes meaningfully to the field of GovTech benchmarking, inspires further research and dialogue, and advances the way we work with benchmarks, making them more timely and, eventually, more detailed and tailored to policymakers' needs.

Thank you for taking the time to engage with my work, I wish you much reading pleasure.

Corné Snoeij
Gouda, June 14, 2024

Executive Summary

This research explores the use of Artificial Intelligence (AI), specifically Large Language Models (LLMs), into the operationalization of Government Technology (GovTech) benchmarks to increase their utility for policymakers. Research and practice consistently highlight persistent challenges in GovTech benchmarking, such as resource-intensive methodologies that provide retrospective rather than real-time analysis, a lack of complexity that overlooks digital infrastructures and emerging technologies in favor of simpler metrics, and improper levels of aggregation that can render results less useful.

This study addresses these issues by employing LLMs to mitigate inherent challenges of timeliness, complexity, and data aggregation in benchmarks. The societal relevance lies in enabling near real-time, more detailed insights at appropriate aggregation levels into GovTech developments, aiding policymakers in making better informed decisions. Considering that benchmarks can significantly influence political outcomes and shape the development of GovTech services, refining benchmarking methodologies using LLMs potentially improves the responsiveness and relevance of government actions that better serve societal needs.

Scientifically, this novel application of LLMs to GovTech benchmarking contributes to the academic discourse on digital government assessment, offering a novel approach to monitoring GovTech advancements. Since the beginning of this century, the same criticisms keep recurring again and again: retrospective insights rather than real-time analysis due to lengthy processes, too many supply-oriented benchmarks, and the importance of context and measuring regional levels. This study is the first to explore the use of LLMs for GovTech benchmarking, thereby contributing a new perspective that responds to the ongoing need for suitable and advanced GovTech benchmarking methods.

Using Design Science Research, an artifact is developed that combines an LLM with Retrieval-Augmented Generation (RAG), fine-tuning and prompt-engineering. Activity Theory facilitates the integration of LLMs into the GovTech benchmarking ecosystem by identifying their potential roles: as artefacts, community enablers, or autonomous subjects. Consistent with the Design Science Research Methodology of this study, which aims to develop an artefact that meets specific objectives, LLMs are predominantly positioned as artifacts. In this role, LLMs serve as advanced analytical tools that aid benchmarkers in processing and analyzing data, thereby effectively operationalizing benchmarks.

Results show that LLMs can operationalize the GovTech Maturity Index (GTMI) benchmark by the World Bank with varying degrees of accuracy, depending on the model configuration. Manual evaluations reveal that some LLM configurations achieve up to 29% accuracy across the full benchmark and 48% for multiple-choice questions, significantly surpassing the 37% accuracy expected from random guessing on multiple-choice questions. However, when assessed using exact match and edit similarity metrics, these models often exhibit lower accuracies. This indicates that while LLMs can provide responses that are contextually relevant, they frequently fall short of perfectly matching the ground truth data. Although the achieved accuracies are not particularly impressive, they are understandable in light of the complex and context-specific domain.

In conclusion, LLMs improve the utility of GovTech benchmarks for policymakers in the Netherlands. By reducing the data collection phase from months to minutes, LLMs enable faster operationalization of benchmarking frameworks, providing policymakers with up-to-date information. This faster processing capability also holds potential to handle more complex data and diverse aggregation levels, which are often restricted by existing time and resource limitations. Broader implications for the GovTech benchmarking process include more responsive policies, a potential reduction in subjectivity due to the consideration of multiple sources with RAG, and increasingly fair and sensitive policies by incorporating a broader range of parameters.

This study acknowledges several limitations in both the artefact and the research. Artefact limitations include using a modest 7B parameter base model, the application of LoRA, performance gaps in the non-English model used, and a limited dataset for RAG. Research limitations arise from the use of Design Science Research Methodology, which neglects the environment in which the artefact will be implemented, and Activity Theory, in its disregard for the broader social and political context and its lack of emphasis on ethical considerations. Additionally, manual evaluations may have introduced subjectivity, and focusing solely on the Dutch context and the GTMI benchmark might limit the generalizability of the findings.

Future research recommendations include ways to improve the model accuracy by tackling artefact limitations, broadening the geographical scope, and operationalizing other benchmarks to increase the generalizability of the findings. Policy recommendations for the Dutch government include advice to continue its proactive stance on AI, actively engaging with AI technologies and experimenting with their use in benchmarking contexts. Finally, the development of a benchmark specifically tailored for operationalization by LLMs is proposed, with a preliminary design for an AI-Supported GovTech Index (AGTI) outlined.

Contents

List of Figures

List of Tables

List of Abbreviations

AGTI	AI-supported GovTech Index
AI	Artificial Intelligence
CRUD	Create, Read, Update, Delete
DESI	Digital Economy and Society Index
DGI	Digital Government Index
DPO	Direct Preference Optimization
DSR	Design Science Research
DSRM	Design Science Research Methodology
EGDI	E-Government Development Index
GovTech	Government Technology
GPT	Generative Pre-training Transformer
GTMI	GovTech Maturity Index
IDA	Intelligent Digital Assistent
KM	Knowledge Management
KVK	Kamer van Koophandel (Dutch Chamber of Commerce)
LLM	Large Language Model
LM	Language Model
LoRA	Low-Rank Adaptation
NLP	Natural Language Processing
NN	Neural Networks
PEFT	Parameter-Efficient Fine-Tuning
RAG	Retrieval Augmented Generation
SFT	Supervised Fine-Tuning
SQuAD	Stanford Question Answering Dataset
TL	Transfer Learning
UNDESA	United Nations Department of Economic and Social Affairs

1 Introduction

1.1 Evaluating Government Operations: a Complex Necessity

Evaluating government operations is essential yet complex. It is essential because it allows policymakers to understand the outcomes and impacts of their actions and investments. This insight enables them to adjust or maintain specific policies, ensuring that government operations are aligned with the public good. However, evaluation is far from straightforward. Many of us may remember receiving an unexpectedly unfair grade during our school days, highlighting that evaluation is not simple math but involves interpretation and personal preferences.

Similarly, traditional methods for assessing government operations are rarely straightforward. A common approach is to analyze financial outcomes, questioning whether operations are cost-effective or revenue-generating and if they are executed efficiently. Typically, these analyses are conducted through Cost-Benefit Analyses. However, this financial perspective often overlooks the broader social, cultural, and environmental impacts (Mishan & Quah, 2020). Another frequently employed method involves conducting public surveys and gathering feedback through for example social media, which aids in evaluating public satisfaction and the perceived impact of governmental activities. Nevertheless, this method might not capture the full complexity and depth of issues within government operations, and response biases can skew results (Ceron & Negri, 2016).

In light of these limitations, benchmarking offers as a structured alternative that allows for a more nuanced comparison of performance and drives continuous improvement in public administration. In the next section, this benchmarking method is further explored.

1.2 Benchmarking

Benchmarking, initially used by civil engineers to compare against a set standard, is now a common method for making structured comparisons in many areas. In particular, it can be used to measure against higher standards with a view to enabling learning about how to close the gap (Papaioannou et al., 2006). Benchmarking usually involves two aspects: firstly, comparing performance levels to pinpoint discrepancies and potential learning sources from leading organizations; secondly, analyzing the strategies of top performers to potentially adapt and implement their successful practices. Benchmarking is characterized by Stapenhurst (2009) as "a method of measuring and improving our organization by comparing ourselves with the best".

The concept of benchmarking was developed by Xerox Corporation in 1979 and has since become a well-established tool for improving organizational performance and competitiveness, predominantly within the private sector. While the public sector operates under different success metrics, notably beyond mere profit margins, benchmarking has proven equally beneficial. It aids public organizations in evaluating the effectiveness of their services and assessing them efficiently (Gunasekaran, 2005). With an increasing focus on performance management and continuous improvement in government, benchmarking is nowadays also widely used within the public sector in the US and Europe (Triantafillou, 2007).

In the context of Government Technology (GovTech), the need for adequate evaluation is particularly pressing. This is shown clearly by the number of prominent benchmarks like the GovTech Maturity Index (GTMI) (Dener et al., 2021), the Digital Economy and Society Index (DESI) (European Commission, 2022), the E-Government Development Index (EGDI) (United Nations, 2022), the Digital Government Index (DGI)

(Ubaldi & Okubo, 2020), along with numerous national assessments (Public, 2021). All these benchmarks are designed to assess government's use of digital technology in public services, thereby keeping citizens informed and enabling public policy makers to make informed decisions and strategize effectively. However, what exactly is GovTech, and how does it relate to e-government? This will be addressed in the next section.

1.3 What is GovTech?

In the contemporary era of rapidly advancing technologies, the intersection of governance and technology has given rise to a field known as GovTech, an acronym of Government Technologies. In Europe, many GovTech initiatives are being launched regularly to accelerate technology adoption across the full spectrum of public services (Kuziemski et al., 2022). Illustrative examples include cyber-trust services for secure authentication, AI-enabled digital assistants for professionals assessing social benefit applications, voice assistants, and data wallets for citizen convenience.

Despite its prevalence, a widely accepted definition for GovTech is still in its early stages of development, with academic definitions being relatively scarce (Bharosa, 2022). For this research, the definition originating from the World Bank is used: “GovTech is a whole-of-government approach to public sector modernization that promotes simple, efficient, and transparent government, with citizens at the centre of reforms.” (Dener et al., 2021). This definition is selected for its inclusive perspective, encompassing concepts like e-Government and Digital Government, as opposed to other definitions used within the field (Bharosa, 2022; Filer, 2019; Yoshida & Thammetar, 2021). As this research focuses on GovTech benchmarking, it will be important to consider all relevant aspects of public sector modernization, ensuring a full understanding of this evolving intersection of governance and technology.

In the next section, the current GovTech benchmarking process is detailed, providing a better understanding of the actors involved, the benchmarking process and the impact it has on citizens.

1.4 GovTech Benchmarking in Practice

This section presents the GovTech benchmarking process in practice, drawing on insights from interviews with practitioners. While section 2.4.2 provides insight into the methodology, the goal here is to offer a practical perspective on the use of benchmarks.

1.4.1 Actors Involved

In the GovTech benchmarking process, two primary types of actors are typically involved: the entity conducting the benchmarking (the *benchmarking* actor) and those being assessed (the *benchmarked* actors). The benchmarked actors are always more than one, as the essence of benchmarking involves making comparisons among different entities. If the benchmark is published, the benchmarking actor is usually not included among the benchmarked actors to ensure objectivity and impartiality in the comparisons.

Commonly, the benchmarking actor is a significant international organization, such as the United Nations or the World Bank, with the benchmarked entities often being various countries. However, the role of a benchmarking actor is not limited to such large international organizations alone; individual countries can also undertake this role, for example when assessing and comparing the GovTech maturity of their own public ministries or departments.

Moreover, the benchmarking process is not exclusive to public sector organizations. Private companies, research institutes, and even citizens and civic organizations can participate. For instance, smaller private firms may contribute to data collection or validation, while larger corporations such as

Capgemini, Gartner, or Accenture often conduct their own benchmarks. Additionally, research institutes and academic partners can assist in developing methodologies, validating data, and more.

1.4.2 GovTech Benchmarking Process

The GovTech benchmarking process, while varying slightly across different benchmarks, generally follows the following set of steps:

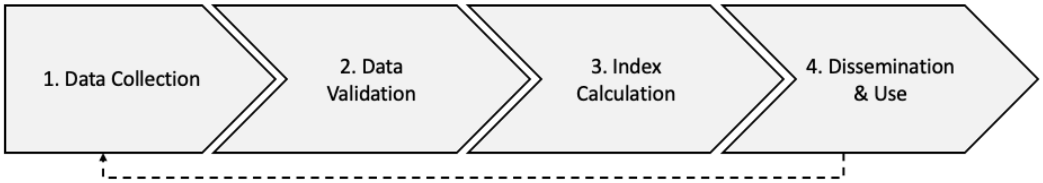


Figure 1.1: GovTech Benchmarking Process

The first step in the GovTech benchmarking process involves extensive data collection, performed by the benchmarking organization. For instance, in case of the GTMI benchmark this would be the World Bank. The data collected can come from a variety of sources, including government reports, public databases, surveys, direct submissions from government agencies and interviews with public officials. Key metrics often include indicators of digital infrastructure, online services, digital literacy among citizens, and the level of e-participation.

Once the data is collected, it undergoes a validation process to ensure its accuracy and reliability. This is done by sending the data to each of the benchmarked actors, who then validate the data found by the benchmarking organization. This validation varies among benchmarked actors, but often involves extensive communication between government organizations, retrieving relevant sources, and cross-checking the data for consistency and accuracy. Detailed reviews and consultations are often necessary to address any discrepancies or errors identified. Once validated, the data is sent back to the benchmarking organization. In some cases, for extra robustness, a second validation process is conducted, repeating the same steps to further ensure data integrity.

After validation, the data is used to calculate various indices that represent different aspects of GovTech performance. These indices are often aggregated into aggregate scores that provide an overall picture of a government's digital maturity and effectiveness. The results are then published in detailed reports, which may include rankings, comparative analyses, and case studies of best practices.

The final step involves the dissemination of the results by the benchmarking organization and their use by the benchmarked organizations. The published reports are distributed widely and made accessible to policymakers, stakeholders, and the general public. The publication of data varies by benchmarking organization; some release the complete dataset while others only publish aggregate scores. A dashed arrow in the diagram highlights the iterative nature of benchmarking, signaling how the outcomes influence policies and, in turn, shape subsequent rounds of data collection. The specifics of this impact will be explored in the following section.

1.4.3 Impact of GovTech Benchmarking Process

The influence of GovTech benchmarks on policy and public services is subtle and varied. Given the multitude of factors that shape policymaking, interviewed practitioners indicate that it is challenging to directly attribute specific policies to benchmark results. Nevertheless, several scholarly works have demonstrated a significant influence of benchmarking on political and economic outcomes and public performance (Bannister, 2007; de Goede et al., 2016; Kunstelj & Vintar, 2005; Magd & Curry, 2003; Muravu, 2023). Broadly, the impact of benchmarking can be understood through a two-step process. Initially, benchmark results provide policymakers with critical data that aids in informed decision-making. Subsequently, these informed decisions have a tangible impact on public service delivery, ultimately improving both the quality and accessibility of services for companies and citizens.

For policymakers, the benefits of benchmarking are twofold. Firstly, benchmarks provide a clearer picture of performance on various fronts. For instance, if benchmarks consistently indicate a country's poor performance in digital inclusion compared to others, it signals a need for improvement. Policymakers might respond by allocating increased funding to improve accessibility features on government platforms, such as adding voice command functionality to e-services to assist users with disabilities. Secondly, benchmarks serve as a learning tool. Seeing a neighboring country excel in digital inclusion might prompt policymakers to study and possibly adopt similar measures, leading to improvement through imitation.

The effect of these policy decisions, guided by benchmarking, are directly experienced by companies and citizens. To use the same example, improvements in digital inclusion can significantly improve access for individuals with disabilities, making digital platforms more user-friendly and accessible. On the other hand, the allocation of government money towards one area, such as digital inclusion, might mean less funding for other critical e-services, such as an online job portal, which could adversely affect job seekers Additionally, consider the impact on the company originally asked to develop this online job portal, which may now lose work to an innovative startup developing an AI voice assistant for improving the digital inclusion. Even though, as mentioned before, it is not possible to fully attribute these decisions to the benchmark results, they do influence policymaking and thus have a significant impact on society.

This effect also introduces important ethical considerations that must be carefully managed throughout the benchmarking process. Firstly, the fairness in resource allocation poses a significant ethical question, as prioritizing one area such as digital inclusion might divert funds from other critical services like employment portals, potentially disadvantaging certain groups of citizens. This raises concerns about how to balance improvements in one sector against the needs in another, ensuring equitable distribution of resources. Additionally, the integrity of the data collection process is an important ethical issue, necessitating transparent and accurate data collection. Misinterpretations or biases in data can lead to misguided policies that might increase existing inequalities or create new ones. Furthermore, the potential for benchmarks to be used in politically motivated ways to justify specific policy decisions or to improve governmental reputations without genuine improvements in services also needs consideration.

These ethical challenges stress the necessity for a thoughtful approach to benchmarking, ensuring that ethical considerations are integrated into the process. However, beyond ethical concerns, numerous other criticisms on benchmarking have been raised by scholars and practitioners, which will be explored in the next section.

1.5 Useful Tools?

Even though in recent years many different benchmarks have been developed and used, benchmarks continue to be heavily criticized by scholars and practitioners. Existing literature, as early as the early 2000s, has consistently highlighted weaknesses in the methods used for measuring e-government (A. Jansen, 2005; Peters et al., 2004). A detailed overview of the literature on GovTech benchmarking is presented in chapter 3. Here, three primary challenges are noted:

Firstly, benchmarks often fail to grasp the full complexity of GovTech. According to Waksberg-Guerrini & Aibar (2007), benchmarks are neglecting deeper transformations governments might be undergoing with intensive use of ICT. Heidlund & Sundberg (2022) similarly observe that evaluation measures on complex digital infrastructures are lacking, with most measures focusing instead on government websites and e-service provision.

Secondly, the aggregation level of benchmarks is often inflexible and overly aggregated, limiting their utility. Berntzen & Olsen (2009) point out that benchmarks often focus on electronic services at the national level, whereas many such services are actually managed by lower levels of government. Even if such services are removed from the analysis, this introduces a considerable source of errors in the assessments.

Lastly, benchmarking is resource-intensive, rendering results outdated and less useful results. Hujran et al. (2022) note that benchmarks often rely on outdated data due to lengthy processes, providing retrospective insights rather than real-time analysis. Berntzen & Olsen (2009) also highlight that as the number of e-government services grows, data collection becomes increasingly challenging. To address this issue, they suggest automatic assessment as a viable solution.

A recent review by Skargren (2020), covering the period from 2003 to 2016 stresses the persistence of certain challenges, stating, "Although change has taken place, many things have a remarkable way of remaining the same. The same criticism, for example, keeps recurring again and again: too many supply-oriented benchmarks, the importance of context and measuring regional levels, and the lack of not measuring back-office processes”. Acknowledging the persistence of these challenges, the next section explores the potential of AI in improving GovTech benchmarking.

1.6 AI: a Potential Remedy?

The rapid advancement in artificial intelligence (AI) offers new ways to overcome the shortcomings of existing measures and benchmarks in monitoring GovTech. Especially, Large Language Models (LLMs), with their capability to understand and generate human language (Chang et al., 2023), show great promise for improving GovTech benchmarking. LLMs can continuously analyze large amounts of textual information, updating insights in near-real time and thus overcoming issues of outdated information that affects current benchmarks. Also, by analyzing large amounts of documents and data, LLMs not only provide basic metrics but also provide more insight into complexities in technical infrastructures and back-office processes, thereby allowing benchmarks to include more complexity.

Already, LLMs are proven useful in many cases and domains. Examples of applications involving LLMs include Psy-LLM (Lai et al., 2023), an AI-based assistive tool using LLMs for question-answering in psychological consultation settings to ease the demand for mental health professions. Another example is FINGPT (Yang et al., 2023), an open-source LLM tailored for the financial sector. Remarkably, public organizations in the Netherlands are already experimenting LLMs. The Province of South-Holland, for example, is experimenting with a LLM using data from their own organization to ensure readily available information to their employees (T. van Grevenbroek, personal communication). Another noteworthy example is Postbus 42, serving as an online portal for questions about the Dutch government (SWIS, n.d.). Considering the promising use of LLMs across different contexts, this research suggests their potential to improve GovTech benchmarking. The next section will define the research problem and specify the research objectives.

1.7 Research Problem and Objective

Research consistently points to enduring challenges in GovTech benchmarking approaches. Key issues include: the resource intensive nature of current methodologies, which leads to retrospective insights rather than real-time analysis (Hujran et al., 2022); a lack of complexity in assessments, which often overlook digital infrastructures, back-office processes, and emerging technologies in favor of simpler website and e-service metrics (Heidlund & Sundberg, 2022; Skargren, 2020); and the methods’ failure to properly adjust the level of aggregation, often resulting in either excessive aggregation or excessive disaggregation rendering it less useful (Berntzen & Olsen, 2009).

The societal relevance of this research lies in its approach to addressing the three primary issues in GovTech benchmarking - timeliness, complexity, and data aggregation – through the use of LLMs. By mitigating these challenges, LLMs could enable benchmarks that offer real-time, detailed insights into GovTech developments, thereby aiding policymakers in making more informed decisions. Given that benchmarks can significantly influence both political and economic outcomes (Bannister, 2007) and shape the development of GovTech services (Kunstelj & Vintar, 2005), refining benchmarking methodologies using LLMs potentially improves the responsiveness and relevance of government actions that better serve societal needs.

The scientific relevance of this research comes from its novel application of LLMs to GovTech benchmarking. Despite the surge in initiatives and applications of LLMs including the field of GovTech, scientific research on the application of LLMs to the GovTech field is lacking. Additionally, since the beginning of this century, the same criticisms keep recurring again and again: retrospective insights rather than real-time analysis due to lengthy processes, too many supply-oriented benchmarks, and the importance of context and measuring regional levels. This study is the first to explore the use of LLMs for GovTech benchmarking, thereby contributing a new perspective that responds to the ongoing need for suitable and advanced GovTech benchmarking methods.

In conclusion, the objective of this research is to address the challenges in GovTech benchmarking- namely timeliness, complexity, and data aggregation - by introducing a new solution that uses LLMs for the operationalization of GovTech benchmarks.

1.8 Research Questions

Main RQ: How do LLMs operationalizing GovTech benchmarks mitigate inherent challenges of timeliness, complexity, and data aggregation, increasing their utility for policymakers?

Sub-questions:

1. Which are practical limitations of current GovTech benchmarks that affect their utility for policymakers?

2. Which AI-technologies are capable of mitigating the limitations of timeliness, lack of complexity, and lack of suitable aggregation within current GovTech benchmarking methodologies?

3. How does Activity Theory guide the placement of LLMs within the GovTech benchmarking ecosystem?

4. What is required from a solution architecture supporting LLMs to operationalize GovTech benchmarks?

5. How does the accuracy of LLMs in operationalizing GovTech benchmarks compare to that of conventional methods?

1.9 Structure of Thesis

This research adopts Design Science Research Methodology (DSRM) as its foundational framework. The next chapter, Chapter 2 - Methodology, introduces DSRM in more detail, and substantiates why it is a suitable methodology for this research. As presented in Figure 1.2 below, subsequent chapters three to seven each align with a step in the six-step Design Science Research Methodology.

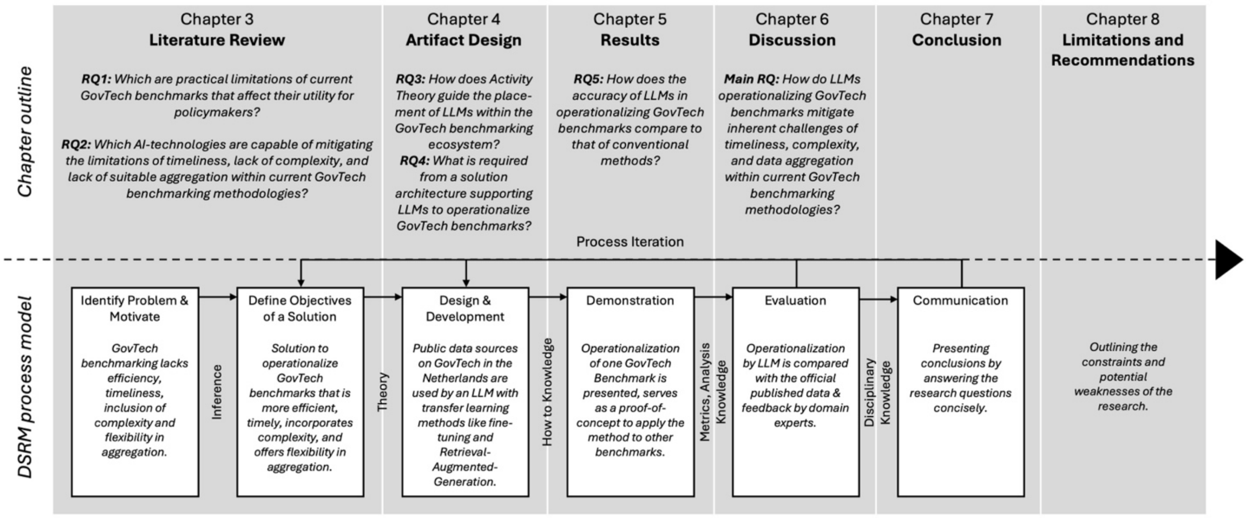

Chapter outline
Chapter 3
Literature Review
RQ1: Which are practical limitations of current GovTech benchmarks that affect their utility for policymakers?
RQ2: Which AI-technologies are capable of mitigating the limitations of timeliness, lack of complexity, and lack of suitable aggregation within current GovTech benchmarking methodologies?
Chapter 4
Artifact Design
RQ3: How does Activity Theory guide the place-ment of LLMs within the GovTech benchmarking ecosystem?
RQ4: What is required from a solution architecture supporting LLMs to operationalize GovTech benchmarks?
Chapter 5
Results
RQ5: How does the accuracy of LLMs in operationalizing GovTech benchmarks compare to that of conventional methods?
Chapter 6
Discussion
Main RQ: How do LLMs operationalizing GovTech benchmarks mitigate inherent challenges of timeliness, complexity, and data aggregation within current GovTech benchmarking methodologies?
Chapter 7
Conclusion
Chapter 8
Limitations and Recommendations
Process Iteration
DSRM process model
Identify Problem & Motivate
GovTech benchmarking lacks efficiency, timeliness, inclusion of complexity and flexibility in aggregation.
Inference
Define Objectives of a Solution
Solution to operationalize GovTech benchmarks that is more efficient, timely, incorporates complexity, and offers flexibility in aggregation.
Theory
Design & Development
Public data sources on GovTech in the Netherlands are used by an LLM with transfer learning methods like fine-tuning and Retrieval-Augmented-Generation.
How to Knowledge
Demonstration
Operationalization of one GovTech Benchmark is presented, serves as a proof-of-concept to apply the method to other benchmarks.
Metrics, Analysis Knowledge
Evaluation
Operationalization by LLM is compared with the official published data & feedback by domain experts.
Disciplinary Knowledge
Communication
Presenting conclusions by answering the research questions concisely.
Outlining the constraints and potential weaknesses of the research.

Figure 1.2: Visual outline thesis

Chapter 2 Methodology, begins by detailing Design Science Research (DSR) to provide the broader context for the Design Science Research Methodology (DSRM). The chapter then outlines the steps involved in the DSRM and explains why this methodology is appropriate for this research. Finally, it describes the validation procedure implemented to ensure the rigor and reliability of the research. This sets the stage for the initiation of the DSRM steps, which starts with an identification of the problem, to be conducted through a literature review in the next chapter.

Chapter 3 Literature Review, examines the development of benchmarking methodologies, highlighting criticisms, challenges, and persistently unresolved issues. This review addresses research question 1 by exploring the practical limitations of current GovTech benchmarks that affect their utility for policymakers. By analyzing these limitations, the review not only sheds light on the gaps within current methodologies but also sets the stage for the development of an innovative solution. Based on the identified limitations, the chapter defines a description of requirements and objectives for a proposed solution. It concludes with a literature search on which AI-technologies are capable of mitigating the identified limitations in GovTech benchmarking methodologies and meeting the defined objectives for a proposed solution, addressing research question 2. This analysis prepares for the design of the artefact using the identified AI-technologies in the next chapter.

Chapter 4 Artefact Design presents the design of the artefact. First, Activity Theory is employed to determine the most effective integration point within the GovTech ecosystem for the proposed solution, addressing research question 3. Subsequently, a detailed description of the solution's design is provided, including the data sources, vector database, embeddings model, base-LLM, prompts and selected benchmark to operationalize. The chapter explains how these components are combined within a solution architecture to optimally achieve the defined objectives, thereby addressing research question 4. This artefact design and its solution architecture sets the stage for the next chapter, where the artefact is demonstrated by operationalizing the selected benchmark.

Chapter 5

Results, applies the designed artefact to operationalize the selected GovTech benchmark. This demonstration is important, as it serves as evidence of the artefact's capability to address the identified issues in GovTech benchmarking and meet the defined objectives. The results of this operationalization are presented and described, offering a clear understanding of the solution's capabilities compared to conventional methodologies, thereby addressing research question 5. While this chapter focuses on simply presenting the data of the operationalization, the next chapter will explore the broader implications of these results, including their impact on the utility for policymakers.

Chapter 6 Discussion, critically analyzes the results presented in Chapter 5, focusing on how LLMs can increase the utility of GovTech benchmarks for policymakers, thereby addressing the main research question. This analysis includes interpreting the results, validating the model against a random chance baseline, and exploring insights from expert interviews. It also examines the contributions to the GovTech benchmarking process, discusses remaining challenges despite the artefact's implementation. While this chapter provides a detailed overview of the implications of the results, the next chapter will present these findings more concisely.

Chapter 7 Conclusion, addresses all the research questions posited in the introduction, summarizing the findings concisely and emphasizing the research's implications. It begins by answering each sub-question in order, ultimately leading to answering the main research question.

Chapter 8 Limitations and Recommendations, discusses the limitations encountered during the research. It categorizes these limitations into two primary types: artefact-related limitations, which are related to the inherent constraints of the developed artefact, and methodological limitations, which are associated with the use of Activity Theory and the Design Science Research Methodology (DSRM). By analyzing these limitations, the chapter provides a clear understanding of the research's boundaries and areas for improvement. Then, recommendations are presented for both future research and policy. Finally, the chapter ends by outlining the contours of a new GovTech benchmarking framework that considers the use of AI: an AI-Supported GovTech Index (AGTI).

2 Methodology

This research adopts Design Science Research Methodology (DSRM). Unlike traditional empirical methodologies that attempt to explain and predict phenomena based on observational activities, Design Science Research (DSR) aims to create an artefact to be an innovative solution for a particular issue. In the following sections, DSR is first introduced (2.1), the rationale for adopting DSR is explained (2.2), and subsequently applied to this context of this research (2.3). In the last section (2.4), the validation process for the artefact is detailed.

2.1 Design Science Research (DSR)

2.1.1 Research Framework

Although Design Science Research (DSR) is a relatively new research methodology, it has received significant attention from researchers in the last decade. With its development still ongoing, there is nonetheless a solid understanding of its core principles (Peffers et al., 2007). DSR focuses on creating and testing IT artefacts designed to address specific problems within organizations. This approach is characterized by a detailed process that includes devising solutions for recognized problems, contributing to research, evaluating these solutions, and sharing the outcomes with relevant audiences. The artefacts may include constructs, models, methods, and instantiations (Hevner et al., 2004). In short, the definition includes any designed object that incorporates a solution for a clearly identified research problem (Peffers et al., 2007).

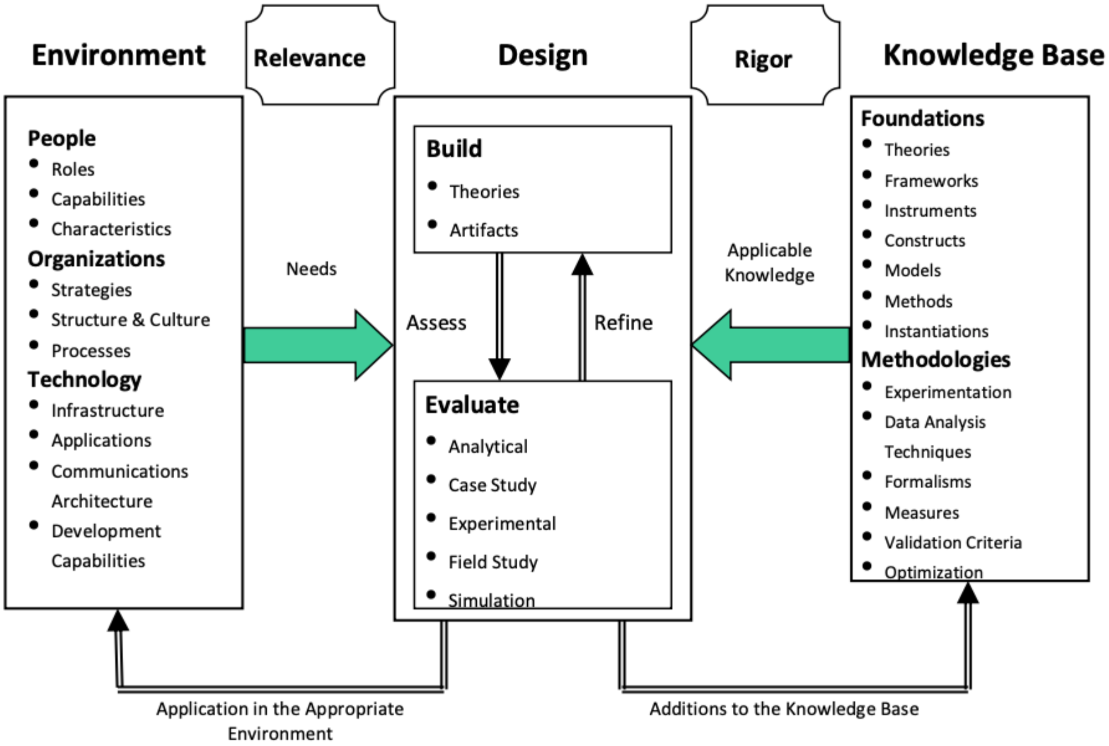


Figure 2.1: Design Science Research Framework (Originally from (Hevner et al., 2004)*, adapted by* (vom Brocke et al., 2020))

Figure 2.1 presents a conceptual framework for understanding, executing, and evaluating DSR. The environment, consisting of people, organizations, and technology, serve as the context for identifying research problems based on the organizational needs. These needs define the relevance of the research problem. The knowledge base provides foundations (theories, models) for designing artefacts and methodologies for their evaluation (Hevner et al., 2004). In case knowledge is already available to solve a problem identified, this knowledge can be applied following "routine design", which does not constitute DSR (vom Brocke et al., 2020). DSR aims to innovate by refining and expanding upon known solutions, engaging in iterative "build" and "evaluate" processes. All put together, this framework connects research to real-world organizational needs. These real-world organizational needs are further met by process models based on DSR, one of which is presented in the next section.

2.1.2 Design Science Research Methodology (DSRM)

The effectiveness of DSR projects often relies on a wide range of process models. One of the mostly widely referenced model is one proposed by Peffers et al. (2007), which will be used within this research. Their Design Science Research Methodology (DSRM), shown in Figure 2.2, outlines a six-step DSR process: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. Additionally, it identifies four potential starting points for initiating a project, thereby offering flexibility based on the specific context of the research. The subsequent section will provide the rationale for adopting DSRM in this study, thereby explaining its relevance and applicability to the research objectives and context.

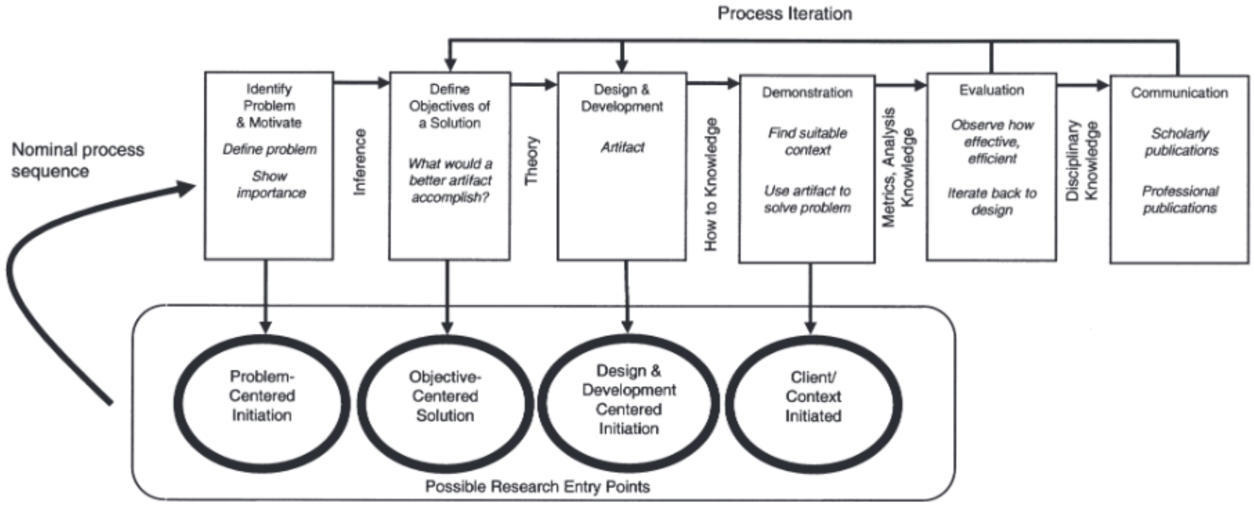

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 2.2: DSRM Process Model (Peffers et al., 2007)

2.2 Rationale for adopting DSR

The choice of Design Science Research (DSR) for this study aligns perfectly with its main objective of developing a practical solution to address existing GovTech benchmarking challenges. Unlike traditional empirical methodologies, which predominantly focus on explaining and predicting phenomena through observation, DSR is tailored towards creating and evaluating artefacts designed to solve specific problems (Peffers et al., 2007). This approach is particularly suited to addressing the current challenges in GovTech benchmarking, such as complexity, timeliness, and flexibility in data aggregation.

By implementing the six-step Design Science Research Methodology (DSRM) of Peffers et al. (2007), this research aims to design and implement a solution that overcomes the limitations of current benchmarking methodologies. DSRM facilitates iterative refinement and validation, ensuring the developed artefact aligns with real-world needs. This structured yet flexible approach allows the research to systematically address specific benchmarking challenges by using LLMs to provide more detailed, timely, and adaptable benchmarking outcomes. In the next section, these steps will be contextualized specifically for this research, providing a detailed and well-founded structure for the study.

2.3 DSRM Applied to LLMs for Benchmark Operationalization

In this section the Design Science Research Methodology (DSRM) is applied to the context of using LLMs for GovTech benchmark operationalization. As described in section 1.9, chapters three to seven each align with a step in the six-step DSRM process. These methodology's steps, as used in this research, are visually presented in Figure 2.3, following the process of Peffers et al. (2007).

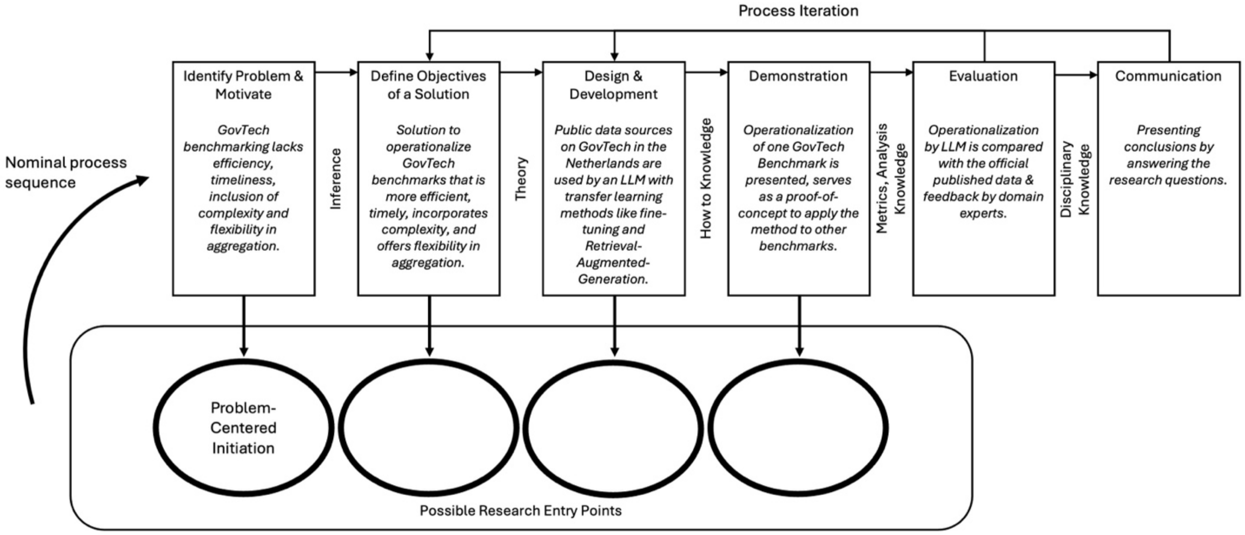

Process Iteration
Identify Problem & Motivate
GovTech benchmarking lacks efficiency, timeliness, inclusion of complexity and flexibility in aggregation.
Inference
Define Objectives of a Solution
Solution to operationalize GovTech benchmarks that is more efficient, timely, incorporates complexity, and offers flexibility in aggregation.
Theory
Design & Development
Public data sources on GovTech in the Netherlands are used by an LLM with transfer learning methods like fine-tuning and Retrieval-Augmented-Generation.
How to Knowledge
Demonstration
Operationalization of one GovTech Benchmark is presented, serves as a proof-of-concept to apply the method to other benchmarks.
Metrics, Analysis Knowledge
Evaluation
Operationalization by LLM is compared with the official published data & feedback by domain experts.
Disciplinary Knowledge
Communication
Presenting conclusions by answering the research questions.
Nominal process sequence
Problem-Centered Initiation
Possible Research Entry Points

Figure 2.3: DSRM Process Model for GovTech benchmarking, adapted from (Peffers et al., 2007)

Figure 2.3 illustrates that this research has a problem-centered initiation, specifically addressing the challenges encountered in existing GovTech benchmarking methodologies. In step 1, this problem is clearly identified and motivated through a literature review, which highlights the development and challenges of current benchmarking methodologies, especially in the GovTech sector. This review pinpoints the necessity for a more effective benchmarking solution. Step 2 defines the objectives of the solution, drawing from the gaps identified earlier, and sets specific goals for overcoming the shortcomings in GovTech benchmarking. Step 3 involves the design and development of the proposed solution, using Activity Theory to integrate the artefact in the benchmarking ecosystem. Step 4 demonstrates the application of this solution on a selected GovTech benchmark, providing evidence of its capability to meet the defined objectives and address the benchmarking issues. Step 5 evaluates the proposed artefact; this process is described in detail in the next paragraph (2.4). Lastly, step 6 involves communicating the research findings through this thesis, by presenting the conclusions through answering the research questions posited in section 1.8.

2.4 Validation

The evaluation step within the DSRM process model focuses on how well the artefact supports a solution to the problem (vom Brocke et al., 2020). Specifically, this research will examine the accuracy with which LLMs can operationalize GovTech benchmarks, aiming to mitigate practical limitations that currently restrict their utility for policymakers. These limitations will be detailed in the next chapter 3.

2.4.1 Rationale for Validation over Verification

In traditional computational modeling and simulation, verification and validation serve as the principal methods for evaluating the models' accuracy and reliability (Oberkampf & Trucano, 2008). Verification involves checking the software for correctness and ensuring that the numerical solutions are accurate within the context of the specified computational model. On the other hand, validation examines the physical accuracy of a computational model by comparing simulation results with experimental data. During verification, the connection of the simulation to the real world is not considered relevant. However, in validation, the critical concern is the correlation between the computational outcomes and real-world (experimental) data.

However, language models, particularly Large Language Models (LLMs), significantly differ from traditional computational and simulation models in terms of system scale and their non-deterministic nature. This distinction necessitates a different approach to model evaluation. In this research, the focus shifts away from verification to concentrate only on validation. This decision is substantiated by several considerations. First, the fundamental differences between conventional models and LLMs render traditional verification techniques either less effective or unsuitable for assessing LLMs, as noted by Huang et al. (2023). Furthermore, it is assumed that any functional inadequacies of the system would become apparent during validation. In essence, if the system does not perform as intended, this will be evident from the validation outcomes, suggesting that if the validation is deemed satisfactory, then the verification is implicitly adequate. Additionally, in the context of benchmarking, the emphasis is on comparing the performance of systems against each other rather than verifying each system against its specifications. Thus, the research prioritizes validation over verification, proceeding on the premise that successful validation confirms the artefact’s functionality, aligning with the objectives of benchmarking.

2.4.2 Validation Methodology

The validation phase will be implemented through a combination of expert interviews and quantitative analyses. Qualitative insights are obtained through interviews with two types of experts to ensure a thorough assessment of the LLM-based operationalizations of GovTech benchmarks. The first type consists of Dutch GovTech experts tasked with evaluating the correctness of the model’s answers, focusing specifically on the accuracy of the outputs given the Dutch context. The second type includes (international) experts in the application and methodology of existing GovTech benchmarks, who examine the practicality of the model’s outputs, such as their adherence to the required data formats and usability in operational contexts. Engaging these distinct types of experts provides qualitative insights into both the accuracy and practical applicability of the LLM outputs in real-world scenarios, facilitating a thorough validation process.

Quantitatively, the validation process involves comparing the LLM's operationalized indicators with official data from the GovTech Benchmark (ground truth). This comparison produces a numerical accuracy score, indicating how closely the LLM’s outputs match the official benchmark data. This score is a direct measure of the LLM’s effectiveness in operationalizing benchmarks. This comparative analysis is conducted through four distinct methods:

1. **Manual Evaluation:** Each output from the LLM is manually evaluated by comparing the LLM's operationalized indicators to the ground truth. Outputs are then categorized into one of five categories: correct and following data format; correct and not following data format; no answer; incorrect and following data format; and incorrect and not following data format. This classification not only measures the model's accuracy but also its practical usability by determining its adherence to the prescribed data formats essential for benchmark operationalization. To calculate the precise accuracy for each model configuration, an answer is deemed accurate if it is correct, regardless of whether it adheres to the prescribed format. The formula then looks like:

Equation 2.1

$$Accuracy = \frac{\sum(correct,\ following\ format) + \sum(correct,\ not\ following\ format)}{\sum(incorrect,\ not\ following\ format) + \sum(incorrect,\ following\ format) + \sum(no\ answer)}$$

2. **Exact Match:** An algorithm calculates the proportion of the LLM's outputs that exactly match the ground truth data. This metric evaluates the model's precision in reproducing the exact answers expected in the benchmark.

3. **Edit Similarity:** Using the Levenshtein distance, this method calculates the edit similarity by determining the number of insertions, deletions, and substitutions needed to transform the model's prediction into the ground truth, normalized by the length of the longest word involved. This established method, referenced in the work of Zhang & Zhang (2020), assesses the semantic similarity between the model's responses and the actual data, offering a more nuanced insight into the accuracy of the LLM's outputs.

Equation 2.2

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

Where, for answer *p* and *g,* for character positions *i* and *j,*

Equation 2.3

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

4. **Random Chance Comparison:** Considering there are multiple choice questions within GovTech Benchmarks, it is important to determine whether random guessing might surpass the model's performance. This evaluation is conducted by calculating the expected accuracy of random guesses for the multiple-choice questions. Assuming a uniform probability distribution, where each of the n answer options in m indicators has an equal chance of being selected, the probability of a correct guess for each indicator varies by the number of options available. The general formula to calculate the expected random accuracy for multiple-choice questions, where each group of indicators has a different number of answer options, is given by:

Equation 2.4

$$Random\ Accuracy = \frac{1}{N} \sum_{i=1}^{k} (m_i \cdot \frac{1}{n_i})$$

Where:

- *N* is the total number of indicators.
- *k* is the number of different groups of indicators, each group having a different number of answer options.
- m_i *is* the number of indicators with n_i answer options.
- n_i *is* the number of answer options for the *i*-th group.

Together, these four methods provide the quantitative validation of the LLM's performance in operationalizing GovTech benchmarks. The combination of manual evaluation, exact match comparison, edit similarity, and random chance comparison ensures a robust assessment of both the accuracy and practical usability of the model's outputs. This approach not only measures how closely the LLM's predictions align with official benchmark data but also evaluates the model's adherence to required data formats and its effectiveness in real-world applications. With this validation methodology in place, the next chapter explores the specific limitations that the artefact seeks to overcome, which compromise the utility of the benchmarks as an aid for decision-making among policymakers.

3 Literature Review

This chapter conducts a two-part literature review. First, it examines GovTech benchmarking methodologies, emphasizing the criticisms, challenges, and unresolved issues that answer research question 1: *"Which are practical limitations of current GovTech benchmarks that affect their utility for policymakers?"* This section concludes by defining objectives for a solution to address these limitations. Subsequently, the review assesses which AI technologies are most effective at overcoming these identified challenges, responding to research question 2: *"Which AI-technologies are capable of mitigating the limitations of timeliness, lack of complexity, and lack of suitable aggregation within current GovTech benchmarking methodologies?"* The chapter is organized as follows: Section 3.1 details the unresolved issues within GovTech benchmarking methodologies, Section 3.2 outlines the objectives for a solution, and Section 3.3 identifies suitable AI technologies from the literature.

3.1 Unresolved Issues in GovTech Benchmarking Methodologies

3.1.1 Search Strategy

To conduct a literature review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process was adopted. The PRISMA flow diagram is presented in Figure 3.1.

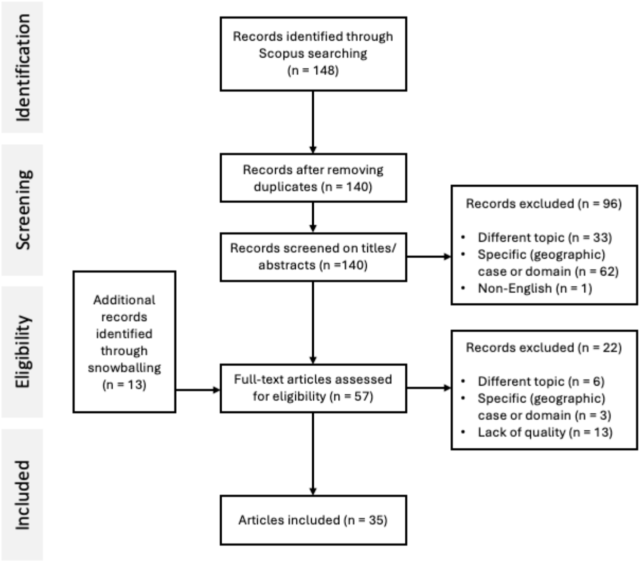

Identification
Screening
Eligibility
Included
Records identified through Scopus searching (n = 148)
Records after removing duplicates (n = 140)
Records screened on titles/ abstracts (n =140)
Records excluded (n = 96)
• Different topic (n = 33)
• Specific (geographic) case or domain (n = 62)
• Non-English (n = 1)
Additional records identified through snowballing (n = 13)
Full-text articles assessed for eligibility (n = 57)
Records excluded (n = 22)
• Different topic (n = 6)
• Specific (geographic) case or domain (n = 3)
• Lack of quality (n = 13)
Articles included (n = 35)

Figure 3.1: PRISMA Flow Diagram

Scopus was selected due to its extensive range of applied public policy publications, high-quality peer-reviewed papers, and advanced targeted search capabilities. The search strategy was formulated with the following query:

(benchmarking OR measuring OR analysing) AND (govtech OR egovernment OR e-government)

Given the broad application of these terms across various situations and contexts, a substantial number of articles, totaling 1376, were initially identified. As the research exclusively concentrates on the measurement methodology, a targeted search within article titles was conducted, yielding a refined set of 148 documents.

After removing duplicates, titles and abstracts of 140 records were screened, leading to the exclusion of 96 records for various reasons: some were off-topic, others had a pronounced focus on a specific (geographic) case or domain rather than on the method of monitoring GovTech itself, and one article was non-English.

Subsequently, the eligibility of the remaining articles was assessed. This step led to the exclusion of an additional 22 articles. Moreover, 13 additional articles were identified through snowballing or citation chaining, which involved reviewing the references of the selected papers to find relevant studies. Consequently, the final number of articles included in this literature review stands at 35.

3.1.2 Development of GovTech Benchmarking Methodologies

From the late 1990s onward, the landscape of e-government has been influenced by a relatively large number of international organizations and multinational companies that claim to be "benchmarking e-government" (Skargren, 2020). Noteworthy early benchmarks include the Gartner's Four Phases of e-Government Model (Baum & Di Maio, 2000) and the United Nations E-Government Development Index (EGDI) (United Nations, 2022). Prominent academic contributions include the Four Stages Growth Model towards e-Government (Layne & Lee, 2001) and the Three Ring Model (Koh & Prybutok, 2003).

Starting from 2003, scholars have begun studying this phenomenon by comparing different benchmarks and frameworks for e-government. Pioneering research on measuring e-government impact was carried out by D. Janssen et al. (2004) and Peters et al. (2004). Their analyses shows that many measurement instruments then adopted too simplistic perspectives, focusing only on easily measurable factors. D. Janssen et al. (2004) therefore recommended a more nuanced interpretation of these benchmarks to avoid policy decisions that focus merely on achieving higher scores rather than genuine improvements in e-government services. Peters et al. (2004), on the other hand, linked this shortfall to the lack of a robust theoretical framework for measuring the impact of e-government. Hence, they proposed the development of standardized measurement instruments to address this gap, enabling governments to effectively compare different e-government approaches.

In the years after, a broad range of such frameworks were developed and used. Early suggestions were put forth by Brüggemeier et al. (2005), concentrating on innovation arenas, and by Kim et al. (2005), which propose a g-CSI model customer satisfaction for e-government suitable to the Internet environment. Notable scientific frameworks include one grounded in structuration theory by Senyucel (2007), frameworks for measuring citizen or user satisfaction (Park, 2008; Verdegem & Hauttekeete, 2007), and a comparative assessment of online services delivery progress through multi-criteria analysis by Giannakopoulos & Manolitzas (2009).

Waksberg-Guerrini & Aibar (2007) were among the first to raise critiques, stating that assessment frameworks in e-government research have largely neglected the deeper transformations governments might be undergoing with the intensive use of ICT. The authors argue that existing assessments, primarily focused on web portals and online services, fail to capture the essence of the potential shift towards a network-like organization, overlooking indicators related to efficiency, public value generation and interactivity. Additionally, another common criticism was the excessive focus on measuring supply-side factors and technologies rather than user needs. Similar critiques were also found and confirmed by Snijkers et al. (2007) and Salem (2007).

Berntzen & Olsen (2009) not only confirm the supply-oriented nature of benchmarks but also offer several other noteworthy critiques. They point out that the benchmarks they analyzed focus on electronic services at the national level, whereas many such services are actually managed by lower levels of government. Even if such services are removed from the analysis, this introduces a considerable source of errors in the assessments. Moreover, they highlight that as the number of e-government services grows, data collection becomes increasingly challenging. To address some of these issues, they suggest automatic assessment as a viable solution to some of the problems experienced by current benchmarking studies.

In the years after, again a diverse range of frameworks and indicators were proposed for evaluating e-government. As these proposals were formulated in response to criticisms, it is possible to discern two distinct strategies for addressing these critiques. The first strategy involves the ongoing addition of new factors, resulting in frameworks becoming more extensive and complex. This complexity, combined with time and resource constraints, has led to a gradual aggregation of indicators, making the frameworks challenging to apply for assessments at lower governmental levels. For instance, the growing complexity made it impractical for the frameworks to be effectively used for assessing municipalities due to the overwhelming amount of work involved. Examples are the multidimensional framework by Chircu (2008), the contextual benchmark method by J. Jansen et al. (2010), the Digital Economy and Society Index (DESI) (European Commission, 2022), the Digital Government Index (DGI) (Ubaldi & Okubo, 2020) and the GovTech Maturity Index (GTMI) (Dener et al., 2021).

A second trend is the development of frameworks tailored to specific domains or use cases. Even though these frameworks capture the complexity and multiplicity of situations well, the disadvantage is however that their applicability might be limited when dealing with broader or interdisciplinary contexts. Specialized frameworks, while effective within their designated domain, may lack the adaptability required for broad assessments across diverse governmental functions. Examples are an approach that focuses on the back-end of e-government (M. Janssen, 2010), methods for measuring the performance of local e-governments (Batlle-Montserrat et al., 2014; de Juana-Espinosa & Tarí, 2012), specific e-government projects (Jukić et al., 2013), or the efficiency of public administrations (Rodríguez-Bolívar, 2014).

In recent years, the field of e-government benchmarking has seen relatively few novel insights. The literature predominantly revisits and confirms previously identified limitations, rather than uncovering novel flaws or critiques of existing benchmarks. For instance, Rorissa et al. (2011) revisits the issue of benchmarks failing to distinguish between basic static websites and more complex, interactive portals. They also assessed the strengths and limitations of six frameworks used to compute e-government indexes. However, there have been a few significant new criticisms. Notably, Scott et al. (2016) introduced Public Value theory as a novel framework to assess e-government success. More recently, Przeybilovicz et al. (2023) highlighted the need for e-government benchmarking to incorporate local contexts and perspectives. They critique the Eurocentric and Global North biases prevalent in current benchmarking practices, advocating for a decolonization approach. This approach questions the metrics established by major organizations, suggesting they may channel governmental efforts into narrowly defined paths and overlook broader, locally relevant developmental goals.

Figure 3.2 presents a concise overview of the literature from the year 2000 onwards, detailing both the frameworks and indices proposed, as well as the critiques identified in literature. The next section will use recent literature reviews to identify the key challenges in the field of GovTech benchmarking.

Year	Selection of Proposed Frameworks & Indices	Selection of Criticism in Literature
2000	Gartner's Four Phases of e-Government Model (Baum & Di Maio, 2000) Four stages growth model towards e-government (Layne & Lee, 2001)	
2004	The United Nations e-Government Development Index (EGDI) launched in 2003 (United Nations, 2022) Innovation arena model for e-government (Brüggemeier et al., 2005) g-CSI model customer satisfaction for e-government suitable to the Internet environment (Kim et al., 2005)	Frameworks adopt simplistic perspectives, focusing only on easily measurable factors (D. Janssen et al., 2004; Peters et al., 2004)
2008	Framework grounded in structuration theory (Senyucel, 2007) Framework for user satisfaction (Verdegem & Hauttekeete, 2007) Assessment of online services delivery progress through multi-criteria analysis (Giannakopoulos & Manolitzas, 2009)	Frameworks are neglecting deeper transformations governments might be undergoing with intensive use of ICT (Waksberg-Guerrini & Aibar, 2007)
2012	Performance of local governments (Batlle-Montserrat et al., 2014; de Juana-Espinosa & Tarí, 2012) Framework for specific e-government projects (Jukić et al., 2013)	Maturity models represent development through distinct stages whereas, in practice stages are not linear but rather interconnected and can occur simultaneously (Andersen et al., 2012)
2016	Digital Economy and Society Index (DESI) launched in 2014 (European Commission, 2022)	Most models are merely restructurings or adjustments of existing ones (Nielsen, 2016)
2020	Digital Government Index (DGI) 2019 by the OECD (Ubaldi & Okubo, 2020) GovTech Maturity Index (GMTI) 2020 by the World Bank (Dener et al., 2021)	E-government maturity models lack the inclusion of emerging technologies (Lemke et al., 2020) Too many supply-oriented benchmarks, missing context and measuring at regional levels, lack of measuring back-office processes (Skargren, 2020)
2024		No real-time insights due to their time-consuming processes (Hujran et al., 2022) No evaluation measures of complex digital infrastructures, only measurable factors like government websites and e-service provision (Heidlund & Sundberg, 2022) Decolonization approach (Przeybilovicz et al., 2023)

Figure 3.2: Overview of proposed frameworks and criticisms in literature

3.1.3 Recurring Criticisms

Over the past few years, three literature reviews have explored the landscape of e-government evaluation methodologies (Heidlund & Sundberg, 2022; Hujran et al., 2022; Skargren, 2020). Each of these reviews offers its unique perspective on the evolution of the research field concerning benchmarking e-government, of which a concise version was provided in the previous section. Furthermore, these reviews shed light on several unresolved issues and challenges inherent in current GovTech benchmarking methodologies.

Skargren (2020) states that "The same criticism, for example, keeps recurring again and again: too many supply-oriented benchmarks, the importance of context and measuring regional levels, and the lack of not measuring back-office processes." A similar like issue is put forth by Heidlund & Sundberg (2022) who write: “We expected the more recent highly cited research to investigate evaluation measures of complex digital infrastructures, but to our surprise, many papers were concerned with government Web sites and e-service provision. Hujran et al. (2022) find that “existing e-government maturity models lack the inclusion of emerging and modern technologies”. Therefore, benchmarking methodologies should be revisited, developed, and extended to also include those emerging technologies. The next section defines objectives for a solution that could overcome these unresolved issues.

3.2 Definition of Objectives for a Solution

Based on the identified unresolved issues in the previous section, it can be concluded that there is a significant demand for an effective methodology capable of benchmarking and monitoring the status of GovTech, meeting the following criteria.

First, the method should possess the capability to stay up to date, acknowledging the difficulty of keeping pace with the rapid advancements in technology, as highlighted by Hujran et al. (2022). Additionally, as indicated by Skargren (2020) and Heidlund & Sundberg (2022), the methodology should involve the capability to encompass not only websites and e-services but also more complex digital infrastructures, even though this often means substantial resource and financial investments. Lastly, it should be possible to change the level of aggregation to suit the purpose. As observed by Berntzen & Olsen (2009), existing methods frequently suffer from unsuitable aggregation levels.

This research suggests employing Artificial Intelligence (AI) to address these unresolved issues, meeting the established criteria. Therefore, the next section will explore literature to identify AI-technologies that are best suited to overcome the limitations of timeliness, lack of complexity, and lack of suitable aggregation within current GovTech benchmarking methodologies.

3.3 Selection Suitable AI Technology

3.3.1 AI for the Public Sector

In an extensive literature review on AI and the public sector, Wirtz et al. (2019) identify several potential AI applications for the public sector, two of which closely match the goals set to be achieved in this research. The first is AI-Based Knowledge Management (KM) Software, where the use of neural networks enables the generation, systematization, analysis, distribution, and sharing of knowledge with others. For GovTech monitoring, such a neural network could be helpful in processing vast amounts of data generated by public organizations on the use of their technologies, identifying patterns, and extracting valuable insights that contribute to informed decision-making. The second application, an Intelligent Digital Assistant (IDA), provides an intuitive interface between a user and a system/device to search for information or complete simple tasks. In the context of benchmarking GovTech, an IDA could serve as a user-friendly interface for accessing relevant information and answering inquiries.

In Figure 3.3, Corea (2019) presents a structured overview of the AI technology landscape, categorizing technologies into two main groups: AI Paradigms (Symbolic, Sub-symbolic, and Statistical approaches) along the X-axis, and AI Problem Domains (Reasoning, Knowledge, Planning, Communication, and Perception) along the Y-axis. In the context of GovTech benchmarking, knowledge-based tools stand out as the most appropriate AI paradigm, given their strengths in knowledge representation and understanding. Choosing a problem domain of knowledge is evident, as an AI for benchmarking GovTech should have the ability to represent and process sector-specific information and insights. However, also the incorporation of perception would prove highly beneficial, as the AI system should be able to respond flexibly to a variety of user prompts and tasks. Therefore, enabling the system to understand raw input in natural language is important. For instance, if a user wishes to inquire about the state of GovTech in a specific domain or requests a particular aggregation, the AI should be proficient in understanding and processing such natural language inputs.

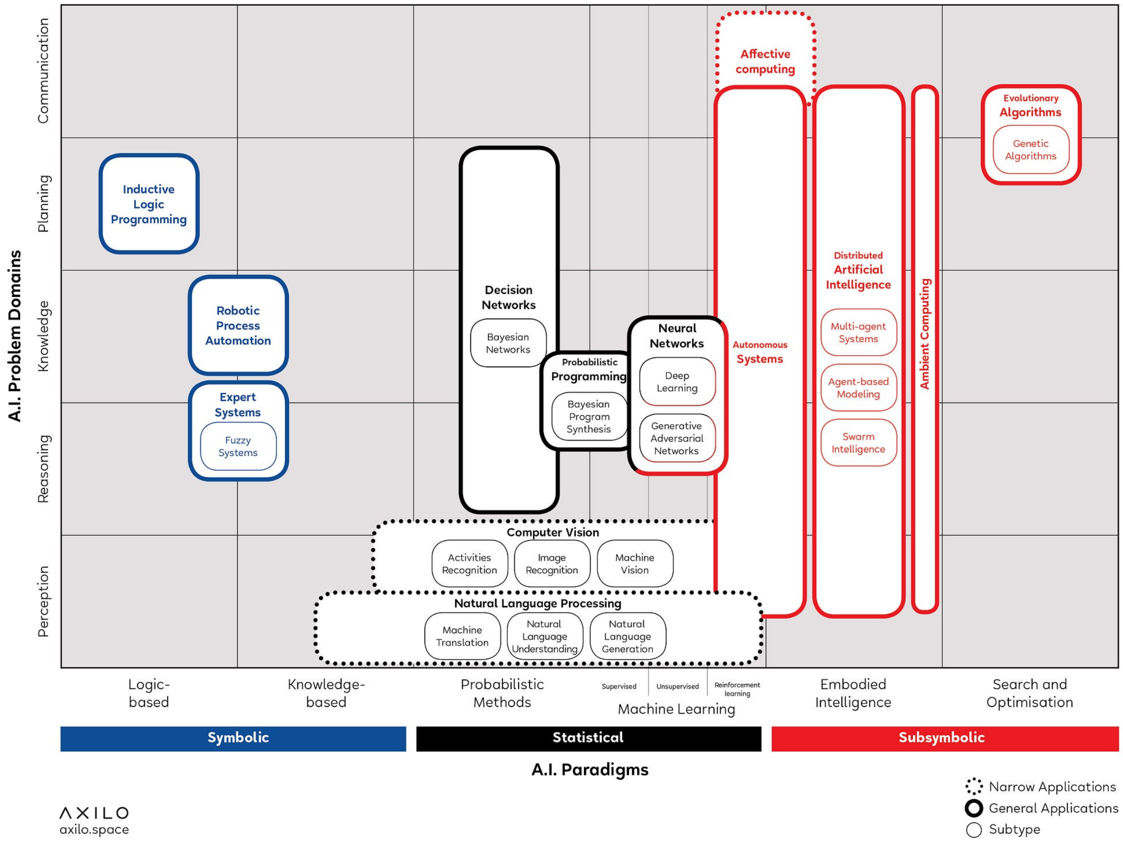


Figure 3.3: AI Knowledge Map (Corea, 2019)

Following the framework, two potential technologies that emerge are neural networks (NN) and natural language processing (NLP). Fortunately, a significant breakthrough has occurred with the recent development of transformer models, enabling the integration of these two technologies. Specifically, Large Language Models (LLMs) now possess the capability to effectively combine neural networks and natural language processing. This technology suits the purpose of an AI that can effectively benchmark GovTech developments very well. The technology behind LLMs is covered in the next section.

3.3.2 Large Language Models

AI models known as Language Models (LMs) are computational models that have the capability to understand and generate human language (Chang et al., 2023). LMs can predict the chance of word sequences or create new text from a given input. The most widespread type, N-gram models, estimate word probabilities using the context before them. Yet, LMs encounter challenges like dealing with rare words, overfitting, and grasping complex language aspects. Researchers are consistently improving LM structures and training methods to tackle these issues.

Large Language Models (LLMs) are advanced LMs known for their massive parameter sizes and impressive learning capabilities. The central component in many LLMs like GPT-4 (OpenAI et al., 2023) is the self-attention module in Transformer, which serves as the basic building block for language modelling tasks. Transformers have revolutionized Natural Language Processing (NLP) by efficiently handling sequential data, enabling parallelization, and capturing long-range dependencies in text. A key feature of LLMs is in-context learning, where the model is trained to generate text based on a given context or prompt. This strengthens LLMs' ability to produce more coherent and contextually relevant responses, making them suitable for interactive and conversational applications (Chang et al., 2023). The adaptability of LLMs is further improved by different transfer learning methods, which are covered in the next section.

3.3.3 Transfer Learning Methods

Transfer Learning (TL) is a vital machine learning technique that aims at improving the performance of models in target domains by transferring the knowledge contained in different but related source domains. In this way, the dependence on a large number of target-domain data can be reduced for model training. Due to the wide application prospects, transfer learning has become a popular and promising area in machine learning. An extensive survey article by Zhuang et al. (2021) reviews more than forty representative transfer learning approaches, especially homogeneous transfer learning approaches, from the perspectives of data and model.

There are three sub-settings of TL strategies, categorized as: inductive TL, transductive TL and unsupervised TL (Hosna et al., 2022). Inductive TL requires the source and target domains to be the same, though the specific tasks the model is working on are different. Transductive TL is used in scenarios where the domains of the source and target tasks are not the same but interrelated. Unsupervised TL is like Inductive TL, but the difference is that the algorithms focus on unsupervised tasks and involve unlabeled datasets both in the source and target tasks.

In the context of this research, an LLM is needed with the following two capabilities: proficiency in question-answering in Dutch and specialized knowledge in GovTech. This model is nowhere available. However, there is a Dutch model trained for question-answering. Given the absence of such a model, the adaptation of an existing Dutch question-answering model through domain adaptation presents a viable solution. Domain adaptation is one of the approaches of transductive transfer learning, in which the task remains the same, but the source and destination have different domains or distributions (Pan & Yang, 2010). Initial tests will assess to what extent the selected base-model has knowledge about the GovTech situation in the Netherlands. It is anticipated, however, that domain-specific knowledge will not be inherently present, and thus domain adaptation will be necessary to improve the LLM’s capabilities. Two techniques to apply domain adaptation are fine-tuning and Retrieval-Augmented-Generation (RAG).

Fine-tuning allows for the base-model, which is already proficient in general language Dutch question and answering, to adapt and specialize in the GovTech domain by training on a smaller, domain-specific dataset. This method ensures that the model not only retains its ability to answer questions but also acquires specialized knowledge relevant to GovTech, thereby facilitating more accurate and contextually appropriate responses. Incorporating Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA, optimizes this process by adjusting only a small subset of the model's parameters, reducing resource usage while effectively adapting the model to new tasks (Hu et al., 2021).

Complementing fine-tuning, Retrieval-Augmented Generation (RAG) improves this method by giving the fine-tuned model a special ability to look up and use extra information related to GovTech when needed. This combination means the model doesn't just rely on the fixed knowledge it was trained with; it can also fetch new and relevant information from outside sources. This ability makes the model much more effective and flexible, especially useful in the fast-changing world of GovTech. The efficacy of RAG in domain-specific contexts has been demonstrated by Siriwardhana et al. (2022), who used RAG in three domains: COVID-19, News, and Conversations, and achieved significant performance improvements compared to the original model.

Lastly, while not a transfer learning method, prompt engineering is an essential technique to consider when using LLMs. Prompt engineering involves carefully designing and optimizing the prompts given to the model to elicit the most accurate and relevant responses. By refining the way queries are presented, the performance of the model can be significantly improved without the need for additional training. This method uses the model's existing knowledge more effectively, ensuring that the generated responses are as precise and useful as possible in the context of GovTech benchmarks (B. Chen et al., 2023).

4 Artefact Design

This chapter presents the design of the artefact. First, Activity Theory is used to guide the integration of the artefact in the GovTech benchmarking ecosystem, addressing research question 3: "*How does Activity Theory guide the placement of LLMs within the GovTech benchmarking ecosystem?".* Following this theoretical application, the chapter proposes a solution architecture that supports the operationalization of GovTech benchmarks using an LLM, responding to research question 4: "*What is required from a solution architecture supporting LLMs to operationalize GovTech benchmarks?".* The structure of the chapter is organized as follows: Section 4.1 applies Activity Theory to determine the most effective integration point within the GovTech ecosystem for the proposed solution. Section 4.2 then provides a detailed description of the solution's architecture and design.

4.1 Integrating LLMs in GovTech Benchmarking Ecosystem

Having identified a fine-tuned LLM with RAG and prompt-engineering as an AI-technology capable of mitigating challenges within GovTech benchmarking in section 3.3, it is important to determine how the LLM can be integrated into the GovTech Benchmarking Ecosystem. This integration process is guided by using Activity Theory, outlined in the next section 4.1.1. Subsequently, section 4.1.2 will explore the various roles that the LLM can take within the ecosystem. The selection of the most fitting role based on the research objectives, is described in 4.1.3.

4.1.1 Activity Theory

Activity Theory, rooted in the early 20th-century work of Soviet psychologists L.V. Vygotsky and A.N. Leontiev, offers a robust framework for examining human activities within their socio-cultural contexts. In its original form, the relationship between *subject* (human doer) and *object* (the thing being done) forms the core of an *activity.* The *object* of an *activity* encompasses the activity's focus and purpose while the *subject*, a person or group engaged in the *activity*, incorporates the subject's various *motives* (Hasan & Kazlauskas, 2014). Another classic element in Activity Theory, the *tool*, is the mediation mechanism or device through which the *subject* aims to achieve the *object*.

Later, Engeström (1987) popularized Activity Theory by providing additional elements of analysis to Vygotsky's and Leontiev's orginal theory: *rules* (social constructs that help determine how subjects can act), *division of labor* (distribution of actions among a community of coworkers), *community* (all partners directly involved in the activity) and *outcome* (what the activity system produces, desired or undesired). Together, these elements these elements form a generic activity system, as represented in Figure 4.1.

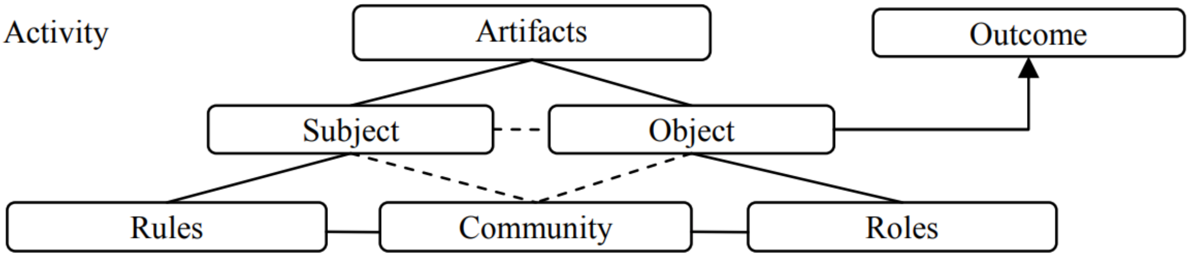


Figure 4.1: Activity Representation in Activity Theory, adapted from (Ojo et al., 2011)

Activity Theory has been used in a wide variety of domains and applications, including e-government benchmarking by Ojo et al. (2011). They emphasize the theory's relevance to GovTech benchmarking for two key reasons. First, benchmarking is inherently contextual, meaning it must be performed within a specific setting. Second, when e-government systems are designed, they often incorporate assumptions about their context of use, which may not align with the reality of their deployment, potentially leading to discrepancies and failures.

Ojo et al. (2011) have operationalized this connection by considering that the benchmarking activity (Activity) is carried out by a benchmarker (Subject); using a certain benchmarking approach (Artefacts); subject to certain benchmarking rules (Rules); and involving benchmarking partners (Community) with their commitments and roles (Roles); to achieve a certain benchmarking purpose (Object) and eventually the expected benchmarking results (Outcome). This model is shown in Figure 4.2.

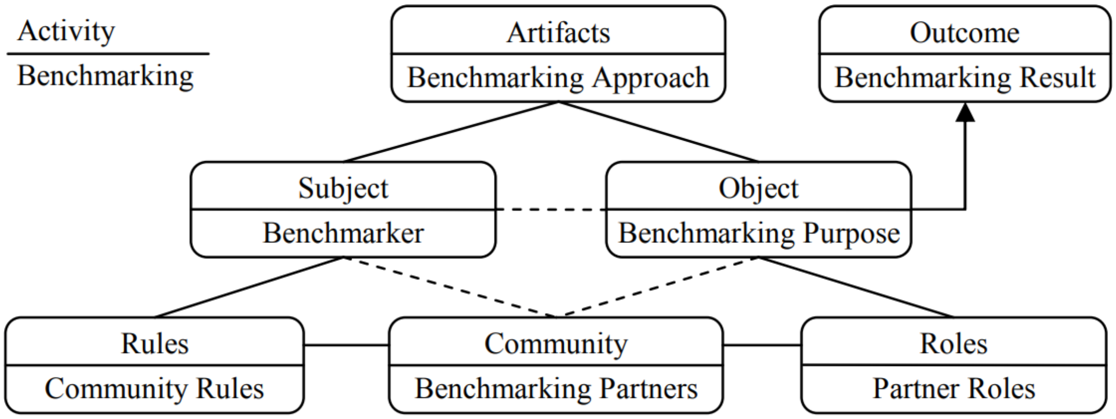


Figure 4.2: Activity Theory-Based Benchmarking Model (Ojo et al., 2011)

In the following, the role of LLMs within the Activity Theory-Based Benchmarking Model is discussed, demonstrating the various ways in which LLMs can be effectively incorporated and used.

4.1.2 Roles of LLMs in the GovTech Benchmarking Ecosystem

4.1.2.1 LLMs as Artefacts

In the most intuitive way, LLMs are positioned as tools or methods employed by the *subject* (benchmarker) to facilitate the benchmarking process. They act as advanced analytical tools that the benchmarker uses to find, process, and analyze data, generating insights that can be used to perform the benchmarking Activity. According to the benchmarking model by Ojo et al. (2011) this would be visualized as presented in Figure 4.3. The artefact is according to Activity Theory the mediation mechanism or device through which the *subject* aims to achieve the *object*, which means that the LLM will be used to achieve the object. According to Ojo et al. (2011) the object is the benchmarking purpose, e.g. "to determine the source of good practice for citizen-focused mobile services". Finally, this results in a benchmarking result, like an EGOV ranking or benchmarking report.

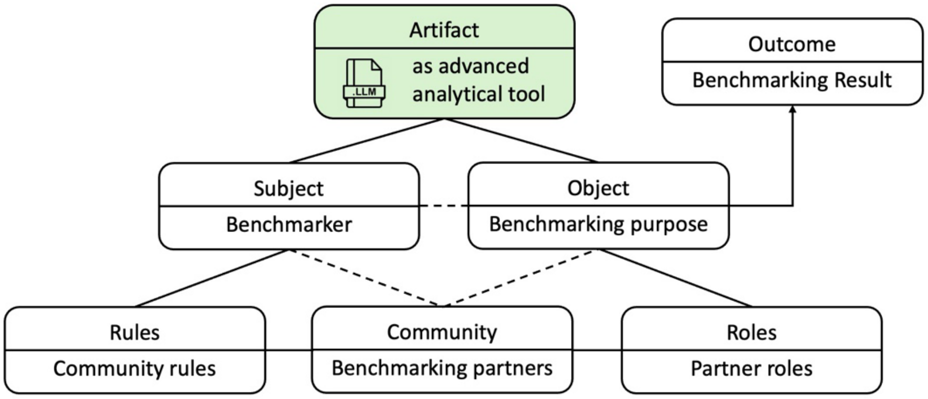


Figure 4.3: LLM as Artefact

4.1.2.2 LLMs as Community Enablers

Beyond serving as tools, LLMs could play an important role in stimulating community engagement and collaboration. By making the LLM publicly accessible, a broader spectrum of stakeholders, including policymakers, researchers, and practitioners, gains easy access to advanced analytical capabilities. This accessibility empowers these groups to contribute more effectively to the benchmarking process, enhancing the collective effort to improve GovTech services. The democratization of knowledge and analytical tools through LLMs thereby strengthens the ecosystem's inclusivity and collaborative potential.

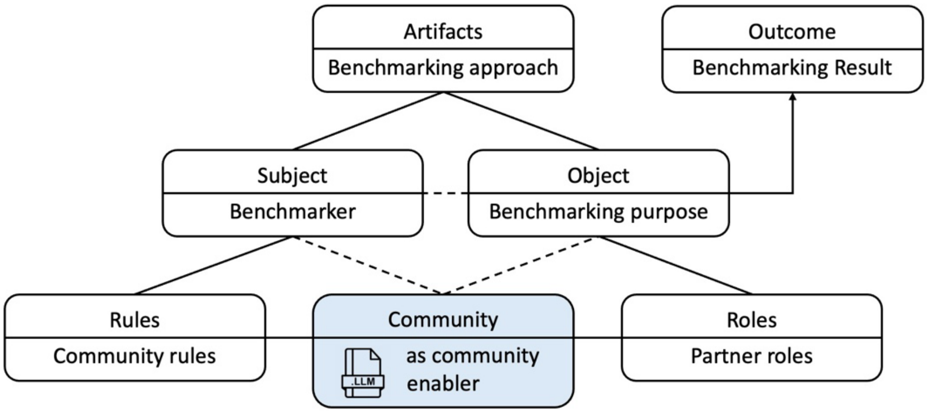


Figure 4.4: LLM as Community Enabler

4.1.2.3 LLMs as Autonomous Subjects

Perhaps the most revolutionary aspect of LLMs in the context of GovTech benchmarking is the potential to function autonomously as subjects, undertaking the benchmarking process independently. In this role, LLMs would not only analyze and collect relevant data, but also identify the useful framework, operationalize it, compare performance metrics, and even suggest improvements without direct human oversight. This self-sufficiency could transform benchmarking processes, offering a scalable, efficient means to continually assess and improve GovTech services, ultimately driving innovation and excellence in public service delivery. However, granting LLMs this degree of agency raises complex semantic, legal and liability concerns regarding safety (Burton et al., 2020), particularly within the governmental context and considering the early stages of AI integration in public systems.

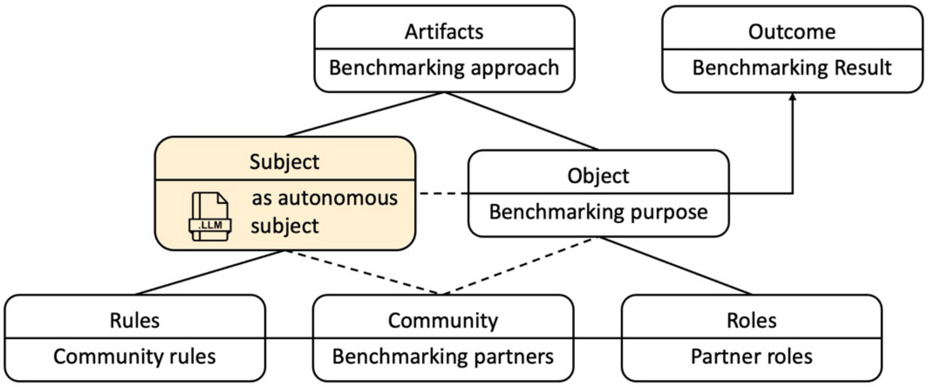


Figure 4.5: LLM as Autonomous Subject

4.1.3 Selection of LLM’s Role within the GovTech Benchmarking Ecosystem

Within the ecosystem of GovTech Benchmarking, this research designates the role of the LLM as an artefact, as depicted in Figure 4.3. This choice is made for two primary reasons. First, positioning the LLM as an artefact is in harmony with the methodology adopted in this research, which aims to develop an artefact that meets the defined objectives in 3.2. In this capacity, the LLM is used as a sophisticated tool for the operationalization of benchmarks, aligning perfectly with the research's objective to create a functional artefact.

Secondly, considering the novelty and innovative aspect of employing LLMs within this domain, their use as artefacts presents the most feasible approach. It situates LLMs as advanced analytical instruments that the benchmarker uses to discover, process, and analyze data, thus facilitating the generation of actionable insights for benchmarking activities. This initial exploration sets a foundation upon which future research could expand, exploring alternative roles LLMs might occupy within the ecosystem.

4.2 Solution Architecture

The subsequent subsections describe the technical structure of the proposed solution, as visualized in Figure 4.6. The full code of the artefact, including the data and results are available on the GitHub-repository of the project[1]. To provide a clear understanding, Section 4.2.1 presents the data pipeline, Section 4.2.2 the prompting pipeline, and Section 4.2.3 the evaluation pipeline.

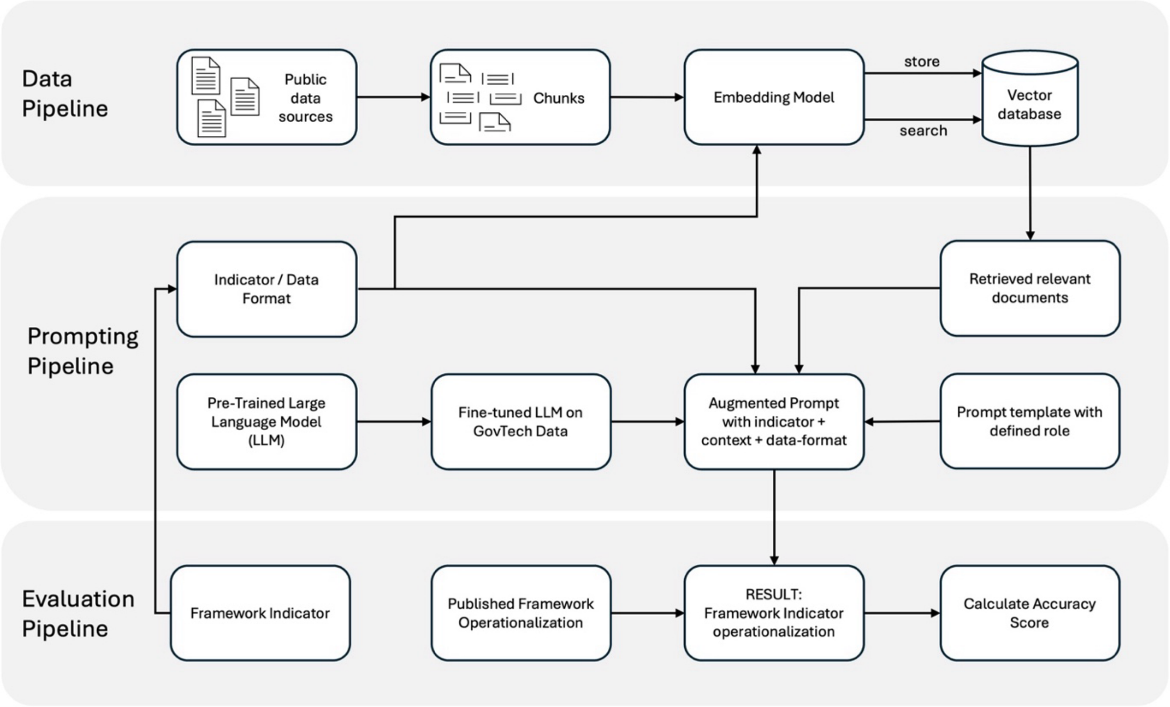

Data Pipeline
Public data sources
Chunks
Embedding Model
store
search
Vector database
Prompting Pipeline
Indicator / Data Format
Retrieved relevant documents
Pre-Trained Large Language Model (LLM)
Fine-tuned LLM on GovTech Data
Augmented Prompt with indicator + context + data-format
Prompt template with defined role
Evaluation Pipeline
Framework Indicator
Published Framework Operationalization
RESULT: Framework Indicator operationalization
Calculate Accuracy Score

Figure 4.6: Overview Solution Architecture

4.2.1 Data Pipeline

The data pipeline is designed to process publicly available data by first dividing it into manageable segments or chunks. These segments are then transformed through an embedding model, resulting in their storage within a vector database. This enables the use of the embedding model to locate data that is closely related or similar. The subsequent sections detail this process further: Section 4.2.1.1 outlines the criteria used for choosing the data sources. Section 4.2.1.2 proceeds to identify and select the relevant data sources based on these criteria. Section 4.2.1.3 explores the embedding models, discussing the rationale behind their selection and deployment. Finally, Section 4.2.1.4 describes the vector database employed to store the chunks.

4.2.1.1 Criteria Definition Data Sources

To ensure high-quality results are generated by the LLM, the data used for transfer learning methods should be of high quality. Data criteria are defined following the data quality framework by Wang & Strong (1996), which categorizes data quality into four primary categories: intrinsic, contextual, representational and accessibility. At least one dimension from each of these categories is considered. The defined criteria are presented in Table 4.1.

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

Table 4.1: Data Quality Criteria (Adapted from (Wang & Strong, 1996))

Intrinsic	Reputable	*Data must come from published sources or those controlled by reputable organizations.*
Contextual	Current	*The data needs to be frequently updated to mirror the most recent advancements. Alternatively, each piece of data should have a date mark to allow for selecting data from different time periods.*
	Relevant	*The selected sources must directly contain or be related to data about GovTech in the Netherlands. For broader datasets, mechanisms should be in place to filter and extract domain-specific information.*
	Substantial	*Data sources must at least contain 500 data entries.*
Representational	Interpretable	*Data should be presented in formats that allow for straightforward understanding and interpretation. Preferred formats include TXT and PDF.*
Accessibility	Accessible	*The sources need to be readily accessible for research purposes.*
	Authorized	*Considering the legal issues that have arisen with the use of data by LLMs, it's important to get permission to use this data. Doing this reduces legal risks and follows ethical standards for research.*

This carefully selected set of criteria ensures that the data supporting the study is both technically accurate and ethically gathered, providing a basis for valuable contributions to the GovTech benchmarking field. In the following section this set of criteria is used to identify data sources.

4.2.1.2 Selection Data Sources

After extensive desk research the data sources found are presented in the following table. Based on the defined criteria, several data sources are not included.

Table 4.2: Selection Data Sources

	Reputable	Current	Relevant	Substantial	Interpretable	Accessible	Authorized
Binnenlands Bestuur	✓	✓	✓	✓	✓	✓	✗
Dutch Government Open Data Portal	✓	✓	✓	✓	✓	✓	✓
iBestuur	✓	✓	✓	✓	✓	✓	✓
KVK open dataset	✓	✓	✗	✓	✓	✓	✓
Tendernet	✓	✓	✓	✓	✓	✓	✓
Woogle	✓	✓	✗	✓	✓	✓	✓
GovTech Today	✓	✓	✗	✓	✓	✓	✗
…							

Binnenlands Bestuur and iBestuur are prominent Dutch platforms focused on government and policy topics within the Netherlands. iBestuur specifically concentrates on governance relating to information technology, aligning closely with GovTech themes. In contrast, Binnenlands Bestuur covers a broader range of topics including finance and the environment, yet it efficiently categorizes its content, making it easy to pinpoint articles relevant to GovTech. Accessing content on both platforms requires authorization, and formal permission was sought to use their articles and information.

Additionally, the Woogle platform, which archives documents released following Open Government Act requests, was initially considered as a potential data source. However, its utility was limited as the required documents were already accessible through the Dutch Government Open Data Portal, rendering Woogle unnecessary for our purposes.

Moreover, platforms like the KVK (Dutch Chamber of Commerce) and Tendernet offer more technical details; KVK provides information about all Dutch companies and startups, including IT-companies. However, the data only included basic contact details which proved too limited. Conversely, Tendernet, which details government tenders, was deemed valuable as it offers insights into IT projects' scopes, the involved companies, and associated costs, hence it was included as a data source.

International platforms such as GovTech Today, despite offering a wealth of GovTech information, were not used due to their lack of specific focus on the Dutch context, stressing the importance of regional relevance in our research scope. Lastly, the presence of an empty row in the table suggests the possibility of incorporating more data sources in the future, should they satisfy the established criteria.

4.2.1.3 Splitting & Embeddings Model

The conversion of raw data sources into useable embeddings begins by breaking down the data into smaller segments, essential for both manageable and efficient processing. This segmentation is achieved using a recursive character splitter, which divides the text at specific characters: [".", "!", "?", "\n"]. Operating from left to right, the splitter continues until it produces segments that are suitably sized (up to 128 characters). This method of splitting not only facilitates easier handling but also increases the specificity of the embeddings, as each one represents a smaller slice of information.

The data chunks are converted into embeddings using a sentence-transformer model, which generates semantically meaningful sentence embeddings from text strings (Reimers & Gurevych, 2020). For this process, the *paraphrase-multilingual-MiniLM-L12-v2*[2] is used. This model, based on the BERT architecture (Devlin et al., 2018), contains 118M parameters and has been fine-tuned on 50+ languages, including Dutch. The model returns a 384-dimension embedding. The creation of the embeddings is performed on a personal computer equipped with an Apple M1 8-core GPU.

4.2.1.4 Vector Database

The embeddings are stored in a Chroma DB vector database. Chroma DB is an AI-native, open-source vector database designed to improve applications using large language models by efficiently managing embedding vectors (Chroma, n.d.). It simplifies the embedding and indexing of data for machine learning models, supporting a variety of operations like CRUD (Create, Read, Update, Delete), similar to traditional databases. With functionalities that allow the storage of metadata and the use of various similarity metrics for precise querying, Chroma DB offers a powerful yet user-friendly platform for developers and researchers. It integrates seamlessly with Python environments, facilitating rapid prototyping and robust application development.

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

In Chroma vector databases, various distance functions can be used to measure the similarity or dissimilarity between vectors, which are important for efficient retrieval of data. For the artefact design, the Squared L2 distance function will be used for its computational benefits and effectiveness in emphasizing larger discrepancies between data points. This distance function calculates the sum of the square of differences between corresponding elements of two vectors, following the formula:

$$d = \sum (A_i - B_i)^2$$

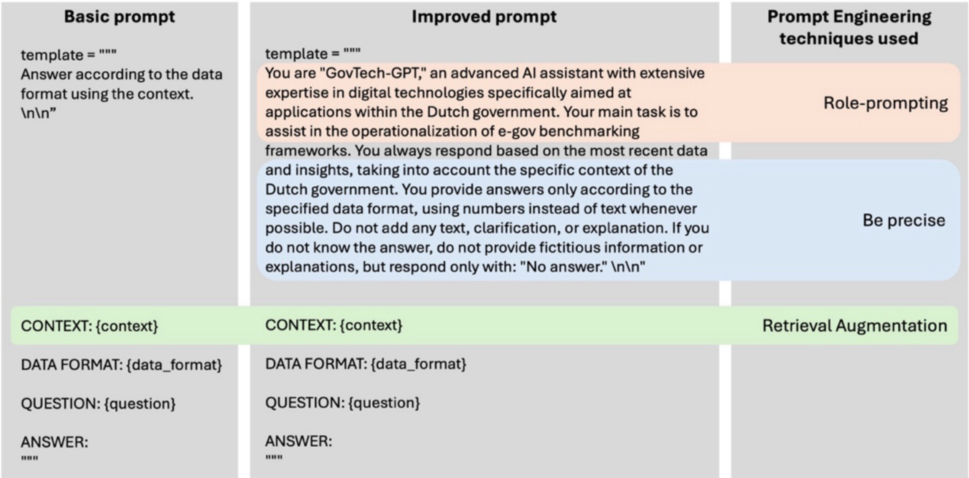
By squaring the differences, the Squared L2 distance accentuates significant deviations, which is essential in applications like outlier detection. This method omits the square root calculation required in standard Euclidean distance, thus reducing computational load, and improving performance in large-scale vector database operations.

4.2.2 Prompting Pipeline

4.2.2.1 Prompt Engineering

The prompting pipeline is designed to refine interactions with the LLM, ensuring the generation of contextually relevant responses. At the core of this process is a specific benchmark indicator with its specified data format, both retrieved from the existing GovTech benchmark. The context, sourced from the sentence-transformer model embeddings by using the Squared L2 formula explained in the previous section 4.2.1.4, enriches the prompt with domain-specific context. These three elements are then added to a prompt template which contains a predefined role, which helps to contextualize the response expected form the LLM. The role remains constant across prompts, ensuring uniformity in the LLM's perspective. How this prompt template looks like, is a whole area of research called 'prompt engineering'.

A paper by Chen et al. (2023) was used to implement the most common methods to improve the prompt template to optimize the usefulness of the LLM's output, namely being precise, role-prompting and retrieval augmentation. Other techniques mentioned like one-shot and few-shot prompting, chain of thought, tree of thoughts and graph of thoughts were not used due to their increased complexity. Figure 4.7 details the basic prompt template, and the prompt template after several prompt engineering techniques mentioned by Chen et al. (2023) were deployed.



Basic prompt	Improved prompt	Prompt Engineering techniques used
template = """ Answer according to the data format using the context. \n\n”	template = """ You are "GovTech-GPT," an advanced AI assistant with extensive expertise in digital technologies specifically aimed at applications within the Dutch government. Your main task is to assist in the operationalization of e-gov benchmarking	Role-prompting
	frameworks. You always respond based on the most recent data and insights, taking into account the specific context of the Dutch government. You provide answers only according to the specified data format, using numbers instead of text whenever possible. Do not add any text, clarification, or explanation. If you do not know the answer, do not provide fictitious information or explanations, but respond only with: "No answer." \n\n"	Be precise
CONTEXT: {context}	CONTEXT: {context}	Retrieval Augmentation
DATA FORMAT: {data_format}	DATA FORMAT: {data_format}	
QUESTION: {question}	QUESTION: {question}	
ANSWER: ""	ANSWER: ""	

Figure 4.7: Prompt template

4.2.2.2 Model Finetuning

The pre-trained base model in this setup is *BramVanroy/GEITje-7B-ultra-sft*, a Dutch instruction/chat model ultimately based on Mistral and aligned with AI feedback via Direct Preference Optimization (DPO). It is a DPO continuation of the Supervised Fine-Tuning (SFT) trained *BramVanroy/GEITje-7B-ultra-sft*, which in turn is based on *Rijgersberg/GEITje-7B*, which in turn is based on *Mistral 7B* and further pretrained on Dutch data.

For this research, the model was further finetuned on a dataset of question-answer pairs on the Dutch government (Rijksoverheid, n.d.) to achieve several goals: 1) increase its knowledge of the Dutch government, 2) improving phrasing and style to better match official Dutch government communications, 3) increase its ability to understand and interpret sector-specific terminology and jargon related to the Dutch GovTech context. The finetuning was conducted on an Amazon Web server equipped with an NVIDIA A100 GPU. The finetuned model, named *GovLLM-7B-ultra*, has been made publicly available on Hugging Face[3]. Figure 4.8 illustrates the loss progression during the finetuning process, which displays a clear plateau, indicating that the model has reached the limits of its learning capacity from the available data.

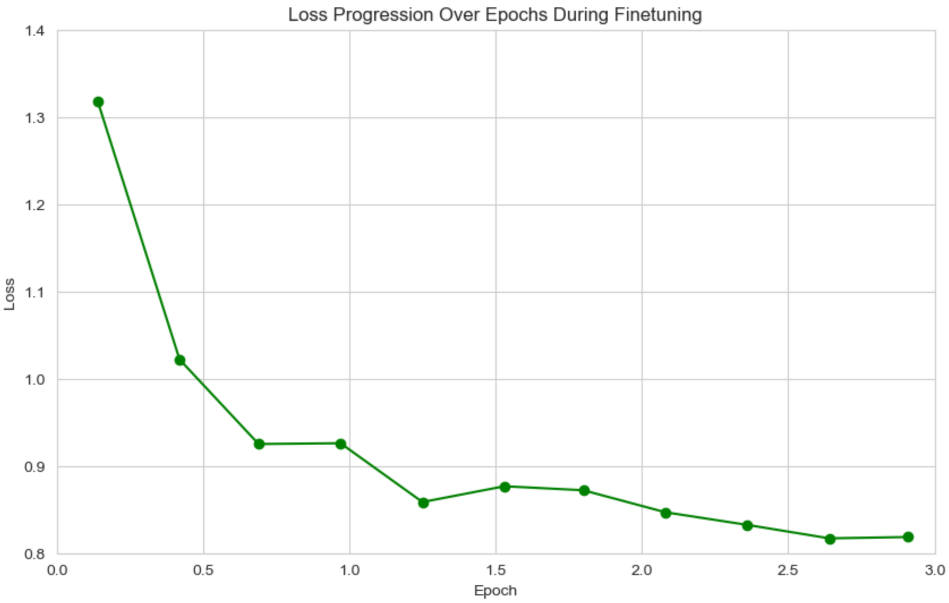


Figure 4.8: Loss Progression Over Epochs During Finetuning

4.2.3 Evaluation Pipeline

The evaluation pipeline facilitates the validation process, which is explained in section 2.4. For each operationalized benchmark indicator, it is essential to use the ground truth, which contains the officially published data from the benchmark organization. Given time and resource limitations, only one GovTech benchmark will be operationalized by the artefact. The selection of an appropriate GovTech benchmark is discussed in Section 4.2.3.1, while the corresponding experimental setup is described in Section 4.2.3.2.

[3] https://huggingface.co/Nelis5174473/GovLLM-7B-ultra

4.2.3.1 Benchmark Selection

To select an appropriate GovTech benchmark to operationalize it with an LLM, it's important to select a benchmark using a set of defined criteria. The benchmarks under consideration include the Digital Economy and Society Index (DESI) (European Commission, 2022), the E-Government Development Index (EGDI) (United Nations, 2022), the Digital Government Index (DGI) (Ubaldi & Okubo, 2020) and the GovTech Maturity Index (GTMI) (Dener et al., 2021), along with numerous national assessments (Public, 2021).

Firstly, evaluating the relevance of topics, the DESI and EGDI, although detailed, heavily focus on statistical data about internet connectivity and usage that are not feasible for LLM analysis, as these require traditional statistical methods and data gathering that an LLM cannot perform. DGI, although relevant, is limited by its partial openness regarding country data, which can hinder the validation of the LLM's operationalization. The GTMI provides an international perspective with a publicly available methodology and relatively open data access, making it a viable option for LLM-based analysis. National assessments, while valuable, exhibit high variability in methodology and scope, which can complicate standardization and comparability in a global context.

Given these observations, the selection process favors the GTMI due to its alignment with the selection criteria: it is international in scope, has an openly available methodology, offers accessible data, and covers relevant GovTech topics in detail. This makes GTMI particularly suitable for this research. Table 4.3 below, summarizes the evaluation based on the defined criteria:

Table 4.3: Criteria benchmark selection

	Scope	Methodology Availability	Data Availability	Relevance of topics
DESI	European	Public	Open	Partially relevant
EGDI	International	Public	Partially closed	Partially relevant
DGI	International	Public	Partially closed	Highly relevant
GTMI	International	Public	Open	Highly relevant
National Assessments	National	Varies	Varies	Varies

4.2.3.2 Experimental Setup

Section 2.4.2 details four validation methods: manual evaluation, exact match, edit similarity and random chance comparison. However, the selected GTMI benchmark can be operationalized by the designed artefact using various configurations. Specifically, it can be implemented using the base model in combination with prompt-engineering and the two transfer learning models described in Section 3.3.3: fine-tuning and Retrieval-Augmented-Generation. To thoroughly address research question 5, *"How does the accuracy of LLMs in operationalizing GovTech benchmarks compare to that of conventional methods?"*, it is essential to evaluate the different model configurations. This evaluation follows a full factorial design, a systematic experimental approach that tests all combinations of factors and their respective levels to determine their effects and interactions on a response variable. In this context, the factors under consideration are base model versus fine-tuned model, the inclusion or exclusion of Retrieval-Augmented-Generation (RAG), and the application or omission of prompt engineering. The resulting configurations are shown in Table 4.4.

Table 4.4: Full factorial design

Combinations	Prompt engineering (P)	RAG (R)	Fine-tuning (F)
ooo	0	0	0
Poo	1	0	0
oRo	0	1	0
ooF	0	0	1
PRo	1	1	0
PoF	1	0	1
oRF	0	1	1
PRF	1	1	1

With this experimental setup, the GTMI benchmark was operationalized by the artefact. The results are presented in the following chapter.

5 Results

This chapter presents the results, thereby aiming to answer research question 5: *"How does the accuracy of LLMs in operationalizing GovTech benchmarks compare to that of conventional methods?".* This is done using the four ways of assessing the accuracy of the model as explained in Section 2.4.2, following the experimental setup as presented in Section 4.2.3.2. The structure of the chapter is organized as follows: Section 5.1 discusses the manual evaluation including the evaluation of multiple choice questions, Section 5.2 presents the exact match results, Section 5.3 shows the results by edit similarity, and Section 5.4 combines the results from these three methods to assess the overall accuracy of the model configurations.

5.1 Manual Evaluation

Each output from the LLM is manually evaluated by comparing its operationalized indicators to the ground truth. In Section 5.1.1, the distribution of answers across manual categories is presented. Section 5.1.2 discusses the accuracy based on manual evaluation, while Section 5.1.3 focuses on the accuracy of multiple-choice questions.

5.1.1 Answer Distribution

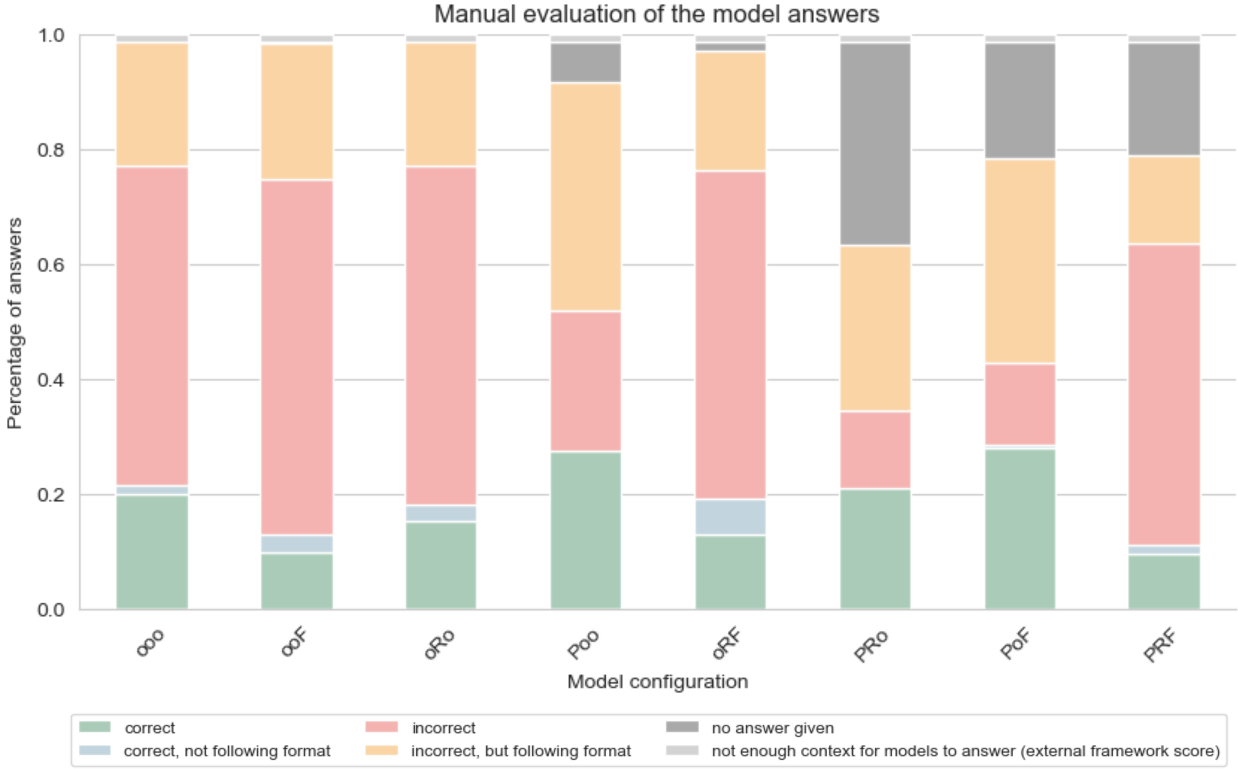


Figure 5.1: Answer distribution manual evaluation

By manually comparing the operationalized indicators to the ground truth, model outputs are categorized into one of five categories: correct; correct but not following data format; no answer; incorrect but following data format; and incorrect with incorrect data format. Figure 5.1 shows a stacked bar plot, showing how accurate each model configuration performs on operationalizing the GTMI benchmark.

Moreover, it shows that some indicators could not be answered because they were derived from external framework scores from other benchmarks. The percentage of these external framework scores is consistent across all model configurations. The 'no answer given' category is much more prevalent in configurations with prompt engineering. This is due to the additional instructions included in the prompts, explicitly directing the model to respond with 'no answer' when no data was available, rather than making up a response.

5.1.2 Accuracy by Manual Evaluation

The accuracy per model configuration, based on manual evaluation, is calculated using Equation 2.1 as described in Section 2.4.2. This means that an answer is deemed accurate when the correct answer is in the output, regardless of the data format. The results are presented in Figure 5.2 below.

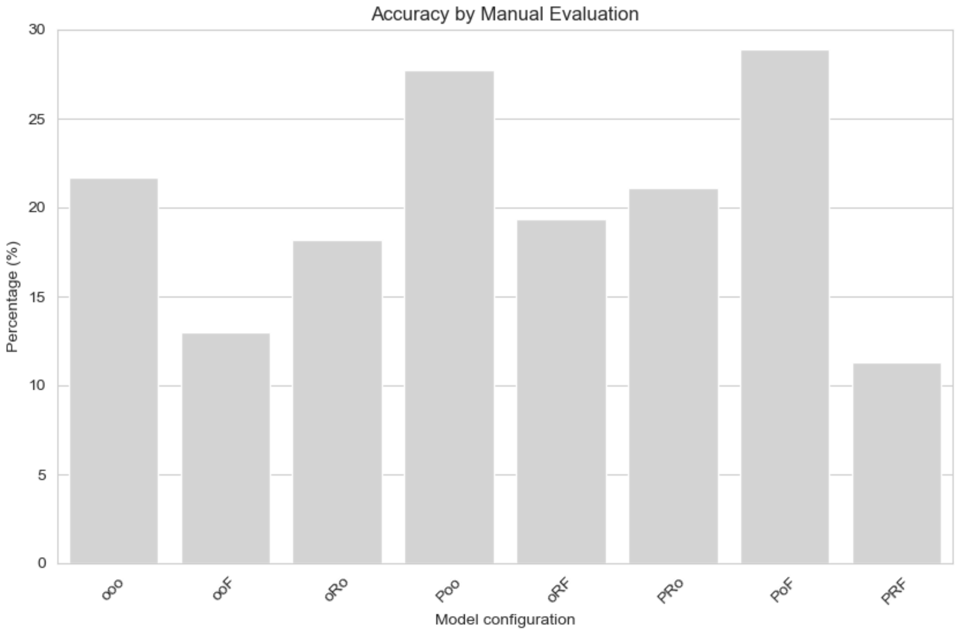


Figure 5.2: Accuracy by Manual Evaluation

The model configuration achieving the highest accuracy, based on manual evaluation, is PoF. This configuration involves a fine-tuned model with prompt engineering and without RAG. The second highest accuracy is achieved by Poo, which is the base model with prompt engineering. The configurations with the lowest accuracy are PRF and ooF, both of which use a fine-tuned model.

5.1.3 Multiple Choice Accuracy

The GTMI benchmark includes a total of 351 indicators and subindicators. Among these, 200 indicators and subindicators require a multiple-choice answer according to the prescribed data format. The multiple-choice questions range from 2-answer to 5-answer options. The accuracy of the model configurations in answering these questions, as determined by manual evaluation, is illustrated in Figure 5.3 below. The model configurations that achieve the highest accuracy on multiple choice questions, based on manual evaluation, PoF, Poo and PRo, in descending order of accuracy. All of these configurations use prompt engineering.

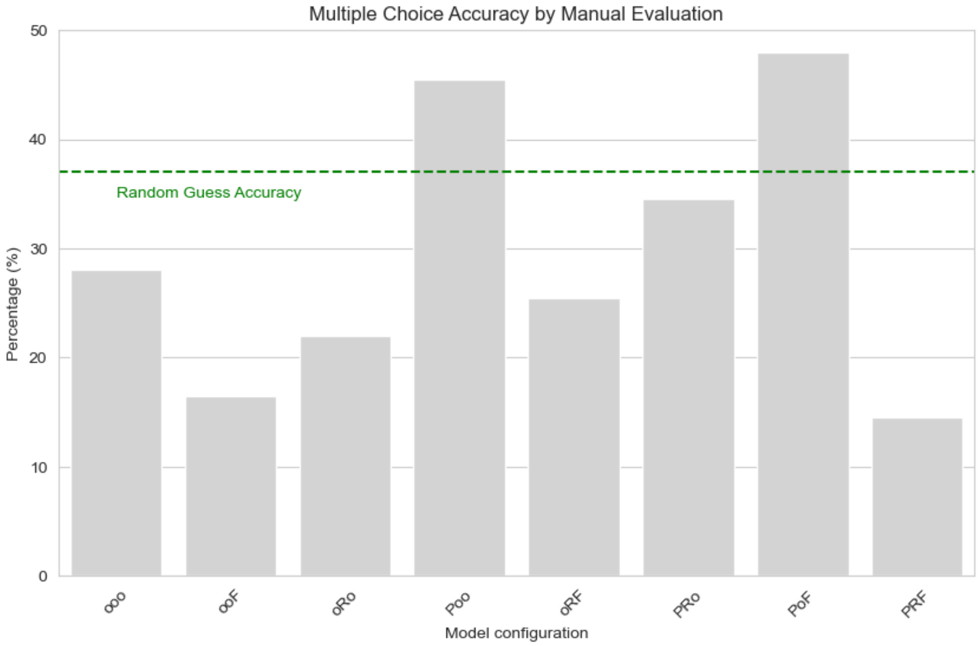


Figure 5.3: Multiple Choice Accuracy by Manual Evaluation

To evaluate whether random guessing outperform these achieved accuracies, an evaluation is conducted by calculating the expected accuracy of random guesses for the multiple-choice questions following Equation 2.4. Using the data in Table 5.1, the expected random accuracy, weighted by the number of indicators in each category is 37.0%. This value is illustrated by the green line in Figure 5.3. This makes it clear that only two model configurations outperform random chance, namely Poo with an accuracy of 45.5% and PoF with an accuracy of 48.0%.

Additionally, the accuracy for multiple-choice questions is calculated separately for each number of answer options (ranging from 2 to 5). The results are presented in Table 5.1 below.

Table 5.1: Multiple Choice Accuracy by Number of Answer Options

Accuracy (%) Multiple Choice options	ooo	ooF	oRo	Poo	oRF	PRo	PoF	PRF
2 (n=65)	40.00	20.00	32.31	46.15	36.92	30.77	47.69	24.62
3 (n=97)	27.84	16.49	20.62	49.48	25.77	36.08	46.39	12.37
4 (n=31)	3.23	9.68	6.45	3.23	6.45	35.48	54.84	0.00
5 (n=7)	28.57	14.29	14.29	42.86	0.00	42.86	42.86	14.29

Assuming that having more options makes it harder to choose the correct answer, it is expected that accuracy will decrease as the number of choices increases. Generally, this pattern holds true. However, there are notable exceptions in the last row, where multiple-choice questions with five options show surprisingly high accuracies, reaching up to 42.86%.

5.2 Exact Match

The Exact Match algorithm determines the proportion of the LLM's outputs that perfectly align with the ground truth data. Due to the stringent requirement for exact correspondence, the achieved percentages are significantly lower. The results are displayed in Figure 5.4.

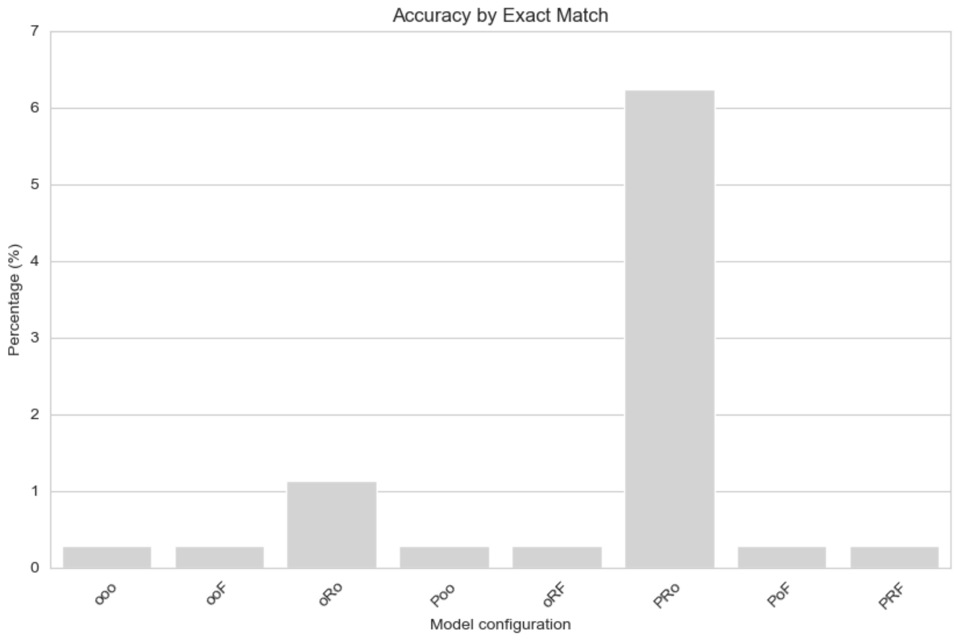


Figure 5.4: Accuracy by Exact Match

The model configuration that achieves the highest accuracy based on exact match is PRo, which employs the base model with prompt engineering and RAG. The second highest accuracy is achieved by the oRo configuration, which also uses the base model with RAG but without prompt engineering. All other model configurations score below 0.5% accuracy and show no significant differences from one another.

5.3 Edit Similarity

The final method for assessing accuracy is Edit Similarity. As described in Section 2.4.2, this method uses the Levenshtein distance to calculate edit similarity by counting the number of insertions, deletions, and substitutions needed to transform the model's prediction into the ground truth, normalized by the length of the longest word involved. The accuracies are shown below in Figure 5.5.

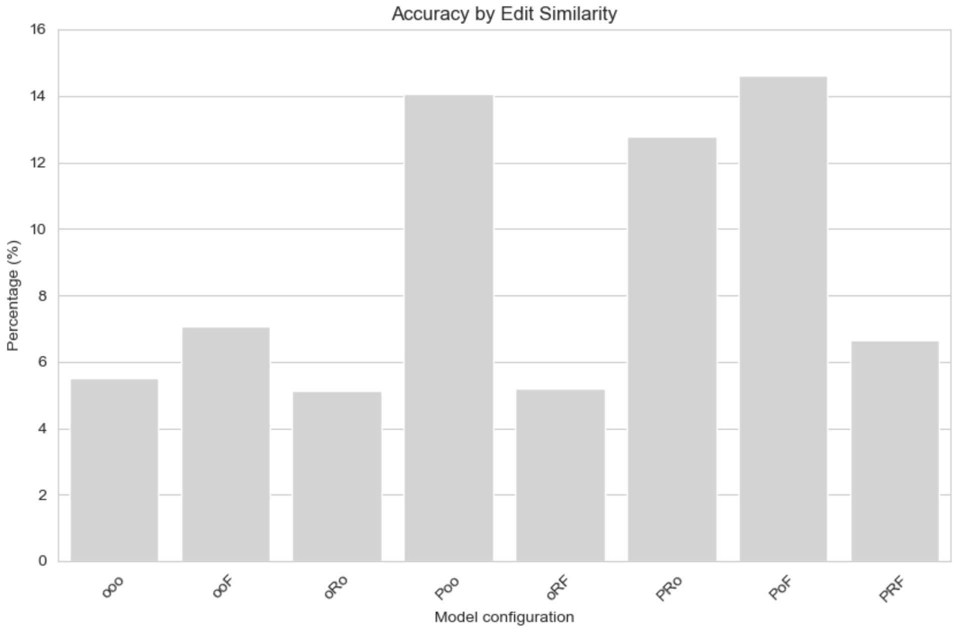


Figure 5.5: Accuracy by Edit Similarity

The model configurations achieving the highest accuracies, with values between 12 and 15 percent, are Poo, PRo, and PoF. These configurations all incorporate prompt engineering. In contrast, the other model configurations without prompt engineering show no significant differences from each other and score between 4 and 6 percent.

5.4 Combined Accuracy Assessment

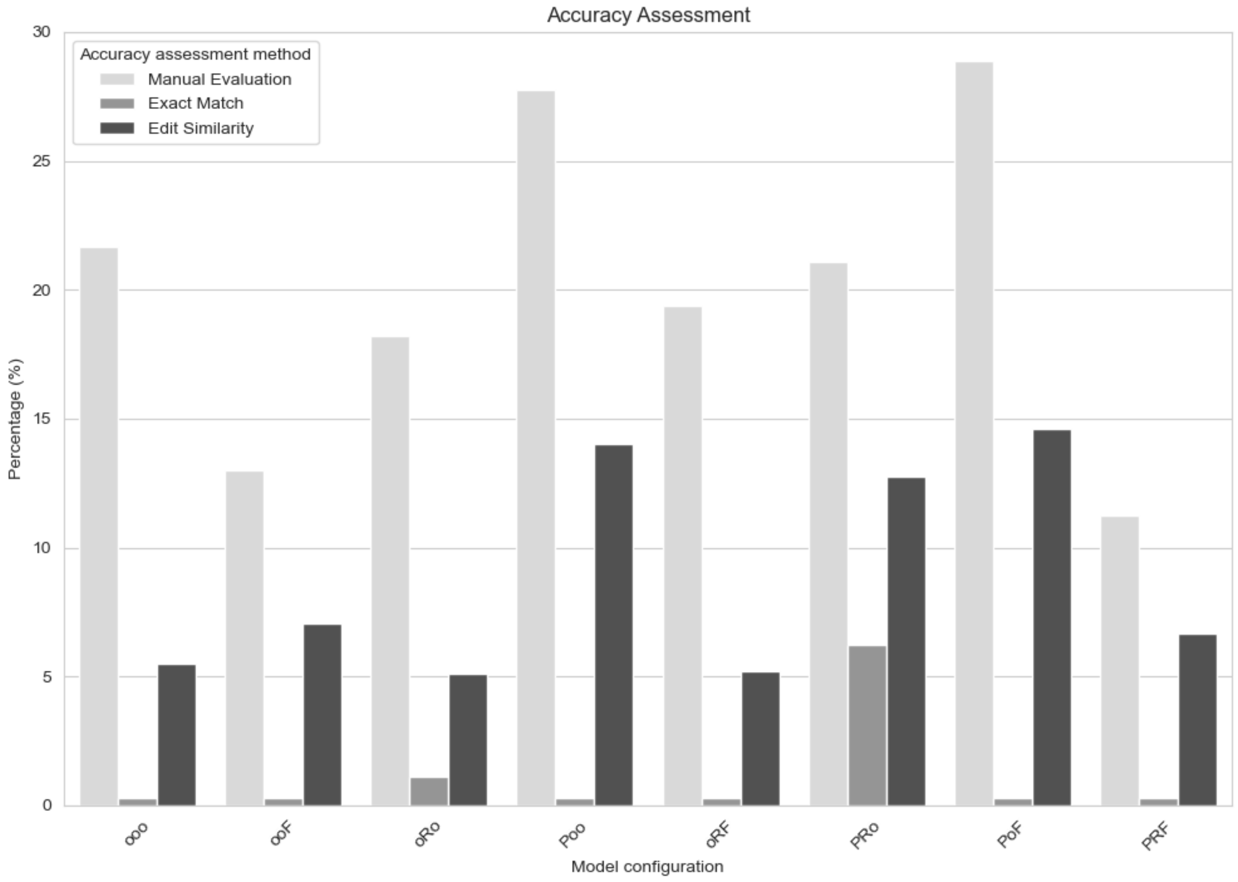


Figure 5.6: Full Accuracy Assessment

The accuracies from the three different methods for calculating accuracy are presented together in a single plot. This plot illustrates that manual evaluation yields the highest values, followed by edit similarity, and finally, exact match. This trend is expected, as manual evaluation considers good answers that may not strictly adhere to the data format as accurate. In contrast, the two automated methods do not account for such variations. Additionally, the exact match method is highly stringent, only marking an answer as accurate if it exactly matches the ground truth.

Having presented the accuracy data from different model configurations and assessment techniques, the next chapter will provide a detailed interpretation of these results. It will explore the implications of these accuracies for the utility of benchmarks for policymakers.

6 Discussion

This chapter critically analyzes the results presented in the previous chapter, focusing on the implications for GovTech benchmarking when using LLMs for benchmark operationalization. The structure of this chapter is as follows: Section 6.1 interprets the results from the previous chapter to assess the accuracy of LLMs in benchmarking operationalization, a prerequisite for LLMs to potentially increase the utility of benchmarks used by policymakers in decision-making. Section 6.2 explores how LLMs can address the inherent challenges of timeliness, complexity, and data aggregation. Subsequently, Section 6.3 evaluates the implications of this study for the GovTech Benchmarking process, while Section 6.4 identifies the remaining challenges despite the artefact's implementation. Collectively, these four sections aim to answer the main research question: *"How do LLMs operationalizing GovTech benchmarks mitigate inherent challenges of timeliness, complexity, and data aggregation, increasing their utility for policymakers?".*

6.1 Interpretation of Findings

6.1.1 Accuracy as a Prerequisite

To determine whether the artefact can potentially mitigate the inherent challenges of benchmarks, as identified in the literature review in Chapter 3, it is essential to first assess the model's accuracy. Specifically, if the model lacks sufficient accuracy, it would be useless to address other challenges related to benchmarking, as the model's output must possess a certain degree of reliability and utility. Accurate operationalization is therefore a critical requirement, without which the artefact cannot contribute meaningfully to overcoming issues such as timeliness, complexity, and data aggregation. Therefore, establishing the model's accuracy is a prerequisite for evaluating whether the artefact can contribute to mitigating the inherent challenges of timeliness, complexity, and data aggregation, thereby increasing the utility for policymakers using these benchmarks as aids in decision-making. This assessment is conducted in the next section.

6.1.2 Assessing Model Accuracies

When reviewing the results from Chapter 5, one might initially consider that a maximum accuracy of 30% suggests that even random guessing could outperform the model. This assertion, however, is quickly refuted when considering the complexity of open-ended questions, where making a correct guess is significantly more challenging. Nonetheless, as outlined in Section 5.1.3, there are 200 indicators and sub-indicators that require answers in a multiple-choice format according to specified data standards. To better validate the accuracy of the artefact, it is important to determine whether random guessing might actually surpass the model's performance. This validation was performed and illustrated in Figure 5.3. With a average random guess accuracy of 37.0%, two model configurations outperform random chance, namely model configurations Poo with an accuracy of 45.5% and PoF with an accuracy of 48.0%. This suggests that when implementing prompt-engineering, the model is better configured to deal with the complexity of the indicators.

When further examining the different transfer learning methods applied, it appears that model configurations using prompt engineering, outperform other configurations for all three validation techniques: manual evaluation, exact match and edit similarity. For example, PoF and Poo, demonstrated the highest accuracy in manual evaluations of multiple-choice questions, with PoF reaching an accuracy of 48.0% and Poo 45.5%, both surpassing the baseline accuracy of random guessing set at 37.0%. This finding is supported by recent literature (White et al., 2023). By refining input queries to align with the model's internal knowledge, this method improves output accuracy and relevance without further training. Additionally, prompt engineering has been proven to have a significant positive impact in various domain-specific contexts, as evidenced by findings from Heston & Khun (2023) and Yu et al. (2023).

Despite these configurations outperforming random chance, the achieved accuracies are not particularly impressive, especially when comparing to other studies. For instance, in the context of open-domain question answering, like the Stanford Question Answering Dataset (SQuAD 1.1), models such as BERT (Devlin et al., 2018) and T5 (Raffel et al., 2019) commonly achieve accuracies above 80%. However, context specific domains like the legal sector present unique challenges that are not reflected by high accuracies. For example, state-of-the-art models can only achieve about 28% accuracy on the legal JEC-QA question-answer dataset (Zhong et al., 2020). Similarly, in clinical question answering, models tested against the LongHealth benchmark exhibit a wide range of accuracies, from 32% to 77% (Adams et al., 2024). These examples show the nuanced nature of performance metrics in context-specific domains, emphasizing that such accuracies, while lower, are reflective of the specialized requirements and complexities inherent in specific domains.

Lastly, some inaccuracies in the ground truth data used for the model assessment also affected the results. This data contained a few obvious errors, such as indicator I-33, which will be detailed in Section 6.3.1. Additionally, some indicators could be interpreted in various ways. For instance, indicator I-1 of the GTMI asks whether there is a shared cloud platform available for all government entities. The ground truth data answers affirmatively, referencing ODC-Noord, one of the four Dutch Governmental Data Centers. However, the artefact's output cited the 2022 Letter to Parliament on government-wide cloud policy (Van Huffelen, 2022), which allows public organizations to use cloud services like Amazon Web Services under specific conditions. Consequently, the model concluded that a shared cloud platform for *all* government entities does not exist. This example illustrates the subjective nature of some indicators, resulting in lower accuracy despite the model providing a valid and useful answer. Such discrepancies, occurring occasionally, contributed to a reduced accuracy.

In conclusion, the assessment of the model's accuracy reveals the artefact's potential to address the inherent limitations of benchmarks. While initial results might suggest that random guessing could outperform the model, deeper analysis shows that with the right configurations, particularly those using prompt engineering, the model can operationalize the benchmark with a sufficient degree of reliability and utility, which is a critical requirement to be of use for policymakers. Moreover, given the time and resource constraints for this project, several limitations of the artefact, as described in Section 8.1, remain and could be improved, potentially improving the accuracies significantly. For now, the achieved accuracy stresses the model's capability to meaningfully contribute to overcoming issues like timeliness, complexity, and data aggregation in benchmarking processes. The next section will cover how the artefact might contribute to the mitigation of these challenges.

6.2 Mitigating Inherent Challenges

6.2.1 Improving Timeliness

One of the most significant improvements that LLMs can contribute to the GovTech benchmarking process is the reduction in time required for data collection. This aspect was highlighted as an unresolved challenge in the literature review in Chapter 3. Furthermore, all interviewees confirmed this during their interviews. Mark Pryce (Appendix B) specifically highlighted the DGI-process as a prime example, where initial data collection is followed by an extensive validation period. Typically, after the data collection phase, there is a validation period of around 15 months during which individual countries validate their data with the OECD. The final benchmark was released in February of this year, although data collection had concluded in the summer of 2022. This sequence of events highlights the lengthy timelines involved in current benchmarking methods. Felipe González-Zapata from the OECD and responsible for the DGI acknowledges this (Appendix C) but stresses the fact that it requires time to ensure the data and the results are extremely accurate.

In contrast, with LLM integration, the time required to populate the framework ranged from 20 to 30 minutes depending on the model configuration. This significant reduction in time shows the potential of LLMs to streamline complex processes, allowing for more efficient and timely data handling in governmental benchmarking. However, this does not imply that the full process can be shortened to 20 to 30 minutes. As will be discussed in Section 6.4.2, a validation process is still necessary to ensure the accuracy and reliability of the data used within the benchmark.

6.2.2 Improving Inclusion of Complexity

A second challenge in GovTech benchmarking, as highlighted in the literature review in Chapter 3, is the insufficient handling of complexity. Current benchmarks predominantly focus on readily quantifiable aspects such as websites and services, rather than analyzing the more complex back-office structures and processes. The current model, as discussed in the results, has not successfully addressed this issue. In fact, the lower accuracy levels indicate a failure to capture certain complexities.

However, practitioners mention time and resource constraints as one of the primary reasons for not including more complexity in benchmarks. For instance, Nicky Tanke (Appendix B) notes that data formats in the DGI typically require a choice between options A *or* B, neglecting scenarios where both may apply. However, expanding the range of answer options to include this complexity would necessitate additional time and resources, further prolonging the process. Additionally, many questions now intersect multiple topics. Mark Pryce illustrates the time it takes to deal with more complex questions with the following scenario: "When you talk about, for example, what are we all doing to make digitalization activities by the government more sustainable? Well, we then have to request this from all government organizations (…) So that makes it quite laborious" (Appendix B).

These examples show that the inclusion of complexity is limited due to the significant time and resources required. Consequently, the current laborious data collection process hinders the inclusion of more complexity. However, as seen in the previous section, LLMs face significantly fewer time and resource constraints. Moreover, considering the existing time and resource constraints which have led to several limitations of the artefact, detailed in Section 8.1, there also remains significant potential for the artifact to improve in accuracy. Therefore, if the artifact were to be improved for greater accuracy, potentially much more complexity can be included, as the artefact requires fewer resources and operates at a much higher speed.

6.2.3 Improving Flexibility in Data Aggregation

The third challenge in GovTech benchmarking, as highlighted in the literature review in Chapter 3, is the inflexibility in aggregation levels. The artefact has so far only demonstrated the operationalization of a benchmark at national level and has therefore not directly addressed this issue. However, similar to the challenge of complexity, the lack of diverse aggregation levels is primarily due to the extensive time and resources required.

Mark Pryce (Appendix B) notes that the primary reason for the lack of diverse aggregation levels is the extensive effort required even at the national level. Yet, there is recognized potential and value in developing benchmarks for specific sectors, domains, or other government levels. Mark Pryce suggests that ministries such as Education or Health would greatly benefit from benchmarks that identify which countries perform well and could provide valuable learning opportunities.

Therefore, although the artefact does not directly address the challenge of data aggregation, its ability to operationalize benchmarks at a much higher speed and with significantly fewer resources presents a substantial opportunity. By using these efficiencies, the artefact has the potential to address the inflexibility in aggregation levels.

Next to the challenges of timeliness, complexity, and data aggregation, there are also more contextual contributions on social and ethical level, which will be discussed in the next section.

6.3 Implications for the GovTech Benchmarking Process

Beyond the technical contributions of using LLMs for benchmark operationalization in addressing challenges of timeliness, complexity, and data aggregation, the use of LLMs will also have broader social and ethical impact. The following subsections explore three of these implications.

6.3.1 Mitigating Subjectivity by Providing Alternative Sources

One of the issues highlighted by practitioners and observed in the ground truth data is the tendency to report data in a way that boosts a country's ranking. For instance, indicator I-33 in the GTMI asks whether there is a national GovTech institution. According to the ground truth data for the Netherlands, this should be the 'GovTech Institute', citing NLDigital as the source, which is actually the collective of the Dutch digital sector. By reporting the existence of such an institution, the Netherlands would score higher on the benchmark. However, in reality, no such 'GovTech Institute' exists, and NLDigital is something entirely different. It is unclear whether this data was intentionally misreported to achieve a higher score or if it was an honest mistake. Despite both Mark Pryce (Appendix B) and Felipe González-Zapata (Appendix C) asserting that the methods used are rigorous, both acknowledge that some countries try to maximize their scores.

As described in Section 1.4.3, the integrity of the data collection process is an important ethical issue. Mistakes or biases in data can lead to misguided policies that might increase existing inequalities or create new ones. The perverse incentive to maximize scores could result in harmful consequences for citizens and companies due to misplaced priorities and unfair allocation of resources, based on misleading benchmark results. For instance, if a benchmark reports a country to have a well-functioning and advanced e-job portal in place, no funding might be allocated to improve it, even if the portal is actually ineffective. This misrepresentation can severely impact jobseekers and companies who rely on such services. This issue is confirmed by Felipe González-Zapata from the OECD who is involved in both the data collection and validation process (Appendix C). He states that "If the performance of a country in a specific area would be inaccurately represented, this mismatch could lead to countries prioritizing or deprioritizing topics based on incorrect data, resulting in significant policy and political implications. That is a very sensitive issue."

The use of LLMs could mitigate this issue in two significant ways. Firstly, an LLM does not have the inherent bias to score as high as possible, leading to more objective operationalization. Secondly, by using RAG, multiple data sources are used to generate one answer, and these sources can be verified. This approach prevents the cherry-picking of sources that would otherwise be used to artificially boost scores. As a result, the use of an LLM could lead to more accurate and reliable benchmarking, reducing the impact of biased data reporting.

6.3.2 Responsive Policies

Another implication for the GovTech benchmarking process is a result of the capability of LLMs to operationalize benchmarks in less than 30 minutes, as opposed to current benchmarks that often rely on outdated data. This rapid processing enables policymakers to react more promptly to the outcomes of benchmarks. For instance, during a public health crisis like the COVID-19 pandemic, having access to near real-time data could allow policymakers to implement targeted interventions, such as distributing resources to the development of specific functionalities for a COVID-19 app used for contact tracing.

However, this increased speed in policy formulation also brings ethical and social considerations. Rapid data processing can pressure policymakers to make quick decisions without thorough deliberation, possibly overlooking long-term implications and minority perspectives. For example, a prompt decision based on near real-time economic data within the benchmark might favor short-term economic gains over sustainable development, thereby neglecting environmental or social equity considerations. Additionally, relying too heavily on LLM operationalized benchmarks could reduce human oversight and critical evaluation needed for balanced governance.

Therefore, policymakers must weigh the benefits of swift responsive policies against the need for thorough analysis and inclusive dialogue to ensure that policies are well-considered and maintain public trust. Balancing these factors is essential to use LLMs for benchmarks effectively while upholding democratic values and ethical standards.

6.3.3 Fair & Sensitive Policies

A final implication for the GovTech benchmarking process is the enablement of fair and sensitive policies. Benchmark results can significantly influence policy outcomes, such as the allocation of funds and resources to specific areas, impacting both citizens and companies, as described in Section 1.4.3. The selection of indicators included in the benchmark, determining 'what is important to measure', inevitably steers governmental efforts in particular directions (Przeybilovicz et al., 2023). This focus is especially critical when considering that local and regional contexts can vary widely, potentially leading to the oversight of minority groups and unique local needs. By incorporating a broader range of parameters, made feasible through the efficiency of LLMs, there is potential for the development of policies that are more equitable and attuned to the diverse conditions within different regions.

However, the inclusion of additional parameters also brings ethical challenges. While the capability of LLMs to handle extensive data could improve the fairness of policymaking, it requires careful consideration to ensure that the data is representative and free from biases. However, automated systems inevitably make biased decisions (Mittelstadt et al., 2016) and could thus lead to unfair outcomes. Therefore, while LLMs offer the promise of more fair and sensitive policies, their deployment must be accompanied by thorough ethical oversight, ensuring that the technology serves to improve rather than undermine social equity.

Beyond these contributions to the GovTech benchmarking process, it is important to recognize that the integration of LLMs, although promising, does not solve all challenges within benchmarking. There are also challenges within the benchmarking process that will undoubtedly persist. These challenges will be discussed in the next section.

6.4 Addressing Persisting Challenges

6.4.1 Transparency and Data Privacy Concerns

One of the principal challenges in integrating AI into governmental operations concerns transparency and data privacy. Mark Pryce articulates this caution well, stating: “At this moment, we are very cautious, primarily because, although we can see the potential benefits for our operations similar to consultancy firms and businesses, we lack confidence in several key areas. We need sufficient transparency about how models are trained, and the data used. Specifically, we are concerned about the handling of personal data included in the training sets. Many large models are trained on the Common Crawl, which contains vast amounts of personal data. Additionally, there are unresolved issues regarding the management of intellectual property during the training of these models. Until these questions are adequately addressed, our approach remains cautious" (Appendix B). This perspective aligns with the Government-wide vision on generative AI of the Netherlands (Ministry of the Interior and Kingdom Relations, 2024), which prescribes strict guidelines on the usage of Generative AI.

During the development of the artefact, efforts were made to adhere closely to these guidelines. Permissions for data use were secured prior to its inclusion in the models, ensuring compliance with copyright laws. The use of RAG improves transparency by not only improving the accuracy and reliability of LLMs with facts fetched from external sources but also by revealing the specific documents and data sources that contributed to each response. This visibility allows users to trace back the origins of the information provided, thereby making the system's decision-making process more transparent. Furthermore, the entire codebase is open-sourced and accessible through Github[4].

Nevertheless, a significant limitation was our dependency on a pre-trained Mistral model. Due to time and resource constraints, it was necessary to use an off-the-shelf model. This model was selected for its open-source availability under the Apache 2.0 license, allowing for unrestricted use. Despite this, the lack of detailed information on the specific datasets and the full training process impedes total transparency. The model includes data from broad-ranging sources like Common Crawl.

To address these challenges, the Dutch initiative, GPT-NL, once available presents a potential solution (TNO, 2023). GPT-NL aims to comply with the EU-values for trustworthy AI (European Commission, 2020), one of which is transparency: the data, system and AI models should be transparent. This is done by fully disclosing training datasets and processes, allowing for the verification of data integrity and security. Using this model could mitigate the limitations associated with the use of commercial off-the-shelf AI models, providing a solution that adheres to strict governance and ethical standards in governmental AI applications.

6.4.2 Integration into Public Decision-Making

Another important persisting challenge when using the artefact for operationalizing GovTech benchmarks, is the integration of the artefact into the process of GovTech benchmarking. As Mark Pryce puts it: “Consider that the actual published data are the result of not only data itself but from an analytical process undertaken by public officials in various countries. These officials request or receive data from colleagues and then formulate a response. This response subsequently undergoes a review process, involving a colleague, a manager, and ultimately, likely approval from a director. Therefore, the published data also embody a significant amount of experience from the involved officials" (Appendix B).

While the artefact is designed to speed up the benchmarking process, the validation by public officials and the benchmarking organization remains crucial. Felipe González-Zapata from the OECD (Appendix C) confirms this and stated this multiple times, as for him “The credibility point is really, really important”. Therefore, the data must be validated extensively. According to Felipe, this includes “honest conversations with countries, which involves not only emails but also calls and meetings to explain our position. Towards the end of the process, we have meetings with most countries to wrap up and explain why certain data cannot be validated. Countries often further explain their perspective, and we try to

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

find a common understanding. If their explanation aligns with our criteria, we work towards a common understanding to accept the evidence”.

Therefore, although Mark Pryce and Nicky Tanke (Appendix B) can envision a future where AI agents handle all interactions between countries and benchmarking organizations with minimal human intervention, their current perspective is more conservative. Just like Felipe González-Zapata (Appendix C), they view AI primarily as a tool for decision support—automating preliminary tasks yet still requiring the oversight and involvement of policy officers and department heads.

7 Conclusion

In this chapter, the findings of the research are synthesized by first addressing each sub-question and then integrating these insights to answer the main research question. The conclusion drawn from these answers provides a thorough understanding of the use of LLMs for GovTech benchmark operationalization.

1. Which are practical limitations of current GovTech benchmarks that affect their utility for policymakers?
The examination of practical limitations in current GovTech benchmarks reveals significant issues that undermine their utility for policymakers. Three primary limitations are identified: resource-intensive methodologies that provide retrospective rather than real-time analysis, a lack of complexity that overlooks digital infrastructures and emerging technologies in favor of simpler metrics, and improper levels of aggregation that render results less useful. These shortcomings result in benchmarks that provide less meaningful insights for decision-making, thereby compromising their utility for policymakers.

2. Which AI-technologies are capable of mitigating the limitations of timeliness, lack of complexity, and lack of suitable aggregation within current GovTech benchmarking methodologies?
AI technologies that can mitigate the limitations of timeliness, complexity, and data aggregation in GovTech benchmarking include neural networks and natural language processing (NLP), particularly Large Language Models (LLMs). LLMs can process large datasets quickly and handle complex information structures. Techniques like prompt-engineering, fine-tuning and Retrieval-Augmented Generation (RAG) improve their accuracy and contextual relevance. These capabilities make LLMs ideal for mitigating the identified limitations within current GovTech benchmarking methodologies.

3. How does Activity Theory guide the placement of LLMs within the GovTech benchmarking ecosystem?
Activity Theory facilitates the integration of LLMs into the GovTech benchmarking ecosystem by identifying their potential roles: as artefacts, community enablers, or autonomous subjects. Consistent with the Design Science Research Methodology of this study, which aims to develop an artefact meeting specific objectives, LLMs are predominantly positioned as artifacts. In this capacity, LLMs function as advanced analytical tools that aid benchmarkers in processing and analyzing data, effectively operationalizing benchmarks. Additionally, Activity Theory provides a theoretical foundation for future research to further investigate and expand upon the potential roles of LLMs within the ecosystem.

4. What is required from a solution architecture supporting LLMs to operationalize GovTech benchmarks?
The essential components of a solution architecture supporting an LLM-based operationalization of GovTech benchmarks include a data pipeline, a prompting pipeline, and an evaluation pipeline. The data pipeline manages the segmentation and transformation of data into embeddings stored in a vector database, facilitating the efficient retrieval of related data. This involves selecting high-quality, relevant data sources based on defined criteria such as accuracy, relevance, and accessibility. The prompting pipeline refines interactions with the LLM to ensure contextually relevant responses by integrating specific benchmark indicators with domain-specific context. Lastly, the evaluation pipeline validates the operationalized benchmarks against official data to assess the accuracy of the LLM in real-world applications. These components collectively ensure that the LLM can effectively process, analyze, and provide valuable operationalizations for GovTech benchmarks.

5. How does the accuracy of LLMs in operationalizing GovTech benchmarks compare to that of conventional methods?
LLMs can operationalize GovTech benchmarks with varying degrees of accuracy, depending on the model configuration. Manual evaluations reveal that some LLM configurations achieve up to 29% accuracy across the full benchmark and 48% for multiple-choice questions, surpassing the 37% accuracy expected from random guessing. However, when assessed using exact match and edit similarity metrics, these models often exhibit lower accuracies. This indicates that while LLMs can provide responses that are contextually relevant, they frequently fall short of perfectly matching the ground truth data. While the achieved accuracies are not particularly impressive, they are understandable in light of the complex and context-specific domain.

Main RQ: *How do LLMs operationalizing GovTech benchmarks mitigate inherent challenges of timeliness, complexity, and data aggregation, increasing their utility for policymakers?*

The integration of LLMs into the operationalization of GovTech benchmarks improves their utility for policymakers, by addressing the challenge of timeliness and creating opportunities to mitigate the challenges of complexity and data aggregation. Firstly, LLMs significantly speed up the data collection process, which traditionally spans several months, reducing it to minutes. This rapid processing capability ensures that benchmarks are updated quickly and can respond directly to evolving policy needs, thereby addressing the challenge of timeliness.

Moreover, although the model configurations used in this research have not fully addressed the issue of complexity, it was found that the primary reason for not including more complex questions is the time and resource constraints associated with operationalizing the benchmarks. With LLMs supporting fast and easy operationalization, this opens possibilities for developing more detailed GovTech benchmarks. These benchmarks could extend beyond easily measurable aspects such as websites and services to include more complex back-office processes and architectures. This expansion will improve the utility for policymakers, providing them with a more realistic and complete understanding of GovTech activities across different countries, rather than a superficial view.

The increased speed of operationalization with LLMs also potentially allows for quicker adaptation across different levels of data aggregation. Traditionally, the extensive time and resources needed have restricted the use of varied aggregation levels. With the ability to operationalize faster, LLMs can support more dynamic data aggregation strategies, facilitating a more tailored and nuanced analysis that can better meet the specific needs of different policymaking contexts.

Overall, the adoption of LLMs in operationalizing GovTech benchmarks represents a considerable improvement in their utility for policymakers. By enabling quicker updates, thereby enabling more complex analyses and allowing for flexible data aggregation, LLMs can transform the landscape of GovTech benchmarking. This improvement leads to more responsive and relevant government actions, ultimately better serving societal needs.

8 Limitations and Recommendations

This chapter explores the limitations faced during this study and aims to provide a clear understanding of the study's constraints and their impact on the results. Section 8.1 covers the artefact limitations, specifically the constraints related to using LLMs for GovTech benchmarks. Section 8.2 addresses broader research limitations, including methodological and contextual factors that affected the findings. Section 8.3 outlines the recommendations encompassing both future research directions and policy suggestions. Lastly, Section 8.4 outlines a preliminary framework for the development of an AI-Supported GovTech Index (AGTI).

8.1 Artefact Limitations

It is important to acknowledge the time and resource constraints that impacted this thesis project, as these limitations influenced the development and effectiveness of the artefact used. This section outlines four key areas where the artefact could be significantly improved to potentially improve both the accuracy and overall outcomes of the research.

Firstly, the base model employed in this research is a relatively modest 7B parameter model. There exists an evident scaling effect in language models: larger model/data sizes and more training compute typically lead to an improved model capacity (Hoffmann et al., 2022; Kaplan et al., 2020). This limitation in model size and capacity may have restricted the depth and breadth of analysis possible within this project.

Secondly, the fine-tuning process was conducted using LoRA, a Parameter-Efficient Fine-Tuning (PEFT) method, chosen for its lightweight nature and low resource cost. While PEFT fine-tuning can be effective, it is not the most sophisticated technique available. Literature indicates that most PEFT methods underperform compared to full fine-tuning in high-resource settings (G. Chen et al., 2022). More advanced methods, such as full model retraining or extensive hyperparameter optimization, could potentially yield a more accurate and stable model. The choice of fine-tuning technique is important for ensuring the model adapts well to the specific requirements and nuances of the GovTech context.

Additionally, there is a notable performance gap in non-English models, as highlighted by (Csaki et al., 2024). The Dutch model used in this project, due to this performance gap, may have contributed to suboptimal accuracy results. This disparity stresses the need for further advancements and improvements in models designed for languages other than English.

Lastly, the dataset compiled for the RAG was also limited to less than one GB. A more extensive and diverse dataset could enable the RAG to provide more accurate and relevant information, thereby improving the model's overall performance.

Considering these factors, there is considerable potential to improve the artefact's capabilities. With more resources and the opportunity to employ a more advanced model, the accuracy of the operationalization could be substantially improved.

8.2 Research Limitations

This section outlines the primary limitations encountered in this research, focusing on both methodological and contextual factors. Understanding these limitations is essential for contextualizing the findings and guiding future research in this domain. In the following, the limitations of Design Science

Research Methodology (8.2.1), Activity Theory (8.2.2), Accuracy Assessment (8.2.3), and Context and Scope (8.2.4) will be addressed.

8.2.1 Design Science Research Methodology

The use of the Design Science Research Methodology (DSRM) in this study presents several limitations. A key limitation becomes evident when comparing the original Design Science Research Framework by Hevner et al. (2004) with the methodology developed by Peffers et al. (2007). The original framework emphasizes the environment, including the roles, capabilities, and characteristics of people, as well as organizational strategies, structures, cultures, and processes. However, after adopting a problem-centered initiation approach, the methodology by Peffers et al. (2007) tends to overlook these environmental aspects. This can result in an overly optimistic focus on artefact development, neglecting the broader context in which the artefact will be implemented.

This oversight is particularly significant when using LLMs in the benchmarking process, where the environment is critical. As discussed in Section 1.4, the process involves multiple actors, each with their own strategies, which do not always align with the goal of perfecting a benchmark and the data. For instance, as highlighted in Section 6.3.1, countries may have perverse incentives to push cherry-picked data that boost scores for a better image, a factor not adequately addressed by DSRM. Another related question that arises is: for whom are we actually developing the artefact? Who will be the end-users of the LLM? The current DSRM steps does not adequately address these considerations.

Additionally, DSRM typically involves multiple iterative cycles of design, implementation, and evaluation. Due to time and resource constraints, this study was limited to a single iteration, which may have restricted the depth and soundness of the artefact's development and evaluation.

8.2.2 Activity Theory

In Chapter 4, Activity Theory (AT) is used to guide the integration of the artefact in the GovTech benchmarking ecosystem. It demonstrates the various roles LLMs can assume, initially positioning them as artefacts but also paving the way for investigating their potential as community enablers or autonomous agents. This approach not only tests the flexibility of AT in accommodating new technologies but also increases our understanding of if and how these technologies can be integrated into existing socio-technical systems with the use of AT.

However, this approach encountered significant limitations, particularly in its disregard for the broader social and political context. While AT is useful in analyzing actions and interactions within a local system, it does not account for the broader implications of these actions. Specifically, the activity mapping by Ojo et al. (2011) focuses on the outcome as the benchmark result, such as an “EGOV ranking or benchmarking report prepared by a government agency for a supervisory office”. However, in reality this is not the only outcome and boundary of the benchmarking activity. Benchmarking is embedded in a much broader ecosystem with political contexts. This critique is also found in literature, where AT is critiqued for its failure to adequately account for macro-social and political contexts, alongside its under-theorization of power and social structures (Martin & Peim, 2009). This oversight reduces the theory's potential for capturing the full complexity of GovTech benchmarking activities, consequently limiting its use in providing a structured approach for integrating the LLM within the benchmarking ecosystem.

A second limitation of AT is its traditional view of artefacts (Karanasios et al., 2021). AT typically focuses on tools like wrenches and hammers, which mediate activities by providing mechanical advantages and amplifying human intentions (Engeström, 1987). In this context, the artefact is the mediation mechanism or device through which the *subject* aims to achieve the *object*. However, digital tools such as LLMs represent a significant departure from these traditional artefacts. Unlike traditional tools that serve singular, well-defined purposes, LLMs are multifunctional. They can not only help the benchmarker to achieve the benchmarking purpose but also assist in defining an appropriate benchmarking purpose. Additionally, based on benchmarking purpose, LLMs can reflect and provide criticisms, influencing the benchmarker (subject) to change the benchmarking purpose (object). This hindered an easy integration by assigning LLMs the role of artefact and thus resulted into an exploration of the multiple integration points proposed in chapter 4, where LLMs function not only as artefacts but also as community enablers and even autonomous subjects. This necessitates a rethinking of AT to accommodate this extended potential of digital artefacts, recognizing their capacity to create new types of agency and significantly impact human behavior and broader social structures.

Another significant limitation of AT in this research is its lack of emphasis on ethical considerations. This issue is closely related to the previously mentioned limitation regarding the scope of benchmarking activities. The boundary of the benchmarking activity, as mapped by Ojo et al. (2011), is the presentation of the benchmark results. However, this approach neglects to address the potential impacts of these results and therefore ethical implications. As outlined in Section 1.4.3, benchmark results can significantly influence policy outcomes, which in turn have effects on society. Despite the importance of these effects, AT does not address whether the resulting impacts are fair or equitable. Put simply: the theory's framework lacks mechanisms to assess the outcome of the activity benchmarking.

8.2.3 Accuracy Assessment

Another methodological limitation of this research is the partial reliance on manual evaluation to assess the accuracy of LLMs in operationalizing GovTech benchmarks. Manual evaluation is inherently subjective and susceptible to biases, as evaluators' perspectives and interpretations can significantly influence the results. This subjectivity can compromise the reliability and validity of the accuracy assessments, necessitating careful consideration of these potential biases when interpreting the findings. Although automated accuracy assessment methods were also employed, as outlined in Section 2.4.2, the manual assessments may have inadvertently skewed the outcomes towards more favorable results. Consequently, this limitation could have affected the overall conclusions drawn from the research.

8.2.4 Context and Scope

The study's focus on the Dutch context presents another limitation. Benchmarking is often influenced by cultural and regional factors, and the results obtained in the Dutch context may not be fully applicable to other countries with different governmental structures and cultural norms. This limitation highlights the importance of considering cultural diversity in benchmarking studies to ensure broader applicability and relevance (Przeybilovicz et al., 2023).

Additionally, the research was limited to evaluating a single benchmark, the GTMI, which includes a significant number of multiple-choice questions. This specific focus may not fully capture the challenges associated with benchmarks that predominantly feature open-text responses. Benchmarks with different formats and question types could present unique challenges that were not addressed in this study. Furthermore, as noted by Felipe González-Zapata (Appendix C), a key difference between the GTMI and the DGI benchmark is that the DGI focuses on practices, whereas the GTMI focuses more on systems. These distinctions between benchmarks could potentially limit the generalizability of the findings to other benchmarking frameworks.

In conclusion, a considerable number of limitations could have impacted the findings of this study. Therefore, the next section outlines recommendations, both for future research directions and policy-oriented recommendations.

8.3 Recommendations

8.3.1 Future Research Recommendations

Firstly, future research should address the artefact limitations identified in this study to improve the robustness and applicability of findings within the GovTech benchmarking ecosystem. This includes several key aspects:

- **Employing Larger and More Advanced Language Models:** Future studies should explore the use of more powerful and larger language models to assess whether such advancements can significantly improve the accuracy of operationalizing GovTech benchmarks. This includes using state-of-the-art models that have demonstrated high performance in various domains.
- **Exploring Full Model Retraining and Hyperparameter Optimization:** Researchers should investigate the potential benefits of full model retraining and extensive hyperparameter optimization on domain-specific data. Moving beyond the use of LoRA, which is a parameter-efficient fine-tuning method, could result in more accurate and domain-adapted models, thus improving the utility of LLMs within the benchmarking process.
- **Systematic Collection and Incorporation of High-Quality Datasets:** There is a critical need for the systematic collection and integration of detailed, high-quality datasets, especially for languages other than English. These datasets are important for developing models that are robust and accurate across diverse linguistic and cultural contexts. Additionally, domain-specific datasets related to GovTech should be thoroughly explored and used, either for training the models directly or in relation with RAG techniques. Both approaches can significantly improve the detail and usefulness of the operationalization, ultimately improving the benchmarking process.

Additionally, regarding the research and methodological limitations, future studies should focus on validating or challenging the implications identified in this study regarding the use of LLMs for GovTech benchmarking. Key areas for future research include:

- **Expending the Geographical Scope:** Future research should broaden its geographical scope to include various cultural and regional contexts beyond the Netherlands. This expansion is important to increase the generalizability of the findings and understanding how different socio-political environments impact the GovTech benchmarking process.
- **Investigating Other GovTech Benchmarks:** Researchers should explore the use of LLM-based operationalization with benchmarks beyond the GTMI. Specifically, attention should be given to benchmarks that incorporate various formats, such as those with more open-text responses, and those that focus on practices like the DGI compared to those that focus on systems like the GTMI. This investigation will help identify unique challenges and opportunities associated with different types of benchmarks.
- **Conducting Ethical Analyses:** Future research must also undertake a thorough analysis of the ethical considerations involved, focusing on two key aspects:
  - **Current GovTech Benchmarking Process:** Researchers should analyze the existing benchmarking process to understand its societal implications, assessing the fairness, transparency, and accountability of current practices. Section 1.4.3 provides a starting point, but a much more detailed analysis is necessary to identify and address the full spectrum of ethical concerns. This includes evaluating how benchmarks impact various stakeholders and ensuring these impacts align with ethical standards and societal values.

  - **Integration of AI into Benchmarking:** Researchers should examine the ethical implications of incorporating AI technologies into the benchmarking process. This involves assessing how these technologies influence outcomes, ensuring results remain fair, unbiased, and transparent. Key questions include the potential biases introduced by AI, the impact on data privacy and security, and the effects on transparency and accountability. Section 6.3 can serve as a starting point, but again a much more extensive analysis is needed.
- **Developing a New Benchmarking Framework:** Future research should focus on developing a new benchmarking framework specifically tailored for AI and LLM applications. This could involve proposing new metrics, methodologies, and guidelines that address the challenges and opportunities posed by AI technologies. Section 8.4 provides an initial suggestion for this development of an AI-Supported GovTech Index (AGTI).

8.3.2 Policy Recommendations

Based on this research, several policy recommendations can be drafted for both national governments and benchmarking organizations. First Section 8.3.2.1 outlines the policy recommendations for the Dutch national government. Even though these recommendations are tailored for the Dutch context, other national governments can deploy similar measures. Then in Section 8.3.2.2 policy recommendations are presented for the World Bank.

8.3.2.1 Policy Recommendations for the Dutch Government

1. Proactive Government Role in the Integration of AI in Benchmarking

The Dutch Government should continue its proactive stance on AI, as demonstrated by its published government-wide vision on Generative AI (Ministry of the Interior and Kingdom Relations, 2024). This involves extending these principles to AI integration within benchmarking processes, whether conducted internally or through independent organizations. This proactive approach ensures that as benchmarking organizations increasingly turn to AI, government standards and oversight are already in place.

- **Strategic Vision and Clear Guidelines:** The Dutch Government should develop a clear policy document articulating their stance and strategy on AI usage in GovTech benchmarking. This document should build upon the government-wide vision on Generative AI and should address:
  - **Ethical Standards:** Governments need to conduct thorough ethics assessments of existing benchmarking processes and the proposed integration of AI. This includes evaluating how benchmarks impact various stakeholders and ensuring these impacts align with ethical standards and societal values. For the proposed integration of AI, frameworks like the Requirements of Trustworthy AI can be used (European Commission, 2020).
  - **Risks Management:** Identify the potential risks associated with AI in benchmarking and establish a framework for addressing these risks. Specify roles, such as a dedicated AI-officer within the Ministry of the Interior and Kingdom Relations, responsible for risk management.
  - **Value sensitive design:** Initiate a design process for AI-technologies to contribute to the benchmarking process, that incorporates ethical findings and stakeholder values, building on the ethical and risks assessment. Engage for instance with the Rathenau Institute to incorporate insights on ethical technology design.

2. Experimentation and Adaptation

Governments should not only prepare for but actively engage with AI technologies by experimenting with their use in benchmarking contexts. Learning from direct experience can help mitigate risks and maximize the benefits of AI integration.

- **Pilot Programs:** Launch pilot programs to explore the use of AI in different aspects of the benchmarking process. These programs should be designed to test the efficiency, accuracy, and impact of AI applications under controlled conditions. One such pilot could focus on developing a benchmark specifically tailored for AI and LLM applications, with an initial outline provided in Section 8.4. Collaboration with institutions like TNO or technical universities can facilitate these pilots and drive innovation.
- **Feedback Loops:** Establish mechanisms for ongoing feedback during pilot programs to refine AI applications according to real-world challenges and results.
- **Scalability Assessments:** Evaluate the scalability of successful pilot programs to larger operations, ensuring that the transition from small-scale tests to broader applications maintains integrity and effectiveness.

3. Community Building and Stakeholder Engagement

To effectively integrate AI into benchmarking processes, governments must invest in community building and engage with key stakeholders. This includes creating a platform or community for collaboration between government agencies, academia, AI companies, ethical experts, and civil society to ensure that all perspectives are considered in the development and implementation of AI benchmarking tools. Innovation labs such as Digicampus could facilitate this collaboration.

By adopting these policy recommendations, the Dutch government can more effectively oversee the integration of AI into GovTech benchmarking processes, ensuring that it happens ethical, transparent, and beneficially for all stakeholders.

8.3.2.2 Policy Recommendations for the World Bank

Based on this research, three specific recommendations are outlined for benchmarking organizations. While countries themselves can be benchmarking entities too, recommendations for countries were presented in the previous section. This section focuses on three targeted recommendations explicitly for the World Bank. Although tailored to the World Bank, these recommendations may also benefit other international benchmarking bodies like the OECD and the United Nation.

1. Start by Including AI-Supported Tools for Data Collection

Implementing LLMs can significantly speed up the data collection process, improving the timeliness of the GTMI and increasing its utility for policymakers. Starting with data collection is advisable because it poses lower risks compared to other stages of the benchmarking process, as the data validation phase remains unchanged. Additionally, this initial step allows organizations to gain experience with AI in a controlled setting, providing a foundation for further integration of AI into more complex tasks.

2. Gradually Integrate AI into Data Validation

Learn incrementally how to use LLMs and maintain high accuracy in data validation. Start with simple validation tasks and progressively assign more complex roles to AI tools as their reliability and effectiveness are proven. This process should be accompanied by robust oversight that ensures AI is used ethically and accurately in the data validation process. This includes setting up protocols for human oversight, continuous monitoring, and iterative improvement based on feedback and performance metrics.

3. Continuously Reassessment and Feedback Loops

As outlined in Section 1.4.3, benchmark results can significantly influence policy outcomes, which in turn have effects on society. Therefore, making sure the data quality is of the highest level, and a thorough validation process is in place, even with AI tools in place. Also create systems for continuous feedback from stakeholders, including the benchmarked countries, to refine and improve the benchmarking process. This involves incorporating stakeholder insights and addressing any ethical and social implications that arise. As demonstrated in Section 6.3, ensure that ethical and social considerations are integral to the benchmarking process. This includes transparency in AI use, addressing potential biases, and ensuring that the benchmarks do not inadvertently cause harm. Collaborate closely with benchmarked countries and other stakeholders to maintain a focus on these implications.

By following these steps, the World Bank can improve the timeliness, and subsequently start improving on limitations associated with the lack of complexity and inflexibility in data aggregation. This approach ensures that benchmarks remain useful tools for informing policy and driving positive societal outcomes.

8.4 Proposing the AI-Supported GovTech Index (AGTI)

This section outlines a preliminary design of the AI-Supported GovTech Index (AGTI), a benchmark specifically tailored to be operationalized by LLM, as recommended in earlier sections. The requirements are derived from the experiences and challenges encountered during the operationalization of the GTMI benchmark with LLMs.

1. **Clarity of Indicators:** Each indicator within the AGTI must be clear and understandable on its own, without necessitating cross-references to other indicators. During the operationalization of the GTMI benchmark, it was observed that the LLM struggled with contextual ambiguities when indicators were interdependent. To streamline the process and improve efficiency, indicators in the AGTI should be designed to stand alone, thereby simplifying their interpretation and application by LLMs.
2. **Diverse Data Sources for Evidence**: Evidence for each indicator should not be limited to a single source. This was the case with the GTMI framework and resulted in lower accuracies. Using RAG, the AGTI can dynamically integrate diverse and relevant data sources, improving indicator robustness and providing a richer, more nuanced analysis. This approach ensures the reliability and credibility of the benchmarking results by mitigating source-specific biases.
3. **Adaptable Question Templates**: The AGTI should feature dynamically adaptable question templates that can be easily customized according to the specific level of government, or the administrative and geographical scope being assessed. For instance, a question such as, “Has the government released any mobile app for the citizens’ access to public services?” (I-19.4 GTMI) could be adjusted to “Has [specified entity] released any mobile app for the citizens’ access to public services?” based on the entity being evaluated. This flexibility would facilitate the rapid reconfiguration of the benchmark across different governmental levels, from municipal to national, ensuring broad applicability and ease of use.

4. **Standardized Response Format:** To maintain consistency across different LLM operationalizations, for each indicator there should be a clear response format defined. This standardization ensures that outputs are uniform, regardless of the underlying model configuration or operational context.
5. **Openness and Transparency:** The design and operational methodologies of the AGTI must be fully open and transparent. Accessibility to the model's structure, the data it uses, and its operational protocols is essential for gaining trust and stimulating widespread adoption. Transparency not only increases the credibility of the AGTI but also enables continuous improvement and validation by other community actors.
6. **Integration of External Indices:** The AGTI should include a dedicated mechanism for incorporating external indices from other benchmarks, utilizing methods such as APIs or specialized scripts. The GTMI framework includes several indicators that reference external indices from other benchmarks, which often posed a challenge for the LLM due to difficulties in precisely referencing the specific indices required. To address this, the AGTI should either eliminate the need for external indices or implement a robust solution for their integration, ensuring accurate referencing.
7. **Standardized Interpretation of Indicators:** The AGTI should incorporate clear definitions or explanations for each indicator to prevent varying interpretations by different LLMs. As noted in Section 6.1.2, LLMs occasionally interpret indicators differently than intended, as shown by the truth data. To ensure consistent operationalization across various LLM configurations, each indicator's definition or explanation should be explicitly provided within the context when the LLM is tasked with operationalizing the indicator. This approach will improve the uniformity and accuracy of the responses generated by the LLMs.

These requirements for the AGTI are proposed as initial guidelines to aid in the development of a robust and effective benchmarking tool. Stakeholders and developers are encouraged to engage with these guidelines to further refine and expand the AGTI into a mature and widely usable benchmark that effectively integrates AI for improved benchmarking.

Literature

Adams, L., Busch, F., Han, T., Excoffier, J.-B., Ortala, M., Löser, A., Aerts, H. JWL., Kather, J. N., Truhn, D., & Bressem, K. (2024). *LongHealth: A Question Answering Benchmark with Long Clinical Documents*.

Andersen, K. N., Henriksen, H. Z., & Medaglia, R. (2012). Maturity models in the age of digital diversity: Beyond the Layne & Lee legacy. *Innovation and the Public Sector*, *19*, 205–220. https://doi.org/10.3233/978-1-61499-137-3-205

Bannister, F. (2007). The curse of the benchmark: an assessment of the validity and value of e-government comparisons. *International Review of Administrative Sciences*, *73*(2), 171–188. https://doi.org/10.1177/0020852307077959

Batlle-Montserrat, J., Blat, J., & Abadal, E. (2014). Benchmarking Municipal E-Government Services. *International Journal of Electronic Government Research*, *10*(4), 57–75. https://doi.org/10.4018/ijegr.2014100103

Baum, C., & Di Maio, A. (2000). *Gartner's Four Phases of E-Government Model. Gartner Group Report No. TU-12-6113.* .

Berntzen, L., & Olsen, M. G. (2009). Benchmarking e-Government - A Comparative Review of Three International Benchmarking Studies. *2009 Third International Conference on Digital Society*, 77–82. https://doi.org/10.1109/ICDS.2009.55

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

Brüggemeier, M., Dovifat, A., & Kubisch, D. (2005). Analyse von Innovationsprozessen im Kontext von E-Government. *Wirtschaftsinformatik*, *47*(5), 347–355. https://doi.org/10.1007/BF03251475

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, *279*, 103201. https://doi.org/10.1016/j.artint.2019.103201

Ceron, A., & Negri, F. (2016). The "Social Side" of Public Policy: Monitoring Online Public Opinion and Its Mobilization During the Policy Cycle. *Policy & Internet*, *8*(2), 131–147. https://doi.org/10.1002/poi3.117

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). *A Survey on Evaluation of Large Language Models*.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2310.14735

Chen, G., Liu, F., Meng, Z., & Liang, S. (2022). Revisiting Parameter-Efficient Tuning: Are We Really There Yet? *Arxiv*.

Chircu, A. M. (2008). E-government evaluation: towards a multidimensional framework. *Electronic Government, an International Journal*, *5*(4), 345. https://doi.org/10.1504/EG.2008.019521

Chroma. (n.d.). *Chroma Docs*. Retrieved May 11, 2024, from https://docs.trychroma.com/

Corea, F. (2019). *AI Knowledge Map: How to Classify AI Technologies* (pp. 25–29). https://doi.org/10.1007/978-3-030-04468-8_4

Csaki, Z., Li, B., Li, J., Xu, Q., Pawakapan, P., Zhang, L., Du, Y., Zhao, H., Hu, C., & Thakker, U. (2024). SambaLingo: Teaching Large Language Models New Languages. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2404.05829

de Goede, M., Enserink, B., Worm, I., & van der Hoek, J. P. (2016). Drivers for performance improvement originating from the Dutch drinking water benchmark. *Water Policy*, *18*(5), 1247–1266. https://doi.org/10.2166/wp.2016.125

de Juana-Espinosa, S., & Tarí, J. J. (2012). Benchmarking Local e-Government. In *Handbook of Research on E-Government in Emerging Economies* (pp. 624–640). IGI Global. https://doi.org/10.4018/978-1-4666-0324-0.ch032

Dener, C., Nii-Aponsah, H., Ghunney, L. E., & Johns, K. D. (2021). *GovTech Maturity Index*. The World Bank. https://doi.org/10.1596/978-1-4648-1765-6

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1810.04805

Engeström, Y. (1987). *Learning by expanding: an activity-theoretical approach to developmental research*. Orienta-Konsultit.

European Commission. (2020). *Requirements of Trustworthy AI*. https://doi.org/doi:10.2759/002360

European Commission. (2022). *Digital Economy and Society Index (DESI) 2022 Methodological Note*. https://ec.europa.eu/newsroom/dae/redirection/document/88557

Filer, T. (2019). *Thinking about GovTech A Brief Guide for Policymakers*. https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2020/12/Thinking_about_Govtech_Jan_2019_online.pdf

Giannakopoulos, D., & Manolitzas, P. (2009). Proposing a tool for the measurement of e-government. *Proceedings of the 3rd European Conference on Information Management and Evaluation, ECIME 2009*, 167–178.

Gunasekaran, A. (2005). Benchmarking in public sector organizations. *Benchmarking: An International Journal*, *12*(4). https://doi.org/10.1108/bij.2005.13112daa.001

Hasan, H., & Kazlauskas, A. (2014). Activity Theory: who is doing what, why and how. In *Being Practical with Theory: A Window into Business Research* (pp. 9–14). THEORI. https://ro.uow.edu.au/buspapers/403

Heidlund, M., & Sundberg, L. (2022). Evaluating e-Government: Themes, trends, and directions for future research. *First Monday*. https://doi.org/10.5210/fm.v27i12.12526

Heston, T., & Khun, C. (2023). Prompt Engineering in Medical Education. *International Medical Education*, *2*(3), 198–205. https://doi.org/10.3390/ime2030019

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., De Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Van Den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). Training Compute-Optimal Large Language Models. *ArXiv*.

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., & Azim, M. A. (2022). Transfer learning: a friendly introduction. *Journal of Big Data*, *9*(1), 102. https://doi.org/10.1186/s40537-022-00652-w

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*.

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*.

Hujran, O., Alarabiat, A., & AlSuwaidi, M. (2022). Analysing e-government maturity models. *Electronic Government*, *19*(1), 1–21. https://doi.org/10.1504/EG.2022.10040036

Jansen, A. (2005). Assessing E-government progress-why and what. *NOKOBIT*. http://www.afin.uio.no/om_enheten/folk/ansatte/jansen.html

Jansen, J., de Vries, S., & van Schaik, P. (2010). The Contextual Benchmark Method: Benchmarking e-Government services. *Government Information Quarterly*, *27*(3), 213–219. https://doi.org/10.1016/j.giq.2010.02.003

Janssen, D., Rotthier, S., & Snijkers, K. (2004). If you measure it they will score: An assessment of international eGovernment benchmarking. *Information Polity*, *9*(3,4), 121–130. https://doi.org/10.3233/IP-2004-0051

Janssen, M. (2010). Measuring and Benchmarking the Back-end of E-Government: A Participative Self-assessment Approach. In *Electronic Government: 9th International Conference, EGOV 2010, Lausanne, Switzerland, August 29 - September 2, 2010, Proceedings (Lecture Notes in Computer Science, 6228)* (pp. 156–167). Springer. https://doi.org/10.1007/978-3-642-14799-9_14

Jukić, T., Vintar, M., & Benčina, J. (2013). Ex-ante evaluation: Towards an assessment model of its impact on the success of e-government projects. *Information Polity*, *18*(4), 343–361. https://doi.org/10.3233/IP-130320

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2001.08361

Karanasios, S., Nardi, B., Spinuzzi, C., & Malaurent, J. (2021). Moving forward with activity theory in a digital world. *Mind, Culture, and Activity*, *28*(3), 234–253. https://doi.org/10.1080/10749039.2021.1914662

Kim, T. H., Im, K. H., & Park, S. C. (2005). Intelligent Measuring and Improving Model for Customer Satisfaction Level in e-Government. *Lecture Notes in Computer Science*, 38–48. https://doi.org/10.1007/11545156_4

Koh, C. E., & Prybutok, V. R. (2003). The three ring model and development of an instrument for measuring dimensions of e-government functions. *Journal of Computer Information Systems*, *43*(3), 34–39.

Kunstelj, M., & Vintar, M. (2005). Evaluating the progress of e-government development: A critical analysis. *Information Polity*, *9*(3,4), 131–148. https://doi.org/10.3233/IP-2004-0055

Kuziemski, Maciej., Mergel, I., Ulrich, P., & Martinez, A. (2022). *GovTech practices in the EU: a glimpse into the European GovTech ecosystem, its governance, and best practices.* (EUR 30985 EN). Publications Office of the European Union. https://doi.org/10.2760/74735

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). *Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models*.

Layne, K., & Lee, J. (2001). Developing fully functional E-government: A four stage model. *Government Information Quarterly*, *18*(2), 122–136. https://doi.org/10.1016/S0740-624X(01)00066-1

Lemke, F., Taveter, K., Erlenheim, R., Pappel, I., Draheim, D., & Janssen, M. (2020). *Stage Models for Moving from E-Government to Smart Government* (pp. 152–164). https://doi.org/10.1007/978-3-030-39296-3_12

Magd, H., & Curry, A. (2003). Benchmarking: achieving best value in public-sector organisations. *Benchmarking: An International Journal*, *10*(3), 261–286. https://doi.org/10.1108/14635770310477780

Martin, D., & Peim, N. (2009). Critical perspectives on activity theory. *Educational Review*, *61*(2), 131–138. https://doi.org/10.1080/00131910902844689

Ministry of the Interior and Kingdom Relations. (2024). *Government-wide vision on Generative AI of the Netherlands*. https://www.government.nl/documents/parliamentary-documents/2024/01/17/government-wide-vision-on-generative-ai-of-the-netherlands

Mishan, E. J., & Quah, E. (2020). *Cost-Benefit Analysis*. Routledge. https://doi.org/10.4324/9781351029780

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Muravu, N. (2023). Strategic Performance Measurement And Management: Theory And Practice Of Public Sector Benchmarking. *IOSR Journal of Business and Management* , *25*(12), 26–52. https://doi.org/http://dx.doi.org/10.9790/487X-2512092652

Nielsen, M. M. (2016). E-governance and stage models: analysis of identified models and selected Eurasian experiences in digitising citizen service delivery. *Electronic Government, an International Journal*, *12*(2), 107. https://doi.org/10.1504/EG.2016.076132

Oberkampf, W. L., & Trucano, T. G. (2008). Verification and validation benchmarks. *Nuclear Engineering and Design*, *238*(3), 716–743. https://doi.org/10.1016/j.nucengdes.2007.02.032

Ojo, A., Janowski, T., & Estevez, E. (2011). Building Theoretical Foundations for Electronic Governance Benchmarking. In *LNCS* (Vol. 6846). www.scopus.com

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*.

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Papaioannou, T., Rush, H., & Bessant, J. (2006). Benchmarking as a policy-making tool: from the private to the public sector. *Science and Public Policy*, *33*(2), 91–102. https://doi.org/10.3152/147154306781779091

Park, R. (2008). Measuring Factors that Influence the Success of E-Government Initiatives. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 218–218. https://doi.org/10.1109/HICSS.2008.244

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

Peters, R. M., Janssen, M., & van Engers, T. M. (2004). Measuring e-government impact. *Proceedings of the 6th International Conference on Electronic Commerce - ICEC '04*, 480. https://doi.org/10.1145/1052220.1052281

Przeybilovicz, E., Cunha, M. A., & Ribeiro, M. M. (2023). Decolonizing e-government benchmarking. *Proceedings of the 24th Annual International Conference on Digital Government Research*, 570–582. https://doi.org/10.1145/3598469.3598534

Public. (2021). *GovTech in The Netherlands: building a leading GovTech nation*. https://www.government.nl/documents/reports/2021/06/30/govtech-in-the-netherlands

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1910.10683

Reimers, N., & Gurevych, I. (2020). *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*.

Rijksoverheid. (n.d.). *Open data: VAC's*. Retrieved June 12, 2024, from https://www.rijksoverheid.nl/opendata/vac-s

Rodríguez-Bolívar, M. P. (2014). The Need for Analyzing e-Government Efficiency: An Introduction. In *Measuring E-government Efficiency: The Opinions of Public Administrators and Other Stakeholders* (Vol. 5, pp. 1–7). Springer. https://doi.org/10.1007/978-1-4614-9982-4_1

Rorissa, A., Demissie, D., & Pardo, T. (2011). Benchmarking e-Government: A comparison of frameworks for computing e-Government index and ranking. *Government Information Quarterly*, *28*(3), 354–362. https://doi.org/10.1016/j.giq.2010.09.006

Salem, F. (2007). Benchmarking the e-government bulldozer: beyond measuring the tread marks. *Measuring Business Excellence*, *11*(4), 9–22. https://doi.org/10.1108/13683040710837892

Scott, M., Delone, W., & Golden, W. (2016). Measuring eGovernment success: A public value approach. *European Journal of Information Systems*, *25*(3), 187–208. https://doi.org/10.1057/ejis.2015.11

Senyucel, Z. (2007). Assessing the impact of e-government on providers and users of the IS function. *Transforming Government: People, Process and Policy*, *1*(2), 131–144. https://doi.org/10.1108/17506160710751968

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2022). *Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering*.

Skargren, F. (2020). What is the point of benchmarking e-government? An integrative and critical literature review on the phenomenon of benchmarking e-government. *Information Polity*, *25*(1), 67–89. https://doi.org/10.3233/IP-190131

Snijkers, K., Rotthier, S., & Janssen, D. (2007). Critical review of e-government benchmarking studies. In *Developments in e-Government: A Critical Analysis* (Vol. 13). IOS Press.

Stapenhurst, T. (2009). *The Benchmarking Book: A How-to-Guide to Best Practice for Managers and Practitioners*.

SWIS. (n.d.). *Wat is Postbus42.nl?* Retrieved May 26, 2024, from https://www.postbus42.nl/wat-is-postbus-42

TNO. (2023, November). *Nederland start bouw GPT-NL als eigen AI-taalmodel*. https://www.tno.nl/nl/newsroom/2023/11/nederland-start-bouw-gpt-nl-eigen-ai/

Triantafillou, P. (2007). BENCHMARKING IN THE PUBLIC SECTOR: A CRITICAL CONCEPTUAL FRAMEWORK. *Public Administration*, *85*(3), 829–846. https://doi.org/10.1111/j.1467-9299.2007.00669.x

Ubaldi, B., & Okubo, T. (2020). *OECD Digital Government Index (DGI): Methodology and 2019 Results*. https://doi.org/https://doi.org/10.1787/b00142a4-en

United Nations. (2022). *E-Government Survey 2022*. https://publicadministration.un.org/en/

Van Huffelen, A. C. (2022, September 29). *Rijksbreed cloudbeleid 2022*.

Verdegem, P., & Hauttekeete, L. (2007). User centered E-government: Measuring user satisfaction of online public services. In P. Kommers (Ed.), *Proceedings IADIS International Conference e-Society, ES 2007 - Part of the IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2007* (pp. 63–71). IADIS Press.

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In J. vom Brocke, A. Hevner, & A. Maedche (Eds.), *Design Science Research. Cases* (pp. 1–13). Springer. https://doi.org/10.1007/978-3-030-46781-4_1

Waksberg-Guerrini, A., & Aibar, E. (2007). Towards a Network Government? A Critical Analysis of Current Assessment Methods for E-Government. In *Electronic Government* (pp. 330–341). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74444-3_28

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Yang, H., Liu, X.-Y., & Wang, C. D. (2023). *FinGPT: Open-Source Financial Large Language Models*.

Yoshida, M., & Thammetar, T. (2021). Education Between GovTech and Civic Tech. *International Journal of Emerging Technologies in Learning (IJET)*, *16*(04), 52. https://doi.org/10.3991/ijet.v16i04.18769

Yu, F., Quartey, L., & Schilder, F. (2023). Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks. *Findings of the Association for Computational Linguistics: ACL 2023*, 13582–13596. https://doi.org/10.18653/v1/2023.findings-acl.858

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

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). JEC-QA: A Legal-Domain Question Answering Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9701–9708. https://doi.org/10.1609/aaai.v34i05.6519

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Appendix A

Semi-structured Interview Guide

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.

Section 1: 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?

Section 2: Specific Challenges in Benchmarking

Timeliness Issues:

- How significant is the challenge of timeliness in the current GovTech / e-Government benchmarks? Can you provide an example?
- What impact does delayed benchmarking 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 overcomplicate the issues?

Aggregation of Data:

- What are the challenges with data aggregation in the current benchmarking 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?

Section 3: Role of AI in Addressing Benchmarking Challenges

Potential of AI Solutions:

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

Section 4: 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?

Section 5: Future of LLMs in GovTech / e-Government Benchmarking
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 benchmarkers / 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 benchmarking process?

Conclusion
Summarize 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.