Developing the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI): A Multidimensional Framework for Evaluating Large Language Models

Paul Andrew Bourne, PhD, DrPH

Assistant Professor, University of the Commonwealth Caribbean, Kingston, Jamaica, WI

DOI: https://doi.org/10.47772/IJRISS.2026.100700494

Received: 24 July 2026; Accepted: 30 July 2026; Published: 05 August 2026

ABSTRACT

The rapid advancement of generative artificial intelligence (GenAI) and large language models (LLMs) has transformed sectors such as healthcare, education, finance, law, and public administration. Despite significant improvements in computational capability, concerns regarding hallucinations, misinformation, inconsistency, bias, opacity, and governance have exposed important limitations in existing benchmark-based evaluation methods. Current frameworks primarily assess task performance and reasoning ability but provide limited insight into the broader characteristics that determine whether AI systems can be trusted in high-stakes environments. This study addresses this gap by developing the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI), a multidimensional conceptual framework for evaluating AI trustworthiness beyond traditional benchmark metrics.

Using a conceptual research design, the study integrates Information Quality Theory, Trust Theory, the Technology Acceptance Model, Measurement Theory, Composite Index Theory, Artificial Intelligence Governance Theory, and Institutional Theory to establish an interdisciplinary foundation for trustworthy AI evaluation. GAITAI conceptualises Generative Artificial Intelligence Trustworthiness as a second-order latent construct comprising four first-order dimensions: Accuracy, Reliability, Robustness, and Transparency, operationalised through twenty-eight observable indicators.

The study contributes theoretically by redefining AI trustworthiness as a multidimensional construct and methodologically by extending psychometric measurement and composite indicator principles to AI evaluation. Practically, GAITAI provides organisations, AI developers, policymakers, and regulators with a structured framework for assessing AI trustworthiness and supporting responsible AI governance. Although conceptual, the framework establishes a foundation for future empirical validation, comparative benchmarking, and international standardisation. GAITAI advances the development of scientifically rigorous methods for evaluating trustworthy generative artificial intelligence and supports evidence-based AI governance across diverse organisational and societal contexts.

Keywords: Generative artificial intelligence; Large language models; Trustworthy artificial intelligence; Artificial intelligence governance; Composite index; Psychometric measurement; AI evaluation; Trustworthiness assessment.

INTRODUCTION

Generative artificial intelligence (GenAI) has emerged as one of the most transformative technological innovations of the twenty-first century, fundamentally reshaping knowledge creation, dissemination, and utilisation across sectors including healthcare, education, finance, law, scientific research, journalism, software engineering, and public administration (Bommasani et al., 2021; Brown et al., 2020; OpenAI, 2023; Vaswani et al., 2017). Built upon transformer architectures, large language models (LLMs) have demonstrated unprecedented

capabilities in natural language understanding and generation, enabling applications such as document summarisation, language translation, software development, clinical decision support, and evidence-based policymaking (Bubeck et al., 2023; Ouyang et al., 2022). Their increasing integration into organisational workflows reflects substantial advances in machine learning, computational power, and large-scale pre-training techniques, simultaneously raising important concerns regarding trustworthiness, reliability, transparency, and factual accuracy (Bommasani et al., 2021; NIST, 2023; OECD, 2019). As governments, businesses, educational institutions, and healthcare organisations increasingly rely on these technologies for high-stakes decision-making, the development of rigorous methodologies capable of evaluating the quality and trustworthiness of generative AI systems has become an important research and governance priority.

Despite their impressive capabilities, contemporary LLMs continue to exhibit significant limitations that constrain their safe and responsible deployment. One of the most widely recognised challenges is the generation of hallucinations, whereby models produce fabricated information, fictitious references, inaccurate numerical values, or logically inconsistent responses, presenting them with a high degree of linguistic confidence (Ji et al., 2023). Such inaccuracies are particularly problematic within medicine, law, finance, and scientific research, where erroneous outputs may result in inappropriate decisions with serious ethical, legal, and societal consequences (Bubeck et al., 2023; OpenAI, 2023). Furthermore, the probabilistic nature of LLMs results in variability across identical prompts because of contextual dependencies, model updates, and stochastic text generation, thereby challenging conventional notions of software reliability and reproducibility (Brown et al., 2020; Ouyang et al., 2022). Evaluating generative AI requires multidimensional approaches that extend beyond predictive accuracy to include reliability, robustness, transparency, explainability, and contextual appropriateness, reflecting the growing complexity of AI systems and their applications (NIST, 2023; OECD, 2019).

Current approaches to AI evaluation rely primarily on benchmark datasets such as MMLU, TruthfulQA, HELM, BIG-bench, GSM8K, and HumanEval, which have significantly advanced comparative assessment of language understanding, reasoning, mathematical problem-solving, coding, and factual knowledge (Bommasani et al., 2021; Hendrycks et al., 2021; Liang et al., 2023). Although these benchmarks provide valuable standardised measures of technical capability, they generally assess isolated aspects of model performance rather than offering comprehensive evaluations of overall trustworthiness. Models may perform exceptionally well on benchmark tasks while simultaneously generating fabricated citations, inconsistent information, or insufficiently transparent responses (Ji et al., 2023). This limitation has become increasingly important as generative AI expands into education, healthcare, scientific publishing, government, and other sectors where accountability, public trust, and regulatory compliance are essential (European Union, 2024; Nature, 2023; NIST, 2023; World Health Organization, 2021). International organisations, including the OECD, ISO, NIST, the European Union, and the World Health Organization, consistently emphasise that trustworthy AI should demonstrate characteristics such as accuracy, robustness, reliability, transparency, accountability, explainability, and appropriate human oversight; these frameworks do not provide a unified quantitative instrument capable of integrating these dimensions into a single interpretable measure (European Union, 2024; ISO/IEC, 2023; NIST, 2023; OECD, 2019; World Health Organization, 2021).

This study addresses that methodological gap through the development of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI), a comprehensive multidimensional framework designed to evaluate the trustworthiness of large language models. Drawing upon the well-established use of composite indices in fields such as economics, governance, healthcare, education, and environmental science, GAITAI integrates benchmark evidence with quantitative and qualitative indicators across four interrelated domains—Accuracy, Reliability, Robustness, and Transparency—to provide a unified assessment of AI trustworthiness. The study therefore seeks to (1) examine the conceptual and theoretical foundations of generative AI evaluation, (2) develop a multidimensional framework for measuring AI trustworthiness, (3) propose a mathematical formulation for constructing a composite trustworthiness index adaptable across application domains, and (4) outline a validation strategy for future empirical investigation. By providing a theoretically grounded and methodologically rigorous framework, GAITAI contributes to the emerging science of AI measurement, supporting responsible AI deployment, regulatory oversight, organisational governance, international standardisation, and evidence-based evaluation of generative artificial intelligence.

LITERATURE REVIEW

Generative artificial intelligence (GenAI) has fundamentally transformed artificial intelligence from systems primarily designed for prediction and classification into technologies capable of generating human-like text, images, software code, and other digital content across diverse domains (Bommasani et al., 2021). Early artificial intelligence was dominated by symbolic reasoning and expert systems that relied upon explicitly programmed rules and demonstrated limited adaptability when confronted with ambiguity or dynamic environments (Russell & Norvig, 2021). The emergence of statistical machine learning and subsequent advances in deep learning enabled computational systems to learn complex patterns directly from large datasets, thereby significantly improving performance in language processing, computer vision, and speech recognition (Goodfellow et al., 2016). A breakthrough occurred with the introduction of the transformer architecture by Vaswani et al. (2017), which revolutionised natural language processing through self-attention mechanisms capable of modelling long-range contextual relationships. This innovation established the foundation for contemporary large language models (LLMs), including GPT, Claude, Gemini, and Llama, whose remarkable capabilities have accelerated the integration of generative AI into healthcare, education, finance, law, scientific research, and public administration (Brown et al., 2020; OpenAI, 2023). Scholarly attention has increasingly shifted from evaluating computational capability alone towards examining whether these systems can be trusted to operate responsibly, transparently, and consistently within high-stakes environments (European Union, 2024; NIST, 2023; OECD, 2019).

The rapid adoption of LLMs has simultaneously intensified concerns regarding the quality and dependability of AI-generated information. Although reinforcement learning from human feedback and other alignment techniques have substantially improved model usefulness, concerns remain regarding hallucinations, factual inaccuracies, algorithmic bias, reproducibility, and transparency (Bubeck et al., 2023; Ji et al., 2023; Ouyang et al., 2022). These challenges demonstrate that increasingly sophisticated computational performance does not necessarily guarantee trustworthy outcomes, particularly in domains where inaccurate information may produce significant ethical, legal, or societal consequences. International organisations, including the OECD, NIST, ISO, and the European Union, increasingly emphasise that trustworthy artificial intelligence should demonstrate transparency, accountability, robustness, reliability, explainability, and appropriate human oversight in addition to technical excellence (European Union, 2024; ISO/IEC, 2023; NIST, 2023; OECD, 2019). This evolution has created a growing demand for multidimensional evaluation frameworks capable of integrating technical performance with broader governance principles.

Accuracy has traditionally served as one of the principal indicators of artificial intelligence performance; its interpretation within generative AI differs substantially from that used in conventional supervised machine learning (Goodfellow et al., 2016; Russell & Norvig, 2021). Rather than producing binary classifications, large language models generate probabilistic responses that combine information, contextual interpretation, logical reasoning, and linguistic fluency (OpenAI, 2023). Evaluating AI accuracy extends beyond simple correctness to encompass factual validity, logical coherence, numerical precision, contextual appropriateness, citation authenticity, and adherence to user instructions (Ji et al., 2023; Liang et al., 2023; Wang & Strong, 1996). Importantly, fluent and persuasive language should not be confused with factual correctness because models may generate coherent responses containing fabricated references or inaccurate scientific, legal, or historical information (Bubeck et al., 2023). Contemporary evaluation frameworks increasingly conceptualise accuracy as a multidimensional construct reflecting the overall quality and integrity of AI-generated information.

The interpretation of accuracy also depends upon the context in which AI-generated outputs are applied. In high-risk sectors such as healthcare, finance, law, and scientific research, even minor factual inaccuracies may have serious implications for patient safety, legal decisions, financial stability, or scientific integrity (European Union, 2024; World Health Organisation, 2021). Furthermore, trustworthy AI systems should communicate uncertainty appropriately when definitive evidence is unavailable rather than presenting speculative responses with unwarranted confidence (NIST, 2023). Confidence calibration has therefore emerged as an important characteristic of trustworthy AI because acknowledging uncertainty frequently enhances rather than diminishes user confidence. Accordingly, the proposed Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) conceptualise accuracy as a higher-order construct composed of multiple measurable indicators reflecting factual integrity, logical consistency, evidentiary support, numerical correctness, and contextual responsiveness, thereby providing a more comprehensive assessment than conventional benchmark scores alone.

Trustworthiness has become one of the central concepts within contemporary artificial intelligence research because it extends AI evaluation beyond computational performance to include ethical, technical, organisational, and societal considerations (European Union, 2024; NIST, 2023; OECD, 2019). As LLMs increasingly influence decision-making across healthcare, education, finance, law, public administration, and scientific research, organisations require assurance that these systems consistently produce outputs that are accurate, reliable, transparent, safe, and aligned with human values. Trustworthiness has evolved from an abstract ethical aspiration into a measurable characteristic influencing public confidence, regulatory compliance, organisational adoption, and long-term technological sustainability. International governance frameworks developed by organisations such as the OECD, NIST, ISO, and the European Union consistently recognise that trustworthy AI should support human well-being through transparency, accountability, robustness, fairness, explainability, and effective human oversight (European Union, 2024; ISO/IEC, 2023; NIST, 2023; OECD, 2019).

The multidimensional nature of trustworthiness is reinforced by both governance and behavioural research. The OECD AI Principles and the NIST Artificial Intelligence Risk Management Framework emphasise that trustworthy AI requires continuous lifecycle management rather than one-time technical evaluation because AI systems evolve through updates, retraining, and changing operational environments (NIST, 2023; OECD, 2019). Similarly, the European Union Artificial Intelligence Act establishes legally enforceable obligations concerning transparency, documentation, cybersecurity, and post-market monitoring for high-risk AI systems (European Union, 2024). Collectively, these frameworks demonstrate that trustworthiness depends upon the interaction of technical performance, organisational accountability, regulatory compliance, and public legitimacy rather than any single performance indicator. Multidimensional evaluation frameworks have become increasingly necessary for supporting responsible AI governance.

Reliability represents another essential dimension of trustworthy artificial intelligence because it reflects the consistency, stability, and dependability of system performance across repeated interactions and varying operational conditions (ISO/IEC, 2023; NIST, 2023). Unlike deterministic software systems, large language models generate probabilistic outputs in which identical prompts may produce different responses owing to contextual variation, sampling strategies, and model updates (Brown et al., 2020; Vaswani et al., 2017). Reliability therefore extends beyond system availability to encompass response consistency, reproducibility, operational stability, and sustained informational integrity over time (Liang et al., 2023). These characteristics are particularly important within healthcare, finance, education, law, and scientific research, where inconsistent recommendations may undermine decision-making, organisational confidence, and public trust (World Health Organization, 2021).

Recognising these challenges, the proposed Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) positions Accuracy, Reliability, Trustworthiness, and the broader principles of responsible AI governance as interconnected components of a multidimensional evaluation framework. Rather than relying upon isolated benchmark scores, GAITAI integrates complementary indicators capable of assessing factual correctness, response consistency, operational stability, governance readiness, and organisational accountability within a unified composite framework. This multidimensional approach addresses important methodological limitations within existing evaluation systems by providing a theoretically grounded and practically relevant foundation for future empirical validation, comparative benchmarking, regulatory oversight, and international standardisation of trustworthy generative artificial intelligence (European Union, 2024; ISO/IEC, 2023; NIST, 2023; OECD, 2019).

THEORETICAL FOUNDATION

The development of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) requires a robust interdisciplinary theoretical foundation because trustworthy artificial intelligence extends beyond computational performance to encompass informational quality, behavioural confidence, governance, institutional legitimacy, and scientific measurement (European Union, 2024; NIST, 2023; OECD, 2019). Traditional engineering approaches have primarily evaluated artificial intelligence using predictive accuracy, computational efficiency, and benchmark performance (Goodfellow et al., 2016). These technical measures alone cannot adequately explain why users trust AI systems, how organisations determine their suitability for high-risk applications, or how responsible AI should be governed. Accordingly, this study integrates six complementary

theoretical perspectives—Information Quality Theory, the Technology Acceptance Model (TAM), Trust Theory, Measurement Theory and Composite Index Theory, Artificial Intelligence Governance Theory, and Institutional Theory—to establish a comprehensive conceptual foundation for evaluating generative AI. Rather than representing competing perspectives, these theories collectively explain the technical, behavioural, organisational, governance, and methodological dimensions underpinning trustworthy artificial intelligence and provide the intellectual basis for the four principal domains of GAITAI: Accuracy, Reliability, Robustness, and Transparency.

Information Quality Theory provides the principal conceptual foundation for the Accuracy domain by proposing that information should be evaluated according to multiple complementary dimensions rather than factual correctness alone (Wang & Strong, 1996). Within information systems research, high-quality information is characterised by attributes including accuracy, credibility, completeness, consistency, relevance, timeliness, interpretability, and accessibility, all of which influence user confidence and decision-making. The emergence of generative artificial intelligence has substantially increased the relevance of this theory because large language models function primarily as producers of information rather than conventional computational tools (Brown et al., 2020; OpenAI, 2023). AI-generated outputs should be evaluated not only for factual correctness but also for contextual relevance, logical coherence, evidentiary support, credibility, and clarity. Information Quality Theory therefore provides strong theoretical justification for conceptualising accuracy as a multidimensional construct and supports the integration of Accuracy, Reliability, Robustness, and Transparency into a unified framework that extends beyond conventional benchmark evaluation.

The Technology Acceptance Model (TAM) complements this perspective by explaining how perceptions of usefulness, ease of use, and trustworthiness influence the adoption of generative artificial intelligence (Davis, 1989; Venkatesh & Davis, 2000; Venkatesh et al., 2003). Although modern large language models demonstrate remarkable capabilities in reasoning, programming, and content generation, sustained organisational adoption depends not only upon technical sophistication but also upon users' confidence that AI-generated information is accurate, reliable, transparent, and dependable (OpenAI, 2023). Perceived usefulness increasingly reflects trust in the quality of AI outputs; perceived ease of use is strengthened when systems clearly communicate their capabilities, limitations, uncertainty, and appropriate applications. Subsequent developments of TAM further recognise that technology adoption is shaped by organisational context, social influence, output quality, and behavioural confidence, highlighting that trustworthiness is a critical antecedent of successful AI implementation. Accordingly, the multidimensional structure of GAITAI directly supports the behavioural mechanisms through which organisations and individuals evaluate, accept, and integrate generative artificial intelligence.

Collectively, the Introduction, Information Quality Theory, and the Technology Acceptance Model establish the initial conceptual foundation for GAITAI by demonstrating that trustworthy artificial intelligence is inherently multidimensional. Information Quality Theory explains why AI-generated information must satisfy multiple quality characteristics to support effective decision-making, whereas TAM demonstrates that perceptions of informational quality and trustworthiness strongly influence technology adoption and sustained organisational use. Together, these theories justify the inclusion of Accuracy as the principal determinant of informational excellence and establish the behavioural rationale linking Accuracy, Reliability, Robustness, and Transparency with user confidence and organisational acceptance. The proposed framework extends traditional benchmark-oriented AI evaluation by integrating technical performance with information quality and behavioural theory, thereby providing a more comprehensive and theoretically grounded approach to assessing the trustworthiness of generative artificial intelligence.

CONCEPTUAL FRAMEWORK AND METHODOLOGY

Introduction

The preceding section established the theoretical foundation for the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) by integrating Information Quality Theory, the Technology Acceptance Model, Trust Theory, Measurement Theory, Composite Index Theory, Artificial Intelligence Governance Theory, and Institutional Theory. Collectively, these theoretical perspectives demonstrated that trustworthy artificial intelligence is a multidimensional construct that cannot be adequately represented through

isolated benchmark scores or single measures of computational performance. Instead, trustworthiness emerges through the interaction of several complementary dimensions that influence information quality, user confidence, organisational legitimacy, governance compliance, and responsible technological adoption. The theoretical synthesis therefore provides a robust conceptual basis for developing a scientifically rigorous framework capable of evaluating large language models beyond conventional performance metrics. Building upon these foundations, the present section translates the theoretical constructs into an operational conceptual model suitable for future empirical investigation and practical implementation.

The primary purpose of this section is to operationalise the theoretical concepts discussed previously into measurable constructs that collectively form the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI). Whereas Section Three explained *why* the four principal domains—Accuracy, Reliability, Robustness, and Transparency—are theoretically important, the present section explains *how* these domains are defined, structured, measured, and integrated into a unified multidimensional index. The section therefore bridges theoretical reasoning and empirical application by demonstrating the relationships among the latent constructs, their observable indicators, and the overall trustworthiness index. This transition from conceptualisation to operationalisation represents a critical stage in instrument development because it establishes the scientific procedures through which abstract theoretical concepts become measurable variables. This section provides the methodological architecture supporting the future validation and application of GAITAI.

The section adopts a conceptual research design, recognising that the present study is primarily concerned with developing a theoretically grounded evaluation framework rather than testing hypotheses using primary empirical data. Conceptual research seeks to integrate existing theoretical knowledge into coherent models that address important gaps within the literature and provide foundations for future empirical investigation. In the context of artificial intelligence evaluation, conceptual research is particularly appropriate because existing benchmark systems remain fragmented, governance frameworks remain largely normative, and no comprehensive multidimensional index currently integrates the principal determinants of AI trustworthiness into a single evaluation framework. Accordingly, the methodology employed in this study emphasises theoretical synthesis, construct development, indicator identification, and conceptual model specification. This approach is consistent with established practices in framework development within information systems, organisational research, and measurement science.

A central assumption underlying the proposed framework is that Generative Artificial Intelligence Trustworthiness represents a higher-order latent construct comprising four interrelated first-order dimensions: Accuracy, Reliability, Robustness, and Transparency. These dimensions are conceptually distinct yet theoretically complementary, with each capturing a unique aspect of trustworthy AI performance. Accuracy evaluates the factual correctness and contextual appropriateness of AI-generated outputs. Reliability measures consistency, reproducibility, and stability across repeated interactions. Robustness assesses resilience under diverse operational conditions, including ambiguous prompts, adversarial inputs, and evolving knowledge environments. Transparency evaluates explainability, documentation, disclosure of limitations, and communication of uncertainty. Collectively, these dimensions provide a comprehensive representation of trustworthy artificial intelligence that extends beyond traditional benchmark evaluation.

The proposed conceptual model assumes that improvements in each of the four first-order constructs contribute positively to the overall level of AI trustworthiness. Rather than functioning independently, these domains interact to influence users' confidence in AI-generated information and organisations' willingness to adopt AI technologies within critical decision-making environments. For example, highly accurate information that lacks transparency may still generate uncertainty among users; transparent systems producing inconsistent outputs may fail to achieve sustained organisational confidence. Similarly, robust systems demonstrating resilience under challenging operational conditions must also maintain factual accuracy and consistent behaviour to be regarded as genuinely trustworthy. These interrelationships illustrate that trustworthy artificial intelligence emerges through the combined influence of multiple complementary characteristics rather than through excellence in any single dimension. Accordingly, GAITAI conceptualises trustworthiness as an integrated multidimensional construct rather than a collection of isolated performance measures.

Methodologically, the framework adopts principles derived from Measurement Theory and Composite Index

Theory to guide the operationalisation of latent constructs into observable indicators. Each first-order construct is represented by a series of measurable indicators derived from the theoretical and empirical literature on artificial intelligence evaluation, information quality, governance, and trust. These indicators provide the empirical basis for future quantitative assessment, ensuring that each construct remains theoretically grounded. The use of multiple indicators for each construct enhances construct validity by capturing the multidimensional nature of AI trustworthiness, reducing dependence upon any individual measure. Furthermore, the hierarchical structure of the framework facilitates future statistical validation through confirmatory factor analysis and structural equation modelling, thereby supporting rigorous psychometric evaluation of the proposed index.

The conceptual framework also recognises that the practical application of GAITAI extends across multiple organisational and disciplinary contexts. Researchers may employ the framework to compare competing large language models, developers may utilise it to guide model improvement, regulators may apply it to support governance and compliance activities, and organisations may incorporate it into procurement, auditing, and risk management processes. The multidimensional structure of GAITAI therefore enhances its flexibility by allowing evaluation across diverse application domains, maintaining a consistent conceptual foundation. This adaptability distinguishes the proposed framework from existing benchmark systems that frequently focus upon narrow task-specific performance without considering broader governance or trustworthiness characteristics. The framework possesses both scientific relevance and practical utility.

The remainder of this section is organised into several interrelated sections. First, the conceptual research design underpinning the study is presented. Second, the conceptual framework and higher-order model of GAITAI are described in detail. Third, the four principal latent constructs—Accuracy, Reliability, Robustness, and Transparency—are operationally defined and linked to measurable indicators derived from the literature. Fourth, procedures for composite index construction, weighting, aggregation, and future validation are discussed. Finally, the section concludes by presenting the complete conceptual architecture of GAITAI, thereby establishing the methodological foundation for subsequent empirical testing and practical implementation. Through this progression, the section transforms the interdisciplinary theoretical foundations established previously into a coherent methodological framework capable of advancing the scientific evaluation of trustworthy generative artificial intelligence.

Research Design

The present study adopts a conceptual, theory-building research design to develop the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) as a multidimensional framework for evaluating the trustworthiness of large language models. Conceptual research is particularly appropriate for addressing emerging scientific problems that require the integration of existing theories, governance frameworks, and scholarly evidence rather than immediate hypothesis testing or primary data collection (Jaakkola, 2020). Given the fragmented nature of current AI benchmark systems and governance approaches, the study systematically synthesises theories from information systems, behavioural science, organisational studies, governance, and measurement science to establish a coherent framework for trustworthy AI evaluation. Guided by principles of theory-building research (Dubin, 1978; Whetten, 1989), the framework conceptualises Generative Artificial Intelligence Trustworthiness as a higher-order latent construct comprising four interrelated dimensions: Accuracy, Reliability, Robustness, and Transparency, identified through an extensive review of scholarly literature and internationally recognised AI governance frameworks.

The research design further emphasises construct development by recognising that trustworthiness is a latent concept that cannot be directly observed but must be represented through multiple measurable indicators (MacKenzie et al., 2011). Accordingly, each of the four principal dimensions is conceptually defined before identifying observable indicators capable of supporting future empirical measurement and psychometric validation. The multidisciplinary nature of the framework combines insights from information systems, psychology, organisational theory, computer science, and measurement science to capture the socio-technical complexity of generative artificial intelligence. Rather than focusing on algorithm or model development, the study concentrates on creating a comprehensive evaluation framework applicable across diverse large language models irrespective of their computational architecture, thereby enhancing its theoretical coherence, flexibility, and generalisability.

The proposed conceptual research design establishes the scientific foundation for future empirical investigation rather than replacing quantitative validation. Future studies are expected to evaluate the psychometric properties of GAITAI through exploratory and confirmatory factor analysis, structural equation modelling, reliability assessment, and validity testing to refine indicator selection, weighting procedures, and aggregation methodologies. The framework's principal strengths include its integration of fragmented theoretical perspectives into a coherent interdisciplinary model, clear conceptualisation of the four trustworthiness dimensions, alignment with emerging international AI governance frameworks, and applicability across sectors such as healthcare, education, finance, legal services, scientific research, and public administration. Consequently, this research design provides a robust methodological foundation for the subsequent development, validation, and practical implementation of GAITAI as a multidimensional framework for evaluating trustworthy generative artificial intelligence.

Conceptual Framework of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI)

The Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) was developed as a comprehensive conceptual framework for evaluating the trustworthiness of generative artificial intelligence beyond traditional benchmark performance. Existing AI evaluation approaches primarily assess isolated characteristics such as benchmark accuracy, reasoning ability, programming performance, and domain-specific task completion, but they provide limited insight into the broader qualities that determine whether AI-generated outputs can be trusted in organisational, professional, and societal contexts. Grounded in the interdisciplinary theoretical synthesis presented in Section Three, GAITAI integrates Information Quality Theory, the Technology Acceptance Model, Trust Theory, Measurement Theory, Composite Index Theory, Artificial Intelligence Governance Theory, and Institutional Theory to conceptualise trustworthiness as a multidimensional higher-order latent construct. Collectively, these theories demonstrate that trustworthy artificial intelligence emerges from the interaction of behavioural, informational, governance, organisational, and measurement perspectives rather than from technical performance alone.

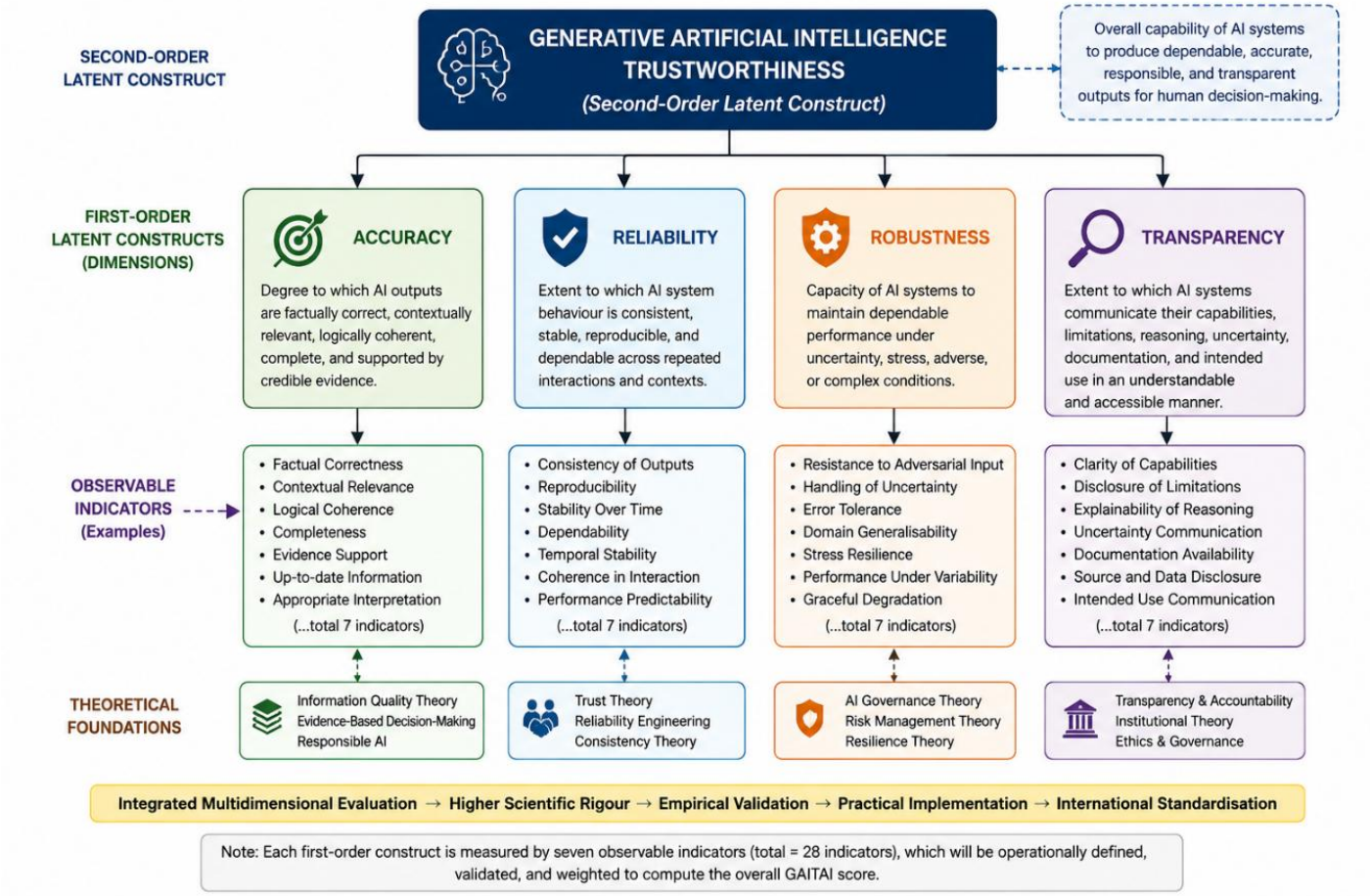
The framework adopts a hierarchical latent variable structure in which Generative Artificial Intelligence Trustworthiness represents the second-order construct measured through four complementary first-order dimensions: Accuracy, Reliability, Robustness, and Transparency (Hair et al., 2022). Accuracy evaluates the factual correctness, contextual relevance, logical coherence, completeness, and informational integrity of AI-generated outputs. Reliability measures the consistency, stability, reproducibility, and dependability of AI behaviour across repeated interactions and changing operational environments, while Robustness assesses the system's capacity to maintain dependable performance under uncertainty, adversarial conditions, and complex real-world situations. Transparency reflects the extent to which AI systems communicate their capabilities, limitations, reasoning processes, uncertainty, documentation, and intended applications, thereby supporting accountability, explainability, regulatory compliance, and informed human oversight (European Union, 2024; NIST, 2023; OECD, 2019). Together, these four interrelated dimensions provide a comprehensive representation of trustworthy generative artificial intelligence.

Although conceptually distinct, the four dimensions are expected to function collectively in determining overall AI trustworthiness, with balanced performance across all domains being essential for responsible AI deployment. Excellence in a single dimension cannot fully compensate for weaknesses in another; highly accurate systems that lack reliability or transparency, for example, may still undermine user confidence and organisational adoption. Consequently, GAITAI conceptualises trustworthiness as an emergent property arising from the interaction of Accuracy, Reliability, Robustness, and Transparency rather than from any single technical attribute. The framework is intentionally designed for broad applicability across researchers, AI developers, organisations, regulators, educators, healthcare professionals, and policymakers, providing a scientifically grounded and flexible methodology for future empirical validation, comparative benchmarking, governance, and responsible implementation of generative artificial intelligence across diverse sectors.

Figure 4.1 presents the conceptual structure underlying the proposed Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI). The figure illustrates the hierarchical relationship between the second-order latent construct of Generative Artificial Intelligence Trustworthiness and its four first-order

dimensions: Accuracy, Reliability, Robustness, and Transparency. Each first-order construct is represented by multiple observable indicators that will be operationally defined in the following section. This conceptual architecture establishes the structural foundation upon which the proposed index will be constructed, validated, and applied in future empirical investigations. By integrating behavioural, informational, governance, organisational, and measurement perspectives into a unified multidimensional model, the framework advances the scientific evaluation of generative artificial intelligence, providing a robust basis for future instrument development and international standardisation.

Figure 4.1. Conceptual Framework of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI)



Operational Definition of the Constructs

The development of a scientifically rigorous evaluation framework requires that each theoretical construct be translated into clearly defined operational dimensions capable of supporting objective measurement. Measurement Theory emphasises that latent constructs cannot be directly observed and therefore must be represented through carefully selected indicators that accurately reflect their conceptual meaning (DeVellis, 2017; Nunnally & Bernstein, 1994). The present study operationalises the four first-order constructs of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI)—Accuracy, Reliability, Robustness, and Transparency—using observable indicators derived from existing scholarly literature, internationally recognised governance frameworks, and established principles of artificial intelligence evaluation. The operational definitions presented in this section establish the conceptual boundaries of each construct, providing the foundation for future empirical measurement, psychometric validation, and practical implementation. These definitions also ensure conceptual consistency by aligning theoretical meaning with measurable evaluation criteria.

The operationalisation process adopted in this study follows internationally accepted procedures for instrument development and construct specification. First, each construct is defined conceptually based upon the theoretical synthesis presented in Section Three. Second, observable indicators representing the essential characteristics of each construct are identified from relevant literature. Third, the indicators are organised into coherent

measurement domains that collectively represent the latent construct. Finally, relationships among the constructs are specified within the broader conceptual framework of GAITAI. This systematic approach strengthens construct validity by ensuring that each indicator reflects established theoretical principles rather than arbitrary measurement decisions. Furthermore, the operational definitions facilitate future quantitative validation through exploratory factor analysis, confirmatory factor analysis, structural equation modelling, and composite reliability assessment.

The operational framework assumes that each of the four constructs contributes positively to the higher-order construct of Generative Artificial Intelligence Trustworthiness. Although each construct represents a distinct theoretical dimension, they collectively influence users' confidence in AI-generated information, organisational adoption of artificial intelligence, regulatory compliance, and responsible decision-making. Accordingly, deficiencies within any individual construct may reduce overall trustworthiness even where other dimensions demonstrate strong performance. This multidimensional perspective reflects the complex nature of trustworthy artificial intelligence supporting comprehensive evaluation across diverse application domains. The following sections present the operational definition of each construct in detail.

Accuracy

Within the proposed GAITAI framework, Accuracy is operationally defined as the extent to which a generative artificial intelligence system consistently produces factually correct, contextually appropriate, logically coherent, evidence-based, and free from material errors or hallucinations. Accuracy encompasses considerably more than traditional benchmark performance because trustworthy information must also demonstrate contextual relevance, internal consistency, appropriate interpretation, and credible sourcing. Large language models increasingly generate information influencing scientific research, healthcare, legal reasoning, financial analysis, education, journalism, and public policy; factual correctness alone cannot adequately represent informational quality. The operational definition therefore recognises that AI-generated information should accurately represent existing knowledge, communicating uncertainty where definitive evidence is unavailable. This broader conceptualisation reflects Information Quality Theory and contemporary scholarship concerning trustworthy artificial intelligence.

The Accuracy construct is represented through several observable indicators. These include factual correctness, referring to the extent to which generated statements correspond with verified knowledge; contextual relevance, measuring the appropriateness of responses within specific user requests or professional domains; logical coherence, evaluating the consistency and rational organisation of generated explanations; citation authenticity, assessing whether referenced sources exist and accurately support generated claims; numerical accuracy, measuring the correctness of quantitative calculations and statistical information; hallucination frequency, evaluating the extent to which fabricated information appears within generated outputs; and completeness of response, assessing whether essential information required for informed decision-making has been omitted. Collectively, these indicators provide a comprehensive representation of informational quality suitable for future empirical evaluation.

Operationally, higher levels of Accuracy indicate that an AI system consistently generates evidence-based information characterised by factual precision, contextual appropriateness, logical reasoning, and minimal hallucinations. Lower levels of Accuracy indicate increased frequencies of factual errors, fabricated references, inconsistent reasoning, misleading interpretations, or incomplete responses that may compromise decision-making. Within GAITAI, Accuracy constitutes one of the primary determinants of trustworthy artificial intelligence because incorrect information may undermine confidence irrespective of excellence in other dimensions. Accuracy represents the foundational informational component of the proposed multidimensional framework.

Reliability

Reliability is operationally defined as the degree to which a generative artificial intelligence system produces stable, consistent, reproducible, and dependable outputs across repeated interactions, equivalent prompts, and varying operational contexts. Reliability reflects the expectation that trustworthy AI systems should behave predictably, maintaining coherent reasoning and comparable performance over time. Unlike Accuracy, which

evaluates correctness at a single point of interaction, Reliability concerns the consistency of system behaviour across multiple observations. This distinction is particularly important because users frequently interact with large language models repeatedly when performing professional tasks requiring dependable information. Reliability constitutes a distinct dimension of trustworthiness rather than a secondary consequence of factual correctness.

Observable indicators representing Reliability include response consistency, assessing similarity of outputs generated from equivalent prompts; reproducibility, evaluating whether repeated evaluations produce comparable results; temporal stability, measuring consistency following software updates or evolving operational conditions; internal consistency, assessing coherence throughout extended conversations; decision consistency, evaluating whether comparable scenarios receive comparable recommendations; error stability, measuring the predictability of system limitations; and operational dependability, assessing overall confidence that the AI system performs consistently across repeated applications. These indicators collectively capture the behavioural consistency expected of trustworthy artificial intelligence systems operating within professional and organisational environments.

Higher Reliability scores indicate AI systems that consistently generate stable and dependable responses, maintaining coherent reasoning across repeated interactions. Lower scores indicate unpredictable behaviour, contradictory recommendations, or inconsistent performance that may reduce user confidence despite acceptable levels of factual accuracy. Accordingly, Reliability contributes directly to organisational trust, behavioural acceptance, and sustained utilisation of artificial intelligence technologies.

Robustness

Within GAITAI, Robustness is operationally defined as the capacity of a generative artificial intelligence system to maintain accurate, reliable, and contextually appropriate performance under diverse operational conditions, including ambiguity, uncertainty, adversarial prompts, incomplete information, and unfamiliar problem domains. Robust systems continue functioning effectively despite environmental variability, thereby reducing operational risk and strengthening user confidence. Because real-world applications rarely resemble controlled benchmark environments, robustness represents a critical determinant of trustworthy artificial intelligence. This construct reflects resilience, adaptability, and dependable performance under realistic operational conditions.

Observable indicators of Robustness include adversarial resilience, measuring resistance to prompt manipulation or malicious inputs; generalisation capability, assessing performance across diverse domains beyond training benchmarks; uncertainty management, evaluating appropriate acknowledgement of incomplete knowledge; error recovery, assessing the ability to respond appropriately following incorrect assumptions or ambiguous requests; operational resilience, measuring stability across varying application environments; domain adaptability, evaluating performance across specialised disciplines; and risk mitigation, assessing the system's ability to minimise potentially harmful outputs during challenging interactions. These indicators collectively capture the operational resilience required for trustworthy AI deployment across complex organisational settings.

Higher Robustness indicates that AI systems continue generating dependable, coherent, and responsible outputs despite uncertainty or changing operational circumstances. Lower Robustness reflects vulnerability to adversarial prompts, inconsistent behaviour under ambiguity, or reduced performance outside narrowly defined benchmark conditions. Robustness therefore contributes directly to responsible AI governance by reducing operational risk and strengthening confidence in AI-supported decision-making.

Transparency

Transparency is operationally defined as the extent to which a generative artificial intelligence system openly communicates its capabilities, limitations, assumptions, sources of uncertainty, reasoning processes, documentation, and intended applications in ways that enable informed human judgement. Transparency does not necessarily require disclosure of proprietary algorithms or model architectures but rather sufficient openness to allow users to understand the appropriate interpretation and application of AI-generated outputs. International governance frameworks consistently identify transparency as a defining characteristic of trustworthy artificial intelligence because it promotes accountability, explainability, auditability, and responsible decision-making. Transparency serves simultaneously as a governance principle, an organisational requirement, and a behavioural

determinant of user trust.

Observable indicators representing Transparency include explainability, assessing the clarity with which AI systems communicate reasoning processes; uncertainty disclosure, measuring acknowledgement of incomplete or uncertain knowledge; documentation quality, evaluating availability and comprehensiveness of technical information; limitation disclosure, assessing openness regarding known weaknesses and operational boundaries; traceability, measuring the ability to identify information sources or reasoning pathways where appropriate; auditability, evaluating the capacity for independent review of AI performance; and communication clarity, assessing whether users receive understandable explanations concerning generated outputs. Collectively, these indicators represent the openness required for responsible AI governance and informed human oversight.

Higher Transparency scores indicate AI systems that communicate openly regarding capabilities, limitations, uncertainty, and appropriate application contexts supporting informed human judgement. Lower scores indicate opaque systems that provide limited explanation, inadequate documentation, or insufficient disclosure regarding operational limitations. Transparency therefore strengthens accountability, governance, regulatory compliance, and user confidence, reducing inappropriate reliance upon AI-generated information.

Integration of the Four Constructs

Although operationally distinct, the four constructs function collectively to represent the higher-order construct of Generative Artificial Intelligence Trustworthiness. Accuracy ensures informational correctness, Reliability establishes behavioural consistency, Robustness promotes resilience under diverse operational conditions, and Transparency supports openness, accountability, and informed human oversight. Improvements across these four dimensions collectively enhance overall AI trustworthiness, whereas substantial weaknesses within any single construct may reduce confidence in AI-generated information regardless of strengths elsewhere. Accordingly, the multidimensional structure of GAITAI reflects the theoretical proposition that trustworthy artificial intelligence cannot be adequately evaluated through isolated technical benchmarks but instead requires comprehensive assessment across multiple complementary domains. The following section identifies the specific measurement indicators and evaluation criteria through which these operational definitions may be translated into a practical multidimensional index suitable for future empirical validation and organisational application.

Measurement Indicators and Evaluation Criteria

The operational definitions presented in the preceding section establish the conceptual meaning of the four first-order constructs underpinning the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI). To facilitate future empirical validation and practical application, these constructs must now be translated into measurable indicators capable of supporting objective and reproducible evaluation. Measurement Theory emphasises that latent variables become scientifically useful only when represented through observable indicators demonstrating conceptual relevance, reliability, and construct validity (DeVellis, 2017; Nunnally & Bernstein, 1994). Similarly, Composite Index Theory recommends that indicators should be theoretically justified, empirically measurable, and sufficiently comprehensive to capture the multidimensional nature of the underlying construct (OECD & Joint Research Centre, 2008). The measurement framework developed in this study identifies observable indicators for each of the four principal dimensions of GAITAI, maintaining consistency with the theoretical foundations established in Sections Two and Three. These indicators collectively provide the basis for future scoring procedures, statistical validation, benchmarking, and comparative evaluation of large language models.

The proposed measurement framework adopts several guiding principles. First, every indicator must possess a clear theoretical justification derived from the literature on information quality, trust, governance, measurement science, or artificial intelligence evaluation. Second, indicators should represent observable characteristics that can be assessed consistently across different large language models and application domains. Third, indicators should minimise subjectivity by relying upon explicit evaluation criteria wherever possible. Fourth, the measurement system should remain sufficiently flexible to accommodate future developments in artificial intelligence without requiring fundamental modification of the conceptual framework. Finally, the indicators should collectively provide comprehensive coverage of the higher-order construct of Generative Artificial

Intelligence Trustworthiness, ensuring that no important dimension is omitted from the evaluation process. These principles strengthen both the scientific integrity and practical applicability of the proposed index.

The measurement indicators are organised according to the four first-order constructs comprising GAITAI. Each construct contains multiple indicators representing complementary aspects of the broader latent variable. Rather than relying upon a single measure, the framework adopts a multidimensional approach in which individual indicators collectively capture the complexity of trustworthy artificial intelligence. This strategy enhances construct validity by reducing dependence upon isolated performance measures, permitting future psychometric analyses to identify the relative contribution of each indicator to the broader construct. Furthermore, multiple indicators facilitate greater flexibility because organisations may adapt evaluation procedures according to specific operational contexts, preserving the overall conceptual integrity of the framework.

Indicators for Accuracy

The Accuracy construct is measured through indicators evaluating the factual correctness, contextual appropriateness, informational completeness, and logical integrity of AI-generated outputs. These indicators reflect the principles of Information Quality Theory and contemporary AI evaluation research, recognising that trustworthy information extends beyond simple factual correctness.

The proposed Accuracy indicators include:

Indicator	Operational Description
Factual Correctness	Degree to which generated information corresponds with verified evidence and established knowledge.
Contextual Relevance	Extent to which responses appropriately address the user's question or task.
Logical Coherence	Internal consistency and rational organisation of explanations or arguments.
Citation Authenticity	Accuracy and existence of references or supporting sources where citations are provided.
Numerical Accuracy	Correctness of calculations, statistical values, quantitative reasoning, and mathematical outputs.
Completeness	Inclusion of essential information necessary for informed understanding or decision-making.
Hallucination Frequency	Frequency with which fabricated facts, references, events, or unsupported claims appear.

Collectively, these indicators evaluate whether AI-generated information demonstrates sufficient informational quality to support professional, organisational, and academic decision-making. High scores indicate evidence-based, coherent, and contextually appropriate outputs characterised by minimal hallucinations and strong factual integrity.

Indicators for Reliability

The Reliability construct evaluates the consistency and reproducibility of AI performance across repeated interactions and operational contexts. Reliable systems should behave predictably, maintaining stable reasoning and dependable performance over time.

The proposed Reliability indicators include:

Indicator	Operational Description
Response Consistency	Similarity of responses generated for equivalent prompts administered repeatedly.
Reproducibility	Ability to generate comparable outputs under similar evaluation conditions.
Temporal Stability	Maintenance of consistent performance following software updates or over extended

Indicator	Operational Description
	periods.
Conversation Consistency	Internal coherence maintained throughout prolonged interactions.
Decision Consistency	Consistency of recommendations across comparable scenarios.
Operational Dependability	Overall predictability of AI performance during routine use.
Error Stability	Consistency in recognising and appropriately handling known limitations.

These indicators collectively evaluate whether users may reasonably expect similar outputs when interacting repeatedly with the same AI system. High Reliability indicates dependable system behaviour, whereas low Reliability reflects unpredictable performance capable of reducing user confidence and organisational trust.

Indicators for Robustness

The Robustness construct measures the resilience of artificial intelligence systems when confronted with uncertainty, ambiguity, adversarial prompts, or unfamiliar operational environments. Robust systems maintain dependable performance despite changing circumstances and imperfect information.

The proposed Robustness indicators include:

Indicator	Operational Description
Adversarial Resistance	Ability to resist manipulation through malicious or misleading prompts.
Generalisation Capability	Capacity to perform effectively across multiple domains beyond benchmark tasks.
Uncertainty Management	Appropriate acknowledgement of incomplete knowledge and uncertain information.
Operational Resilience	Stability of performance under diverse environmental conditions.
Error Recovery	Ability to recognise mistakes and provide appropriate corrective responses.
Domain Adaptability	Effectiveness across specialised disciplines requiring different forms of expertise.
Risk Mitigation	Capacity to minimise potentially harmful, misleading, or unsafe outputs.

These indicators evaluate the resilience of AI systems operating within complex real-world environments where uncertainty and variability are inevitable. High Robustness demonstrates dependable performance across diverse circumstances, reducing operational risk.

Indicators for Transparency

The Transparency construct evaluates the openness with which AI systems communicate capabilities, limitations, uncertainty, documentation, and reasoning processes. Transparency enables informed human oversight and strengthens accountability within responsible AI governance.

The proposed Transparency indicators include:

Indicator	Operational Description
Explainability	Clarity of explanations accompanying generated responses or recommendations.
Uncertainty Disclosure	Communication of confidence levels, ambiguity, or limitations where appropriate.
Documentation Quality	Availability and comprehensiveness of technical documentation describing system capabilities and constraints.
Limitation Disclosure	Openness regarding known weaknesses, biases, and operational boundaries.
Traceability	Ability to identify supporting evidence or reasoning processes where feasible.
Auditability	Extent to which AI performance can be independently evaluated and reviewed.
Communication Clarity	Ease with which users understand AI outputs and accompanying explanations.

These indicators collectively assess whether AI systems provide sufficient transparency to support responsible use, regulatory compliance, and informed human decision-making. High Transparency reduces inappropriate reliance upon AI, strengthening organisational accountability and public trust.

Multidimensional Measurement Structure

The measurement framework proposed in this study consists of 28 observable indicators distributed across the four principal latent constructs of GAITAI. Rather than treating individual indicators independently, the framework assumes that each indicator contributes to its respective first-order construct, which in turn contributes to the higher-order construct of Generative Artificial Intelligence Trustworthiness. This hierarchical measurement structure is consistent with established principles of psychometric instrument development and higher-order latent variable modelling (Hair et al., 2022). Future empirical studies may evaluate the dimensionality of these indicators through exploratory factor analysis and confirmatory factor analysis, assessing construct reliability using Cronbach's alpha, composite reliability, and average variance extracted.

An important strength of the proposed measurement framework is its flexibility. Although the indicators are presented as a comprehensive set suitable for broad AI evaluation, future empirical investigations may refine, expand, or adapt individual indicators according to specific application domains. For example, healthcare applications may incorporate additional indicators relating to clinical safety, diagnostic reasoning, and evidence-based medicine, whereas legal applications may include indicators addressing statutory interpretation, legal citation accuracy, and procedural consistency. Educational contexts may place greater emphasis upon pedagogical appropriateness, academic integrity, and instructional clarity. This adaptability enables GAITAI to remain theoretically stable, accommodating the evolving capabilities and applications of generative artificial intelligence.

Overall, the measurement indicators presented in this section translate the theoretical constructs developed throughout the preceding sections into observable variables capable of supporting objective evaluation. The framework integrates principles from Information Quality Theory, Trust Theory, Artificial Intelligence Governance Theory, Measurement Theory, and Composite Index Theory into a coherent multidimensional measurement system. By identifying explicit indicators representing Accuracy, Reliability, Robustness, and Transparency, the proposed framework establishes the methodological foundation for constructing a scientifically rigorous composite index capable of evaluating the trustworthiness of large language models across diverse organisational, professional, and societal contexts. The following section describes the procedures through which these indicators may be standardised, weighted, aggregated, and combined to produce the overall Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI).

Although the present study does not empirically operationalise the proposed indicators, future applications of GAITAI require transparent measurement procedures. Table 4.5 presents illustrative approaches through which selected indicators may be evaluated. These examples are intended to demonstrate the operational feasibility of the framework rather than prescribe mandatory measurement procedures. Future empirical investigations may refine these methods according to the application domain, available data sources, and validation requirements.

The measurement methods presented in Table 4.5 are illustrative rather than prescriptive and are included to demonstrate the operational feasibility of the proposed framework. Future empirical investigations may refine these procedures and develop domain-specific scoring protocols according to the characteristics of the AI system under evaluation.

Table 4.5. Illustrative Measurement Methods for GAITAI Indicators

Indicator	Example Measurement Method	Data Source	Proposed Scale
Factual correctness	Expert comparison against verified reference	Human evaluators	1–5
Citation authenticity	Percentage of valid citations	Automated verification	0–100
Logical coherence	Independent expert rating	Human judges	1–5
Numerical accuracy	Comparison with correct answer	Automated scoring	Percentage

Indicator	Example Measurement Method	Data Source	Proposed Scale
Hallucination frequency	False statements ÷ total statements	Human + automated	Percentage
Response consistency	Repeated prompt similarity	Repeated testing	0–100
Explainability	Expert evaluation rubric	Human	1–5

Development of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI)

Proposed Operational Scoring Protocol

Although GAITAI is introduced as a conceptual framework, the proposed scoring procedure provides a transparent methodology that may guide future empirical implementation.

Step 1: Each observable indicator is evaluated independently using predefined evaluation criteria.

Step 2: Each indicator receives a raw score on a five-point scale:

Raw Score	Interpretation
1	Very Poor
2	Poor
3	Acceptable
4	Good
5	Excellent

Step 3: Raw scores are standardised to a common 0–100 metric:

$$S_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \times 100$$

Step 4: Domain scores are calculated as the arithmetic mean of the standardised indicators:

$$D_j = \frac{\sum S_i}{n}$$

Step 5: Overall trustworthiness is calculated as:

$$GAITAI = \sum_{j=1}^{4} w_j D_j$$

where

$$\sum w_j = 1.$$

Equal weighting is proposed during conceptual development, although future empirical studies may estimate optimal weights statistically.

The principal objective of the present study is the development of a scientifically rigorous composite index capable of evaluating the trustworthiness of generative artificial intelligence through the integration of multiple complementary dimensions. The preceding sections identified the conceptual framework and measurable indicators underpinning GAITAI; the present section explains how these indicators may be combined into a unified multidimensional index suitable for systematic evaluation, benchmarking, governance, and future empirical validation. Composite indices have become widely recognised as effective tools for measuring complex

multidimensional phenomena because they integrate diverse indicators into a single interpretable measure preserving the conceptual richness of the underlying construct (OECD & Joint Research Centre, 2008). Well-established examples include the Human Development Index, Global Innovation Index, Environmental Performance Index, and Corruption Perceptions Index, each demonstrating how multiple dimensions may be synthesised into meaningful measures supporting evidence-based decision-making. The proposed GAITAI extends this established methodological tradition into the rapidly evolving field of generative artificial intelligence evaluation.

The development of a composite index involves several sequential methodological stages, including construct specification, indicator selection, normalisation, weighting, aggregation, validation, and interpretation (OECD & Joint Research Centre, 2008). These stages ensure that the resulting index possesses conceptual coherence, methodological transparency, and empirical reliability. In the present study, construct specification was established through the interdisciplinary theoretical framework presented in Section Three; indicator selection was described in the preceding section. The remaining stages concern the transformation of multiple indicators into a single composite measure representing Generative Artificial Intelligence Trustworthiness. Adhering to internationally recognised principles of composite indicator construction enhances the scientific credibility of GAITAI, facilitating future empirical validation and cross-model comparison.

The conceptual architecture of GAITAI assumes a hierarchical measurement structure. At the lowest level are the 28 observable indicators identified in Section 4.5. These indicators contribute to four first-order latent constructs: Accuracy, Reliability, Robustness, and Transparency. Each first-order construct subsequently contributes to the second-order latent construct of Generative Artificial Intelligence Trustworthiness, which is represented by the overall GAITAI score. This hierarchical structure reflects established practices in psychometric modelling and structural equation modelling, where complex latent variables are represented through multiple interrelated dimensions (Hair et al., 2022). Such an approach strengthens theoretical coherence, permitting future statistical validation using higher-order confirmatory factor analysis.

An important methodological consideration concerns the standardisation of measurement indicators. Because individual indicators may initially be measured using different scales, scoring procedures, or evaluation methods, standardisation is necessary to permit meaningful comparison and aggregation. Composite Index Theory recommends transforming indicators onto a common measurement scale before combining them into composite scores (OECD & Joint Research Centre, 2008). Future empirical implementation of GAITAI may therefore employ standardisation techniques such as min-max normalisation, z-score standardisation, or percentile transformation depending upon the nature of available data. Standardisation ensures that no individual indicator exerts disproportionate influence simply because of differences in measurement units or scale ranges. The resulting index more accurately reflects theoretical relationships rather than measurement artefacts.

Following standardisation, individual indicators are aggregated to produce scores for each of the four first-order constructs. Each construct score represents the combined performance of its constituent indicators, preserving the conceptual distinction among Accuracy, Reliability, Robustness, and Transparency. This intermediate level of analysis enables organisations to identify specific strengths and weaknesses within AI systems rather than relying exclusively upon an overall trustworthiness score. For example, an AI model may demonstrate exceptional Accuracy, exhibiting comparatively weaker Transparency, thereby informing targeted model improvement or governance interventions. This multidimensional reporting structure enhances the practical utility of GAITAI by supporting diagnostic evaluation in addition to comparative benchmarking. Both overall and construct-specific scores provide valuable information for researchers, developers, regulators, and organisational decision-makers.

The determination of indicator weights represents another important consideration in composite index development. Existing composite indices employ various weighting strategies, including equal weighting, expert judgement, statistical weighting based upon factor loadings or principal component analysis, and policy-driven weighting reflecting contextual priorities (OECD & Joint Research Centre, 2008). At the conceptual stage of the present study, equal weighting across the four first-order constructs provides the most appropriate initial approach because each dimension is theoretically justified as an essential component of trustworthy artificial intelligence. Equal weighting also enhances transparency, interpretability, and methodological neutrality during early

framework development. Nevertheless, future empirical investigations may evaluate alternative weighting schemes using confirmatory factor analysis, structural equation modelling, Delphi expert panels, or multicriteria decision analysis to determine whether differential weighting improves predictive validity or practical applicability across specific sectors.

Under an equal-weighting approach, the four first-order constructs contribute equally to the overall trustworthiness score. Conceptually, the overall GAITAI score may therefore be represented by the following expression:

$$\text{GAITAI} = \frac{\text{Accuracy} + \text{Reliability} + \text{Robustness} + \text{Transparency}}{4}$$

where each construct score represents the aggregated standardised performance of its constituent indicators. This formulation reflects the theoretical proposition that trustworthy artificial intelligence depends upon balanced performance across multiple complementary dimensions rather than exceptional performance in a single domain. Deficiencies within one construct cannot be fully compensated by excellence in another, thereby reinforcing the multidimensional nature of AI trustworthiness established throughout this study.

Future empirical validation may extend this conceptual model by introducing weighted formulations where theoretical or statistical evidence supports differential contributions of the four constructs. A more general expression may therefore be represented as:

$$\text{GAITAI} = w_1(\text{Accuracy}) + w_2(\text{Reliability}) + w_3(\text{Robustness}) + w_4(\text{Transparency})$$

subject to the constraint:

$$w_1 + w_2 + w_3 + w_4 = 1$$

where w_1, w_2, w_3, and $w_4$represent empirically determined weighting coefficients. This formulation provides methodological flexibility, maintaining the conceptual integrity of the framework. Sector-specific implementations may subsequently adjust weighting according to operational priorities without altering the theoretical structure of GAITAI.

To facilitate interpretation, the composite index may also be categorised into qualitative trustworthiness levels. Although empirical validation is required before establishing definitive thresholds, an illustrative interpretation framework may classify overall scores as follows:

GAITAI Score	Interpretation
90–100	Excellent Trustworthiness
80–89	High Trustworthiness
70–79	Good Trustworthiness
60–69	Moderate Trustworthiness
50–59	Low Trustworthiness
Below 50	Very Low Trustworthiness

These categories are presented solely as conceptual examples and should be refined through future empirical validation involving multiple large language models, expert evaluation panels, and psychometric analysis. Nevertheless, such classification enhances the practical usability of GAITAI by enabling straightforward interpretation for policymakers, organisations, researchers, developers, and end-users.

The proposed index possesses several methodological advantages compared with existing benchmark-based evaluation systems. First, it integrates technical, behavioural, organisational, and governance dimensions into a unified multidimensional framework rather than focusing exclusively upon computational performance. Second, it supports diagnostic evaluation by reporting both construct-specific and overall trustworthiness scores. Third,

its hierarchical structure facilitates advanced statistical validation using higher-order latent variable modelling. Fourth, the framework remains sufficiently flexible to accommodate future developments in generative artificial intelligence without requiring substantial conceptual revision. Finally, the proposed methodology aligns closely with internationally recognised principles governing composite index development and responsible artificial intelligence, thereby strengthening both scientific credibility and practical applicability.

Overall, the development process described in this section establishes the methodological foundation for transforming the theoretical constructs of Accuracy, Reliability, Robustness, and Transparency into a comprehensive multidimensional index capable of evaluating trustworthy generative artificial intelligence. The framework integrates established principles from Measurement Theory and Composite Index Theory, maintaining consistency with the behavioural, informational, organisational, and governance perspectives developed throughout the preceding sections. By combining observable indicators into a scientifically structured higher-order construct, GAITAI provides a practical methodology for benchmarking, governance, organisational decision-making, and future empirical investigation. The following section presents the proposed procedures for psychometric validation, reliability assessment, construct validation, and future empirical testing required to establish the scientific robustness of the index.

Validation of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI)

The development of a scientifically rigorous measurement instrument extends beyond theoretical conceptualisation and indicator identification to include systematic validation procedures that establish the reliability, validity, and practical applicability of the proposed framework. Measurement Theory emphasises that newly developed instruments should demonstrate strong psychometric properties before being adopted for widespread research or organisational use (DeVellis, 2017; Nunnally & Bernstein, 1994). Accordingly, although the present study is conceptual in nature, it is important to specify the methodological procedures through which the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) may be empirically validated in future investigations. Such validation will determine whether the proposed framework accurately measures the latent construct of Generative Artificial Intelligence Trustworthiness, ensuring consistency, reproducibility, and generalisability across different large language models and application contexts. This section therefore presents a comprehensive validation framework consistent with internationally accepted practices in psychometrics, information systems research, and composite index development.

Validation is a continuous scientific process rather than a single statistical procedure. Modern measurement theory recognises several complementary forms of validity, including content validity, face validity, construct validity, convergent validity, discriminant validity, criterion-related validity, and predictive validity (Hair et al., 2022). Similarly, reliability must be assessed through multiple complementary procedures evaluating internal consistency, temporal stability, and reproducibility. Collectively, these validation procedures establish scientific confidence that the proposed instrument accurately measures the intended construct, producing stable results across different contexts. The future empirical validation of GAITAI should adopt a comprehensive psychometric strategy rather than relying upon isolated statistical indicators.

The first stage of validation concerns content validity, which evaluates the extent to which the proposed indicators adequately represent the conceptual domain of trustworthy artificial intelligence. Content validity is generally established through systematic literature review and expert evaluation rather than statistical analysis (DeVellis, 2017). In the present study, content validity has been strengthened by deriving the four principal constructs—Accuracy, Reliability, Robustness, and Transparency—from well-established theoretical frameworks, including Information Quality Theory, Trust Theory, Measurement Theory, Artificial Intelligence Governance Theory, and internationally recognised governance standards developed by the OECD, NIST, ISO, and the European Union. Future empirical studies may further strengthen content validity by convening multidisciplinary Delphi panels comprising experts in artificial intelligence, computer science, information systems, governance, psychometrics, ethics, healthcare, law, and organisational management. Such expert evaluation will ensure that the indicators comprehensively represent the multidimensional concept of AI trustworthiness, identifying opportunities for refinement before large-scale empirical implementation.

Closely related to content validity is face validity, which assesses whether the proposed framework appears

logically appropriate to experts, practitioners, regulators, and intended users. Although face validity does not constitute formal statistical evidence, it represents an important preliminary indicator of practical credibility because users are more likely to adopt evaluation frameworks perceived as conceptually meaningful and operationally relevant. Future pilot studies involving AI developers, organisational leaders, researchers, policymakers, and governance specialists may therefore examine whether the proposed constructs, indicators, and terminology accurately reflect their understanding of trustworthy artificial intelligence. Positive expert feedback would strengthen confidence that GAITAI possesses practical relevance in addition to theoretical coherence.

The next stage involves assessment of construct validity, which determines whether the observable indicators accurately represent the latent constructs they are intended to measure (Cronbach & Meehl, 1955). Construct validity is particularly important because Generative Artificial Intelligence Trustworthiness represents an abstract higher-order construct that cannot be observed directly. Future empirical investigations should initially employ Exploratory Factor Analysis (EFA) to determine whether the proposed indicators naturally group into the four hypothesised dimensions of Accuracy, Reliability, Robustness, and Transparency. Following exploratory analysis, Confirmatory Factor Analysis (CFA) should be conducted to evaluate the adequacy of the proposed measurement model and to determine whether the observed data support the hypothesised hierarchical structure of GAITAI. Goodness-of-fit indices, including the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardised Root Mean Square Residual (SRMR), should be examined to assess model adequacy according to internationally accepted criteria (Hair et al., 2022).

Following confirmation of the factor structure, future research should evaluate convergent validity, which examines whether indicators designed to measure the same construct exhibit strong positive relationships with one another. Convergent validity is commonly assessed through factor loadings, Average Variance Extracted (AVE), and Composite Reliability (CR). Indicators demonstrating strong standardised factor loadings (typically ≥ 0.70), AVE values exceeding 0.50, and composite reliability values above 0.70 provide evidence that the indicators adequately represent their underlying latent construct (Hair et al., 2022). Within GAITAI, strong convergent validity would indicate that indicators assigned to Accuracy, Reliability, Robustness, and Transparency collectively measure their respective dimensions consistently and coherently. Such evidence would provide important support for the conceptual architecture proposed throughout this study.

Complementing convergent validity is discriminant validity, which determines whether conceptually distinct constructs remain empirically distinguishable. Although the four dimensions of GAITAI are expected to exhibit positive correlations because they collectively contribute to AI trustworthiness, each construct should nevertheless capture a unique aspect of the broader latent variable. Future validation studies may assess discriminant validity using the Fornell-Larcker criterion, cross-loading analysis, and the Heterotrait-Monotrait Ratio (HTMT) of correlations (Hair et al., 2022). Demonstrating discriminant validity would confirm that Accuracy, Reliability, Robustness, and Transparency represent complementary rather than redundant dimensions of trustworthy artificial intelligence. Such findings would strengthen the scientific justification for maintaining a multidimensional evaluation framework rather than reducing trustworthiness to a single composite performance measure.

The reliability of GAITAI should also be evaluated through several complementary statistical procedures. Internal consistency reliability may be assessed using Cronbach's alpha, Composite Reliability, and McDonald's Omega, with values of 0.70 or higher generally indicating acceptable measurement consistency (Nunnally & Bernstein, 1994). In addition, test-retest reliability should be examined by applying the framework repeatedly to identical or equivalent large language models over time to determine whether comparable trustworthiness scores are obtained under similar evaluation conditions. Stable results would demonstrate that observed variation reflects genuine differences among AI systems rather than instability within the measurement instrument itself. Where multiple expert evaluators assess identical AI outputs, inter-rater reliability may also be examined using Intraclass Correlation Coefficients (ICC) or Cohen's kappa statistics. Collectively, these reliability assessments would establish the consistency and reproducibility of GAITAI across diverse evaluators and operational environments.

Future empirical validation should further evaluate the predictive validity and criterion-related validity of GAITAI. Predictive validity concerns the ability of the index to forecast meaningful outcomes associated with

trustworthy artificial intelligence, such as organisational adoption, user satisfaction, regulatory compliance, reduced operational risk, or improved decision quality. Criterion-related validity may be examined by comparing GAITAI scores with established benchmark systems, governance assessments, expert evaluations, or independent measures of AI performance. Although GAITAI is conceptually broader than existing benchmark frameworks, statistically significant positive relationships with recognised evaluation measures would provide additional evidence supporting the scientific credibility of the proposed index. Simultaneously, the framework should demonstrate incremental explanatory value beyond existing benchmarks by capturing governance, transparency, and behavioural dimensions not represented within traditional technical evaluations.

The final stage of validation should involve cross-contextual and cross-model validation, recognising the diversity of contemporary generative artificial intelligence systems. Future studies should apply GAITAI across multiple large language models, including both proprietary and open-source systems, as well as across different application domains such as healthcare, education, legal services, finance, public administration, scientific research, and software engineering. Cross-cultural and multilingual validation should also be considered because trustworthy artificial intelligence increasingly operates within globally diverse linguistic, regulatory, and organisational environments. Demonstrating stable measurement properties across different models, industries, and geographical contexts would substantially strengthen the generalisability and international applicability of GAITAI. Such validation would position the framework as a robust instrument suitable for global benchmarking and responsible AI governance.

Overall, the validation procedures proposed in this section provide a comprehensive roadmap for establishing the scientific robustness of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI). By integrating established principles of psychometric validation, structural equation modelling, measurement science, and composite index development, the framework offers a rigorous methodological strategy for future empirical investigation. Successful validation will demonstrate that GAITAI is capable of measuring trustworthy artificial intelligence consistently, accurately, and meaningfully across diverse large language models and application domains. The proposed validation framework strengthens the theoretical contribution of the present study, laying the empirical foundation necessary for transforming GAITAI from a conceptual framework into an internationally recognised instrument for evaluating the trustworthiness and accuracy of generative artificial intelligence.

Limitations

The principal limitation of the present study is that the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) remains a conceptual framework that has not yet undergone empirical validation. Although the proposed four-domain structure was derived from an extensive synthesis of Information Quality Theory, Trust Theory, Composite Index Theory, Artificial Intelligence Governance Theory, and internationally recognised AI governance frameworks, the psychometric properties of the index remain to be established. Consequently, the framework should be interpreted as a theoretically grounded instrument-development study rather than a validated measurement instrument.

A second limitation concerns the proposed twenty-eight observable indicators. While each indicator is supported by theoretical and governance literature, their measurement characteristics, relative contribution to their respective constructs, and cross-domain stability have not yet been empirically examined. Future research should evaluate indicator performance using exploratory factor analysis, confirmatory factor analysis, composite reliability, average variance extracted, and measurement invariance across different classes of generative AI systems.

A further limitation relates to the weighting strategy. Equal weighting is proposed as a transparent and theoretically neutral starting point because no empirical evidence currently exists to justify differential weighting among the four latent domains. Nevertheless, future empirical studies may derive statistically informed weights using structural equation modelling, principal component analysis, Delphi expert consensus, or multi-criteria decision analysis. Such investigations may demonstrate that certain dimensions contribute more strongly to overall trustworthiness within specific application domains such as healthcare, education, finance, or law.

Finally, the proposed scoring framework has not yet been evaluated across multiple large language models, prompting conditions, languages, or organisational contexts. Cross-platform validation, longitudinal assessment, and sector-specific calibration will therefore be essential before GAITAI can be recommended as a standardised international evaluation instrument.

DISCUSSION, IMPLICATIONS, AND FUTURE RESEARCH

Discussion and Implications

The rapid advancement of generative artificial intelligence (GenAI) has transformed knowledge creation, dissemination, and decision-making across healthcare, education, finance, law, scientific research, and public administration (Bommasani et al., 2021; OpenAI, 2023). Although large language models (LLMs) have demonstrated exceptional capabilities, their widespread adoption has also heightened concerns regarding hallucinations, misinformation, bias, inconsistency, transparency, explainability, accountability, and regulatory oversight (Bender et al., 2021; Weidinger et al., 2022). International organisations increasingly recognise that evaluating AI solely through computational performance or benchmark accuracy is insufficient for determining whether AI-generated outputs are trustworthy in real-world environments (European Union, 2024; NIST, 2023; OECD, 2019; UNESCO, 2021). Responding to this challenge, the present study developed the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI), conceptualising trustworthiness as a multidimensional higher-order construct comprising Accuracy, Reliability, Robustness, and Transparency. Through the integration of Information Quality Theory, Trust Theory, the Technology Acceptance Model, Measurement Theory, Composite Index Theory, Artificial Intelligence Governance Theory, and Institutional Theory, the framework extends existing benchmark-oriented approaches by incorporating behavioural, governance, organisational, and informational perspectives into a unified conceptual model.

A major contribution of GAITAI lies in its methodological advancement of AI evaluation. Existing assessment approaches primarily emphasise benchmark testing, reasoning capability, programming performance, and task-specific accuracy, providing limited attention to broader dimensions of trustworthy artificial intelligence (Bommasani et al., 2021; Liang et al., 2023). GAITAI addresses this limitation by applying latent variable measurement principles, conceptualising trustworthiness as a second-order construct measured through four theoretically grounded first-order dimensions (DeVellis, 2017; Hair et al., 2022; Nunnally & Bernstein, 1994). This higher-order measurement architecture aligns AI evaluation with established psychometric and behavioural science principles, providing a more comprehensive representation of AI quality than isolated benchmark scores alone. The framework establishes a scientifically grounded foundation for measuring AI trustworthiness within increasingly complex organisational and societal environments.

The study further advances AI evaluation through the application of Composite Index Theory, extending methodologies widely employed in economics, governance, healthcare, and environmental science to the assessment of generative artificial intelligence (OECD & Joint Research Centre, 2008). Rather than treating evaluation metrics as independent measures, GAITAI systematically aggregates twenty-eight observable indicators into the four dimensions of Accuracy, Reliability, Robustness, and Transparency before combining them into an overall trustworthiness index. This hierarchical structure enables both construct-specific diagnosis and overall trustworthiness assessment, allowing researchers and practitioners to identify strengths and weaknesses across individual dimensions while simultaneously evaluating the broader performance profile of large language models. The framework therefore bridges the longstanding gap between benchmark-oriented AI evaluation and multidimensional measurement science.

Another important methodological contribution is the explicit operationalisation of trustworthiness through clearly defined indicators and comprehensive psychometric validation procedures. Unlike many existing frameworks that discuss concepts such as transparency, accountability, fairness, or explainability in largely normative terms, GAITAI translates these constructs into measurable indicators suitable for empirical investigation (Cronbach & Meehl, 1955; DeVellis, 2017; Hair et al., 2022). The framework further proposes rigorous validation procedures, including content, construct, convergent, discriminant, criterion-related, and predictive validity, together with reliability assessment using Cronbach's alpha, Composite Reliability, McDonald's Omega, test-retest reliability, and inter-rater reliability. These recommendations strengthen

methodological transparency, enhance scientific credibility, and provide a robust foundation for future empirical validation and international standardisation.

Beyond its methodological contributions, GAITAI has significant practical implications for organisations responsible for developing, deploying, governing, and regulating artificial intelligence. As AI becomes increasingly integrated into decision-making, organisations require evaluation frameworks capable of assessing not only computational performance but also behavioural consistency, governance readiness, operational resilience, and transparency (European Union, 2024; NIST, 2023; OECD, 2019). The multidimensional structure of GAITAI enables organisations to evaluate AI systems before implementation, monitor trustworthiness throughout deployment, and identify areas requiring technical improvement or governance intervention. The framework supports enterprise risk management, organisational accountability, regulatory compliance, and evidence-based AI governance by transforming abstract governance principles into measurable indicators suitable for monitoring and continuous improvement.

The framework also offers important practical guidance for AI developers and technology companies. Contemporary AI development has largely prioritised improvements in model scale, benchmark performance, reasoning capability, and computational efficiency (Bommasani et al., 2021; OpenAI, 2023). Persistent concerns regarding hallucinations, misinformation, bias, and limited transparency demonstrate that technical excellence alone cannot establish public confidence (Bender et al., 2021; Weidinger et al., 2022). GAITAI encourages developers to adopt a broader development philosophy in which improvements in Accuracy, Reliability, Robustness, and Transparency receive equal emphasis throughout the model lifecycle. By evaluating trustworthiness across multiple complementary dimensions, developers may identify weaknesses overlooked by conventional benchmarking, thereby supporting more responsible, user-centred, and trustworthy AI innovation.

The proposed framework likewise has considerable implications for public policy and sector-specific governance. International initiatives, including the European Union Artificial Intelligence Act, the NIST Artificial Intelligence Risk Management Framework, the OECD AI Principles, and the UNESCO Recommendation on the Ethics of Artificial Intelligence, consistently emphasise transparency, accountability, robustness, and responsible governance as fundamental requirements for trustworthy AI (European Union, 2024; NIST, 2023; OECD, 2019; UNESCO, 2021). GAITAI contributes to these initiatives by providing regulators with a structured, quantitative framework capable of assessing compliance through observable indicators rather than relying exclusively on qualitative judgement. The framework is similarly applicable across healthcare, education, legal practice, finance, and scientific research, where multidimensional trustworthiness assessment can strengthen patient safety, educational quality, legal reliability, financial governance, research integrity, and organisational decision-making.

Overall, GAITAI extends well beyond a conceptual contribution to artificial intelligence research by providing a comprehensive multidimensional framework capable of supporting scientific evaluation, organisational governance, public policy, and international standardisation. By integrating behavioural theory, information quality, governance principles, psychometric measurement, and composite index methodology into a unified higher-order model, the framework bridges the longstanding divide between benchmark-based AI evaluation and broader trustworthiness assessment. GAITAI offers researchers, developers, regulators, organisational leaders, and policymakers a scientifically grounded methodology for evaluating generative artificial intelligence that is simultaneously rigorous, adaptable, transparent, and suitable for responsible AI deployment across diverse operational environments.

Future Research Agenda

The present study represents the initial stage in the development of the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) as a scientifically rigorous instrument for evaluating the trustworthiness of generative artificial intelligence. Although the framework establishes a theoretically grounded multidimensional model integrating Accuracy, Reliability, Robustness, and Transparency, its long-term contribution depends upon systematic empirical validation and refinement. Future research should therefore adopt a phased programme of instrument development consistent with internationally accepted procedures in psychometric measurement, composite index development, and structural equation modelling. Such an approach

will enable GAITAI to evolve from a conceptual framework into a robust, empirically validated instrument suitable for benchmarking large language models across diverse organisational and societal contexts.

The first phase of future research should focus on establishing the content validity of the proposed framework through expert evaluation. A multidisciplinary Delphi study involving specialists in artificial intelligence, computer science, information systems, psychometrics, governance, ethics, healthcare, education, law, and public policy would provide an appropriate mechanism for evaluating the relevance, clarity, comprehensiveness, and representativeness of the proposed twenty-eight indicators. Iterative rounds of expert review would facilitate refinement of indicator definitions, eliminate redundancy, and ensure that the framework adequately captures the multidimensional nature of trustworthy artificial intelligence across different application domains. Establishing expert consensus at this stage would strengthen the theoretical foundation of GAITAI before large-scale empirical testing.

Following expert validation, the second phase should involve pilot testing of the framework across multiple contemporary large language models representing both proprietary and open-source systems. Applying the proposed scoring protocol to models such as GPT, Claude, Gemini, Llama, and other emerging generative AI systems would enable researchers to examine the practical feasibility, usability, and consistency of the proposed indicators under real-world conditions. Pilot implementation would also provide preliminary evidence regarding the operational characteristics of the scoring framework, identify ambiguities in the evaluation procedures, and inform revisions to the measurement protocol before more comprehensive psychometric assessment.

The third phase should concentrate on evaluating the underlying dimensional structure of GAITAI through exploratory factor analysis (EFA). Using data collected from multiple AI systems and evaluation scenarios, exploratory factor analysis would determine whether the proposed twenty-eight indicators naturally cluster into the four hypothesised latent constructs of Accuracy, Reliability, Robustness, and Transparency. This stage would provide empirical evidence regarding the adequacy of the conceptual structure proposed in the present study and facilitate refinement of the measurement model by identifying poorly performing or redundant indicators. Such analyses are essential for establishing the factorial validity of newly developed multidimensional instruments.

The fourth phase should employ confirmatory factor analysis (CFA) and structural equation modelling (SEM) to validate the higher-order measurement model proposed in this study. Confirmatory analyses would evaluate the extent to which the observed data support the hypothesised hierarchical structure in which the four first-order constructs collectively define the second-order latent construct of Generative Artificial Intelligence Trustworthiness. At this stage, researchers should also assess internal consistency reliability, composite reliability, average variance extracted, discriminant validity, and alternative weighting strategies based on empirical factor loadings or structural path coefficients. Successful confirmation of the measurement model would provide robust psychometric evidence supporting the scientific validity and reliability of GAITAI.

The final phase should extend validation beyond individual models by examining the generalisability of GAITAI across different cultural, linguistic, organisational, and regulatory contexts. Cross-cultural validation and measurement invariance testing would determine whether the proposed framework measures AI trustworthiness consistently across countries, languages, industries, and application domains. In addition, comparative benchmarking studies involving healthcare, education, finance, legal services, scientific research, and public administration would establish the practical utility of GAITAI as a standardised evaluation instrument. Such international validation would provide the empirical foundation necessary for the development of globally recognised benchmarks for trustworthy generative artificial intelligence and would support the adoption of GAITAI as a scientifically rigorous framework for responsible AI governance.

Collectively, this phased research programme provides a clear pathway for advancing GAITAI from a theoretically grounded conceptual framework to a fully validated multidimensional measurement instrument. By integrating expert consensus, pilot implementation, psychometric validation, structural equation modelling, and international benchmarking, future research will strengthen both the scientific credibility and practical applicability of the framework. This progression is consistent with established approaches to instrument development in the behavioural, health, and social sciences and reinforces the present study's contribution as the foundational phase of a broader research programme dedicated to the measurement and governance of

trustworthy generative artificial intelligence.

CONCLUSION

The rapid evolution of generative artificial intelligence (GenAI) has transformed knowledge creation, decision-making, and innovation across numerous sectors, simultaneously raising important concerns regarding trust, transparency, accountability, governance, misinformation, and explainability (Bommasani et al., 2021; European Union, 2024; NIST, 2023). Although large language models demonstrate remarkable capabilities in healthcare, education, finance, law, scientific research, and public administration, benchmark performance alone cannot determine whether these systems are sufficiently trustworthy for high-stakes decision-making (Bender et al., 2021; Liang et al., 2023; OpenAI, 2023). Addressing this limitation, the present study developed the Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) as a multidimensional framework integrating Accuracy, Reliability, Robustness, and Transparency into a higher-order construct of AI trustworthiness. Drawing upon Information Quality Theory, Trust Theory, the Technology Acceptance Model, Measurement Theory, Composite Index Theory, Artificial Intelligence Governance Theory, and Institutional Theory, the framework extends traditional benchmark-oriented evaluation by incorporating behavioural, organisational, governance, informational, and psychometric perspectives into a unified conceptual model.

The study makes significant theoretical and methodological contributions to the emerging literature on trustworthy artificial intelligence. Theoretically, it conceptualises AI trustworthiness as a multidimensional latent construct in which four interrelated dimensions collectively determine the overall trustworthiness of generative AI systems rather than functioning as isolated characteristics. Methodologically, the framework operationalises this construct through twenty-eight observable indicators organised within a hierarchical measurement model, supported by established principles of psychometric measurement, composite index development, and higher-order structural modelling (DeVellis, 2017; Hair et al., 2022; OECD & Joint Research Centre, 2008). Furthermore, GAITAI translates internationally recognised governance principles into measurable evaluation criteria, thereby bridging the gap between normative AI governance frameworks and empirical assessment, providing a scientifically rigorous foundation for future validation, benchmarking, certification, and international standardisation.

Beyond its academic contributions, GAITAI offers important practical implications for organisations, policymakers, developers, regulators, and professionals responsible for implementing artificial intelligence. The framework enables organisations to evaluate AI systems before deployment, monitor trustworthiness throughout their operational lifecycle, and identify areas requiring technical improvement or governance intervention. It also supports evidence-based enterprise risk management, regulatory compliance, quality assurance, and responsible innovation by transforming abstract governance principles into measurable indicators. As governments and international organisations continue developing frameworks for trustworthy AI, GAITAI provides a common conceptual and methodological language capable of supporting organisational governance, public policy, certification programmes, and international cooperation across diverse application domains.

The study further argues that the future of artificial intelligence evaluation must move beyond a narrow emphasis on computational capability toward comprehensive multidimensional assessment. Responsible AI deployment increasingly requires evaluation methodologies capable of measuring not only technical performance but also consistency, resilience, transparency, governance readiness, and organisational accountability (European Union, 2024; NIST, 2023; OECD, 2019; UNESCO, 2021). The interdisciplinary integration of behavioural science, information systems, measurement science, organisational theory, and AI governance demonstrates that no single theoretical perspective sufficiently explains AI trustworthiness. GAITAI establishes a flexible conceptual foundation that can be refined through empirical validation, cross-cultural investigation, sector-specific adaptation, longitudinal assessment, and comparative benchmarking as AI technologies and governance requirements continue to evolve.

In conclusion, this study contends that sustainable progress in generative artificial intelligence depends not only on developing increasingly powerful models but also on establishing equally rigorous methods for evaluating whether those models deserve public trust. technical capability remains essential, trustworthy AI must also demonstrate reliability, robustness, transparency, accountability, and responsible governance to support safe and

effective deployment across society. The proposed Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) represent a theoretically grounded, methodologically rigorous, and practically relevant framework that advances the scientific evaluation of generative artificial intelligence beyond conventional benchmark metrics. By integrating multidimensional trustworthiness assessment with established theoretical and governance principles, GAITAI provides a strong foundation for future research, international standardisation, and evidence-based AI governance, contributing to the responsible development and long-term societal acceptance of generative artificial intelligence.

REFERENCES

1. Acevedo, S., LaFramboise, N., & Wong, J. (2017). *Caribbean tourism in the global marketplace*. In *Unleashing growth and strengthening resilience in the Caribbean*. International Monetary Fund.
2. Bass, B. M. (1985). *Leadership and performance beyond expectations*. Free Press.
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
4. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., et al. (2021). *On the opportunities and risks of foundation models* (CRFM Report). Stanford University.
5. Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
6. Committee on Publication Ethics. (2023). *COPE position statement: Authorship and AI tools*. https://publicationethics.org
7. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957
8. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008
9. DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.
10. DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organisational fields. *American Sociological Review, 48*(2), 147–160.
11. European Parliament and Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union.
12. European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union.
13. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1
14. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50.
15. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
16. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2022). *Multivariate data analysis* (9th ed.). Cengage Learning.
17. ISO/IEC. (2023). *ISO/IEC 42001: Artificial intelligence—Management system*. International Organisation for Standardization.
18. ISO/IEC. (2023). *ISO/IEC 42001:2023 Artificial intelligence—Management system*. International Organisation for Standardization.
19. Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review, 10*(1–2), 18–26.
20. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
21. Kaplan, R. M., & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications, and issues* (9th ed.). Cengage.
22. Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.
23. Liang, P., Bommasani, R., Lee, T., Tsipras, D., et al. (2023). *Holistic evaluation of language models (HELM).* Stanford Center for Research on Foundation Models.
24. MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioural research. *MIS Quarterly, 35*(2), 293–334.

25. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organisational trust. *Academy of Management Review, 20*(3), 709–734.
26. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances. *American Psychologist, 50*(9), 741–749.
27. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* U.S. Department of Commerce.
28. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* U.S. Department of Commerce.
29. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
30. OECD & Joint Research Centre. (2008). *Handbook on constructing composite indicators: Methodology and user guide.* OECD Publishing.
31. OECD. (2019). *OECD principles on artificial intelligence.* Organisation for Economic Co-operation and Development.
32. OpenAI. (2023). *GPT-4 technical report*. arXiv. https://arxiv.org/abs/2303.08774
33. OpenAI. (2023). *GPT-4 technical report. arXiv*. https://arxiv.org/abs/2303.08774
34. Organisation for Economic Co-operation and Development. (2019). *OECD principles on artificial intelligence*. OECD Publishing.
35. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
36. Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
37. Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of machine learning and data mining* (2nd ed.). Springer.
38. Scott, W. R. (2014). *Institutions and organizations: Ideas, interests, and identities* (4th ed.). Sage Publications.
39. Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again.* Basic Books.
40. UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.* United Nations Educational, Scientific and Cultural Organization.
41. UNESCO. (2023). *Guidance for generative AI in education and research.* UNESCO.
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
43. Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.
44. 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.
45. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., et al. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.
46. Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14*(4), 490–495.
47. WHO. (2021). *Ethics and governance of artificial intelligence for health*. World Health Organization.
48. World Economic Forum. (2024). *The future of global financial services in the age of artificial intelligence.* World Economic Forum.
49. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health.* World Health Organization.
50. Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353.

APPENDIX

APPENDIX A

Generative Artificial Intelligence Trustworthiness and Accuracy Index (GAITAI) Evaluation Instrument (Version 1.0)

Section A: General Information

Variable	Response
Date of Evaluation	_____
Evaluator ID	_____
AI Model Evaluated	_____
Version Number	_____
Organisation	_____
Domain Tested	Healthcare / Education / Law / Finance / Research / General
Number of Prompts Evaluated	_____
Evaluation Method	Human / Automated / Hybrid

Accuracy Domain

Indicator	Score (0–100)
Factual Correctness	____
Logical Consistency	____
Citation Validity	____
Numerical Accuracy	____
Instruction Adherence	____

Domain Score

Reliability Domain

Indicator	Score
Response Reproducibility	____
Output Consistency	____
Prompt Sensitivity	____
Inter-session Stability	____
Temporal Consistency	____

Domain Score

Robustness Domain

Indicator	Score
Hallucination Resistance	____
Context Preservation	____
Domain Adaptability	____
Adversarial Resilience	____
Uncertainty Recognition	____

Domain Score

Transparency Domain

Indicator	Score
Explainability	____
Evidence Attribution	____
Disclosure of Limitations	____
Interpretability	____
Auditability	____

Domain Score

Overall GAITAI Score

Interpretation

APPENDIX B

Operational Definitions of All Indicators

Code	Indicator	Definition
A1	Factual Correctness	Degree to which statements correspond with verifiable evidence.
A2	Logical Consistency	Internal coherence of reasoning.
A3	Citation Validity	Accuracy and authenticity of references.
A4	Numerical Accuracy	Correctness of calculations and quantitative information.
A5	Instruction Adherence	Extent to which user instructions are followed.
R1	Response Reproducibility	Consistency across repeated prompts.

Code	Indicator	Definition
R2	Output Consistency	Stability within conversations.
R3	Prompt Sensitivity	Resistance to unnecessary prompt variation.
R4	Inter-session Stability	Consistency across sessions.
R5	Temporal Consistency	Stability across software updates.
RB1	Hallucination Resistance	Avoidance of fabricated information.
RB2	Context Preservation	Maintenance of conversational context.
RB3	Domain Adaptability	Performance across disciplines.
RB4	Adversarial Resilience	Resistance to malicious prompts.
RB5	Uncertainty Recognition	Appropriate communication of uncertainty.
T1	Explainability	Clarity of explanations.
T2	Evidence Attribution	Proper referencing of supporting evidence.
T3	Disclosure of Limitations	Communication of model limitations.
T4	Interpretability	Ease of understanding outputs.
T5	Auditability	Ability to independently verify outputs.

APPENDIX C

Proposed Five-Point Rating Scale

Rating	Interpretation	Equivalent Score
1	Very Poor	20
2	Poor	40
3	Moderate	60
4	Good	80
5	Excellent	100

Alternatively,

Direct

0–100

continuous scoring may be used.

APPENDIX D

Sample Evaluation Worksheet

Prompt

AI Response

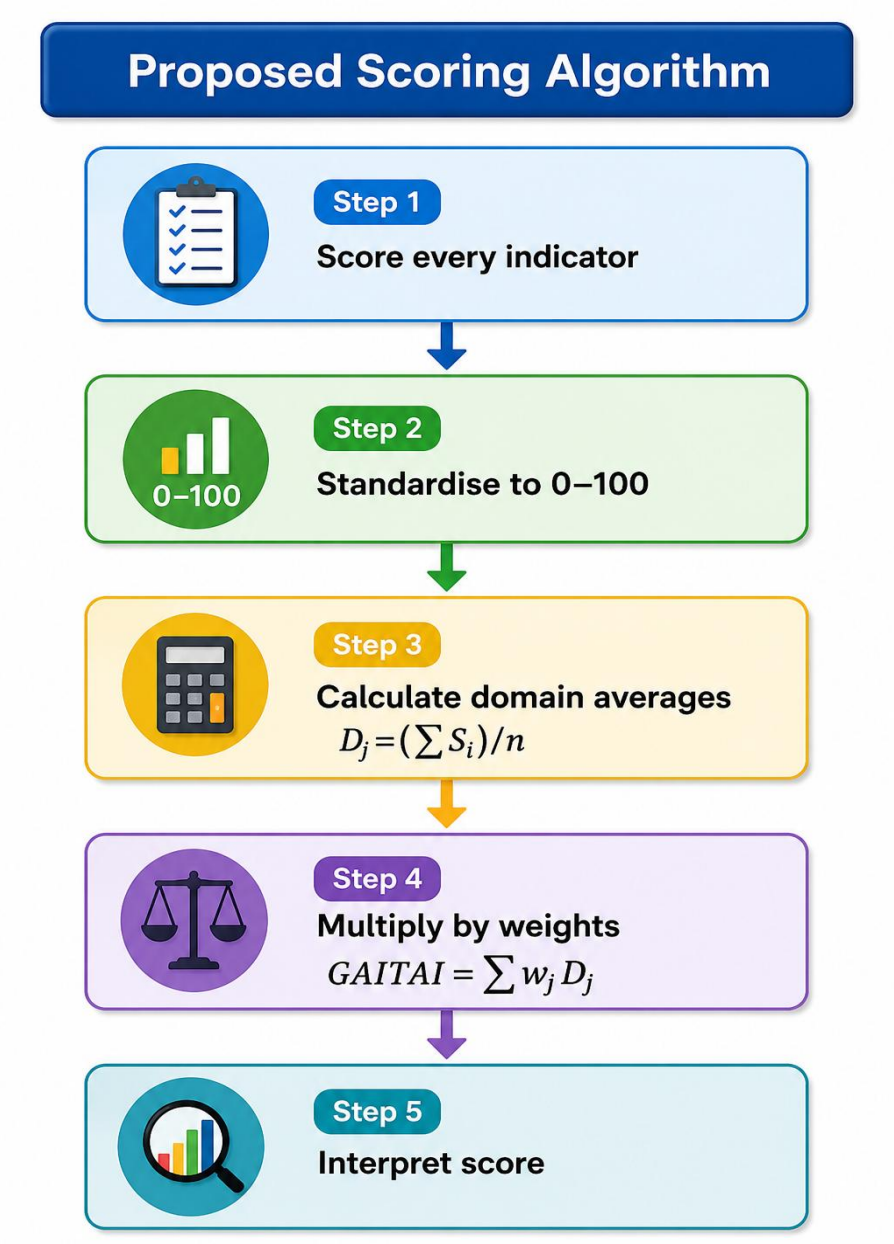
Observed Facts

Evaluation

Indicator	Score	Notes
Factual Correctness		
Logical Consistency		
Citation Validity		
Numerical Accuracy		
Instruction Adherence		

Comments

APPENDIX E



ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII July 2026

APPENDIX F

Interpretation of Score

Score	Interpretation
90–100	Exceptional Trustworthiness
80–89	Very High Trustworthiness
70–79	High Trustworthiness
60–69	Moderate Trustworthiness
50–59	Marginal Trustworthiness
<50	Low Trustworthiness

APPENDIX G

Example Evaluation

Model

GPT-X

Task

Summarise a peer-reviewed article.

Domain	Score
Accuracy	91
Reliability	88
Robustness	83
Transparency	79

Using

$$GAITAI = (0.35 \times 91) + (0.25 \times 88) + (0.25 \times 83) + (0.15 \times 79) = 86.45$$

Interpretation

Very High Trustworthiness

(This example is hypothetical and provided solely to illustrate application of the framework.)

APPENDIX H

Proposed Validation Plan

Phase	Purpose	Method
Phase 1	Content Validity	Expert panel
Phase 2	Face Validity	User feedback
Phase 3	Pilot Study	5–10 AI systems
Phase 4	Reliability	Cronbach's α, McDonald's ω, ICC
Phase 5	Construct Validity	EFA and CFA
Phase 6	Criterion Validity	Comparison with established benchmarks

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII July 2026

Phase	Purpose	Method
Phase 7	Cross-validation	Multiple domains and countries

APPENDIX I

Recommended Expert Panel Composition

Discipline	Number of Experts
Artificial Intelligence	3
Computer Science	2
Information Science	2
Psychometrics	2
Statistics	2
AI Ethics	2
Healthcare AI	1
Education Technology	1
Public Policy	1