Evaluating Government Website Readiness for Generative AI Search Engines: A Framework and Empirical Assessment

Hussein Ali Bahadi
Digital Government Authority, Saudi Arabia

***Abstract*—The rapid adoption of generative artificial intelligence (GenAI) search engines is changing how citizens obtain government information. Unlike conventional search engines, which return ranked links, GenAI systems synthesise answers directly from retrieved sources, positioning government websites as inputs to AI-mediated information ecosystems rather than as destinations. To our knowledge, no systematic framework currently exists for determining whether government websites are prepared for this form of access. This study develops and validates an AI Readiness Index for government websites through a three-phase sequential mixed-methods design. The framework is validated perceptually through a one-round Delphi-informed expert survey (n = 18) and a citizen survey (n = 85) analysed using partial least squares structural equation modelling. The framework is applied to government portals across 14 jurisdictions using a documented scoring protocol. The expert panel confirmed content validity for Data Foundation, Technical Optimisation, and Content Quality; Discoverability met the consensus threshold marginally, and Governance did not reach it (I-CVI = 0.67). The structural model identifies Technical Optimisation and Content Quality as significant predictors of perceived AI readiness, explaining 55 per cent of the variance, while Discoverability and Governance are non-significant. Portal assessment reveals wide disparities, with relative strength in data foundations and technical optimisation and consistent weakness in governance, ethics, and accountability. Because the index measures structural properties that condition AI retrieval rather than testing retrieval performance directly, its scores are proxy indicators of readiness rather than measures of realised AI output quality.**

Keywords*—Generative AI; government websites; AI readiness; E-government; retrieval-augmented generation; SmartPLS; digital transformation; evaluation framework

I. Introduction

Generative artificial intelligence (GenAI) search engines are displacing traditional web browsing, synthesizing answers from government websites without citizens ever visiting them. This shifts the role of a government portal from a curated destination for human readers to a critical, yet often invisible, data input for autonomous systems. In this emergent paradigm, the quality of a citizen's interaction with the state is no longer determined by a website's usability or visual design, but by its machine-readability and semantic coherence. Consequently, the prevailing frameworks for evaluating e-government - focused on human-centric metrics like navigation and accessibility - are rendered obsolete [1]. This study addresses the resulting lacuna by developing and validating a framework to assess whether government websites are structurally prepared to serve as reliable sources for GenAI-driven information retrieval.

The scale is not marginal: the United Kingdom's GOV.UK platform alone hosts over 700,000 pages [2], any of which may be retrieved, excerpted, and synthesised without human review. Yet the standards by which government websites are presently evaluated, such as usability, readability, accessibility, and visual design, all presuppose a human reader. This transformation introduces a new intermediary layer between citizens and government information [3]. Governments around the world are starting to appreciate the possibility and the difficulties of GenAI integration. Applications are being realized in jurisdictions. Recently, the Indiana Department of Transportation (INDOT) created an AI-based workflow that focuses on RAG in line with a 30-day executive order mandating government efficiency reporting. INDOT used Vertex AI Search with Gemini to create a searchable vector index of internal documents and generate draft reports with 98% fidelity across 9 divisions, estimated to save 360 hours of manual search time, and defined a template that 60 other state agencies have adopted [5]. And in the same vein, the UK Government has introduced GOV.UK Chat, a pilot service based on RAG to enable citizens to pose natural language questions through the GOV.UK estate. Initial testing with hundreds of users indicated high potential for enhancing access to government information and unveiled vital insights into the importance of automated quality assurance and how RAG surpasses fine-tuning based on frequent content updates [2].

At the strategic level, countries are formulating structures to evaluate and improve AI preparedness. To measure the institutional readiness of adopting AI technologies in government organizations, the Saudi Data and Artificial Intelligence Authority (SDAIA) has introduced a National AI Index in 2025, which is based on three pillars, seven dimensions, and 23 subcategories [6]. This project is indicative of an increasing awareness that systematic evaluation is a precondition to successful integration of AI. Although this has evolved, there is a gaping hole in the literature and practice. Current AI preparedness measures are done on two levels: national and organizational. On a national scale, national indices like the Government AI Readiness Index of Oxford Insights evaluate the general AI preparedness of a country in terms of policy, infrastructure, and human capital aspects [7]. On the organizational level, as recent studies suggest, dual-level models that integrate Technology-Organization-Environment (TOE)

theories with Unified Theory of Acceptance and Use of Technology (UTAUT) would help determine how the public sector is ready to adopt AI [8].

Nevertheless, the level of AI preparedness at a website level-how well a government's digital presence can be utilized by AI systems created by other parties-has not been studied. Whereas organizational readiness deals with whether an agency is capable of developing and implementing AI systems, website readiness deals with whether an agency is capable of effectively consuming AI systems created by commercial vendors. This difference is essential as citizens can more readily obtain government information via non-governmental AI interfaces that are out of governmental control. Recent research has started looking into related dimensions. Krauze-Maślankowska and Vrabie [9] studied 108 municipal websites in Poland and Romania, discovered the mention of GenAI, and found increasing awareness but at the initial stage of integration, mostly with visual content generation. Divides in e-government research have added a GenAI Gap dimension based on access, task uses, and competence, and have found evidence that GenAI may increase disparities in the short term by increasing a vulnerable group's exposure to higher levels of verification and ethics [10].

In information infrastructure research on AI, there is a focus on multi-actor information infrastructures with important variables in three categories: 1) information-sharing process, systems, and services, 2) information format, and 3) governance structures [11]. These variables should be taken into account early enough to reap the most and minimize the risk of failure. In the same way, the study of AI fusion with Open Government Data has determined eight perfect types of AI applications, such as AI as Portal Curator, Explorer, Linker, and Monitor [12], that offer conceptual bases for understanding how AI interacts with government information systems. However, these insights have not been combined in a systematic and empirically proven framework for assessing the preparedness of individual government websites for GenAI-mediated access. This gap leaves government agencies without guidance on how to prepare their online presence for the GenAI future and researchers without empirical baseline data against which to assess progress.

This study addresses this gap through the following research questions:

RQ1: What dimensions and metrics constitute a comprehensive framework for evaluating government website readiness for generative AI search engines?

RQ2: To what extent do experts agree on the relevance, clarity, and comprehensiveness of the proposed dimensions and metrics, as assessed through a one-round Delphi-informed survey?

RQ3: Which dimensions (Technical Optimisation, Content Quality, Discoverability, Governance) significantly influence citizens' perceived AI readiness of government websites, as tested through a PLS-SEM analysis of citizen survey responses?

RQ4: How do current government websites perform when evaluated against the framework using both an equal-weight index and a perception-weighted index that accounts for the empirical findings?

The remainder of this study is organised as follows. Section 2 reviews three literature streams, namely e-government evaluation, generative AI and information retrieval, and AI readiness assessment, and identifies the website-level gap this study addresses. Section 3 sets out the three-phase sequential mixed-methods design, covering framework development, perceptual validation through expert and citizen surveys, and the portal assessment protocol. Section 4 reports results across all three phases, including expert consensus outcomes, the structural model, and the cross-jurisdictional assessment under both equal and adjusted weighting. Section 5 discusses theoretical and practical implications. Section 6 concludes with limitations and directions for future research.

II. Literature Review

A. The Evolution of E-Government Evaluation

The evaluation of e-government over the last 20 years has changed considerably as a result of the maturity of technology as well as the academic knowledge of what successful digital government entails. The initial solutions revolved around the presence of websites and accessibility of online services. The four-stage model of e-government development proposed by Layne and Lee [13] - cataloguing, transaction, vertical integration, and horizontal integration - provided a framework that would help to determine the evolution of simple information provision to integrated service delivery. This four-phase methodology has impacted later assessment models by defining that e-government maturity is a multi-dimensional and progressive approach.

The model that has had a significant impact on e-government studies is the DeLone and McLean Information Systems Success Model [14]. This model recognizes six dimensions of success that are interrelated: system quality, information quality, service quality, user satisfaction, intention to use, and net benefits. What is strong about the model is that it acknowledges the fact that information systems success is multi-faceted and that these dimensions are not independent of each other but are dependent. This model was confirmed by Wang and Liao [15] in the e-government setting, where the authors discovered that system quality, information quality, and service quality play a significant role in generating user satisfaction and perceived net benefits.

Newer assessment frameworks have added new dimensions based on emerging knowledge of e-government success. Valdés et al. [16] proposed a multidimensional e-government maturity model incorporating technological, organizational, operational, and human-capability dimensions for evaluating public agencies. While this framework provides a comprehensive basis for assessing digital-government capability, it remains oriented toward human-operated service delivery rather than AI-mediated information consumption. Pather and Gomez [17] introduce the dimensions of functional suitability, reliability, and security to the ISO 25010 quality model used on e-government. This quality-driven method offers well-organized systems of assessing the quality of software products, and it can be directly applied to government sites as information systems. But these defined structures have one thing in common; they all assume that the final consumer of a government site is a human being. Usability, readability, and visual design metrics

presuppose human cognition and perception. Navigation structures presuppose human navigation. Patterns of human information processing are presupposed in content organization. This premise is growing less tenable with AI intermediaries being key entry points. The AI system, rather than the citizen, is the direct consumer of the content on the web when the citizen poses a question to ChatGPT regarding government services. The quality of information that the citizen gets depends on the accuracy of the interpretation, retrieval, and synthesis of that content by the AI. This alters the nature of evaluation requirements: now websites are required not just to be readable by humans, but by machines as well, and to be readable and synthesizable by machines.

B. Generative AI and Information Retrieval

Currently, genAI models are trained on large models of language (LLMs) that are trained on large text collections. These models have outstanding language understanding, language generation, and reasoning abilities. Nevertheless, they also have clear-cut constraints, such as the propensity towards hallucination- the production of plausible-sounding, but factually inaccurate information - and impairment of tasks that demand accurate, current, or specialized information [1]. They are particularly consequential in a governmental context, because the correctness of information is not only desirable, but a necessity to fulfill legal needs, to be eligible to obtain benefits, and to possess rights. Having an AI that hallucinates and believes that it requires a passport, or that it has to pay taxes, can actually be harmful. To alleviate these weaknesses, generative AI systems are increasingly adopting retrieval-augmented generation (RAG), a text generation model that combines information retrieval [4]. In RAG architecture, the user can pose a question, and the system will extract the documents or passages within a knowledge base - in this case, the indexed information on government websites. It then feeds these passages that have been retrieved into the LLM, which gives a coherent answer synthesizing the information. This approach will improve accuracy and reduce the amount of hallucinations because the responses are based on retrieved evidence.

RAG has successfully been implemented in government situations. The AI system of Indiana DOT consisted of Vertex AI Search, which generated a vector index of internal documents, and then gave retrieved snippets to Gemini to generate responses with 98 percent fidelity to original documents [5]. In the same way, GOV.UK Chat uses RAG on the 700,000+ pages of GOV.UK, developers observe that "and since GOV.UK content is constantly being updated, Retrieval-Augmented Generation (RAG) is more accurate and relevant in responses than fine-tuning" [2]. RAG architecture has far reaching implication on government website construction and testing. To be effective, RAG systems require the content of websites to meet a number of requirements. Content has to be readable by machines, and it should be indexed and available to crawlers. This requires technical infrastructure that facilitates automated access, including correctly set robots.txt, sitemaps, and not using formats (e.g., image-based PDFs) that are hard to extract text from (Van Donge et al., 2024). The content should be in such a way that retrieval algorithms can recognize the most relevant passages. This involves semantic markup, proper heading structures, and metadata that reflects content correctly [4]. The information must also be factual, as the LLM will merely combine retrieved information without any further verification. This puts the value of accurate content on a higher level than the human-facing scenario, where users may cross-reference a variety of sources. The contents must be recent, as there is no inherent priority of recent information in RAG systems in comparison to that provided by retrieval algorithms. Certain dates of publication and versions are crucial pointers.

C. AI Readiness: From National to Website Levels

The concept of AI readiness has been developed at the dissimilar levels of analysis and with diverse paradigms and measurement plans. On a macro level, Oxford Insights publishes an annual Government AI Readiness Index, which assesses 195 countries across 69 indicators grouped into fourteen dimensions and six pillars: government, technology sector, and data infrastructure [7]. The index takes into account a range of factors such as digital capacity, innovation capability, and availability of data, and provides comparative rankings, which inform policymaking and international benchmarking. Regarding the same, the National AI Index of the Saudi Data and Artificial Intelligence Authority assesses government organizations based on three pillars, seven dimensions, and 23 categories and reports the outcomes that indicate the level of maturity of AI adoption and inform capability improvement [6]. These national indices are similar in that they all revolve around systemic factors, which include policy frameworks, human capital, and infrastructure to enable AI development and deployment across government.

A recent study has employed factor and cluster analysis to assess the extent of AI readiness in the European Union member countries and has identified two focal dimensions: Digital Skills and Electronic Government Engagement, and Transparency and Electronic Government Service Availability [18]. It was observed that there were six unique profiles of preparedness of member states, with significant differences in both infrastructural and citizen-level preparedness, which signifies more general digital divides. A literature gap is a systematic emphasis on AI-readiness at a website level - how well individual government websites are prepared to be fed to third-party AI systems. Whereas organizational readiness contends with the issue of whether an agency can build AI, and infrastructure readiness with the information systems that underlie them, website readiness relates to the digital presence to the public that AI systems actually discover [19].

This is an important gap for a number of reasons. First, AI intermediaries are becoming more and more accessible to citizens who are already using AI to gain access to government information, irrespective of the official adoption of AI by the agencies. The readiness of websites directly affects the results for citizens, irrespective of the AI approach in its structure. Second, the characteristics of the site that are more likely to be involved in the intervention are not the organizational or national factors but the website features. Metadata, content structure, and governance processes can be enhanced using digital service teams, without requiring a policy change or infrastructure redesign. Third, web preparedness provides a similar unit of analysis, which can be empirically measured and compared to other similar units.

However, no existing framework integrates technical requirements of GenAI systems with e-government evaluation approaches to assess website-level readiness for AI-mediated access. Furthermore, no study has validated such a framework against stakeholder perceptions using advanced statistical methods. This gap is increasingly consequential as citizens adopt GenAI tools for government information seeking, regardless of whether government agencies have officially embraced AI technologies, as is evident from the UK public sector [20].

III. METHODOLOGY

This study used a sequential mixed-methods research design to create an operationalised, empirically validated AI Readiness Index of government websites. This design type is especially suitable in the emerging fields of research where the development of concepts should be followed by empirical validation. The study is carried out in three phases.

A. Phase 1: Framework Development

1) Theoretical foundations and literature integration: To create the AI Readiness Index, a structured, concept-centric literature review was conducted following the methodology of Webster and Watson [23]. This approach organises the review around concepts rather than around individual publications, and is appropriate where the objective is theory synthesis for construct development rather than exhaustive evidence aggregation.

The study framework incorporates multi-actor information infrastructure views [11], which focus on the significance of data formats, information-sharing mechanisms and governance organization. This allows the conceptualisation of AI readiness as a socio-technical entity instead of a technical capability. The framework also takes into consideration the latest international guidelines and policy standards to be relevant and aligned with the best practices in the world. According to the United Kingdom Government Digital Service [24], technical optimisation, data quality, organisational context and ethical compliance are identified as crucial pillars of AI-ready datasets. The Australian Government Technical Standard of AI [25] focuses on lifecycle governance, auditability and bias management. The Georgia Technology Authority has also published guidelines for the responsible use of AI in government, emphasising transparency and accountability [26]. Likewise, the Saudi Data and Artificial Intelligence Authority [6] suggests a multi-dimensional index of readiness, and it is important to note that systematic assessment is essential.

2) The AI readiness index: This study synthesizes the theoretical literature and new international frameworks to propose a five-dimensional AI Readiness Index for government websites.

Dimension 1: Data Foundation

This dimension deals with the quality, structure, and provenance of underlying data that drives the content of the websites.

DF1: Data Quality and Completeness – The extent to which data underlying website content is accurate, complete, and fit for purpose. This includes assessment of data gaps, inconsistencies, and errors that would propagate through AI systems.

DF2: Machine-Readable Formats – Content is available in formats that AI crawlers can easily parse, avoiding image-based PDFs and other non-machine-readable formats. This extends beyond basic HTML to structured data formats.

DF3: Data Provenance and Lineage – Clear documentation of data sources, collection methods, and transformation processes, enabling AI systems and users to assess reliability and potential biases.

DF4: Metadata Quality and Completeness – Comprehensive metadata that accurately describes content, including descriptive, structural, and administrative metadata that enables AI systems to understand content meaning and context.

Dimension 2: Technical Optimisation

This dimension addresses the technical infrastructure that enables AI systems to discover, access, and process website content effectively.

TO1: Structured Data Implementation – Use of schema.org or other structured data markup to provide explicit context and relationships, enhancing AI understanding of content meaning. This includes appropriate schema types for government services, organisations, events, and publications.

TO2: API Availability and Design – Programmatic access to data through well-documented APIs that enable direct machine-to-machine data exchange, following best practices for API design and documentation.

TO3: Crawlability and Indexability – Robots.txt and sitemap configurations that facilitate comprehensive AI crawling without blocking legitimate AI agents, including explicit allowances for AI crawlers where appropriate.

TO4: Performance and Reliability – Technical performance characteristics including page load speed, uptime, and response times that affect both human and AI access.

Dimension 3: Content Quality for AI Consumption

This dimension addresses the quality characteristics of website content that determine its utility for AI synthesis.

CQ1: Accuracy and Factual Correctness – Information is factually correct and verified. For AI-mediated access, accuracy is paramount as AI systems synthesise without independent verification, and errors can propagate with authoritative presentation.

CQ2: Currency and Timeliness – Content is regularly updated with clear publication and revision dates, enabling AI systems to prioritise current information.

CQ3: Completeness and Self-Containment – Information is comprehensive and self-contained, not requiring cross-referencing multiple pages to understand key concepts.

CQ4: Clarity and Unambiguity – Content uses plain language and avoids ambiguity that could lead to AI misinterpretation. This includes avoiding vague references, implicit assumptions, and culturally specific knowledge that AI

systems may lack.

Dimension 4: Discoverability and Context

This dimension addresses the extent to which content can be accurately retrieved and appropriately contextualised by AI systems.

DC1: Semantic Structure and Hierarchy – Use of heading tags (H1, H2, H3) and other structural elements that create logical content hierarchies, enabling AI to understand content organisation and relationships.

DC2: Contextual Independence – Individual content units are understandable when retrieved out of context, requiring self-contained explanations and minimal reliance on implicit knowledge from elsewhere on the site.

DC3: Internal Linking and Relationship Mapping – Clear relationships between related content through descriptive links that provide semantic information about the nature of relationships.

DC4: Metadata for Discovery – Descriptive metadata (titles, descriptions, keywords) that accurately represent page content and support accurate retrieval by AI systems.

Dimension 5: Governance, Ethics, and Accountability

This dimension addresses the organisational processes, ethical frameworks, and accountability mechanisms that support reliable AI interpretation and public trust.

GE1: Information Ownership and Accountability – Clear attribution of content to responsible agencies, teams, or individuals, enabling accountability and follow-up.

GE2: Quality Assurance Processes – Evidence of systematic review processes and quality controls, including documentation of reviews, approvals, and updates. This supports the Australian requirement for end-to-end auditability.

GE3: Transparency and Disclosure – Clear explanations of information sources, limitations, and any AI involvement in content creation or curation. This aligns with Georgia's requirement to label AI outputs and broader transparency principles.

GE4: Ethical Framework and Bias Management – Evidence of attention to ethical considerations including bias identification and mitigation, protection of privacy, and consideration of impacts on vulnerable groups.

GE5: Version Control and Audit Trail – Complete version history with change notes, enabling reconstruction of content evolution and supporting accountability.

All metrics are rated on a 3-0 scale that has specific behavioural anchors. The 0-3 scale offers just enough granularity to capture the meaningful difference in readiness and yet is reliable between evaluators. Per metric, the evaluators reviewed the page against the rubric anchors (Appendix A), giving a score according to the most appropriate description. Dimension scores were computed as the sum of metrics in each of the dimensions, and total readiness score as the sum of all metrics. Normalised percentages enabled comparison across websites and dimensions.

B. Phase 2: Index Validation Through Expert Panel and Causal Relationships

1) One-round Delphi-informed expert survey: To determine expert consensus on the significance, clarity, and completeness of the five dimensions and 20 measures of the AI Readiness Index, a single round (informed by the principles of the Delphi) was used. An international sample was purposively sampled to recruit 18 international experts who had a background in academia (n=8), government digital agencies (n=7), and international organisations (n=3) around the world. These were the inclusion criteria: 1) e-government, AI preparedness, or information systems research; 2) 3 years of professional experience in digital government or AI policy.

Experts received a briefing document summarising the theoretical foundations of the AI Readiness Index, the five dimensions, and the twenty metrics with their 0–3 behavioural anchors. For each metric, experts rated three aspects on 7-point Likert scales (1 = strongly disagree/unacceptable, 7 = strongly agree/excellent):

Relevance – How important is this metric for assessing government website AI readiness?

Clarity – Is the metric definition and behavioural anchor unambiguous and operationalisable?

Consensus was defined a priori as:

Median relevance ≥ 6.0 AND interquartile range (IQR) ≤ 1.0 on the 7-point scale.

Item-level Content Validity Index (I-CVI) ≥ 0.78 (proportion of experts rating relevance as 6 or 7).

A single structured round was used rather than the multiple iterative rounds of a classical Delphi. This design choice reflects the study's objective, which was content validation of a predefined framework rather than the generation of consensus from an open problem space. Descriptive statistics (median, IQR, mean, standard deviation) were calculated for each metric. Qualitative comments were analysed thematically to identify refinement suggestions.

Quantitative Perceptual Validation (PLS-SEM)

The expert survey provides content validity that the five dimensions are relevant, clear, and exhaustive to domain experts, so that the index measures what it aims to measure. Nonetheless, content validity by itself does not indicate the perception of the end users, who are citizens and who will eventually depend on AI-mediated access to government information, of the preparedness of these websites. The perceptual validation stage looks at how the site's objective characteristics and citizen perceptions are related. The following hypotheses were tested in the model.

H1: Technical Optimisation positively influences citizens' perceived AI readiness of government websites.

H2: Content Quality positively influences citizens' perceived AI readiness of government websites.

H3: Discoverability positively influences citizens' perceived

AI readiness of government websites.

H4: Governance positively influences citizens' perceived AI readiness of government websites.

H5: Digital Literacy moderates the relationship between Technical Optimisation and Perceived AI Readiness. Digital Literacy was incorporated not as a further determinant of perceived readiness but to test an additional mechanism that could plausibly matter: whether a citizen's own familiarity with AI-mediated retrieval shapes how strongly a website's technical quality translates into perceived readiness. Including this moderation allows the study to check whether the framework's perceptual relationships hold uniformly across users or depend on the respondent's level of digital sophistication, which is an important consideration for the generalisability of the findings.

2) Questionnaire design and measurement instrument: The survey instrument is developed using established guidelines for scale development [27], [28] (Table I). All constructs are modelled as reflective and measured using multiple items on a five-point Likert scale ranging from strongly disagree to strongly agree (Appendix B). To ensure that responses reflected perceptions of a concrete object rather than general attitudes, all respondents evaluated the same designated government website, which was provided to them before completing the survey. Four of the five framework dimensions were carried forward into the citizen perceptual model. Data Foundation was not, and the reason concerns observability rather than importance. This dimension addresses the quality, provenance, and machine-readability of the data underlying published content, properties that are largely invisible at the point of citizen interaction. A citizen can judge whether a page loaded reliably, whether its content appeared accurate and current, whether they could locate what they needed, and whether the responsible agency was identifiable; they are not positioned to assess dataset completeness, lineage documentation, or metadata schema conformance, and survey items asking them to do so would elicit inference rather than experience.

TABLE I. SOURCES FOR THE ESTABLISHMENT OF CONSTRUCTS

Construct	Items	Source
Technical Optimisation	3 items	Adapted from Wang & Liao [15] - validated in e-government context
Content Quality	4 items	Adapted from Wang & Liao [15] and DeLone & McLean [14]
Discoverability	3 items	Adapted from Venkatesh et al. [21], UTAUT (effort expectancy items)
Governance	3 items	Developed for this study based on Van Donge et al. [11] and international frameworks (UK, Australian, Saudi guidelines)
Digital Literacy	3 items	Adapted from Parasuraman & Colby [29] TRI 2.0
Perceived AI Readiness	3 items	Developed for this study based on constructs from Kelly et al. [22], a systematic review of AI acceptance factors

3) Sampling and data collection: The target population includes citizens who have accessed government websites in the last six months. A non-probability snowball purposive sampling method is used. The data collection is carried out in six weeks. Screening questions established the eligibility of respondents. Eighty-five valid responses were retained. The adequacy of this sample was assessed using the inverse-square-root method [31], which bases the minimum sample size on the smallest path coefficient the study aims to detect. At the conventional significance level (0.05) and statistical power (0.80), detecting the weakest hypothesised meaningful effect, corresponding to the smallest significant path in the model ($\beta = 0.437$), requires a minimum of approximately 33 observations, and detecting a path of $\beta = 0.30$ requires approximately 69. The retained sample of 85 therefore provides adequate power for the effect sizes hypothesised as substantively meaningful. It should be noted that the sample is less well powered to detect small effects ($\beta \approx 0.20$ would require approximately 155 observations); the non-significant paths reported below are interpreted with this in mind.

4) Data analysis using SmartPLS: SmartPLS 4.0 was used to conduct data analysis in accordance with the best practices of PLS-SEM [30]. The analysis was done in two phases. The measurement model was initially assessed. Outer loadings were used to evaluate indicator reliability, and a value greater than 0.70 is said to be acceptable. Cronbach's alpha and composite reliability were the two measures used to evaluate internal consistency. Average Variance Extracted (AVE) was considered to assess convergent validity, and a threshold of 0.50 was used. The Fornell-Larcker criterion and the HTMT ratio below 0.85 were used to measure discriminant validity. Second, the structural model was tested using bootstrapping of 5,000 resamples. Path coefficients and their importance were discussed. R^2 was used to determine model explanatory power and Q^2 to determine predictive relevance. The contribution of each construct is calculated by finding its effect size (f^2). Moderation analysis was conducted using the product indicator approach to examine the moderating effect of Digital Literacy.

C. Phase 3: Empirical Assessment of Government Websites

1) Sample selection: The framework was applied to a purposive sample of government websites from each country and policy artefacts from multiple jurisdictions, including Canada, the European Union, the United States, Singapore, New Zealand, Australia, India, Pakistan, Saudi Arabia, and the United Kingdom. The sample includes both operational portals and policy documents to capture variation in implementation and strategic guidance.

2) Evaluation procedure: Each website was evaluated using publicly available information, including website structure, metadata, API documentation, licensing policies, and governance documentation. Evaluators systematically assessed each metric using predefined criteria (as per dimensional criteria described above) and documented supporting evidence. A structured scoring protocol was used to ensure consistency. Where evidence was incomplete, conservative scoring was applied. Fractional scores (e.g., 2.5) were permitted where evidence suggested. Such fractional scoring was used very

conservatively and only in exceptional cases where qualitative assessment indicated that a website substantially met the higher criterion but retained minor limitations. Scores are aggregated to produce dimension-level and overall readiness measures, enabling cross-case comparison.

To support reproducibility, the assessment protocol is specified as follows. The unit of analysis was the national government web presence, operationalised as the primary central-government portal together with a consistent set of high-traffic service and information pages linked from it; where a single national portal did not exist, the principal equivalent gateway was used. Each metric was scored against publicly available evidence only, drawn from four source types: the rendered pages themselves, the underlying page source and structured-data markup, machine-access artefacts (robots.txt, XML sitemaps, and any published APIs and their documentation), and published governance material (data licences, editorial and quality-assurance policies, and accountability statements). Each metric carries a four-point behavioural rubric (0-3) with explicit anchors for each level; the full rubric is provided in the supplementary material (Appendix). Each of the five dimensions comprises several sub-metrics in the rubric (for example, Data Foundation covers data quality, machine-readable formats, provenance, and metadata); the dimension score reported in Table X is the rounded aggregate of its constituent sub-metric ratings, and the accompanying descriptor summarises the most salient factor supporting that score. Where evidence was absent or incomplete, the lower score was assigned, ensuring that unverifiable capability was never credited. Because portal content changes over time, all assessments were conducted within a single defined evaluation window, and the evidence basis was captured at the time of scoring so that scores can be re-verified against archived material. The assessment was performed by a single evaluator, which introduces a risk of subjective bias. To mitigate this, scoring followed the fixed behavioural rubric described above, applied conservatively and with a documented evidence basis for each metric.

The weighting scheme was also subjected to a sensitivity analysis to assess how far the results depend on the specific weights chosen. Because any weighting scheme embeds a judgment about relative importance, the index was tested for sensitivity to that judgment. Four schemes were compared: equal weighting across all five dimensions; the pragmatic scheme adopted here, assigning full weight to Data Foundation, Technical Optimisation and Content Quality and half weight to Discoverability and Governance; a scheme with weights proportional to the expert panel's I-CVI values; and a path-informed scheme downweighting the two dimensions that did not significantly predict perceived readiness in the structural model.

IV. Results

A. Expert Validation Results of the One-Round Delphi-informed Expert Survey

A total of eighteen experts rated the five dimensions of the AI Readiness Index on a seven-point scale for relevance, clarity, and comprehensiveness. Technical Optimisation achieved the strongest consensus, with a median of 7.0, an interquartile range of 0.0, 100 percent agreement (18 out of 18 experts rating 6 or 7), an item-level content validity index (I-CVI) of 1.00, and a mean clarity score of 6.9. Data Foundation also reached consensus, with a median of 7.0, an interquartile range of 0.5, 88.9 percent agreement (16 out of 18), an I-CVI of 0.89, and a mean clarity score of 6.4. Content Quality similarly achieved consensus, with a median of 7.0, an interquartile range of 1.0, 83.3 percent agreement (15 out of 18), an I-CVI of 0.83, and a mean clarity score of 6.6. Discoverability and Context met the consensus criteria but only borderline, with a median of 6.0, an interquartile range of 1.0, 77.8 percent agreement (14 out of 18), an I-CVI of 0.78 (exactly at the acceptable threshold), and a mean clarity score of 6.1. Governance, Ethics and Accountability did not reach consensus, with a median of 6.0, an interquartile range of 1.5, 66.7 percent agreement (12 out of 18), an I-CVI of 0.67 (below the 0.78 threshold), and a mean clarity score of 5.7. These results indicate strong expert support for data, technical, and content dimensions, borderline support for discoverability, and a lack of consensus on governance. Governance, Ethics and Accountability did not meet the a priori consensus threshold (I-CVI = 0.67, below the 0.78 criterion), yet the dimension was retained in the framework. This decision was deliberate and rests on three considerations. First, the dimension is strongly grounded in the reviewed literature and in current government AI policy instruments, several of which make accountability, transparency, and auditability explicit requirements; excluding it would leave the framework silent on properties that policy already treats as essential. Second, the sub-threshold result concerns expert agreement on citizen-facing relevance rather than the dimension's intrinsic importance, and the divergence itself is an informative finding: governance is genuinely harder to assess from public evidence than the more technical dimensions. However, the weighting scheme in the results accordingly assigns Governance a reduced rather than a full weight. Table II reports the expert consensus outcomes.

TABLE II. Expert Consensus on Five Dimensions (N=18)

Dimension	Median	IQR	% rating 6 or 7 (Agreement)	I-CVI	Clarity (mean)	Consensus reached?
Data Foundation	7.0	0.5	88.9%	0.89	6.4	Yes
Technical Optimisation	7.0	0.0	100%	1.00	6.9	Yes
Content Quality	7.0	1.0	83.3%	0.83	6.6	Yes
Discoverability & Context	6.0	1.0	77.8%	0.78	6.1	Yes (borderline)
Governance, Ethics & Accountability	6.0	1.5	66.7%	0.67	5.7	No

B. Data Analysis Using SmartPLS

The collected survey data was analysed using Partial Least Squares Structural Equation Modelling (PLS-SEM) with SmartPLS 4.0. PLS-SEM was selected due to its suitability for exploratory and prediction-oriented research, particularly in studies involving complex models, multiple constructs, and relatively non-normal data distributions [30]. Additionally, the method is appropriate for models that include moderation effects and aim to maximise explained variance in the dependent construct. The data analysis followed a two-stage approach comprising measurement model assessment and structural model evaluation, in accordance with established guidelines.

1) Measurement model assessment: The measurement model was evaluated to assess the reliability and validity of the constructs. As all constructs were modelled as reflective, the following criteria were applied:

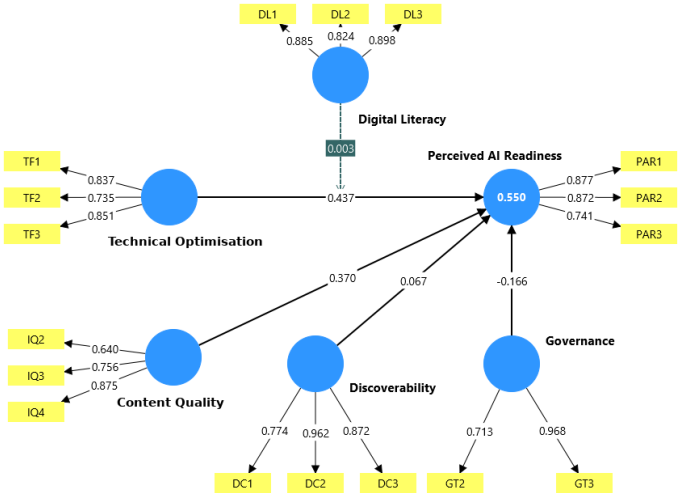

Fig. 1. Measurement model assessment.

Indicator reliability was assessed by examining outer loadings of each item on its respective construct. Loadings of 0.70 or higher were considered acceptable, indicating that the indicator explains at least 50% of the variance in the construct. Items with loadings between 0.60 and 0.70 were retained where theoretically justified. Two items were removed during measurement-model purification: IQ1 (Content Quality) and GT1 (Governance), whose outer loadings fell below the 0.60 threshold. Their removal improved the convergent validity and reliability of their respective constructs without altering the conceptual content of either, and both removals were made before structural-model estimation. The resulting outer loadings are shown in Table III.

Internal consistency was evaluated using both Cronbach's alpha and composite reliability (CR). Values above 0.70 were considered acceptable, indicating satisfactory reliability of the constructs. Convergent validity was assessed using Average Variance Extracted (AVE). An AVE value greater than 0.50 indicates that the construct explains more than half of the variance in its indicators, confirming adequate convergent validity. These reliability and validity statistics are reported in Table IV.

TABLE III. OUTER LOADINGS

Construct	Outer loadings		
DC1 <- Discoverability	0.774	IQ3 <- Content Quality	0.756
DC2 <- Discoverability	0.962	IQ4 <- Content Quality	0.875
DC3 <- Discoverability	0.872	PAR1 <- Perceived AI Readiness	0.877
DL1 <- Digital Literacy	0.885	PAR2 <- Perceived AI Readiness	0.872
DL2 <- Digital Literacy	0.824	PAR3 <- Perceived AI Readiness	0.741
DL3 <- Digital Literacy	0.898	TF1 <- Technical Optimisation	0.837
GT2 <- Governance	0.713	TF2 <- Technical Optimisation	0.735
GT3 <- Governance	0.968	TF3 <- Technical Optimisation	0.851
IQ2 <- Content Quality	0.640	IQ3 <- Content Quality	0.756

TABLE IV. INTERNAL CONSISTENCY RELIABILITY AND CONVERGENT VALIDITY

	Cron bach's alpha	Composite reliability (rho_a)	Composite reliability (rho_c)	Average variance extracted (AVE)
Digital Literacy	0.844	0.890	0.903	0.756
Discoverability	0.873	0.813	0.905	0.762
Governance	0.877	0.912	0.835	0.722
Content Quality	0.719	0.729	0.804	0.582
Perceived AI Readiness	0.777	0.801	0.871	0.693
Technical Optimisation	0.747	0.788	0.850	0.655

Discriminant validity, reported in Table V, was assessed using the Heterotrait–Monotrait (HTMT) ratio, where values below 0.85 indicate satisfactory discriminant validity.

TABLE V. DISCRIMINANT VALIDITY VALUES ASSESSED USING THE HETEROTRAIT-MONOTRAIT (HTMT) RATIO

	DL	DC	GT	CQ	PAR	TO	DL x TO
DL							
DC	0.124						
GT	0.341	0.189					
CQ	0.190	0.168	0.200				
PAR	0.386	0.132	0.206	0.598			
TO	0.077	0.107	0.329	0.258	0.644		
DL x TO	0.088	0.233	0.108	0.153	0.144	0.077	

Table VI reports the variance inflation factor (VIF) values for all predictor constructs, with all values well below the conservative threshold of 5, indicating that multicollinearity is not a concern in this model.

TABLE VI. COLLINEARITY AMONG PREDICTOR CONSTRUCTS WAS ASSESSED USING THE VARIANCE INFLATION FACTOR (VIF). VIF VALUES BELOW 5 INDICATED THAT MULTICOLLINEARITY WAS NOT A CONCERN

Construct	VIF		
DC1	2.298	IQ3	1.513
DC2	2.595	IQ4	1.462
DC3	2.212	PAR1	1.851
DL1	1.925	PAR2	1.913
DL2	2.035	PAR3	1.381
DL3	2.102	TF1	1.471
GT2	1.356	TF2	1.513
GT3	1.356	TF3	1.489
IQ2	1.130	Digital Literacy x Technical Optimisation	1.000

2) Structural model assessment: The structural model was evaluated using bootstrapping with 5,000 resamples to assess the significance of path coefficients. The full structural results are reported in Table VII and illustrated in Fig. 2. The results indicate that Technical Optimisation has the strongest positive effect on Perceived AI Readiness ($\beta = 0.437$, $t = 4.108$, $p < 0.001$). This suggests that the technical infrastructure and system capabilities of government websites are the most critical determinants of their readiness for generative AI. Content Quality also demonstrates a significant positive relationship with Perceived AI Readiness ($\beta = 0.370$, $t = 2.886$, $p = 0.004$). This finding highlights the importance of accurate, complete, and well-structured content in enabling effective AI-based information retrieval. Moreover, Discoverability does not exhibit a significant relationship with Perceived AI Readiness ($\beta = 0.067$, $t = 0.487$, $p = 0.626$). Although discoverability is theoretically important, its effect is not statistically supported in this model. Similarly, Governance shows a negative but non-significant relationship ($\beta = -0.166$, $t = 1.422$, $p = 0.155$). This suggests that governance mechanisms alone may not directly influence perceived AI readiness unless supported by strong technical and informational factors. The moderating effect of Digital Literacy on the relationship between Technical Optimisation and Perceived AI Readiness was also examined. Unexpectedly, the interaction term was found to be non-significant ($\beta = 0.003$, $t = 0.026$, $p = 0.980$), indicating that Digital Literacy does not strengthen or weaken the effect of Technical Optimisation. Although the model also returns a direct Digital Literacy path, this was treated as a control rather than a hypothesised effect; the study's interest in Digital Literacy is confined to its moderating role. Digital Literacy was included in the model specifically to examine whether respondents' own familiarity with AI-mediated retrieval conditions the way website attributes translate into perceived readiness, on the reasoning that more digitally literate citizens might weigh technical quality differently from less experienced users. The non-significant interaction indicates that this conditioning effect was not observed: the influence of Technical Optimisation on perceived readiness did not vary with respondents' digital literacy.

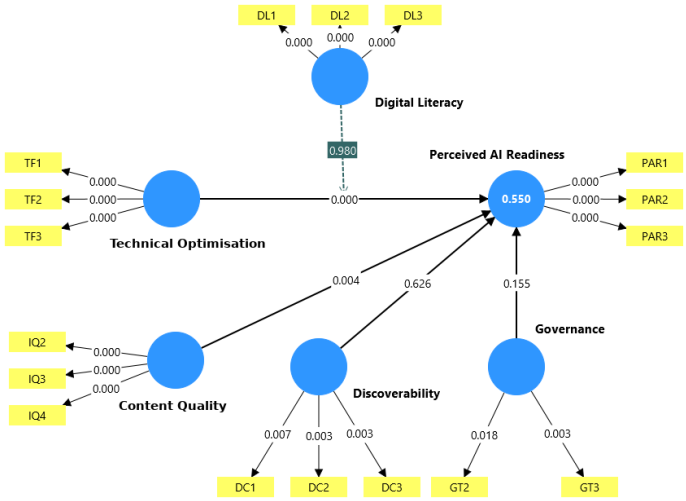

Fig. 2. Structural model.

TABLE VII. STRUCTURAL MODEL ASSESSMENT

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values
DL -> PAR	0.357	0.338	0.091	3.942	0.000
DL x TO -> PAR	0.003	-0.024	0.102	0.026	0.980
DC -> PAR	0.067	0.026	0.137	0.487	0.626
GO -> PAR	-0.166	-0.141	0.117	1.422	0.155
CQ -> PAR	0.370	0.360	0.128	2.886	0.004
TO -> PAR	0.437	0.443	0.106	4.108	0.000

C. Model Explanatory Power

The explanatory power of the model (Fig. 1) was assessed using the coefficient of determination (R^2). The results show that the model explains 55.0% of the variance in Perceived AI Readiness ($R^2 = 0.550$), with an adjusted R^2 value of 0.494. This indicates a moderate level of explanatory power, suggesting that the included constructs collectively provide a meaningful explanation of AI readiness perceptions. Table VIII presents the coefficient of determination.

TABLE VIII. COEFFICIENT OF DETERMINATION (R^2)

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values
PAR	0.550	0.616	0.084	6.563	0.000

D. Predictive Relevance

Predictive relevance, reported in Table IX, was evaluated using the Stone–Geisser Q^2 values obtained through the blindfolding procedure. All Q^2 values were positive (ranging from 0.149 to 0.338), indicating that the model has sufficient predictive relevance. Further assessment using Q^2predict compared the PLS-SEM model against the linear model (LM) benchmark at the indicator level. The PLS-SEM model produced lower RMSE and MAE values for PAR1 and PAR3, while for PAR2 the two approaches performed identically

(RMSE = 0.523 for both). Following Shmueli et al., this pattern, in which the majority but not all indicators favour the PLS-SEM model, indicates medium rather than strong predictive power. Because Q²predict was computed at the indicator level, construct-level predictive relevance is not reported here.

TABLE IX. PREDICTIVE RELEVANCE (Q²PREDICT) AT THE INDICATOR LEVEL

	Q²predict	PLS-SEM_RMSE	PLS-SEM_MAE	LM_RMSE	LM_MAE	IA_RMSE	IA_MAE
PAR1	0.338	0.533	0.424	0.562	0.456	0.656	0.539
PAR2	0.262	0.523	0.415	0.523	0.415	0.609	0.556
PAR3	0.149	0.586	0.443	0.686	0.499	0.636	0.563

The analysis confirms that Technical Optimisation and Content Quality are the primary drivers of perceived AI readiness. The model demonstrates moderate explanatory power and strong predictive capability, supporting the robustness of the proposed framework.

E. Weighting Adjustments for Index

Based on the combined findings of the expert consensus survey and the citizen SmartPLS perceptual validation, a pragmatic weighting scheme was adopted for the empirical website assessment. The expert survey confirmed that all five dimensions are relevant for assessing AI readiness, with Technical Optimisation, Data Foundation, and Content Quality receiving the strongest consensus (I-CVI ≥ 0.83). The SmartPLS analysis further showed that Governance and Discoverability did not significantly predict perceived AI readiness from a citizen perspective (H3 and H4 not supported, with $\beta = 0.067$ and $\beta = -0.166$ respectively, both $p > 0.05$). While these dimensions remain theoretically important, their direct influence on user perceptions is limited. Therefore, the index was modified to assign full weight (1.0) to Data Foundation, Technical Optimisation, and Content Quality, while Governance and Discoverability each received half weight (0.5). This adjustment preserves the comprehensiveness of the framework while reflecting the empirical finding that citizens and experts prioritise technical and content dimensions. The maximum possible weighted score becomes 12 (sum of 3+3+3+1.5+1.5), and each website's weighted score was calculated accordingly and normalised to a percentage.

F. Empirical Assessment of Government Websites

Fig. 3 presents the full matrix of dimension scores across all fourteen portals, and Fig. 4 summarises the cross-portal mean for each dimension. Data Foundation is uniformly the strongest dimension (mean 2.86 of 3.0), while Discoverability and Content Quality are the weakest, and the developing-economy portals show their steepest declines on these two dimensions.

Most of the evaluated sites perform reasonably well on data foundation. The United Kingdom, Canada, United States, Japan, Singapore, India, Pakistan, Kenya, New Zealand, Saudi Arabia, and Australia all provide datasets or policy guidance that emphasises machine-readable formats and high-quality data (Table X). For example, Canada's guidance explicitly recommends tagging content with machine-readable JSON-LD, and Japan's Digital Agency stresses that public data should be accessible under the Basic Act. These commitments reflect a growing recognition that data foundations underpin AI-driven public services. However, South Africa, Brazil, and Uruguay exhibit weaker foundations. Their portals often provide data only in PDF or HTML form, with sometimes restricted mention of machine-readable formats. Brazil's National Data Catalogue guide does call for open, structured data, but implementation across the broader domain appears inconsistent. AI systems depend on high-quality, well-structured data; without consistent machine-readable formats, generative AI will struggle to retrieve accurate information from government portals. Countries with stronger data foundations are better positioned to integrate AI into public services.

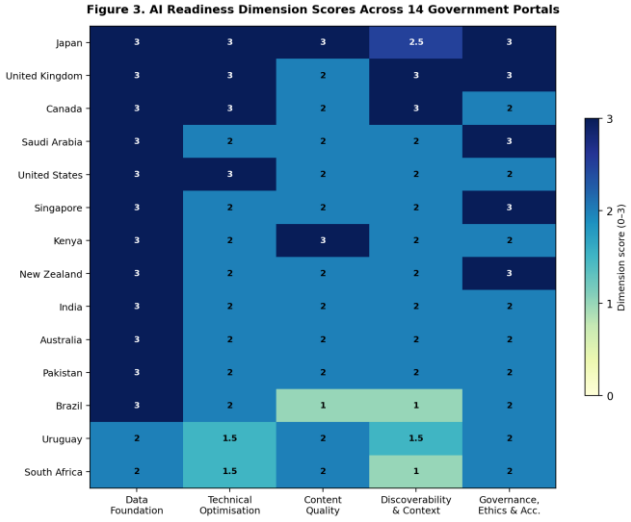

Fig. 3. AI readiness dimension scores across the fourteen assessed government portals.

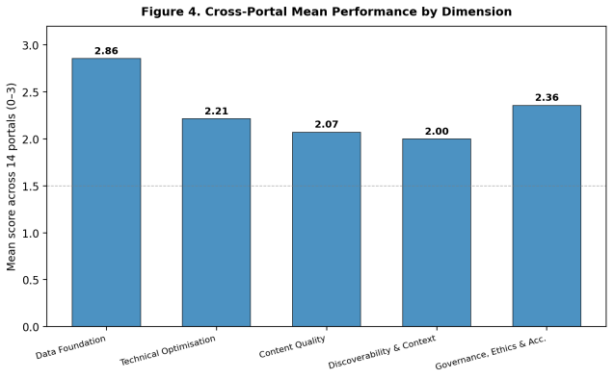

Fig. 4. Cross-portal mean performance by dimension.

High-performing sites not only provide data but also optimise for machine consumption. GOV.UK implements multiple schema.org templates via a shared metadata component, making its pages highly discoverable. Canada's guidance details how to select appropriate schemas and validate JSON-LD. The US federal government uses DCAT-US to standardise metadata for datasets and APIs, and Digital.gov highlights the use of schema.org for Special Announcements and FAQ pages during the pandemic. Japan's portal offers advanced search, visualisations and a communication function,

reflecting strong technical optimisation. In contrast, Singapore, India, Pakistan, Kenya, New Zealand, and Australia demonstrate moderate optimisation: they provide APIs and some metadata guidance but lack comprehensive schema implementation or uniform practices. For example, Singapore's licence permits widespread data reuse but doesn't prescribe technical standards, and Pakistan's playbook advocates APIs but notes that they're not substitutes for bulk data. The lowest scores occur in South Africa, Brazil, and Uruguay, where there is little evidence of structured metadata or standardised APIs. These portals appear to prioritise content publication over machine accessibility. Without structured metadata and standardised APIs, AI models must rely on scraping unstructured HTML, increasing the risk of hallucination and misinterpretation. Countries with strong technical optimisation are better prepared for future AI-driven service delivery.

The quality of content determines the accuracy of information, the up-to-date information, and the contextual independence of the information. High performers (e.g., Japan and the UK) will always keep information updated, and the pages are always meaningful when taken out of context. Canada and the US also have fairly high standards but depend on the individual agencies to be accurate, and hence the variability in departments. Middle tier nations (Singapore, India, Pakistan, Kenya, New Zealand, Australia, and Saudi Arabia) release valuable datasets but occasionally contain out-of-date data or make use of documents that are hard for AI to read (e.g., scanned PDFs). Saudi Arabia's my.gov.sa presents bilingual content (Arabic and English) that is regularly updated and organised by life events, yet content quality varies across 289 contributing agencies. Brazil and Uruguay have problems with context and outdated content.

Portals with high performance invest in search and navigation. Breadcrumb and FAQ templates in GOV.UK help users and crawlers navigate the site, whereas the portal provides cross-sectional search and visualisation in Japan. Canada promotes search engine optimisation using structured data. Interactive charts, maps, or catalogues are available in Kenya, Singapore, New Zealand, Australia, and Saudi Arabia, although navigation remains unreliable across some agencies in the latter. Saudi Arabia's my.gov.sa employs life-event-based navigation and a dedicated FAQ section to aid discovery, but cross-agency metadata consistency varies. Low scores in South Africa, Brazil, Uruguay, and, to a lesser degree, India and Pakistan are a manifestation of portals with some limitations in search functionality, metadata, and not well-developed contextual indicators (breadcrumbs or prominent headings). In the absence of these signals, humans and AI have difficulties in locating and interpreting information.

The governance aspect shows sharp differences. Nations that have developed open data licensing and ethics - including the Open Government Licence in the UK, usage policies (implicit) in Canada, the Public Data Licence in Japan, and the open data licence in Singapore - have high scores due to their definition of reuse rights and focus on transparency. Licences and privacy considerations are also outlined in Kenya, New Zealand and Australia. Saudi Arabia stands out as having a comprehensive AI governance ecosystem: SDAIA's AI Ethics Principles (2023), the Generative AI Guidelines for Government (2024), a Personal Data Protection Law enforced since September 2023, and ISO 42001 AI management systems certification achieved in July 2024, together with the National AI Index measuring government AI readiness. This multi-layered framework earns Saudi Arabia a maximum governance score. On the other hand, South Africa, Brazil and Uruguay have more lax governance structures. The US and Pakistan have moderate results: there are national open licences, yet the ethical control and management of bias with AI are in the early stages. The assessment of Saudi Arabia's Unified National Platform (my.gov.sa) included in Table X and Table XI of this study was conducted using publicly available sources. During the evaluation period, the portal at times returned HTTP 403 (access denied) errors, which prevented complete direct access and analysis of certain technical aspects of the live website.

TABLE X. AI READINESS ASSESSMENT OF VERIFIED GOVERNMENT WEBSITES

Country (Region)	Data Foundation	Technical Optimisation	Content Quality	Discoverability & Context	Governance, Ethics & Accountability	Total (0–15)	%
Japan (Asia) https://www.digital.go.jp	3	3	3	2.5	3	15	97
United Kingdom (Europe) https://www.gov.uk	3	3	2	3	3	14	93
Canada (North America) https://www.canada.ca	3	3	2	3	2	13	87
Saudi Arabia (Middle East) https://my.gov.sa	3	2	2	2	3	12	80
United States https://www.usa.gov	3	3	2	2	2	12	80
Singapore (Asia) https://data.gov.sg	3	2	2	2	3	12	80
Kenya (Africa) https://www.opendata.go.ke	3	2	3	2	2	12	80
New Zealand (Oceania) https://www.data.govt.nz	3	2	2	2	3	12	80
India (Asia) https://www.data.gov.in	3	2	2	2	2	11	73
Australia (Oceania) https://data.gov.au	3	2	2	2	2	11	73
Pakistan (Asia) https://www.pakistan.gov.pk	3	2	2	2	2	11	73
Brazil (South America) https://www.gov.br	3	2	1	1	2	9	60
Uruguay (South America) https://www.gub.uy	2	1.5	2	1.5	2	9	60
South Africa (Africa) https://www.gov.za	2	1.5	2	1	2	8.5	57

G. Recalculated AI Readiness Index (Expert-Weighted)

Applying the Delphi-derived importance weights to the normalised dimension scores obtained in Phase 2, a weighted readiness percentage was computed for each government website. Table XI presents a comparison between the original equal-weight index (each dimension contributing equally) and the perception-weighted index based on expert consensus. The two schemes produce closely aligned rankings (Spearman $\rho = 0.982$, $p < 0.001$), with no website shifting by more than two rank positions. Re-weighting therefore adjusts absolute scores only modestly and leaves the overall ordering essentially unchanged, which supports the robustness of the index to reasonable variation in dimension importance.

High-performing websites such as Japan, the United Kingdom, and Canada retain their leading positions, with weighted scores of 98 %, 92 %, and 88 %, respectively. Their consistent strength across Data Foundation, Technical Optimisation, and Content Quality keeps them at the top under either weighting. Saudi Arabia (my.gov.sa) records a weighted score of 79 %, remaining within the closely grouped mid-tier of portals; its strong governance and data foundation offset comparatively lower discoverability. The mid-tier group, including the United States, Kenya, Singapore, New Zealand, India, Pakistan, and Australia, shifts only slightly under re-weighting, with changes of a few percentage points in either direction and at most a two-position rank movement.

TABLE XI. COMPARISON OF ORIGINAL EQUAL-WEIGHT AND DELPHI-WEIGHTED AI READINESS INDEX SCORES, WITH ABSOLUTE CHANGE (PERCENTAGE POINTS) AND RANK CHANGE FOR EACH GOVERNMENT WEBSITE

Government Website	Original equal-weight score (%)	Delphi-weighted score (%)	Change (p.p.)	Rank change
Japan	97	98	+1	0
United Kingdom	93	92	–1	0
Canada	87	88	+1	0
United States	80	83	+3	0
Singapore	80	79	–1	–2
Saudi Arabia (my.gov.sa)	80	79	–1	–2
Kenya	80	83	+3	0
New Zealand	80	79	–1	–2
India	73	75	+2	0
Pakistan	73	75	+2	0
Australia	73	75	+2	0
Brazil	60	62	+2	0
Uruguay	60	60	0	–1
South Africa	57	58	+1	0

The lower-scoring portals move in the same modest fashion: Brazil, Uruguay, and South Africa remain at the foot of the ranking under both schemes, with weighted scores of 62 %, 60 %, and 58 %, respectively. The absence of large re-orderings indicates that overall AI readiness is driven by broadly consistent quality differences across dimensions rather than by any single heavily weighted component. The expert-weighted index thus offers a citizen-centric emphasis on the dimensions experts consider most important, while yielding an ordering that is substantively consistent with the equal-weight baseline for reliable AI-mediated information access.

V. DISCUSSION

This research aimed to analyze government website preparedness to consume generative AI through the incorporation of a multidimensional assessment framework with the experts' agreement, citizen perception validation based on PLS-SEM, and objective assessment of the websites. The results can bring a number of valuable theoretical and practical insights into the position of governments in a new information ecosystem mediated by AI.

The initial significant contribution is that it helps to change the notion of AI readiness at the organisational and national levels to the level of the website, where the government information is actually read by AI systems and citizens. The single-round Delphi-informed expert survey demonstrated content validity of all five dimensions, but Governance, Ethics and Accountability had only borderline consensus (I-CVI = 0.67, which falls below the 0.78 criterion), and Discoverability only marginally met the criteria (I-CVI = 0.78). Such a shortage of good expert consensus on governance implies that although governance is theoretically significant, experts disagree on how to operationalize it based on publicly accessible features of the websites. Nevertheless, it is because of this that governance was retained, as it is eminent in international policy structures despite the divergence being realized.

The PLS-SEM results of citizen perceptions offered solid empirical evidence of the key role of Technical Optimisation and Content Quality in the process of perceived AI readiness. Technical Optimisation became the most significant predictor, followed by Content Quality. In contrast, Discoverability and Governance did not significantly influence citizens' perceptions. The citizens seem to emphasize the material, technical, and content qualities, and the governance processes are not seen by the citizens unless they are made known.

The analysis of the cross-country websites provides insights that although open data principles have been embraced by many governments, it does not necessarily equate to AI readiness. Examples of countries that do perform well, like Japan, the United Kingdom, and Canada, show that preparation needs not just the availability of data but also standardized metadata, regular content design, and transparent governance structures. On the contrary, weaker countries tend to have a lack of alignment between policy intent and technical execution, so their websites are human-friendly and not machine-friendly.

According to these empirical results, a pragmatic weighting scheme was used on the index. Governance and Discoverability were not important predictors of perceived AI readiness, so they were given half the weight (0.5 each), whereas Data Foundation, Technical Optimisation, and Content Quality were given the full weight (1.0 each). This change maintains the wholeness of the framework and makes the index consistent with the views of citizens. The weighted scores did not cause significant changes in rankings in comparison to the equal-weight index, which

verifies the strength of the original rankings. The websites that remain at the lower end of the ranking under both schemes (e.g. South Africa, Brazil, Uruguay) score consistently lower across Content Quality and Technical Optimisation, supporting the view that these dimensions are the principal drivers of readiness.

A sensitivity analysis comparing the four weighting schemes showed that the portal rankings are highly stable. Spearman rank correlations against the equal-weight baseline were 0.982 for the pragmatic scheme, 0.981 for the I-CVI-proportional scheme, and 0.981 for the path-informed scheme (all $p < 0.001$), and no portal moved more than two positions under any scheme. Absolute scores shifted modestly, most within three percentage points of the equal-weight value, but the ordering of jurisdictions was essentially invariant. Because Data Foundation scores are uniformly high and portals that perform well on one dimension tend to perform well on others, the composite is dominated by consistent cross-dimensional quality differences rather than by any single weighted component. The rankings can therefore be read as robust to reasonable disagreement about dimension importance.

The comparative analysis reveals that there are major differences in AI preparedness by country and region. The developed digital governments are more integrated in terms of technical, informational, and governance elements, which leads to higher readiness scores. Conversely, partial readiness is seen in most developing countries, where the quality of available data is high, but poorly optimised in technical terms and with irregular governance. This implies that an AI readiness gap is developing, with similar traits akin to the traditional digital divides but with novel aspects concerning machine intelligibility and structured information.

The relevance of content quality also points to the relevance of correct, up-to-date, and context-neutral content. Generative AI systems are also very dependent on source data to generate a response, and any lack of content quality can cause misinformation or incomplete results. Thus, it is important to note that governments should not only focus on the availability of information, but on its semantic clarity and reliability. A further and less visible challenge concerns the unevenness of quality within portals rather than merely between them. Even governments that score highly overall show marked variation across their contributing agencies, so that accuracy, currency, and structure can differ from one department to the next. This intra-portal variability means that readiness must be secured at the level of the individual page, since a retrieval system draws on whichever source it locates.

VI. Conclusion

This study contributes to the growing literature on digital government and generative AI by developing and validating a comprehensive framework for assessing the AI readiness of government websites. A one-round Delphi-informed expert survey confirmed content validity for all five dimensions, though governance did not reach the consensus threshold (I-CVI = 0.67, below the 0.78 criterion). A citizen PLS-SEM study revealed that Technical Optimisation and Content Quality are the strongest positive predictors of perceived AI readiness, while Discoverability and Governance are not significant. Based on these findings, a pragmatic weighting scheme was applied to the website assessment, giving half weight to the non-significant dimensions. The weighted index produced rankings largely consistent with the equal-weight version, confirming robustness. The study highlights substantial global disparities in readiness, suggesting the emergence of an AI readiness divide between advanced and developing digital governments. Addressing this divide will require coordinated efforts to improve technical capacity, standardise data practices, and strengthen governance frameworks. Overall, the study underscores that AI readiness is not simply a matter of adopting new technologies, but of restructuring digital information ecosystems to support machine interpretation and interaction. Because the index measures structural properties that condition AI retrieval rather than testing retrieval performance directly, its scores should be read as proxy indicators of readiness rather than as measures of realised generative-AI output. Governments that restructure their information ecosystems accordingly are, on this evidence, better positioned to benefit from AI-mediated access, though realised gains in service delivery and transparency remain to be demonstrated empirically. This study is subject to several limitations. First, the website evaluation is based on a purposive sample and may not fully capture the diversity of government websites globally. Second, the scoring process involves a degree of subjective judgment, although efforts were made to standardise evaluation criteria.

Third, the citizen perceptual model was validated on a modest sample ($n = 85$); while sufficient for PLS-SEM estimation, a larger and more diverse respondent pool would strengthen generalisability.

A. Future Research Directions

Several directions follow from this work. First, the framework could be operationalised as an automated auditing instrument, using crawlers and structured-data validators to score machine-readability, metadata completeness, and content structure at scale, reducing reliance on manual evaluation. Second, longitudinal application would allow the AI readiness divide identified here to be tracked over time, testing whether it widens or narrows as governments respond to generative-AI adoption. Third, as autonomous AI agents increasingly access public services on behalf of citizens, the framework could be extended to assess readiness for agentic interaction, including programmatic authentication, transaction completion, and machine-actionable service interfaces.

References

[1] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, art. 102274, 2023.

[2] Incubator for Artificial Intelligence, GOV.UK Chat: A Pilot Service. London, U.K.: UK Government, 2025.

[3] H. Mehr, Artificial Intelligence for Citizen Services and Government. Cambridge, MA, USA: Ash Center, Harvard Kennedy School, 2024.

[4] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2023.

[5] C. Hoffman and S. Grand, Indiana DOT's AI-Powered Workflow: A Case Study. Indianapolis, IN, USA: Indiana Department of Transportation, 2025.

[6] Saudi Data and Artificial Intelligence Authority, National Data Index. Riyadh, Saudi Arabia: SDAIA, 2025.

[7] Oxford Insights, Government AI Readiness Index 2025, 8th ed. Oxford, U.K.: Oxford Insights, Jan. 2026.

[8] R. Hržica, K. Debelak, and P. Pevcin, "A dual-level model of AI readiness in the public sector: Merging organizational and individual factors using TOE and UTAUT," Systems, vol. 13, no. 8, art. 705, 2025, doi: 10.3390/systems13080705.

[9] I. Krauze-Maślankowska and C. Vrabie, "Generative AI in municipal websites: A comparative analysis," Transylvanian Review of Administrative Sciences, vol. 21, no. 1, pp. 45–62, 2025.

[10] S. Radojičić and D. Vukmirović, "Exploring the impact of generative AI on digital inclusion: A case study of the e-government divide," AI, vol. 6, no. 12, art. 303, 2025, doi: 10.3390/ai6120303.

[11] W. van Donge, M. F. W. H. A. Janssen, and N. Bharosa, "Preparing public agencies for harnessing AI: A study on variables shaping multi-actor information infrastructures," in Electronic Government (EGOV 2024), LNCS vol. 14841. Cham, Switzerland: Springer, 2024, doi: 10.1007/978-3-031-70274-7_16.

[12] A. Simonofski, A. Nikiforova, M. Lnenicka, and N. Bono Rosselló, "Artificial intelligence as a catalyzer for open government data ecosystems: A typological theory approach," in Proc. 58th Hawaii Int. Conf. System Sciences (HICSS), 2025, pp. 2176–2185, doi: 10.24251/HICSS.2025.267.

[13] K. Layne and J. Lee, "Developing fully functional e-government: A four stage model," Government Information Quarterly, vol. 18, no. 2, pp. 122–136, 2001.

[14] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: A ten-year update," Journal of Management Information Systems, vol. 19, no. 4, pp. 9–30, 2003.

[15] Y. S. Wang and Y. W. Liao, "Assessing e-government systems success: A validation of the DeLone and McLean model," Government Information Quarterly, vol. 25, no. 4, pp. 717–733, 2008.

[16] Valdés, Gonzalo, Mauricio Solar, Hernán Astudillo, Marcelo Iribarren, Gastón Concha, and Marcello Visconti. "Conception, development and implementation of an e-Government maturity model in public agencies." Government Information Quarterly 28, no. 2 (2011): 176-187.

[17] S. Pather and R. Gomez, "A framework for e-government evaluation: The application of the ISO 25010 quality model," in Proc. 4th Int. Conf. Theory and Practice of Electronic Governance (ICEGOV), 2010, pp. 375–382.

[18] E. Amaral, M. Naranjo-Zolotov, and F. Bação, "Unequal AI readiness: Institutional and digital disparities in e-government across the European Union," Telematics and Informatics, vol. 107, art. 102400, Jun. 2026.

[19] S. Alsheibani, Y. Cheung, and C. Messom, "Towards an artificial intelligence (AI)-ready public sector: A systematic literature review," in Proc. 24th Pacific Asia Conf. Information Systems, 2020, pp. 1–15.

[20] J. Bright et al., "Generative AI is already widespread in the public sector: Evidence from a survey of UK public sector professionals," Digital Government: Research and Practice, 2024, doi: 10.1145/3700140.

[21] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quarterly, vol. 27, no. 3, pp. 425–478, 2003.

[22] S. Kelly, S. A. Kaye, and O. Oviedo-Trespalacios, "What factors contribute to the acceptance of artificial intelligence? A systematic review," Telematics and Informatics, vol. 77, art. 101925, 2023.

[23] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," MIS Quarterly, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[24] UK Government Digital Service, AI Ready Datasets: Key Pillars. London, U.K.: UK Government, 2026.

[25] Australian Government, Australian Government Technical Standard for AI. Canberra, Australia: Digital Transformation Agency, 2025.

[26] Georgia Technology Authority, Artificial Intelligence Responsible Use (SS-23-002). Atlanta, GA, USA: State of Georgia, 2023.

[27] R. F. DeVellis, Scale Development: Theory and Applications, 4th ed. Thousand Oaks, CA, USA: SAGE, 2017.

[28] S. B. MacKenzie, P. M. Podsakoff, and N. P. Podsakoff, "Construct measurement and validation procedures in MIS and behavioral research," MIS Quarterly, vol. 35, no. 2, pp. 293–334, 2011.

[29] A. Parasuraman and C. L. Colby, "An updated and streamlined technology readiness index: TRI 2.0," Journal of Service Research, vol. 18, no. 1, pp. 59–74, 2015.

[30] Hair Jr, Joe F., Marko Sarstedt, Lucas Hopkins, and Volker G. Kuppelwieser. "Partial least squares structural equation modeling (PLS-SEM)." European business review 26, no. 2 (2014): 106.

[31] N. Kock and P. Hadaya, "Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods," Information Systems Journal, vol. 28, no. 1, pp. 227–261, 2018, doi: 10.1111/isj.12131.

APPENDIX A. WEBSITE ASSESSMENT RUBRIC

Each metric was scored on a four-point scale (0-3) against the behavioural anchors below, using publicly available evidence only. Where evidence was absent or incomplete, the lower score was assigned. Fractional scores (e.g., 2.5) were permitted only in exceptional cases where a portal substantially met the higher anchor but retained a minor limitation.

TABLE XII. WEBSITE ASSESSMENT RUBRIC

Dimension	Code	Metric	Score 0	Score 1	Score 2	Score 3
Data Foundation	**DF1**	**Data Quality and Completeness**	No usable data published, or significant errors/gaps throughout.	Some data available but with visible gaps, inconsistencies, or no disclosed quality process.	Data generally accurate and complete; minor gaps in some areas.	Strong, documented data quality with comprehensive coverage and minimal gaps.
Data Foundation	**DF2**	**Machine-Readable Formats**	Content only in non-machine-readable formats (e.g. scanned/image PDFs).	Some machine-readable formats (CSV/JSON) available but inconsistent or limited.	Machine-readable formats available for most key datasets/content.	Machine-readable formats mandated/standard across the portal (CSV, JSON, XML, etc.).
Data Foundation	**DF3**	**Data Provenance and Lineage**	No documentation of data sources or collection methods.	Minimal or inconsistent provenance information.	Provenance documented for some datasets.	Clear, consistent documentation of sources, methods, and transformations.
Data Foundation	**DF4**	**Metadata Quality and Completeness**	No descriptive metadata.	Sparse or inconsistent metadata.	Reasonable metadata coverage with some gaps.	Comprehensive, standardised metadata (descriptive, structural, administrative).

Technical Optimisation	**TO1**	**Structured Data Implementation**	No schema.org or structured markup.	Isolated or inconsistent use of structured markup.	Structured markup present on some page types.	Comprehensive schema.org implementation across page types (e.g. via a shared component).
Technical Optimisation	**TO2**	**API Availability and Design**	No public API.	Limited or undocumented API.	APIs available with some documentation.	Well-documented APIs following best practice, with broad coverage.
Technical Optimisation	**TO3**	**Crawlability and Indexability**	robots.txt blocks crawling, or no sitemap exists.	Crawlable but with significant gaps or issues.	Generally crawlable, with a sitemap present.	Fully crawlable; explicit sitemaps; no unnecessary blocking of legitimate crawlers.
Technical Optimisation	**TO4**	**Performance and Reliability**	Frequent outages or very slow load times.	Noticeable performance or uptime issues.	Generally reliable, with occasional issues.	Consistently fast and high-uptime.
Content Quality for AI Consumption	**CQ1**	**Accuracy and Factual Correctness**	Significant factual errors found, or no verification process evident.	Some inaccuracies, or an unclear verification process.	Generally accurate, with minor lapses.	High accuracy, with evident verification processes.
Content Quality for AI Consumption	**CQ2**	**Currency and Timeliness**	Content is stale, with no dates shown.	Dates shown but content is significantly outdated.	Mostly current, with some lag.	Regularly updated, with clear publish/revision dates.
Content Quality for AI Consumption	**CQ3**	**Completeness and Self-Containment**	Requires extensive cross-referencing to understand.	Limited self-containment.	Mostly self-contained.	Comprehensive, self-contained content.
Content Quality for AI Consumption	**CQ4**	**Clarity and Unambiguity**	Highly ambiguous or jargon-heavy.	Some ambiguity present.	Generally clear.	Plain language, unambiguous throughout.
Discoverability and Context	**DC1**	**Semantic Structure and Hierarchy**	No heading structure.	Inconsistent use of headings.	Generally good heading hierarchy.	Consistent, logical heading hierarchy site-wide.
Discoverability and Context	**DC2**	**Contextual Independence**	Content is incomprehensible out of context.	Significant reliance on external context.	Mostly understandable independently.	Fully self-contained, context-independent content.
Discoverability and Context	**DC3**	**Internal Linking and Relationship Mapping**	No internal linking or breadcrumbs.	Sparse or inconsistent linking.	Reasonable internal linking.	Comprehensive breadcrumb/related-content linking.
Discoverability and Context	**DC4**	**Metadata for Discovery**	No descriptive metadata (titles/descriptions).	Minimal or inconsistent metadata.	Present but inconsistent.	Comprehensive, accurate discovery metadata.
Governance, Ethics and Accountability	**GE1**	**Information Ownership and Accountability**	No attribution of content to an agency/owner.	Inconsistent attribution.	Generally attributed.	Clear, consistent attribution and accountability.
Governance, Ethics and Accountability	**GE2**	**Quality Assurance Processes**	No evidence of a review process.	Minimal evidence of review.	Some documented review process.	Systematic, documented QA process.
Governance, Ethics and Accountability	**GE3**	**Transparency and Disclosure**	No disclosure of sources or limitations.	Minimal disclosure.	Reasonable disclosure.	Clear disclosure of sources, limitations, and any AI involvement.
Governance, Ethics and Accountability	**GE4**	**Ethical Framework and Bias Management**	No ethical framework.	Minimal or aspirational mention only.	Documented framework, partially implemented.	Comprehensive ethics framework with active bias management.
Governance, Ethics and Accountability	**GE5**	**Version Control and Audit Trail**	No version history.	Minimal or inconsistent version information.	Some version control evident.	Complete version history and audit trail.

Dimension-level application. In practice, the metric anchors informed a holistic 0–3 judgment at the dimension level, which is the level reported in the cross-country assessment. A dimension scored 3 where practice was comprehensive and consistent across its metrics, 2 where practice was reasonable with some gaps, 1 where practice was minimal or inconsistent, and 0 where the relevant capability was absent.

APPENDIX B. SURVEY INSTRUMENT

The instrument administered to citizen respondents comprised nineteen items across six constructs, each measured on a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5).

TABLE XIII. SURVEY INSTRUMENT

Section / Construct	Code	Item (measured on a five-point Likert scale)
A. Technical Optimisation	**TF1**	This website loaded quickly and worked reliably when I used it.
	TF2	This website performed smoothly across the pages and functions I needed.
	TF3	This website was technically well-built and worked well on my device.
B. Content Quality	**IQ1**	The information on this website was accurate.
	IQ2	The information on this website was current and up to date.
	IQ3	The information on this website was complete enough for my needs.
	IQ4	The information on this website was clearly presented and easy to understand.
C. Discoverability	**DC1**	I could easily find the specific information I was looking for on this website.
	DC2	The content on this website was logically structured.
	DC3	The information on this website was easy to navigate to and retrieve.
D. Governance	**GT1**	It was clear which government body was responsible for this website's content.
	GT2	The content on this website appeared to be regularly reviewed and maintained.
	GT3	This website was transparent about how its information is sourced or produced.
E. Digital Literacy	**DL1**	I understand how AI systems retrieve information from websites.
	DL2	I am confident using AI-powered search tools.
	DL3	I can generally tell when online information is well-structured for AI tools to use.
F. Perceived AI Readiness	**PAR1**	This website is well prepared for AI tools to read and use its information accurately.
	PAR2	This website is structured in a way that would let AI tools represent its content faithfully.
	PAR3	Overall, this website is ready to be used effectively by AI-powered search and assistant tools.