Trustworthiness, Accuracy, and Quality in Artificial Intelligence Contribution Assessment: A Conceptual Framework for Evaluating AI–Human Authorship in Written Documents

Paul Andrew Bourne, PhD, DrPH

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

DOI: https://doi.org/10.47772/IJRISS.2026.100700496

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

ABSTRACT

The rapid adoption of generative artificial intelligence (GenAI) has fundamentally transformed academic, scientific, and professional writing by enabling increasingly sophisticated collaboration between humans and intelligent computational systems. Although numerous AI contribution assessment technologies have emerged to distinguish AI-generated from human-written text, existing approaches remain primarily focused on computational detection of linguistic patterns and provide limited evidence regarding intellectual contribution, measurement validity, transparency, or overall trustworthiness. Consequently, the theoretical foundations required to evaluate AI contribution assessment systems remain underdeveloped, limiting their suitability for high-stakes educational, scholarly, organisational, and regulatory decision-making.

The current study addresses the conceptual gap by developing the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI), an interdisciplinary theory-building framework for evaluating the trustworthiness of AI contribution assessment systems. Drawing upon Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, Human–Computer Interaction Theory, the Technology Acceptance Model, and Responsible Artificial Intelligence Governance, the framework conceptualises trustworthy AI contribution assessment as a second-order reflective latent construct represented by eight theoretically derived dimensions: Detection Accuracy, Attribution Accuracy, Reliability, Transparency, Explainability, Human Contribution Recognition, Contextual Understanding, and Measurement Validity. Rather than proposing another AI detection algorithm, HAICATI provides a multidimensional framework for evaluating whether AI contribution assessment systems are scientifically credible, transparent, contextually appropriate, and ethically responsible evidence capable of supporting informed human decision-making.

The principal contribution of this study is the reconceptualisation of AI contribution assessment from a computational classification problem to a multidimensional theory of trustworthy assessment that integrates psychometric measurement, behavioural science, information systems, organisational governance, and responsible artificial intelligence. By providing a coherent theoretical foundation for future scale development, empirical validation, independent benchmarking, and international standardisation, HAICATI advances the emerging literature on trustworthy AI and establishes a foundation for the responsible evaluation and governance of collaborative human–AI authorship.

Keywords: Generative artificial intelligence; AI contribution assessment; AI authorship; trustworthiness; information quality; psychometric measurement; responsible artificial intelligence; explainable artificial intelligence; AI governance; conceptual framework.

INTRODUCTION

The rapid advancement of generative artificial intelligence (GenAI) has fundamentally transformed the production, dissemination, and evaluation of written information across education, scientific research, journalism, business, healthcare, law, and public administration (Bommasani et al., 2021; OpenAI, 2023). Large language models (LLMs), including ChatGPT, Claude, Gemini, Llama, and other generative AI systems, are increasingly

employed to generate, revise, edit, summarise, translate, and enhance written documents with unprecedented speed and sophistication (Brown et al., 2020; OpenAI, 2023). Beyond functioning as automated text generators, these technologies have evolved into collaborative writing partners that support brainstorming, drafting, language refinement, critical editing, and content development, thereby reshaping traditional concepts of originality, authorship, intellectual contribution, and academic integrity. Consequently, educational institutions, publishers, employers, governments, and regulatory agencies are increasingly challenged to determine how AI-assisted writing should be evaluated and how the respective contributions of humans and artificial intelligence should be recognised within academic, professional, and organisational contexts (Erol et al., 2025; Malik & Amjad, 2025; Wakjira et al., 2025). These developments have intensified the need for robust approaches capable of evaluating AI-assisted authorship in ways that are scientifically credible, ethically defensible, and consistent with emerging principles of responsible artificial intelligence governance.

In response to these developments, numerous AI contribution assessment technologies have been introduced to estimate whether written content contains AI-generated or AI-assisted material. Most existing systems employ machine learning and natural language processing techniques to identify statistical linguistic characteristics associated with AI-generated text, with performance typically evaluated using conventional computational metrics such as accuracy, precision, sensitivity, specificity, recall, and false-positive and false-negative rates (Liang et al., 2023; Weber-Wulff et al., 2023). Although these measures provide important information regarding algorithmic classification performance, growing empirical evidence suggests that many AI assessment systems demonstrate inconsistent performance across different writing styles, academic disciplines, languages, and successive generations of large language models (Liang et al., 2023; Weber-Wulff et al., 2023). Independent evaluations have reported substantial false-positive and false-negative classifications, limited reproducibility, and declining performance as generative AI technologies continue to evolve, raising important concerns regarding the scientific reliability and practical application of these systems in high-stakes educational, professional, and legal environments.

The limitations of existing AI contribution assessment systems extend beyond computational accuracy. Contemporary writing increasingly reflects collaborative human–AI workflows in which artificial intelligence supports idea generation, language refinement, summarisation, translation, coding, and editing, while human authors remain responsible for conceptual development, critical reasoning, interpretation of evidence, ethical judgement, and intellectual accountability. Under these circumstances, authorship is more appropriately conceptualised as a continuum of human–AI collaboration rather than a binary distinction between human-written and AI-generated text. Consequently, technologies designed to assess AI contribution should be capable of evaluating not only linguistic patterns but also the broader characteristics that determine whether their outputs are trustworthy, transparent, explainable, reliable, and capable of recognising meaningful human intellectual contribution. However, relatively little research has examined AI contribution assessment from this broader perspective, with most studies concentrating on algorithm development and detection performance rather than on the conceptual foundations necessary for evaluating the quality and trustworthiness of AI assessment systems themselves.

This limitation reflects a broader theoretical gap within the emerging literature on artificial intelligence. Although substantial advances have been made in explainable artificial intelligence, information quality, psychometric measurement, attribution research, human–computer interaction, technology acceptance, and responsible AI governance, these perspectives have largely developed independently and have not been integrated into a unified framework for evaluating AI contribution assessment technologies. Consequently, there remains no comprehensive conceptual model that systematically defines the dimensions of trustworthy AI contribution assessment or provides a theoretically grounded basis for evaluating whether existing systems produce scientifically valid, transparent, reliable, and ethically defensible assessments. Addressing this gap is particularly important given the increasing reliance on AI assessment technologies in decisions affecting academic integrity, publication ethics, professional certification, organisational governance, and public trust.

The present study addresses this gap through the development of the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI), an interdisciplinary conceptual framework designed to evaluate the trustworthiness, accuracy, and quality of AI contribution assessment systems. Drawing upon Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, Human–Computer Interaction Theory,

the Technology Acceptance Model, and Responsible Artificial Intelligence Governance, the framework conceptualises Trustworthy AI Contribution Assessment as a multidimensional higher-order construct reflected by eight first-order dimensions: Detection Accuracy, Attribution Accuracy, Reliability, Transparency, Explainability, Human Contribution Recognition, Contextual Understanding, and Measurement Validity. Rather than proposing another AI-detection algorithm, HAICATI provides a theoretically grounded framework for evaluating the scientific foundations of systems that assess human and artificial intelligence contributions within written documents. The framework is intended to guide future instrument development, empirical validation, governance frameworks, and international standardisation while supporting more transparent, accountable, and scientifically rigorous approaches to collaborative human–AI authorship assessment. By shifting scholarly attention from the detection of AI-generated text to the trustworthiness of AI contribution assessment itself, this study contributes to the growing literature on generative artificial intelligence, information systems, measurement science, authorship attribution, and responsible AI governance.

LITERATURE REVIEW

The Evolution of Generative AI and Human–AI Collaborative Writing

The emergence of generative artificial intelligence (GenAI) has fundamentally transformed written communication by shifting artificial intelligence from a passive writing aid to an active collaborator throughout the writing process. Earlier generations of writing technologies were primarily designed to assist with spelling, grammar, formatting, and basic language correction, leaving idea generation, critical reasoning, conceptual development, and intellectual synthesis almost entirely to human authors (Russell & Norvig, 2021). The introduction of transformer architectures and large language models (LLMs) has substantially expanded these capabilities, enabling AI systems to generate coherent narratives, summarise complex literature, produce computer code, draft policy documents, assist with translation, and support increasingly sophisticated reasoning tasks (Brown et al., 2020; OpenAI, 2023). Rather than functioning solely as automated text generators, contemporary LLMs have become collaborative partners that participate in multiple stages of document development, including brainstorming, drafting, editing, language refinement, and iterative revision. Consequently, the production of written work is increasingly characterised as a collaborative process in which human expertise and artificial intelligence jointly contribute to the final document, challenging long-established concepts of originality, authorship, intellectual ownership, and academic integrity (Floridi & Chiriatti, 2020).

The rapid integration of GenAI into higher education, research, publishing, business, and professional practice has prompted organisations to reconsider policies governing AI-assisted writing. Universities, publishers, and professional bodies have increasingly recognised that responsible AI use is compatible with scholarly practice provided that human authors retain responsibility for conceptualisation, interpretation, ethical decision-making, and the overall integrity of their work (Committee on Publication Ethics [COPE], 2023; UNESCO, 2023). This evolving perspective reflects a growing consensus that AI should be viewed as a tool that supports, rather than replaces, human intellectual activity. Consequently, authorship is increasingly conceptualised as existing along a continuum of human–AI collaboration rather than as a simple dichotomy between entirely human-written and entirely AI-generated text. This conceptual shift has significant implications for the evaluation of written work because assessment technologies must increasingly distinguish varying levels of AI assistance within collaborative writing environments rather than merely classify documents into binary categories.

The evolution of collaborative human–AI writing has also challenged conventional assumptions regarding how intellectual contribution should be recognised and evaluated. Traditional approaches to authorship have generally assumed that linguistic production reflects underlying human cognition and creative effort. However, modern writing frequently involves iterative interactions in which AI contributes to language generation, structural organisation, summarisation, editing, and computational support while human authors provide disciplinary expertise, critical reasoning, contextual interpretation, methodological judgement, and ethical oversight. Under these circumstances, evaluating authorship requires consideration of both the observable textual output and the less visible cognitive and editorial processes that shape the final document. These developments have created a growing need for assessment approaches capable of recognising the complexity of collaborative human–AI authorship while maintaining scientific credibility, fairness, and accountability.

Contemporary AI Contribution Assessment Technologies

The rapid adoption of generative artificial intelligence has stimulated the development of numerous commercial AI contribution assessment technologies designed to estimate the likelihood that written documents contain AI-generated or AI-assisted content. Prominent systems, including GPTZero, Turnitin AI Writing Detection, Copyleaks, Originality.ai, Winston AI, ZeroGPT, and Sapling, have become increasingly integrated into educational institutions, publishing workflows, and professional environments where concerns regarding academic integrity and responsible AI use continue to expand (Liang et al., 2023). Although these platforms differ in their underlying computational architectures, they share the common objective of identifying textual characteristics associated with artificial intelligence-generated language.

Most contemporary AI contribution assessment systems rely on natural language processing, machine learning, and statistical pattern recognition to analyse features such as perplexity, burstiness, lexical diversity, semantic consistency, syntactic regularity, and stylometric characteristics that distinguish AI-generated text from human writing. Rather than observing the writing process directly, these systems infer AI involvement from characteristics present within the completed document. Consequently, their outputs represent probabilistic estimates derived from linguistic analysis rather than direct measurements of cognitive activity, creative reasoning, editorial decision-making, or intellectual ownership. This distinction is particularly important because the increasing integration of AI into iterative writing workflows means that linguistic characteristics alone may not adequately reflect the relative contributions of human authors and artificial intelligence.

The continued evolution of foundation models further complicates AI contribution assessment. Successive generations of LLMs increasingly produce language that closely resembles human writing, while human authors routinely edit, revise, restructure, and integrate AI-generated content into original scholarly work. As a result, the linguistic boundary between human-written and AI-assisted text has become progressively more difficult to identify. Furthermore, most commercial assessment technologies operate using proprietary algorithms, undisclosed training datasets, and limited public documentation regarding validation procedures, calibration methods, or uncertainty estimation (European Union, 2024; NIST, 2023). Although these systems provide valuable support for identifying potential AI involvement, they should be understood as computational tools that estimate linguistic patterns associated with AI-generated text rather than instruments capable of directly measuring intellectual contribution or authorship.

Measurement Challenges in AI Contribution Assessment

The increasing adoption of AI contribution assessment technologies has highlighted the importance of evaluating these systems using established principles of scientific measurement rather than relying solely on computational performance indicators. Although contemporary AI assessment tools are frequently evaluated using metrics such as accuracy, sensitivity, specificity, precision, recall, and F1 scores, these measures primarily assess the performance of classification algorithms and do not necessarily establish whether the underlying construct of AI contribution has been measured validly or reliably (Sokolova & Lapalme, 2009). From a psychometric perspective, measurement requires that the construct of interest be conceptually defined, theoretically justified, and empirically represented through reliable and valid indicators (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022; Furr, 2022). Consequently, demonstrating high classification accuracy alone is insufficient to establish that an AI contribution assessment system accurately captures the complex phenomenon of collaborative human–AI authorship.

A fundamental challenge arises from the absence of an accepted reference standard for AI contribution. Unlike many traditional measurement contexts in which observable outcomes can be compared with objectively verified criteria, AI-assisted writing exists along a continuum of collaboration that encompasses varying levels of human input, AI assistance, editorial revision, and intellectual synthesis. Modern writing frequently involves iterative interactions in which artificial intelligence supports brainstorming, drafting, summarisation, coding, language refinement, and editing, while human authors contribute conceptual development, disciplinary expertise, critical reasoning, interpretation of evidence, and ethical judgement. Under these circumstances, establishing a definitive measure of AI contribution becomes inherently difficult because the observable text represents the cumulative product of multiple interacting processes rather than the output of a single identifiable source (DeVellis & Thorpe,

2022; Furr, 2022). Consequently, the validity of AI contribution assessment depends not only on algorithmic performance but also on whether the assessment meaningfully represents the underlying construct it purports to measure.

Reliability presents an additional challenge for existing AI contribution assessment technologies. In psychometric theory, reliability refers to the consistency and stability of measurement across repeated observations, evaluators, contexts, and time (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022). However, independent evaluations have demonstrated that AI assessment systems may generate different classifications for identical or minimally modified documents, with results varying across software platforms, language models, document formatting, writing style, and system updates (Liang et al., 2023; Weber-Wulff et al., 2023). Such variability raises important concerns regarding the dependability of assessment outcomes, particularly in educational, scientific, legal, and organisational settings where decisions may have significant academic, professional, or reputational consequences. These findings suggest that reliability should be regarded as a central dimension of trustworthy AI contribution assessment rather than merely a secondary characteristic of algorithmic performance.

Construct validity represents perhaps the most significant measurement challenge. AI contribution is frequently inferred from linguistic characteristics embedded within completed documents, yet linguistic similarity does not necessarily correspond to intellectual ownership, originality, conceptual reasoning, or scholarly responsibility. Human authors may extensively revise AI-generated text, while AI systems may assist with language refinement without contributing substantively to the conceptual development of the work. Consequently, linguistic regularities cannot be assumed to provide a direct measure of intellectual contribution or authorship (Cronbach & Meehl, 1955). This distinction underscores the importance of differentiating between the detection of AI-associated textual patterns and the measurement of AI contribution, as the two represent conceptually distinct phenomena requiring different theoretical and methodological approaches. Accordingly, trustworthy AI contribution assessment should be conceptualised as a multidimensional measurement construct encompassing psychometric quality, interpretability, contextual appropriateness, and scientific credibility rather than as a simple classification outcome.

These measurement challenges demonstrate that evaluating AI contribution assessment systems requires a broader conceptual perspective than computational performance alone. Established principles of psychometric measurement emphasise that scientific assessment depends upon the integration of reliability, validity, theoretical coherence, and meaningful construct representation (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022; Furr, 2022). Applying these principles to AI contribution assessment suggests that future evaluation frameworks should extend beyond algorithmic accuracy to incorporate multiple complementary dimensions capable of capturing the complexity of collaborative human–AI writing. This perspective provides an important theoretical foundation for the multidimensional framework proposed in the present study.

Trustworthy AI and Responsible Artificial Intelligence Governance

The rapid deployment of artificial intelligence across education, healthcare, government, business, and scientific research has elevated trustworthiness to a central principle of responsible AI development and implementation. International organisations and regulatory bodies consistently emphasise that AI systems influencing consequential decisions should be transparent, explainable, accountable, fair, robust, and subject to meaningful human oversight (European Commission, 2019; OECD, 2019; UNESCO, 2021). These principles recognise that technical performance alone is insufficient to establish confidence in AI systems, particularly when their outputs influence decisions affecting individuals, institutions, or society. Instead, trustworthiness reflects a broader socio-technical concept that encompasses both the quality of algorithmic performance and the ethical, organisational, and governance structures that support responsible AI use.

Within the context of AI contribution assessment, trustworthiness extends beyond the ability of a system to classify text accurately. Assessment technologies increasingly inform decisions concerning academic integrity, publication ethics, research governance, professional accreditation, employment, and organisational compliance. Consequently, stakeholders must have confidence that these systems produce scientifically credible evidence, communicate uncertainty appropriately, and support informed human decision-making rather than replacing professional judgement. Trust has therefore become a multidimensional concept encompassing confidence in the

consistency, transparency, fairness, and interpretability of AI assessment systems (Glikson & Woolley, 2020; Mayer et al., 1995). From this perspective, trustworthy AI contribution assessment depends not only on computational effectiveness but also on whether users can understand, evaluate, and appropriately interpret the basis for individual assessment outcomes.

Transparency and explainability represent two closely related but conceptually distinct components of trustworthy AI governance. Transparency concerns the extent to which developers disclose relevant information regarding system design, assumptions, data sources, validation procedures, intended applications, and known limitations. Explainability, in contrast, refers to the ability of an AI system to provide understandable reasons for individual classifications and to communicate how specific conclusions were reached (Adadi & Berrada, 2018). Together, these characteristics enhance user confidence by enabling independent evaluation of AI assessment processes and reducing reliance on opaque or proprietary decision-making mechanisms. As AI contribution assessment technologies become increasingly integrated into high-stakes environments, transparency and explainability are likely to become essential requirements for responsible implementation rather than optional system features.

Responsible AI governance further requires accountability, fairness, robustness, and continual adaptation. AI assessment systems should support human decision-makers while recognising the contextual complexity of collaborative human–AI writing and the limitations inherent in probabilistic inference (Floridi et al., 2018; NIST, 2023). Similarly, fairness requires that systems perform consistently across different languages, disciplines, writing styles, educational contexts, and evolving generations of large language models, thereby minimising unintended bias and inequitable outcomes (Mehrabi et al., 2021). Collectively, these governance principles reinforce the need to evaluate AI contribution assessment using a multidimensional framework that integrates technical performance with ethical responsibility, organisational accountability, and scientific measurement, thereby providing a comprehensive basis for trustworthy assessment.

Human Authorship, Attribution, and Intellectual Contribution

The emergence of generative artificial intelligence has fundamentally challenged traditional understandings of authorship by separating the production of language from the creation of knowledge. Historically, authorship has been associated with intellectual responsibility for conceptualisation, methodological design, interpretation of evidence, originality, and scholarly accountability rather than merely the production of textual content (ICMJE, 2024). Although large language models can generate fluent and contextually appropriate language through sophisticated statistical prediction, they do not possess intentionality, independent reasoning, conscious understanding, or ethical responsibility (Bender et al., 2021; Brown et al., 2020). Consequently, the generation of text should not be equated automatically with intellectual authorship, as meaningful scholarly contribution extends beyond linguistic production to include the cognitive and ethical processes that underpin knowledge creation.

Contemporary writing increasingly reflects collaborative interactions in which humans and artificial intelligence contribute differently throughout the writing process. Artificial intelligence may assist with brainstorming, summarisation, language refinement, editing, translation, coding, and structural organisation, whereas human authors remain responsible for defining research questions, interpreting evidence, exercising critical judgement, evaluating the appropriateness of AI-generated outputs, and ensuring the scientific and ethical integrity of the final work (Floridi & Chiriatti, 2020). These complementary roles illustrate that authorship is no longer appropriately viewed as a binary distinction between human-written and AI-generated text but rather as a continuum of collaborative contribution in which varying levels of human oversight and AI assistance coexist. Consequently, evaluating AI contribution requires approaches capable of recognising this complexity rather than relying solely on textual characteristics observed within completed documents.

Existing AI contribution assessment technologies primarily analyse completed textual outputs rather than the cognitive, editorial, and decision-making processes through which documents are developed. As a result, estimates of AI involvement are derived from linguistic characteristics rather than direct observation of intellectual activity or creative contribution. Although such analyses may indicate the likelihood that AI-assisted language appears within a document, they cannot determine the extent to which humans generated original ideas,

interpreted evidence, exercised scholarly judgement, or substantially revised AI-generated material before publication. Accordingly, reported percentages of AI involvement should be interpreted as probabilistic estimates based on textual analysis rather than precise measurements of intellectual ownership or authorship. Recognising this distinction is essential for ensuring that AI contribution assessment remains scientifically credible and ethically appropriate in contexts where assessment outcomes may influence academic, professional, or legal decisions.

Attribution Theory provides an important conceptual perspective for understanding these challenges because it examines how responsibility is assigned when outcomes arise from multiple interacting influences (Heider, 1958; Kelley, 1973). Within collaborative human–AI writing environments, completed documents reflect the combined effects of disciplinary expertise, prompting strategies, iterative revision, editorial refinement, institutional writing conventions, computational assistance, and human critical judgement. The individual contributions of these interacting influences are often inseparable within the final text, making precise attribution methodologically complex. Consequently, AI contribution assessment should acknowledge that attribution represents an inferential process characterised by varying degrees of uncertainty rather than a direct observation of intellectual ownership.

This perspective has important implications for the future development of AI contribution assessment systems. Rather than attempting to assign definitive proportions of human and artificial intelligence contribution, assessment technologies should communicate uncertainty transparently while recognising the collaborative nature of modern writing. Such an approach is more consistent with established theories of authorship, attribution, and scholarly responsibility and provides a stronger conceptual foundation for evaluating AI contribution within increasingly complex writing environments. These considerations further reinforce the need for an interdisciplinary framework capable of integrating measurement science, behavioural theory, governance principles, and information quality into a comprehensive model of trustworthy AI contribution assessment.

Knowledge Gaps and the Need for an Integrated Conceptual Framework

The literature demonstrates that significant progress has been made in the development of AI-generated text detection technologies and in understanding the broader implications of generative artificial intelligence for education, research, publishing, and professional practice. Existing scholarship has contributed valuable knowledge regarding computational detection methods, psychometric measurement, explainable artificial intelligence, information quality, human–computer interaction, attribution processes, technology acceptance, and responsible AI governance. Collectively, these contributions have substantially advanced understanding of individual aspects of AI contribution assessment and have highlighted the importance of ensuring that AI technologies operate responsibly, transparently, and fairly within high-stakes environments.

Despite these advances, important theoretical and methodological gaps remain. First, much of the existing literature continues to emphasise improvements in computational detection performance rather than examining whether AI contribution itself constitutes a scientifically measurable construct. Consequently, concepts such as *AI-generated*, *AI-assisted*, *AI-authored*, and *AI-influenced* writing are frequently used interchangeably despite representing distinct forms of human–AI collaboration with potentially different implications for authorship, responsibility, and assessment. This conceptual ambiguity has limited the development of consistent theoretical foundations for evaluating AI contribution across diverse educational, scientific, and professional contexts.

Second, existing assessment technologies remain predominantly output-oriented, focusing on observable linguistic characteristics while providing comparatively limited insight into the underlying cognitive, editorial, and decision-making processes that shape written work. As collaborative human–AI writing becomes increasingly sophisticated, reliance on textual analysis alone is unlikely to provide a sufficiently comprehensive representation of intellectual contribution, scholarly responsibility, or creative ownership. Furthermore, the widespread use of proprietary algorithms, limited methodological transparency, and restricted opportunities for independent validation continue to present challenges for accountability, explainability, and responsible governance (European Commission, 2019; OECD, 2019; UNESCO, 2021).

Finally, the literature remains fragmented across several academic disciplines. Computer science has concentrated primarily on algorithm development and computational performance; psychology has contributed

theories of attribution and trust; information systems research has examined technology adoption and information quality; measurement science has established principles of reliability and validity; and governance research has emphasised transparency, accountability, fairness, and human oversight. Although each perspective provides valuable insights, relatively little work has integrated these complementary domains into a unified theoretical framework capable of evaluating the overall trustworthiness of AI contribution assessment systems.

These gaps provide the theoretical justification for the development of the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI). Rather than proposing another AI detection methodology, the present study synthesises established theories from information quality, psychometric measurement, attribution, trust, human–computer interaction, technology acceptance, and responsible artificial intelligence governance to conceptualise trustworthy AI contribution assessment as a multidimensional higher-order construct. This interdisciplinary perspective provides the theoretical foundation for the conceptual framework presented in the following section and establishes a basis for future empirical validation, governance development, and international standardisation of AI contribution assessment technologies.

THEORETICAL FOUNDATION

Introduction

The rapid evolution of generative artificial intelligence has substantially outpaced the development of theoretical frameworks capable of explaining how AI contribution assessment systems should be evaluated. Although considerable research has focused on improving machine learning algorithms, natural language processing techniques, and computational detection performance, comparatively little attention has been devoted to the theoretical assumptions underlying systems that claim to distinguish and evaluate human and artificial intelligence contributions within written documents. Consequently, the conceptual foundations of AI contribution assessment remain fragmented across several academic disciplines, including information systems, psychology, measurement science, human–computer interaction, technology adoption, and responsible artificial intelligence governance. This fragmentation has limited the development of scientifically grounded frameworks capable of supporting trustworthy, transparent, and internationally comparable approaches to AI contribution assessment.

The multidisciplinary nature of collaborative human–AI writing necessitates a similarly multidisciplinary theoretical perspective. AI contribution assessment extends beyond the technical task of identifying linguistic characteristics associated with AI-generated text to encompass broader questions concerning information quality, measurement validity, intellectual attribution, user trust, human interaction with intelligent technologies, organisational adoption, and ethical governance. Each of these dimensions represents a distinct aspect of trustworthy AI contribution assessment that cannot be adequately explained by a single theoretical perspective. Accordingly, the present study adopts an integrative approach by synthesising seven complementary theories—Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, Human–Computer Interaction Theory, the Technology Acceptance Model, and Responsible Artificial Intelligence Governance—to establish a comprehensive conceptual foundation for the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI).

Rather than functioning as competing explanations, these theories provide complementary perspectives that collectively explain how trustworthy AI contribution assessment should be conceptualised, evaluated, and implemented. Information Quality Theory explains the characteristics of high-quality assessment information; Trust Theory addresses confidence in automated decision-making; Attribution Theory explains the complexity of assigning intellectual contribution within collaborative environments; Measurement Theory establishes the psychometric principles required for scientific assessment; Human–Computer Interaction Theory explains collaborative relationships between humans and intelligent systems; the Technology Acceptance Model provides insight into organisational adoption and acceptance; and Responsible Artificial Intelligence Governance defines the ethical and regulatory principles necessary for responsible implementation. Collectively, these theories provide a coherent interdisciplinary foundation that supports the central proposition of this study: that trustworthy AI contribution assessment is a multidimensional phenomenon requiring the integration of technical, behavioural, psychometric, informational, organisational, and governance perspectives rather than evaluation based solely on algorithmic performance.

Rationale for Selecting the Seven Theoretical Perspectives

The selection of the seven theoretical perspectives was guided by the conceptual framework development process described in the methodology. During the structured synthesis of the multidisciplinary literature, recurring concepts associated with trustworthy AI contribution assessment were identified and grouped into broader theoretical domains. These concepts consistently reflected issues relating to information quality, trust, attribution of responsibility, scientific measurement, human interaction with intelligent technologies, organisational acceptance, and responsible governance. Rather than selecting theories independently, each was retained because it addressed a specific conceptual domain that was not adequately explained by the others. Collectively, the seven theories provide complementary explanatory mechanisms that encompass the informational, behavioural, psychometric, technological, organisational, and ethical dimensions required to evaluate AI contribution assessment comprehensively.

Information Quality Theory was selected because AI contribution assessment systems fundamentally function as information-producing systems whose outputs must support evidence-based decision-making. Trust Theory explains why users accept or reject assessment outcomes and therefore provides the behavioural foundation for confidence in AI-assisted decision-making. Attribution Theory contributes an understanding of how responsibility and intellectual contribution are assigned within collaborative human–AI environments where multiple interacting influences shape the final document. Measurement Theory establishes the scientific principles of reliability, validity, and construct representation that are essential for developing trustworthy assessment instruments. Human–Computer Interaction Theory explains how humans interact with AI throughout the writing process and highlights the importance of recognising collaborative workflows rather than treating AI as an independent author. The Technology Acceptance Model provides a framework for understanding organisational adoption and institutional acceptance of AI contribution assessment systems, while Responsible Artificial Intelligence Governance establishes the ethical principles of transparency, accountability, explainability, fairness, robustness, and human oversight that increasingly underpin international AI policy and regulation.

The integration of these seven theories reflects the interdisciplinary nature of trustworthy AI contribution assessment. Each theory contributes a unique perspective while reinforcing the others, thereby providing a comprehensive explanation of why trustworthiness cannot be reduced to computational detection accuracy alone. This theoretical synthesis forms the basis for the development of HAICATI as a multidimensional higher-order construct capable of guiding future empirical validation, governance development, and international standardisation.

Table 1. Mapping of Theoretical Foundations to HAICATI Dimensions

Theoretical Perspective	Primary Contribution to HAICATI	HAICATI Dimensions Primarily Supported
Information Quality Theory (Wang & Strong, 1996)	Explains the quality, usefulness, relevance, completeness, and interpretability of assessment information.	Detection Accuracy, Transparency, Explainability, Contextual Understanding, Measurement Validity
Trust Theory (Mayer et al., 1995; Glikson & Woolley, 2020)	Explains confidence in AI systems and the factors influencing acceptance of automated decisions.	Reliability, Transparency, Explainability
Attribution Theory (Heider, 1958; Kelley, 1973)	Explains how responsibility and intellectual contribution are assigned within collaborative human–AI environments.	Attribution Accuracy, Human Contribution Recognition, Contextual Understanding

Theoretical Perspective	Primary Contribution to HAICATI	HAICATI Dimensions Primarily Supported
Measurement Theory (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022; Furr, 2022)	Provides the psychometric foundation for scientific measurement, reliability, and construct validity.	Detection Accuracy, Reliability, Measurement Validity
Human–Computer Interaction Theory	Explains collaborative interactions between humans and intelligent technologies during document creation.	Human Contribution Recognition, Contextual Understanding, Explainability
Technology Acceptance Model (Davis, 1989)	Explains organisational adoption and user acceptance of AI contribution assessment technologies.	Transparency, Explainability, Reliability
Responsible Artificial Intelligence Governance (European Commission, 2019; OECD, 2019; UNESCO, 2021; NIST, 2023)	Provides ethical, regulatory, and governance principles for responsible AI implementation.	**Supports all eight HAICATI dimensions**, particularly Transparency, Explainability, Measurement Validity, and Human Contribution Recognition

Information Quality Theory

Information Quality Theory provides one of the principal theoretical foundations for the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI) because AI contribution assessment systems function fundamentally as information-producing systems rather than merely computational classifiers. Their primary purpose is not simply to identify linguistic characteristics associated with artificial intelligence but to generate information that supports decisions concerning authorship, originality, academic integrity, publication ethics, professional accountability, and organisational governance. Consequently, the value of these systems depends not only on computational performance but also on whether the information they produce is accurate, reliable, relevant, interpretable, complete, credible, and useful for informed decision-making (Wang & Strong, 1996). From this perspective, AI contribution assessment should be evaluated according to the quality of the information it provides rather than solely according to statistical measures of algorithmic performance.

Information Quality Theory proposes that high-quality information possesses four complementary dimensions: intrinsic quality, contextual quality, representational quality, and accessibility (Wang & Strong, 1996). Intrinsic quality concerns the inherent accuracy, credibility, objectivity, and trustworthiness of information independent of its use. Contextual quality evaluates whether information is relevant, timely, complete, and appropriate for the specific decision-making context. Representational quality focuses on clarity, consistency, interpretability, and ease of understanding, while accessibility addresses the extent to which information is available and usable by intended users. Collectively, these dimensions provide a comprehensive framework for evaluating whether information generated by AI contribution assessment systems is sufficiently robust to support consequential academic, scientific, organisational, and legal decisions.

These principles have direct relevance for AI contribution assessment because assessment outputs increasingly influence decisions that extend well beyond simple text classification. Universities, journal editors, employers, funding agencies, professional bodies, and regulatory authorities require evidence that is scientifically credible and capable of supporting transparent, equitable, and defensible decisions regarding AI-assisted writing. Numerical probability scores or binary classifications alone provide limited guidance unless accompanied by meaningful explanations regarding how conclusions were reached, the degree of uncertainty surrounding those conclusions, and the contextual factors that may influence interpretation. Assessment outputs that lack

explanatory information or fail to acknowledge methodological limitations may encourage unwarranted confidence in automated classifications, potentially leading to inappropriate academic sanctions, publication decisions, or professional consequences despite substantial uncertainty regarding the true extent of AI contribution.

Information Quality Theory also emphasises that the usefulness of information depends upon its capacity to facilitate informed human judgement rather than replace it (Wang & Strong, 1996). Within AI contribution assessment, this principle is particularly important because collaborative human–AI writing represents a complex and evolving process that cannot be fully characterised through statistical analysis of completed documents alone. Assessment systems should therefore assist decision-makers by presenting evidence in a manner that is transparent, interpretable, contextually relevant, and appropriately qualified, enabling human reviewers to integrate AI-generated evidence with disciplinary expertise, institutional policies, and professional judgement. This perspective aligns closely with emerging principles of responsible artificial intelligence, which emphasise that AI should augment rather than replace human decision-making.

Within HAICATI, Information Quality Theory provides the principal theoretical foundation for Detection Accuracy, Transparency, Explainability, Contextual Understanding, and Measurement Validity. Detection Accuracy reflects the intrinsic quality of assessment information by evaluating whether system outputs accurately represent the phenomenon under investigation. Transparency and Explainability correspond to representational quality because they determine whether assessment processes and conclusions are communicated clearly and meaningfully to users. Contextual Understanding reflects contextual information quality by recognising that AI contribution must be interpreted within the specific circumstances of document creation, disciplinary conventions, and collaborative writing practices. Measurement Validity further ensures that assessment outputs genuinely represent AI contribution rather than merely identifying statistical linguistic regularities. Collectively, these dimensions reinforce the proposition that trustworthy AI contribution assessment depends fundamentally upon the production of high-quality information capable of supporting scientifically rigorous, transparent, and evidence-based decision-making.

Accordingly, Information Quality Theory serves as one of the foundational pillars of HAICATI by shifting attention from computational performance alone to the broader quality, usefulness, interpretability, and credibility of assessment information. This perspective establishes that trustworthy AI contribution assessment should be evaluated not only according to whether systems correctly classify text but also according to whether the information generated is sufficiently accurate, transparent, contextually appropriate, and scientifically meaningful to support responsible human judgement across educational, scientific, professional, and regulatory settings.

Trust Theory

Trust Theory provides the behavioural foundation for understanding why individuals and organisations accept, question, or reject decisions generated by AI contribution assessment systems. As AI increasingly influences decisions concerning academic integrity, publication ethics, employment, professional accreditation, and organisational governance, users must be confident that these technologies produce evidence that is scientifically credible, ethically defensible, and appropriate for the contexts in which they are applied. Trust therefore represents more than confidence in computational performance; it reflects a broader judgement regarding whether an AI system behaves consistently, transparently, reliably, and in ways that align with users' expectations and institutional values (Mayer et al., 1995; Glikson & Woolley, 2020).

Mayer et al. (1995) conceptualised trust as the willingness of individuals to become vulnerable to the actions of another party based on perceptions of ability, benevolence, and integrity. Although originally developed to explain interpersonal trust, these principles have been widely applied to human interactions with intelligent technologies because users frequently evaluate AI systems according to similar criteria. Ability reflects perceptions that a system performs competently and accurately; benevolence relates to confidence that the system operates in ways that promote users' interests rather than causing harm; and integrity concerns adherence to accepted ethical principles, fairness, and consistency. Together, these characteristics influence whether users consider AI-generated outputs sufficiently trustworthy to support consequential decisions.

Subsequent research has extended Trust Theory to artificial intelligence by demonstrating that confidence in AI systems depends not only on technical performance but also on transparency, explainability, consistency, predictability, and users' understanding of how automated decisions are produced (Glikson & Woolley, 2020). Trust in AI therefore develops through repeated interactions in which systems consistently produce understandable, reliable, and contextually appropriate outputs. Conversely, unexplained variability, opaque decision-making processes, or uncertainty regarding system limitations can undermine confidence, even when computational accuracy remains relatively high. These findings are particularly relevant for AI contribution assessment because users are often expected to interpret assessment outcomes without access to information regarding proprietary algorithms, training datasets, validation procedures, or the confidence associated with individual classifications.

Within educational, scientific, and professional environments, trust is especially important because AI contribution assessment systems increasingly inform decisions with significant academic, legal, and organisational implications. Stakeholders require confidence that assessment technologies produce reliable evidence, communicate uncertainty honestly, and support rather than replace informed human judgement. Trust therefore depends upon a combination of technical competence and procedural transparency, recognising that confidence in AI cannot be established through computational performance alone. Instead, users must understand both the strengths and limitations of assessment technologies before relying upon their outputs in consequential decision-making.

Within the HAICATI framework, Trust Theory provides the principal theoretical basis for Reliability, Transparency, and Explainability. Reliability reflects users' expectations that assessment systems will produce consistent results across repeated analyses and comparable contexts. Transparency strengthens trust by ensuring that users understand the assumptions, limitations, and methodological foundations of assessment technologies, while Explainability enables users to interpret individual assessment outcomes and evaluate the reasoning underlying automated conclusions. Collectively, these dimensions reinforce the proposition that trustworthy AI contribution assessment depends upon sustained confidence in both the quality of assessment outputs and the integrity of the processes through which those outputs are generated.

Trust Theory therefore complements Information Quality Theory by demonstrating that scientifically accurate information alone is insufficient to establish trustworthy AI contribution assessment. Information must also be communicated through systems that users perceive as reliable, understandable, transparent, and ethically responsible. Integrating these behavioural and informational perspectives strengthens the theoretical foundation of HAICATI and reinforces its conceptualisation as a multidimensional framework for evaluating the trustworthiness of AI contribution assessment systems.

Attribution Theory

Attribution Theory provides an essential theoretical foundation for HAICATI because AI contribution assessment ultimately seeks to determine how intellectual responsibility should be assigned within collaborative human–artificial intelligence writing environments. Traditional AI assessment technologies infer AI involvement primarily from linguistic characteristics contained within completed documents. However, the presence of AI-associated language patterns does not necessarily indicate the extent to which artificial intelligence contributed to conceptual development, critical reasoning, methodological decision-making, or scholarly interpretation. Consequently, evaluating AI contribution requires a theoretical perspective capable of explaining how responsibility and intellectual ownership are inferred when multiple interacting influences contribute to a single outcome. Attribution Theory provides this perspective by examining the cognitive processes through which individuals assign responsibility for observed outcomes (Heider, 1958; Kelley, 1973).

Heider (1958) proposed that individuals naturally seek to explain behaviour by attributing outcomes to internal or external causes. Subsequent developments by Kelley (1973) expanded this perspective by demonstrating that attribution involves evaluating multiple sources of evidence, contextual information, consistency across situations, and the interaction of several contributing factors. Rather than assuming that a single observable characteristic provides definitive evidence of causation, Attribution Theory recognises that responsibility often emerges from complex combinations of interacting influences. This perspective is particularly relevant to AI-

assisted writing because completed documents frequently reflect contributions from human expertise, disciplinary knowledge, prompting strategies, iterative revision, editorial judgement, institutional writing conventions, and computational assistance, all of which influence the final product.

The collaborative nature of contemporary writing challenges conventional assumptions regarding authorship because linguistic production and intellectual contribution no longer necessarily originate from the same source. Artificial intelligence may generate fluent text, propose structural organisation, summarise literature, or assist with language refinement, while human authors remain responsible for conceptual development, interpretation of evidence, methodological choices, critical evaluation, and ethical accountability. Consequently, linguistic similarity to AI-generated text cannot be interpreted automatically as evidence that artificial intelligence was primarily responsible for the intellectual content of a document. Instead, attribution requires careful consideration of both observable textual characteristics and the broader collaborative processes through which knowledge is created.

Within AI contribution assessment, Attribution Theory therefore highlights the importance of distinguishing between linguistic production and intellectual contribution. Existing assessment technologies primarily evaluate completed textual outputs and infer AI involvement from statistical linguistic regularities. However, they generally cannot observe prompting strategies, iterative human revision, conceptual refinement, or the cognitive processes that ultimately determine scholarly ownership. Accordingly, AI contribution assessment should recognise that attribution represents an inferential judgement characterised by varying degrees of uncertainty rather than a direct measurement of intellectual responsibility. Communicating this uncertainty transparently is essential for preventing the overinterpretation of assessment results and supporting fair, evidence-based decision-making.

Within HAICATI, Attribution Theory provides the principal theoretical basis for Attribution Accuracy, Human Contribution Recognition, and Contextual Understanding. Attribution Accuracy reflects the extent to which assessment systems distinguish appropriately between human and AI contributions without oversimplifying collaborative writing processes. Human Contribution Recognition acknowledges that intellectual ownership extends beyond linguistic production to encompass conceptual reasoning, methodological judgement, creativity, and scholarly accountability. Contextual Understanding recognises that attribution decisions should be interpreted within the specific circumstances surrounding document development rather than inferred exclusively from observable textual characteristics. Collectively, these dimensions reinforce the proposition that trustworthy AI contribution assessment requires nuanced interpretation of collaborative human–AI authorship rather than binary classification of textual outputs.

Attribution Theory therefore strengthens HAICATI by demonstrating that trustworthy AI contribution assessment should move beyond identifying statistical linguistic characteristics toward evaluating the broader processes through which intellectual contribution is developed, interpreted, and communicated. This perspective provides an important conceptual bridge between behavioural science and measurement theory, reinforcing the need for assessment systems that recognise both the complexity and uncertainty inherent in collaborative human–AI writing.

Measurement Theory

Measurement Theory provides the psychometric foundation of HAICATI by establishing the scientific principles necessary for developing valid, reliable, and interpretable assessment instruments. Although AI contribution assessment technologies frequently report computational performance using measures such as accuracy, precision, recall, sensitivity, specificity, and F1 scores, these statistics evaluate algorithmic classification performance rather than the scientific adequacy of the construct being measured (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022; Furr, 2022). Measurement Theory therefore shifts attention from computational effectiveness to the more fundamental question of whether AI contribution represents a construct that can be defined, operationalised, measured, and interpreted in a scientifically meaningful manner.

Cronbach and Meehl (1955) emphasised that meaningful measurement requires a clearly specified theoretical construct supported by evidence of construct validity. Measurement should therefore demonstrate that observed

indicators genuinely represent the underlying phenomenon rather than unrelated or confounding characteristics. In the context of AI contribution assessment, this distinction is particularly important because linguistic regularities identified by AI detection systems may not necessarily correspond to intellectual ownership, originality, conceptual reasoning, or scholarly responsibility. Consequently, computational indicators alone cannot establish that AI contribution has been measured validly unless they demonstrate meaningful correspondence with the broader theoretical construct of collaborative human–AI authorship.

Measurement Theory further emphasises the importance of reliability, recognising that scientific measurement should produce stable and consistent results across repeated observations, different contexts, and comparable populations (DeVellis & Thorpe, 2022; Furr, 2022). Independent evaluations of AI contribution assessment technologies have demonstrated variability in classifications across software platforms, document formats, writing styles, and successive generations of large language models (Liang et al., 2023; Weber-Wulff et al., 2023). Such findings reinforce the importance of evaluating reliability as a core characteristic of trustworthy AI contribution assessment rather than simply as a desirable property of computational algorithms. Reliable systems provide greater confidence that observed differences reflect genuine variation in AI contribution rather than instability within the assessment process itself.

Measurement Theory also highlights the importance of content validity, construct validity, and criterion-related validity when developing multidimensional assessment instruments (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022; Furr, 2022). These principles are particularly relevant to HAICATI because the framework conceptualises trustworthy AI contribution assessment as a higher-order latent construct reflected through multiple complementary dimensions rather than a single observable characteristic. Future empirical research will therefore require rigorous psychometric validation using techniques such as expert content validation, exploratory factor analysis, confirmatory factor analysis, structural equation modelling, and measurement invariance testing to establish whether the proposed dimensions adequately represent the broader construct of trustworthiness across different institutional, disciplinary, linguistic, and cultural contexts.

Within HAICATI, Measurement Theory provides the principal theoretical foundation for Detection Accuracy, Reliability, and Measurement Validity, while also supporting the conceptual coherence of the higher-order framework itself. Detection Accuracy ensures that assessment systems correctly identify the phenomenon they are intended to measure, Reliability evaluates the consistency and stability of assessment outcomes, and Measurement Validity establishes that the framework genuinely represents trustworthy AI contribution assessment rather than isolated computational performance indicators. Together, these dimensions reinforce the scientific integrity of HAICATI by ensuring that future empirical validation is grounded in internationally recognised principles of psychometric measurement.

Measurement Theory therefore serves as the methodological cornerstone of the proposed framework. It establishes that trustworthy AI contribution assessment should be evaluated according to established standards of scientific measurement rather than solely through algorithmic performance metrics, thereby providing the psychometric foundation necessary for future scale development, empirical testing, and international validation of HAICATI.

Human–Computer Interaction Theory

Human–Computer Interaction (HCI) Theory provides an important theoretical foundation for HAICATI because AI contribution assessment occurs within a broader ecosystem of collaborative human–AI interaction rather than as an isolated computational process. Contemporary generative artificial intelligence has transformed writing from a purely human activity into an interactive process in which users iteratively engage with intelligent systems to generate ideas, refine language, reorganise content, summarise information, and improve document quality. Consequently, AI contribution assessment should not be conceptualised solely as the evaluation of completed textual outputs but as the assessment of collaborative processes involving continuous interaction between human cognition and computational assistance. Human–Computer Interaction Theory provides the conceptual framework for understanding these collaborative relationships and their implications for trustworthy AI contribution assessment (Norman, 2013; Shneiderman et al., 2016).

Human–Computer Interaction Theory emphasises that the effectiveness of intelligent technologies depends upon the quality of interaction between users and computer systems. Effective interaction is characterised by usability, transparency, feedback, learnability, user control, and shared understanding between human users and technological systems (Norman, 2013). These principles are particularly relevant to AI contribution assessment because users must understand how assessment systems function, what their outputs represent, and how assessment results should be interpreted within specific educational, scientific, or professional contexts. Assessment technologies that provide opaque outputs without meaningful explanations or opportunities for human interpretation may reduce users' ability to make informed decisions and undermine confidence in automated assessments.

The evolution of generative AI has further reinforced the importance of collaborative interaction by demonstrating that AI functions increasingly as an intelligent assistant rather than an autonomous author. Human users formulate prompts, evaluate generated content, revise outputs, integrate disciplinary knowledge, and determine the appropriateness of AI-generated material for specific purposes. Consequently, completed documents often represent iterative cycles of interaction rather than independent contributions from either humans or artificial intelligence. This collaborative workflow highlights the importance of recognising that AI contribution cannot be interpreted accurately without considering the broader context of human decision-making, editorial judgement, and intellectual oversight throughout the writing process.

Within AI contribution assessment, Human–Computer Interaction Theory therefore emphasises that assessment systems should support effective collaboration between automated analysis and human expertise rather than replacing professional judgement. Assessment technologies should communicate results in ways that promote understanding, facilitate interpretation, and encourage critical evaluation of AI-generated evidence. This perspective aligns closely with human-centred approaches to artificial intelligence, which advocate designing intelligent systems that enhance rather than diminish human autonomy, responsibility, and decision-making capacity.

Within HAICATI, Human–Computer Interaction Theory provides the principal theoretical foundation for Human Contribution Recognition, Contextual Understanding, and Explainability. Human Contribution Recognition acknowledges the central role of human judgement throughout collaborative writing, while Contextual Understanding recognises that AI contribution should be interpreted within the broader interaction between human expertise and computational assistance. Explainability ensures that assessment systems communicate findings in ways that enable meaningful interaction between users and AI technologies. Together, these dimensions reinforce the proposition that trustworthy AI contribution assessment depends upon supporting productive human–AI collaboration rather than merely identifying statistical characteristics of completed text.

Human–Computer Interaction Theory therefore complements Attribution Theory by shifting attention from the assignment of responsibility to the collaborative processes through which human and artificial intelligence jointly produce written work. This perspective strengthens HAICATI by recognising that trustworthy assessment should reflect the dynamic and interactive nature of contemporary writing rather than treating AI contribution as a static characteristic of completed documents.

Technology Acceptance Model

The Technology Acceptance Model (TAM) provides the organisational and behavioural foundation for understanding whether AI contribution assessment systems are likely to be adopted and used within educational institutions, research organisations, publishing environments, and professional settings. Regardless of their technical sophistication, assessment technologies cannot improve research integrity or organisational governance unless users perceive them as valuable, understandable, and appropriate for their intended purposes. Consequently, trustworthy AI contribution assessment depends not only on scientific validity but also on users' willingness to accept and integrate assessment technologies into routine decision-making processes. The Technology Acceptance Model offers a well-established theoretical framework for examining these issues (Davis, 1989).

Davis (1989) proposed that technology adoption is primarily determined by two interrelated perceptions:

perceived usefulness and perceived ease of use. Perceived usefulness reflects the extent to which users believe that a technology enhances task performance, while perceived ease of use concerns the degree to which the technology can be understood and operated without unnecessary effort. These perceptions influence users' attitudes toward technology, behavioural intentions to use it, and ultimately its successful adoption within organisational settings. The Technology Acceptance Model has subsequently been applied extensively to information systems, healthcare technologies, educational innovations, and artificial intelligence, demonstrating that successful implementation depends upon more than technical capability alone.

Within AI contribution assessment, perceived usefulness depends upon users' confidence that assessment technologies provide meaningful, accurate, and scientifically credible information capable of supporting fair and defensible decisions. Journal editors require systems that facilitate publication integrity, universities require evidence to support academic integrity policies, employers require reliable evaluations of written work, and regulatory agencies require transparent assessment procedures. Technologies that produce inconsistent results, lack methodological transparency, or provide limited explanatory information are less likely to be regarded as useful regardless of reported computational performance. Similarly, perceived ease of use depends upon whether assessment outputs are communicated clearly, interpreted consistently, and integrated effectively into existing institutional workflows.

The Technology Acceptance Model also highlights the importance of trust, transparency, and organisational readiness for successful implementation. Users are more likely to adopt AI contribution assessment technologies when they understand how systems operate, recognise their limitations, and perceive that outputs support rather than replace professional judgement. Consequently, institutional acceptance depends upon the integration of technical performance with effective communication, user education, organisational governance, and ongoing evaluation of system performance. These considerations reinforce the need for multidimensional evaluation frameworks capable of addressing both technological capability and user acceptance.

Within HAICATI, the Technology Acceptance Model primarily supports Transparency, Explainability, and Reliability. Transparent systems strengthen perceptions of usefulness by enabling users to understand assessment methodologies, while Explainability enhances ease of use by facilitating meaningful interpretation of assessment outcomes. Reliability further reinforces organisational confidence by demonstrating that assessment technologies produce consistent and dependable results across different applications and contexts. Collectively, these dimensions contribute to greater institutional acceptance of AI contribution assessment technologies and support their responsible implementation within diverse organisational environments.

The Technology Acceptance Model therefore complements Trust Theory by extending individual confidence in AI systems to the broader processes through which organisations evaluate, adopt, and integrate AI contribution assessment into institutional practice. This perspective reinforces the proposition that trustworthy assessment systems should not only be scientifically rigorous but also usable, understandable, and acceptable to the communities they are intended to serve.

Responsible Artificial Intelligence Governance

Responsible Artificial Intelligence Governance provides the overarching ethical and regulatory foundation for HAICATI by establishing the principles that should guide the design, evaluation, implementation, and oversight of AI contribution assessment systems. As artificial intelligence increasingly influences decisions affecting education, research, publishing, employment, and professional practice, governments, international organisations, and regulatory agencies have emphasised that AI systems must operate in ways that are transparent, accountable, fair, robust, explainable, and subject to meaningful human oversight (European Commission, 2019; OECD, 2019; UNESCO, 2021; NIST, 2023). These principles recognise that technical performance alone is insufficient to ensure responsible AI implementation and that trustworthy systems must also satisfy broader ethical, legal, and societal expectations.

Responsible AI Governance adopts a socio-technical perspective by recognising that artificial intelligence operates within complex organisational and institutional environments where technological decisions have significant human consequences. Within AI contribution assessment, automated classifications may influence

allegations of academic misconduct, publication decisions, employment evaluations, professional accreditation, and legal proceedings. Consequently, assessment systems should be designed and evaluated in ways that minimise bias, communicate uncertainty transparently, facilitate independent scrutiny, and preserve appropriate levels of human judgement throughout decision-making processes. Ethical governance therefore extends beyond algorithm development to encompass accountability for the decisions supported by AI-generated evidence.

International AI governance frameworks consistently identify transparency and explainability as essential characteristics of trustworthy AI. Transparency requires meaningful disclosure regarding system objectives, methodological assumptions, training data, validation procedures, intended applications, and recognised limitations. Explainability requires that AI systems communicate the reasoning underlying individual assessment outcomes in ways that are understandable to relevant stakeholders. Together, these principles enable users to evaluate the credibility of AI-generated evidence, recognise situations in which caution is warranted, and integrate automated assessments appropriately within broader decision-making processes.

Responsible AI Governance also emphasises fairness, robustness, accountability, and human oversight. Fair assessment requires consistent performance across languages, disciplines, demographic groups, writing styles, and evolving generations of large language models, thereby reducing the risk of systematic bias or discriminatory outcomes. Robustness requires that systems maintain dependable performance under diverse operating conditions, while accountability ensures that developers, organisations, and users remain responsible for the consequences of AI-assisted decisions. Human oversight reinforces the principle that AI contribution assessment should support informed professional judgement rather than replacing human responsibility for consequential decisions. Collectively, these governance principles provide the ethical architecture necessary for trustworthy AI contribution assessment.

Within HAICATI, Responsible Artificial Intelligence Governance provides an overarching theoretical foundation that informs all eight dimensions of the framework. Detection Accuracy, Reliability, Transparency, Explainability, Attribution Accuracy, Human Contribution Recognition, Contextual Understanding, and Measurement Validity each contribute to the broader objective of ensuring that AI contribution assessment systems operate responsibly, ethically, scientifically, and transparently. Unlike the preceding theories, which primarily explain specific conceptual domains, Responsible AI Governance integrates these domains within a unified framework for responsible implementation, accountability, and continuous improvement.

Accordingly, Responsible Artificial Intelligence Governance functions as the integrative ethical framework for HAICATI, ensuring that technical excellence is accompanied by scientific rigour, organisational accountability, and responsible human oversight. This perspective positions HAICATI not merely as a framework for evaluating computational performance but as a comprehensive model for promoting trustworthy AI contribution assessment across educational, scientific, professional, and regulatory contexts.

Integrated Theoretical Synthesis

The seven theoretical perspectives underpinning HAICATI collectively demonstrate that trustworthy AI contribution assessment is an inherently multidisciplinary phenomenon that cannot be explained through computational performance alone. Information Quality Theory establishes the characteristics of high-quality assessment information; Trust Theory explains confidence in automated decisions; Attribution Theory clarifies the assignment of intellectual responsibility in collaborative human–AI writing; Measurement Theory provides the psychometric principles necessary for scientific evaluation; Human–Computer Interaction Theory explains collaborative workflows between users and intelligent systems; the Technology Acceptance Model addresses organisational adoption and sustained use; and Responsible Artificial Intelligence Governance provides the ethical and regulatory principles necessary for responsible implementation. Each theory contributes a distinct conceptual perspective while addressing limitations that would remain if any single theory were considered in isolation.

The integration of these theories provides the conceptual basis for modelling trustworthy AI contribution assessment as a higher-order multidimensional construct reflected through eight complementary dimensions: Detection Accuracy, Reliability, Transparency, Explainability, Attribution Accuracy, Human Contribution

Recognition, Contextual Understanding, and Measurement Validity. Rather than representing independent attributes, these dimensions are theorised to operate collectively in shaping overall trustworthiness. For example, accurate detection without transparency may reduce user confidence, while transparent systems lacking measurement validity may fail to produce scientifically credible evidence. Similarly, robust governance depends upon reliable measurement, meaningful explanations, and appropriate recognition of human intellectual contribution. Trustworthiness therefore emerges from the interaction of multiple complementary dimensions rather than from excellence in any single characteristic.

This interdisciplinary synthesis distinguishes HAICATI from existing AI contribution assessment approaches that primarily evaluate computational classification performance. By integrating information systems theory, behavioural science, psychometric measurement, human–computer interaction, organisational adoption, and responsible AI governance, the framework provides a comprehensive theoretical foundation for evaluating AI contribution assessment systems across diverse educational, scientific, professional, and regulatory settings. The theoretical integration presented in this chapter therefore establishes the conceptual basis for the HAICATI framework developed in the following section and provides a foundation for future empirical validation through rigorous psychometric testing and cross-contextual evaluation.

METHODOLOGY

Research Design

This study employed a conceptual theory-building research design to develop a comprehensive framework for evaluating the trustworthiness of artificial intelligence (AI) contribution assessment systems. Unlike empirical studies that seek to test hypotheses using primary or secondary datasets, conceptual research aims to integrate existing knowledge, identify theoretical gaps, synthesise diverse perspectives, and generate new explanatory frameworks capable of guiding future empirical investigation (Jaakkola, 2020; Gilson & Goldberg, 2015). The objective of the present study was not to evaluate the performance of a particular AI-detection system but to develop a theoretically grounded framework capable of assessing the scientific credibility, trustworthiness, and governance of technologies that claim to distinguish human and artificial intelligence contributions within written documents. Accordingly, the study adopted an interdisciplinary conceptual synthesis approach, drawing upon established theories from information systems, behavioural science, psychometrics, attribution research, human–computer interaction, technology adoption, and responsible artificial intelligence governance.

Conceptual framework development is particularly appropriate when existing literature remains fragmented across disciplines, important constructs lack theoretical integration, or emerging technologies evolve more rapidly than corresponding theoretical explanations (Jaakkola, 2020). Although considerable research has evaluated the computational performance of AI-detection systems, relatively little attention has been devoted to whether AI contribution assessment itself represents a scientifically valid, trustworthy, and measurable construct. Existing studies have predominantly focused on improving classification accuracy, reducing false-positive and false-negative rates, or developing more sophisticated natural language processing algorithms, while comparatively little work has examined the broader theoretical foundations required to support trustworthy AI contribution assessment. This conceptual gap provided the rationale for the present investigation.

The methodological approach followed recognised principles for theory construction within information systems and management research. Existing empirical findings, theoretical perspectives, governance frameworks, and psychometric principles were systematically synthesised to identify recurring concepts associated with trustworthy AI assessment. Rather than proposing a completely new theory independent of prior scholarship, the study integrated complementary theoretical traditions into a unified conceptual model capable of explaining trustworthy AI contribution assessment as a multidimensional phenomenon. The resulting Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI) therefore represents an integrative conceptual framework intended to guide future instrument development, empirical validation, institutional implementation, and international standardisation.

Literature Search Strategy

The conceptual framework was developed through a structured review and synthesis of multidisciplinary

literature relating to artificial intelligence, authorship attribution, information quality, psychometric measurement, trust, responsible AI governance, explainable artificial intelligence, and technology acceptance. Because AI contribution assessment spans multiple academic disciplines, literature was intentionally drawn from computer science, information systems, education, psychology, measurement science, management, and AI governance to ensure comprehensive theoretical coverage.

Electronic searches were conducted using Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed, Google Scholar, ScienceDirect, and SpringerLink, thereby capturing peer-reviewed journal articles, conference proceedings, international standards, policy documents, and influential conceptual publications. Additional grey literature was consulted where appropriate, including publications from the National Institute of Standards and Technology (NIST), UNESCO, the Organisation for Economic Co-operation and Development (OECD), the European Commission, the Committee on Publication Ethics (COPE), and documentation produced by major AI assessment software developers. These documents were included because governance frameworks and technical documentation frequently contain information not yet available within the peer-reviewed literature.

Searches employed combinations of controlled vocabulary and free-text terms including "artificial intelligence detection," "AI-generated text," "AI contribution assessment," "AI authorship," "human–AI collaboration," "trustworthy AI," "responsible artificial intelligence," "explainable AI," "information quality," "measurement validity," "trust in AI," "authorship attribution," "algorithmic transparency," "AI governance," "technology acceptance," "human–computer interaction," and "psychometric measurement." Boolean operators (AND, OR) and truncation techniques were used to maximise retrieval while maintaining relevance. Reference lists of key articles were also examined manually to identify additional influential publications not captured during the electronic searches.

Priority was given to publications produced between 2015 and 2025, reflecting the rapid development of generative AI and contemporary AI governance frameworks. Earlier landmark publications were retained where they represented foundational theoretical contributions, including Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, and the Technology Acceptance Model. This strategy ensured that both classical theoretical foundations and recent developments in generative AI were incorporated into the conceptual synthesis.

Study Selection and Eligibility Criteria

Retrieved publications underwent an iterative screening process designed to identify literature directly relevant to trustworthy AI contribution assessment. Publications were eligible for inclusion when they satisfied one or more of the following criteria:

1. Examined AI-generated or AI-assisted writing, authorship attribution, or AI detection technologies.
2. Presented theoretical perspectives relevant to trust, measurement, information quality, attribution, governance, or technology adoption.
3. Evaluated AI systems in educational, professional, scientific, or organisational contexts.
4. Discussed transparency, explainability, accountability, fairness, or responsible AI governance.
5. Addressed psychometric concepts including validity, reliability, construct measurement, or measurement theory.

Studies focusing exclusively on algorithm development without discussing broader issues of trustworthiness, governance, measurement, or AI contribution were considered less informative for conceptual framework development and were therefore used selectively. Similarly, publications describing AI applications unrelated to authorship or contribution assessment were excluded from the conceptual synthesis.

Rather than attempting statistical aggregation of empirical findings, the review emphasised conceptual relevance and theoretical richness. Consistent with conceptual research methodology, inclusion decisions prioritised

explanatory value over study design hierarchy, allowing integration of empirical investigations, theoretical papers, international guidelines, standards documents, and influential policy reports.

Framework Development Procedure

The development of HAICATI followed a four-stage conceptual synthesis process.

Stage 1: Identification of Core Concepts. Relevant publications were examined to identify recurring constructs associated with trustworthy AI assessment, including detection accuracy, attribution, reliability, transparency, explainability, human oversight, contextual interpretation, fairness, accountability, and measurement validity. Similar concepts appearing under different terminologies were grouped during preliminary coding.

Stage 2: Theoretical Integration. The identified concepts were subsequently mapped against established theoretical perspectives to determine their underlying explanatory foundations. Seven complementary theories—Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, Human–Computer Interaction Theory, the Technology Acceptance Model, and Responsible Artificial Intelligence Governance—were retained because collectively they provided comprehensive coverage of the informational, behavioural, psychometric, technological, organisational, and governance dimensions required to explain trustworthy AI contribution assessment.

Stage 3: Dimension Development. Through iterative comparison of concepts across theoretical perspectives, eight multidimensional domains were identified as representing the principal characteristics of trustworthy AI contribution assessment systems. These dimensions comprised Detection Accuracy, Attribution Accuracy, Reliability, Transparency, Explainability, Human Contribution Recognition, Contextual Understanding, and Measurement Validity. Operational definitions for each construct were refined through repeated comparison with the theoretical literature to maximise conceptual distinctiveness while maintaining coherence within the overall framework.

Stage 4: Framework Integration. The final stage involved synthesising the eight dimensions into the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI). Rather than conceptualising AI assessment as a single computational outcome, HAICATI proposes that trustworthiness emerges from the interaction of multiple complementary dimensions operating simultaneously. The resulting framework was specified as a **provisional higher-order conceptual model** intended to guide future scale development, confirmatory factor analysis, structural equation modelling, and cross-cultural validation. At this stage, the framework should be interpreted as a theory-building model requiring empirical verification rather than a validated measurement instrument.

Comparative Assessment of Existing AI Contribution Assessment Systems

To demonstrate the conceptual utility of HAICATI, a comparative analysis of leading commercial AI contribution assessment systems was undertaken. The systems included Turnitin AI Writing Detection, GPTZero, Copyleaks, Originality.ai, Winston AI, and ZeroGPT because they are among the most widely recognised commercial tools used within higher education, publishing, and professional environments.

The comparison relied on publicly available evidence, including peer-reviewed evaluation studies, technical documentation, product white papers, user documentation, and policy statements. Assessment focused on whether each system publicly demonstrated evidence corresponding to the eight HAICATI dimensions rather than attempting to evaluate proprietary algorithms or independently verify vendor performance claims. Particular attention was given to publicly documented evidence relating to transparency, explainability, human contribution recognition, contextual sensitivity, reliability, and measurement validity.

Accordingly, the comparative analysis should be interpreted as a conceptual benchmarking exercise rather than an empirical performance evaluation. Because commercial AI assessment technologies employ proprietary algorithms, undisclosed training datasets, and confidential validation procedures, independent verification of internal system performance was beyond the scope of the present study. Nevertheless, the comparative analysis illustrates how HAICATI may be used as a structured framework for evaluating existing and future AI contribution assessment technologies using consistent theoretical criteria.

CONCEPTUAL FRAMEWORK OF THE HUMAN–ARTIFICIAL INTELLIGENCE CONTRIBUTION ASSESSMENT TRUSTWORTHINESS INDEX (HAICATI)

Introduction

The increasing integration of generative artificial intelligence (GenAI) into academic, scientific, professional, and organisational writing has fundamentally altered traditional conceptions of authorship, originality, and intellectual contribution. Large language models (LLMs) now participate in multiple stages of document development, including brainstorming, drafting, summarisation, editing, coding, translation, and language refinement, resulting in writing that increasingly reflects collaboration between humans and artificial intelligence rather than the work of a single author (Bommasani et al., 2021; Brown et al., 2020; Floridi & Chiriatti, 2020; OpenAI, 2023). This transformation has created substantial challenges for universities, publishers, employers, funding agencies, regulators, and professional organisations that must determine the extent to which AI has contributed to written work while simultaneously preserving academic integrity, intellectual accountability, and public trust. As AI-assisted writing becomes increasingly sophisticated and commonplace, assessment systems capable of evaluating AI contribution must themselves demonstrate scientific credibility, transparency, fairness, and trustworthiness before they can be relied upon in high-stakes decision-making.

Most existing AI assessment technologies have been developed primarily from computational and natural language processing perspectives, with emphasis placed on detecting statistical linguistic patterns associated with AI-generated text (Liang et al., 2023; Weber-Wulff et al., 2023). Consequently, evaluation has focused largely on conventional machine-learning performance metrics such as accuracy, sensitivity, specificity, precision, recall, and false-positive or false-negative rates. Although these indicators provide valuable information regarding classification performance, they offer only a partial assessment of whether an AI contribution assessment system produces scientifically valid, ethically defensible, and practically useful evidence concerning collaborative human–AI authorship. A system may demonstrate excellent classification accuracy while simultaneously lacking transparency, providing limited explanations for its conclusions, failing to recognise substantive human intellectual contribution, or producing inconsistent results across different languages, disciplines, writing styles, and successive generations of artificial intelligence. Such limitations suggest that computational performance alone is insufficient for evaluating the overall credibility of AI contribution assessment technologies.

The literature reviewed in the preceding sections demonstrates that trustworthy AI contribution assessment is inherently multidimensional. Information Quality Theory emphasises the quality, completeness, relevance, and interpretability of assessment information (Wang & Strong, 1996); Trust Theory highlights confidence in automated decision-making (Mayer et al., 1995; Glikson & Woolley, 2020); Attribution Theory explains the complexity of assigning responsibility for collaborative outcomes (Heider, 1958; Kelley, 1973); Measurement Theory establishes the psychometric principles necessary for valid scientific measurement (Cronbach & Meehl, 1955; DeVellis & Thorpe, 2022; Furr, 2022); Human–Computer Interaction Theory recognises the evolving nature of collaborative human–AI workflows; the Technology Acceptance Model explains organisational acceptance of intelligent technologies (Davis, 1989); and Responsible Artificial Intelligence Governance emphasises transparency, accountability, explainability, fairness, robustness, and meaningful human oversight (European Commission, 2019; OECD, 2019; UNESCO, 2021; NIST, 2023). Collectively, these theoretical perspectives indicate that AI contribution assessment extends far beyond the computational task of detecting machine-generated language and should instead be understood as a complex socio-technical system requiring simultaneous consideration of informational, behavioural, psychometric, technological, ethical, and governance-related factors.

Building upon the structured conceptual framework development methodology presented in Section 2, this study synthesises these complementary theoretical perspectives into the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI). Rather than proposing another AI text-detection algorithm, HAICATI provides a comprehensive conceptual framework for evaluating the trustworthiness of systems that assess AI contribution within written documents. The framework shifts scholarly attention away from the narrow question of *whether text appears to have been generated by AI* toward the broader question of *whether the assessment technology itself produces trustworthy evidence capable of supporting fair, transparent, scientifically credible, and accountable decision-making*. In doing so, HAICATI repositions AI contribution assessment from

a predominantly computational classification problem to a multidimensional measurement and governance challenge requiring integration across multiple academic disciplines.

Accordingly, HAICATI conceptualises Trustworthy AI Contribution Assessment as a provisional second-order reflective latent construct represented through eight theoretically derived first-order dimensions: Detection Accuracy, Attribution Accuracy, Reliability, Transparency, Explainability, Human Contribution Recognition, Contextual Understanding, and Measurement Validity. These dimensions emerged through the iterative synthesis of the multidisciplinary literature and the theoretical mapping process described in the methodology rather than through arbitrary selection. Each dimension represents a distinct yet complementary manifestation of the broader latent construct of trustworthiness, and together they provide a comprehensive conceptual foundation for evaluating AI contribution assessment systems across educational, scientific, governmental, legal, and professional contexts. At this stage, the framework should be regarded as a theory-building model intended to guide future scale development, confirmatory factor analysis, structural equation modelling, and international validation rather than as a fully validated measurement instrument.

The remainder of this section explains the conceptual development of HAICATI, justifies the selection of its eight dimensions, describes the theoretical relationships among the proposed constructs, and presents the higher-order conceptual model that forms the basis for future empirical investigation. By explicitly linking the framework to established theories of information quality, trust, attribution, psychometric measurement, technology acceptance, human–computer interaction, and responsible AI governance, this study provides a rigorous interdisciplinary foundation for advancing research on trustworthy AI contribution assessment and for informing the development of internationally recognised standards for evaluating collaborative human–artificial intelligence authorship.

Development of the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI)

The Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI) was developed through the structured conceptual synthesis described in the methodology section rather than through the adaptation of an existing measurement instrument. The objective was not to create another AI-detection algorithm, but to establish a theoretically grounded framework capable of evaluating whether AI contribution assessment technologies produce scientifically credible, ethically defensible, and practically useful evidence concerning collaborative human–AI authorship. Existing commercial systems have largely evolved from advances in computational linguistics and machine learning, with comparatively limited integration of behavioural science, psychometric measurement, information quality, governance theory, and organisational decision-making. Consequently, the development of HAICATI sought to bridge this interdisciplinary gap by integrating complementary theoretical perspectives into a unified conceptual model that could serve as the foundation for future empirical validation.

The framework development process began with the identification of recurring concepts within the multidisciplinary literature relating to AI-generated text detection, authorship attribution, trustworthy artificial intelligence, information quality, measurement science, explainable AI, human–computer interaction, technology acceptance, and responsible AI governance (Adadi & Berrada, 2018; Arrieta et al., 2020; European Commission, 2019; Liang et al., 2023; OECD, 2019; Wang & Strong, 1996; Weber-Wulff et al., 2023). During the initial synthesis, numerous overlapping concepts were identified, including accuracy, reliability, transparency, explainability, fairness, accountability, contextual awareness, validity, interpretability, human oversight, confidence, and attribution. Because several of these concepts represented closely related theoretical domains, an iterative process of comparison and conceptual refinement was undertaken to eliminate redundancy while preserving conceptual completeness. This process resulted in the consolidation of related concepts into eight theoretically distinct dimensions that collectively capture the multidimensional nature of trustworthy AI contribution assessment.

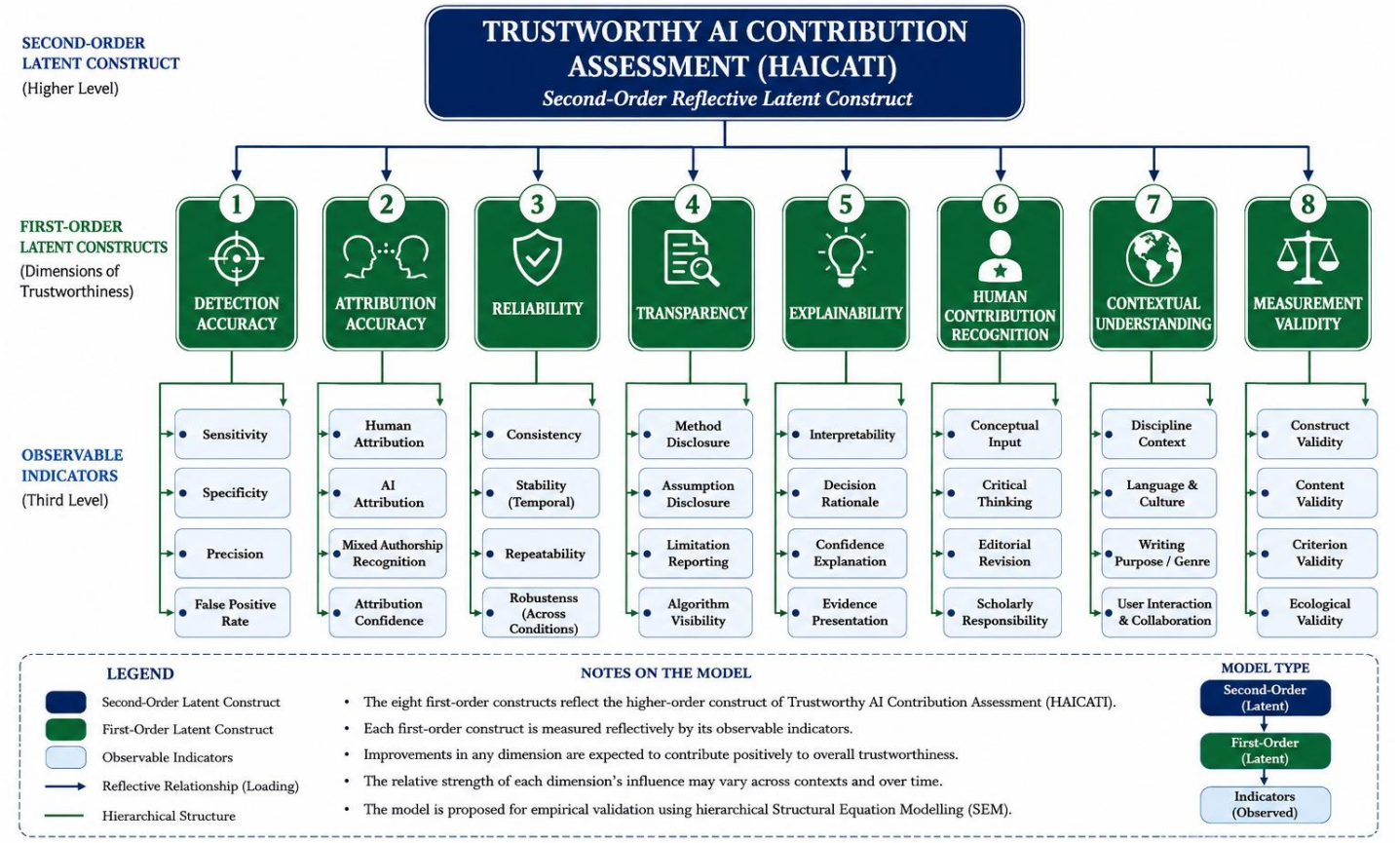


Figure 4.1. Proposed Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI) Framework.
Note: The figure illustrates a hierarchical reflective measurement model.

DISCUSSION

Introduction

The rapid expansion of generative artificial intelligence has transformed writing into an increasingly collaborative process involving both human cognition and computational assistance. Although numerous AI contribution assessment technologies have been developed to distinguish AI-generated from human-written text, existing approaches remain primarily focused on computational detection of linguistic patterns. As demonstrated throughout this study, such approaches provide useful operational capabilities but offer only limited evidence regarding intellectual contribution, measurement validity, or the broader trustworthiness of assessment systems in high-stakes educational, scientific, organisational, and regulatory contexts.

To address this conceptual gap, this study introduced the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI) as an interdisciplinary framework for evaluating AI contribution assessment systems. Rather than conceptualising assessment as a classification problem alone, HAICATI positions trustworthiness as the central evaluative construct by integrating theories from information quality, trust, attribution, psychometric measurement, human–computer interaction, technology acceptance, and responsible AI governance. The discussion therefore focuses on the framework's principal theoretical contributions and its implications for future research, policy, and practice.

Theoretical Contributions

The primary contribution of HAICATI is the reconceptualisation of AI contribution assessment as a multidimensional theory of trustworthiness rather than a problem of computational detection. Existing research has predominantly evaluated AI assessment technologies using performance metrics such as accuracy, precision, recall, sensitivity, and classification error rates. While these indicators remain essential for assessing algorithmic performance, they provide limited evidence regarding whether assessment systems generate scientifically credible, ethically defensible, and decision-relevant evaluations of collaborative human–AI authorship. By integrating perspectives from information systems, behavioural science, psychometric measurement, and

responsible AI governance, HAICATI extends existing scholarship beyond algorithm evaluation toward a broader conceptualisation of trustworthy assessment.

A second contribution is the distinction between AI detection and AI contribution assessment. Contemporary AI detection technologies generally infer AI involvement from linguistic characteristics embedded within completed documents. HAICATI demonstrates that identifying statistical language patterns should not be interpreted as equivalent to measuring intellectual contribution or scholarly responsibility. Human authors contribute conceptual development, theoretical reasoning, methodological judgement, critical evaluation, and ethical accountability that cannot be inferred reliably from textual outputs alone. Distinguishing linguistic detection from intellectual attribution therefore addresses a significant conceptual limitation within the existing AI assessment literature and provides a stronger theoretical basis for evaluating collaborative authorship.

The framework also advances measurement theory by conceptualising Trustworthy AI Contribution Assessment as a second-order reflective latent construct supported by multiple theoretically derived dimensions. Unlike existing commercial systems that have evolved primarily from computational linguistics and machine learning, HAICATI explicitly incorporates principles of construct validity, reliability, and higher-order latent variable modelling. Consequently, the framework provides a coherent theoretical foundation for future psychometric validation using Confirmatory Factor Analysis and Structural Equation Modelling, thereby moving AI contribution assessment closer to an evidence-based measurement science.

A further theoretical contribution lies in the interdisciplinary integration of seven complementary theoretical perspectives. Rather than relying on a single explanatory model, HAICATI synthesises Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, Human–Computer Interaction Theory, the Technology Acceptance Model, and Responsible Artificial Intelligence Governance into a unified conceptual framework. This synthesis demonstrates that trustworthy AI contribution assessment emerges from the interaction of informational, behavioural, psychometric, organisational, and governance processes rather than from computational performance alone. In doing so, the study provides one of the first comprehensive theoretical models for evaluating AI contribution assessment systems across multiple disciplinary contexts.

Comparison with Existing AI Contribution Assessment Systems

The comparative analysis demonstrates that HAICATI differs fundamentally from existing AI contribution assessment technologies because it evaluates the trustworthiness of assessment systems rather than the classification of textual outputs. Contemporary commercial platforms—including Turnitin AI Writing Detection, GPTZero, Copyleaks, Originality.ai, Winston AI, and ZeroGPT—primarily analyse linguistic features to estimate the likelihood of AI-generated text. These technologies provide valuable operational tools but remain focused on observable textual characteristics rather than the broader construct of trustworthy AI contribution assessment.

HAICATI instead provides a conceptual framework through which existing and future AI assessment technologies may be evaluated systematically. Rather than functioning as another detection algorithm, the framework establishes theoretical criteria for assessing whether AI contribution assessment systems produce scientifically credible, transparent, reliable, and contextually appropriate evidence capable of supporting high-stakes decision-making. This distinction shifts attention from the performance of individual algorithms to the quality and trustworthiness of the assessment systems themselves.

The comparison further illustrates that many characteristics increasingly emphasised within responsible AI governance—including transparency, explainability, measurement validity, and human oversight—remain only partially represented within existing commercial platforms. HAICATI integrates these principles into a unified evaluation framework, providing a conceptual foundation for future benchmarking, independent validation, procurement standards, and regulatory oversight. Consequently, the framework should be viewed as complementary to existing detection technologies rather than as a replacement for them.

Practical Implications

The HAICATI framework has important implications for the development, evaluation, and implementation of AI contribution assessment technologies across educational, scientific, organisational, and professional settings. By

conceptualising trustworthiness as a multidimensional construct, the framework encourages stakeholders to evaluate AI assessment systems using broader scientific and governance criteria rather than relying exclusively on reported detection accuracy. This perspective shifts emphasis from algorithmic performance alone toward the quality, credibility, and responsible use of assessment evidence in practice.

For educational institutions, HAICATI provides a conceptual basis for evaluating AI contribution assessment technologies before they are incorporated into academic integrity policies. Assessment outputs should be regarded as one component of a broader evidentiary process rather than as definitive proof of inappropriate AI use. Decisions concerning academic misconduct should therefore incorporate complementary sources of evidence, including drafts, revision histories, oral explanations, writing portfolios, and contextual information regarding the intended learning outcomes. Such an approach is more consistent with principles of procedural fairness and responsible AI implementation.

For scholarly publishers and research organisations, the framework reinforces the distinction between linguistic detection and scholarly contribution. Editorial decisions should recognise that AI-assisted writing exists along a continuum of human collaboration and that responsibility for the scientific content of manuscripts ultimately remains with human authors. Consequently, AI contribution assessment should complement existing editorial processes that emphasise transparency, disclosure, research integrity, and accountability rather than functioning as an independent determinant of publication decisions.

Software developers and AI vendors may also benefit from HAICATI by using the framework as a guide for future system development. The framework highlights the importance of improving transparency, explainability, measurement validity, contextual sensitivity, and independent validation alongside computational performance. Incorporating these broader dimensions may enhance user confidence, facilitate responsible deployment, and support greater acceptance of AI contribution assessment technologies across diverse institutional environments.

Policy and Governance Implications

Beyond its practical applications, HAICATI provides a conceptual foundation for the development of policies and governance frameworks supporting responsible AI contribution assessment. As AI-assisted writing becomes increasingly common across education, research, government, and professional practice, institutions require scientifically defensible standards for evaluating assessment technologies before they are used in consequential decision-making. HAICATI contributes to this need by translating broad principles of trustworthy AI into a structured framework capable of informing organisational policy, procurement, evaluation, and oversight.

The framework supports governance approaches that prioritise transparency, accountability, explainability, fairness, and meaningful human oversight. AI contribution assessment should function as advisory evidence supporting informed professional judgement rather than replacing human decision-making. Organisations should therefore establish policies requiring clear documentation of system capabilities, recognised limitations, validation procedures, uncertainty estimates, and appropriate opportunities for review and appeal. Such safeguards are particularly important where assessment outcomes may influence academic progression, publication decisions, employment, or regulatory compliance.

HAICATI also provides a useful reference for organisations responsible for procuring or evaluating AI assessment technologies. Rather than selecting systems primarily on the basis of advertised detection performance, procurement processes should consider broader evidence relating to scientific validation, reliability, governance, transparency, and contextual applicability. Independent audits, periodic validation studies, and ongoing monitoring should become routine components of institutional governance to ensure that assessment technologies remain appropriate as generative AI continues to evolve.

Finally, the framework contributes to international discussions concerning trustworthy artificial intelligence by aligning AI contribution assessment with principles articulated by the OECD, UNESCO, NIST, and the European Commission. Rather than introducing an alternative governance model, HAICATI operationalises these widely recognised principles within the specific context of collaborative human–AI authorship. In doing so, it offers a conceptual foundation for future international standards capable of promoting scientifically rigorous, ethically responsible, and globally comparable approaches to AI contribution assessment.

Limitations of the Study

The Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI) is a conceptual framework and therefore should be interpreted as a theory-building contribution rather than a validated measurement instrument. Although the framework is grounded in an interdisciplinary synthesis of information systems, behavioural science, psychometric measurement, human–computer interaction, technology acceptance, and responsible AI governance, its proposed constructs and relationships remain provisional until confirmed through empirical investigation. Future studies are required to evaluate the dimensionality, reliability, construct validity, and generalisability of the framework across diverse contexts.

A second limitation concerns the operationalisation of the proposed dimensions. Although the constructs are theoretically defined, observable indicators and validated measurement scales have not yet been developed. Some dimensions may demonstrate conceptual overlap, particularly Transparency and Explainability or Attribution Accuracy and Human Contribution Recognition, making future assessments of convergent and discriminant validity essential. Similarly, the specification of HAICATI as a second-order reflective construct represents a theoretical proposition that should be compared empirically with alternative reflective, formative, or hybrid measurement models.

The framework is further constrained by the rapid evolution of generative AI technologies and AI contribution assessment systems. Advances in language models, provenance technologies, multilingual capabilities, writing-process analytics, and watermarking may alter the relative importance of individual dimensions over time. Moreover, the comparative analysis relied primarily on publicly available information because proprietary algorithms, training datasets, and internal validation procedures remain inaccessible. Finally, the framework focuses exclusively on written documents and may require adaptation for multimodal outputs such as software, images, audio, video, or scientific simulations, where different indicators of human and AI contribution are likely to be required.

Future Research Directions

The immediate priority for future research is the empirical operationalisation and validation of HAICATI as a multidimensional measurement framework. Although the proposed constructs are theoretically grounded, validated indicators and measurement scales remain to be developed. Future studies should therefore generate observable items through expert consultation, stakeholder engagement, and qualitative methods involving educators, researchers, publishers, software developers, psychometricians, employers, regulators, and AI ethicists. Content validation, cognitive interviewing, and pilot testing should establish the clarity, relevance, and representativeness of proposed indicators before large-scale empirical investigation.

Subsequent research should evaluate the psychometric properties of HAICATI using established measurement approaches, including Exploratory Factor Analysis, Confirmatory Factor Analysis, and Structural Equation Modelling. Particular attention should be given to construct reliability, convergent validity, discriminant validity, measurement invariance, and alternative higher-order specifications, including reflective, formative, hybrid, and bifactor models. Criterion-related validity should also be examined by comparing HAICATI with independent indicators such as expert evaluations, user trust, organisational adoption, decision quality, and documented assessment outcomes.

Future investigations should extend beyond psychometric validation to examine the practical application of HAICATI across diverse educational, scientific, organisational, and regulatory environments. Comparative studies involving different languages, academic disciplines, professional sectors, and cultural contexts will be essential for establishing the generalisability of the framework. Longitudinal research will also be necessary because advances in generative AI, writing practices, and assessment technologies are likely to influence both the interpretation and relative importance of individual dimensions over time.

Finally, future scholarship should investigate how HAICATI can inform technology design, procurement decisions, organisational governance, and international standards for responsible AI contribution assessment. Research should explore how the framework can guide the development of next-generation assessment systems incorporating multimodal evidence, writing-process analytics, provenance technologies, and transparent

reporting mechanisms. Extensions of HAICATI to multimodal AI outputs—including software, images, audio, video, and scientific simulations—represent another important direction for investigation. Collectively, these research priorities position HAICATI as the foundation of an interdisciplinary research programme aimed at advancing the scientific validity, governance, fairness, and responsible implementation of AI contribution assessment technologies.

CONCLUSION

The rapid integration of generative artificial intelligence into writing has fundamentally transformed traditional understandings of authorship, originality, and intellectual contribution. Although substantial advances have been made in AI-generated text detection, existing assessment technologies remain primarily focused on computational classification of linguistic characteristics. This emphasis has limited the development of scientifically grounded approaches capable of evaluating whether AI contribution assessment systems themselves are trustworthy for use in educational, scientific, professional, and regulatory decision-making.

This study addressed that gap through the development of the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI), an interdisciplinary conceptual framework that reconceptualises AI contribution assessment as a multidimensional higher-order construct. Drawing upon Information Quality Theory, Trust Theory, Attribution Theory, Measurement Theory, Human–Computer Interaction Theory, the Technology Acceptance Model, and Responsible Artificial Intelligence Governance, the framework demonstrates that trustworthy AI contribution assessment extends beyond computational performance to encompass scientific measurement, transparency, explainability, contextual interpretation, recognition of human intellectual contribution, organisational acceptance, and ethical governance.

Unlike existing commercial AI detection technologies, HAICATI is not intended to function as another AI classifier. Instead, it provides a conceptual framework for evaluating the trustworthiness of AI contribution assessment systems themselves. This distinction represents the principal theoretical contribution of the study by shifting attention from the detection of AI-generated language to the broader scientific, behavioural, organisational, and governance requirements necessary for responsible AI contribution assessment. In doing so, the framework establishes a foundation for future psychometric validation, independent benchmarking, institutional governance, and international standardisation.

As generative artificial intelligence continues to evolve, assessment systems will require ongoing refinement to reflect new technologies, collaborative writing practices, and societal expectations. The framework presented here should therefore be regarded as a theory-building contribution that provides a structured foundation for future empirical investigation rather than as a completed measurement instrument. Its ultimate value will depend upon rigorous validation, cross-disciplinary evaluation, and continued collaboration among researchers, educators, publishers, software developers, policymakers, and regulators.

Ultimately, the significance of HAICATI lies not in proposing another method for identifying AI-generated text but in redefining how AI contribution assessment should be conceptualised, evaluated, and governed. By integrating perspectives from information systems, behavioural science, psychometric measurement, and responsible AI governance, the framework provides a comprehensive theoretical foundation for advancing trustworthy, transparent, and scientifically credible approaches to assessing collaborative human–artificial intelligence authorship in an increasingly AI-enabled world.

REFERENCES

1. Ajzen, I. (1991). The theory of planned behaviour. *Organisational Behaviour and Human Decision Processes, 50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J.,

Bosselut, A., Brunskill, E., & Liang, P. (2021). *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models.

4. Copyleaks. (2025). *AI detector*. https://copyleaks.com/ai-detector
5. Copyleaks. (2025). *AI text detection API documentation*. https://docs.copyleaks.com/concepts/products/ai-text-detection-api/
6. 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
7. DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research, 3*(1), 60–95. https://doi.org/10.1287/isre.3.1.60
8. DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19*(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748
9. Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Pattern Recognition, 140*, 109555.
10. Erol, G., Ergen, A., Erol, B. G., Ergen, Ş. K., Bora, T. S., Çölgeçen, A. D., Araz, B., Şahin, C., Bostancı, G., Kılıç, İ., Macit, Z. B., Sevgi, U. T., & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica, 167*(1), 214. https://doi.org/10.1007/s00701-025-06622-4
11. GPTZero. (2025). *GPTZero*. https://gptzero.me
12. Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.
13. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
14. Malik, M. A., & Amjad, A. I. (2025). AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini? *Journal of Applied Learning & Teaching, 8*(1), 91–101. https://doi.org/10.37074/jalt.2025.8.1.9
15. McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce. *Information Systems Research, 13*(3), 334–359.
16. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
17. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework
18. OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
19. OECD. (2019). *What are the OECD Principles on AI?* OECD Publishing. https://doi.org/10.1787/6ff2a1c4-en
20. Originality.ai. (2025). *Originality.ai*. https://originality.ai
21. Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioural research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.
22. Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.
23. Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.
24. Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J.-H., Ting, H., Vaithilingam, S., & Ringle, C. M. (2019). Predictive model assessment in PLS-SEM. *European Journal of Marketing, 53*(11), 2322–2347.
25. Turnitin. (2025). *AI writing detection model*. https://guides.turnitin.com/hc/en-us/articles/28294949544717-AI-writing-detection-model
26. Turnitin. (2025). *AI Writing Report*. https://guides.turnitin.com/hc/en-us/articles/22774058814093-Using-the-AI-Writing-Report
27. UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137
28. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

29. Wakjira, T. G., Tijani, I. A., Alam, M. S., Mashal, M., & Hasan, M. K. (2025). Can We Trust AI Content Detection Tools for Critical Decision-Making? *Information*, *16*(10), 904. https://doi.org/10.3390/info16100904
30. Winston AI. (2025). *Winston AI*. https://gowinston.ai
31. ZeroGPT. (2025). *ZeroGPT*. https://www.zerogpt.com

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

APPENDIX

APPENDIX A

Definitions of the HAICATI Constructs

Construct	Definition
Trustworthy AI Contribution Assessment	The extent to which an AI contribution assessment system produces scientifically valid, reliable, transparent, explainable, contextually appropriate, and ethically defensible evaluations of the respective intellectual contributions of humans and artificial intelligence within collaboratively produced written documents.
Detection Accuracy	The extent to which an assessment system correctly identifies linguistic characteristics associated with AI-generated or AI-assisted writing while minimising classification errors.
Attribution Accuracy	The extent to which an assessment system correctly identifies and attributes the intellectual contributions of human authors and artificial intelligence throughout the writing process.
Reliability	The consistency and reproducibility of assessment outcomes across repeated evaluations, document revisions, users, AI models, and operational settings.
Transparency	The extent to which an assessment system openly communicates its methodologies, assumptions, limitations, uncertainty, and decision processes.
Explainability	The extent to which users can understand the reasons underlying individual assessment outcomes.
Human Contribution Recognition	The ability of an assessment system to recognise human conceptual, methodological, analytical, and editorial contributions despite AI assistance.
Contextual Understanding	The extent to which an assessment system considers disciplinary, institutional, linguistic, cultural, and writing-context differences when interpreting AI contribution.
Measurement Validity	The extent to which the assessment system genuinely measures AI contribution rather than merely identifying statistical linguistic characteristics associated with machine-generated text.

APPENDIX B

Proposed Measurement Dimensions and Example Indicators

Detection Accuracy

- Correctly identifies AI-generated text
- Minimises false-positive classifications
- Minimises false-negative classifications
- Performs consistently across AI models
- Performs consistently after human editing

Attribution Accuracy

- Correctly identifies human intellectual contribution
- Distinguishes drafting from conceptual development
- Recognises collaborative authorship
- Represents uncertainty appropriately

Reliability

- Stable across repeated testing

- Stable across document versions
- Stable across users
- Stable across software updates

Transparency

- Methodology disclosed
- Decision rules described
- Known limitations reported
- Confidence levels communicated

Explainability

- Individual results are understandable
- Reasons for classifications are provided
- Influential document characteristics identified
- Uncertainty explained clearly

Human Contribution Recognition

- Recognises conceptual input
- Recognises methodological reasoning
- Recognises critical interpretation
- Recognises editorial judgement

Contextual Understanding

- Accounts for discipline
- Accounts for language
- Accounts for institutional policy
- Accounts for writing purpose

Measurement Validity

- Measures AI contribution rather than linguistic similarity
- Supported by empirical validation
- Demonstrates construct validity
- Demonstrates criterion validity
- Demonstrates ecological validity

APPENDIX C

Proposed Research Model

Second-order Construct

Trustworthy AI Contribution Assessment (HAICATI)

↓

Reflects:

1. Detection Accuracy
2. Attribution Accuracy
3. Reliability
4. Transparency
5. Explainability

6. Human Contribution Recognition
7. Contextual Understanding
8. Measurement Validity

Each first-order construct is reflected by multiple observable indicators suitable for future Confirmatory Factor Analysis and Structural Equation Modelling.

This study proposes a reflective second-order hierarchical latent variable model to be evaluated empirically in future research.

APPENDIX D

Proposed HAICATI Measurement Instrument

Section A. Respondent Information

Please complete the following demographic information.

1. Age (years): _____

2. Gender:

- ☐ Male
- ☐ Female
- ☐ Non-binary
- ☐ Prefer not to say
- ☐ Other

3. Country of residence:_____

4. Occupation

- ☐ Student
- ☐ Lecturer/Faculty
- ☐ Researcher
- ☐ Journal Editor
- ☐ Software Developer
- ☐ Government Employee
- ☐ Private Sector Professional
- ☐ Other _____

5. Highest academic qualification

- ☐ Bachelor's

- ☐ Master's
- ☐ Doctorate
- ☐ Professional qualification
- ☐ Other:_____

6. Primary discipline:_____
7. Years of professional experience_____ years
8. Experience using Generative AI

- ☐ None
- ☐ Less than one year
- ☐ 1–2 years
- ☐ 3–5 years
- ☐ More than 5 years

9. Experience using AI contribution assessment software

- ☐ Never
- ☐ Occasionally
- ☐ Monthly
- ☐ Weekly
- ☐ Daily

10. Which AI assessment systems have you used? (Check all that apply.)

☐ Turnitin

☐ GPTZero

☐ Copyleaks

☐ Originality.ai

☐ Winston AI

☐ ZeroGPT

☐ Other _____

Instructions to Respondents

The following statements relate to your perceptions of an Artificial Intelligence (AI) contribution assessment system (e.g., Turnitin AI Detection, GPTZero, Copyleaks, Originality.ai, Winston AI, or another comparable system). Please indicate the extent to which you agree with each statement.

Response Scale

Response	Score
Strongly Disagree	1
Disagree	2
Somewhat Disagree	3
Neutral	4
Somewhat Agree	5
Agree	6
Strongly Agree	7

Section B. Detection Accuracy (DA)

The AI contribution assessment system accurately identifies AI-generated content.

DA1. The system correctly identifies AI-generated text.

DA2. The system accurately distinguishes AI-generated from human-written content.

DA3. The system produces few false-positive results.

DA4. The system produces few false-negative results.

DA5. The system performs accurately across different writing contexts.

Section C. Attribution Accuracy (AA)

AA1. The system correctly distinguishes AI assistance from human intellectual contribution.

AA2. The system recognises when humans substantially revise AI-generated content.

AA3. The system reflects the true level of AI involvement.

AA4. The system accurately attributes intellectual ownership.

AA5. The assessment reflects actual collaborative human–AI writing.

Section D. Reliability (REL)

REL1. The assessment system produces consistent results for repeated analyses.

REL2. Similar documents receive similar assessment results.

REL3. The system remains consistent after software updates.

REL4. The system performs consistently across different document types.

REL5. I trust the stability of the assessment results.

Section E. Transparency (TR)

TR1. The system clearly explains how assessment decisions are made.

TR2. The methodology used by the system is adequately described.

TR3. The limitations of the system are clearly communicated.

TR4. The system reports uncertainty appropriately.

TR5. The assessment process is sufficiently transparent.

Section F. Explainability (EX)

EX1. The reasons for each assessment result are easy to understand.

EX2. The assessment output is meaningful to non-technical users.

EX3. The system explains why particular sections were flagged.

EX4. The explanations support informed decision-making.

EX5. I can understand how the final assessment was produced.

Section G. Human Contribution Recognition (HCR)

HCR1. The system recognises substantial human intellectual effort.

HCR2. The system recognises conceptual contributions made by human authors.

HCR3. The system recognises critical revision performed by humans.

HCR4. The system distinguishes editing from genuine intellectual contribution.

HCR5. Human creativity is adequately reflected in the assessment.

Section H. Contextual Understanding (CU)

CU1. The system considers the purpose of the document.

CU2. The system considers disciplinary differences.

CU3. The system recognises legitimate AI use in different contexts.

CU4. The system considers institutional AI policies.

CU5. The assessment reflects the writing context appropriately.

Section I. Measurement Validity (MV)

MV1. The assessment system measures AI contribution rather than merely AI-like language.

MV2. The assessment accurately reflects the construct it claims to measure.

MV3. The assessment provides meaningful evidence of AI contribution.

MV4. The system measures collaborative authorship appropriately.

MV5. Overall, the assessment represents a valid measure of AI contribution.

Scoring Procedure

Each first-order construct is calculated as the arithmetic mean of its respective items.

Construct	Number of Items
Detection Accuracy	5
Attribution Accuracy	5
Reliability	5
Transparency	5
Explainability	5
Human Contribution Recognition	5
Contextual Understanding	5
Measurement Validity	5

The overall **Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI)** score is calculated as a second-order latent construct estimated using Confirmatory Factor Analysis or Structural Equation Modelling. Simple summation of the eight dimensions is not recommended until the measurement model has been empirically validated.

Total proposed measurement items: 40 reflective indicators across eight first-order constructs.

APPENDIX E

Proposed Validation Procedure

The proposed validation of HAICATI should proceed through the following stages:

1. Literature review and construct specification.
2. Generation of an initial item pool.
3. Expert review for content validity.
4. Cognitive interviews.
5. Pilot testing.
6. Exploratory Factor Analysis.
7. Confirmatory Factor Analysis.
8. Reliability assessment (Cronbach's alpha, McDonald's omega, and Composite Reliability).
9. Convergent validity assessment using Average Variance Extracted.
10. Discriminant validity assessment using the Fornell–Larcker criterion and the Heterotrait–Monotrait ratio.
11. Structural Equation Modelling of the second-order construct.
12. Cross-cultural and measurement invariance testing.
13. Longitudinal validation.
14. Continuous refinement of the measurement instrument.

APPENDIX F

Comparative Evaluation of Existing AI Contribution Assessment Systems Using the HAICATI Framework

Table F1 demonstrates that existing commercial AI assessment technologies primarily concentrate on **Detection Accuracy**, with comparatively limited attention to **Attribution Accuracy**, **Human Contribution Recognition**, **Contextual Understanding**, and **Measurement Validity**. Several systems have improved transparency and explainability through sentence-level highlighting, writing-process analytics, or explanatory reporting. However, no currently available platform has demonstrated a comprehensive multidimensional framework for evaluating trustworthy AI contribution assessment. Consequently, HAICATI extends existing technologies by integrating

computational, behavioural, psychometric, governance, and contextual dimensions into a unified evaluation framework.

Table F1

Comparison of Commercial AI Contribution Assessment Systems Across the Eight HAICATI Dimensions

HAICATI Dimension	Turnitin AI Detection	GPTZero	Copyleaks	Originality.ai	Winston AI	ZeroGPT
Detection Accuracy	✓	✓	✓	✓	✓	✓
Attribution Accuracy	✗	Limited	✗	✗	✗	✗
Reliability	Partial evidence	Partial evidence	Partial evidence	Partial evidence	Limited evidence	Limited evidence
Transparency	Limited	Moderate	Moderate	Moderate	Limited	Limited
Explainability	Moderate	Moderate	High	Moderate	Limited	Limited
Human Contribution Recognition	✗	Limited	✗	✗	✗	✗
Contextual Understanding	Limited	Limited	Limited	Limited	Limited	Limited
Measurement Validity	Not demonstrated	Not demonstrated	Not demonstrated	Not demonstrated	Not demonstrated	Not demonstrated

Note.

✓ = Strong evidence publicly available.

Limited = Some evidence or partial functionality.

Partial evidence = Performance claims reported but insufficient psychometric evidence.

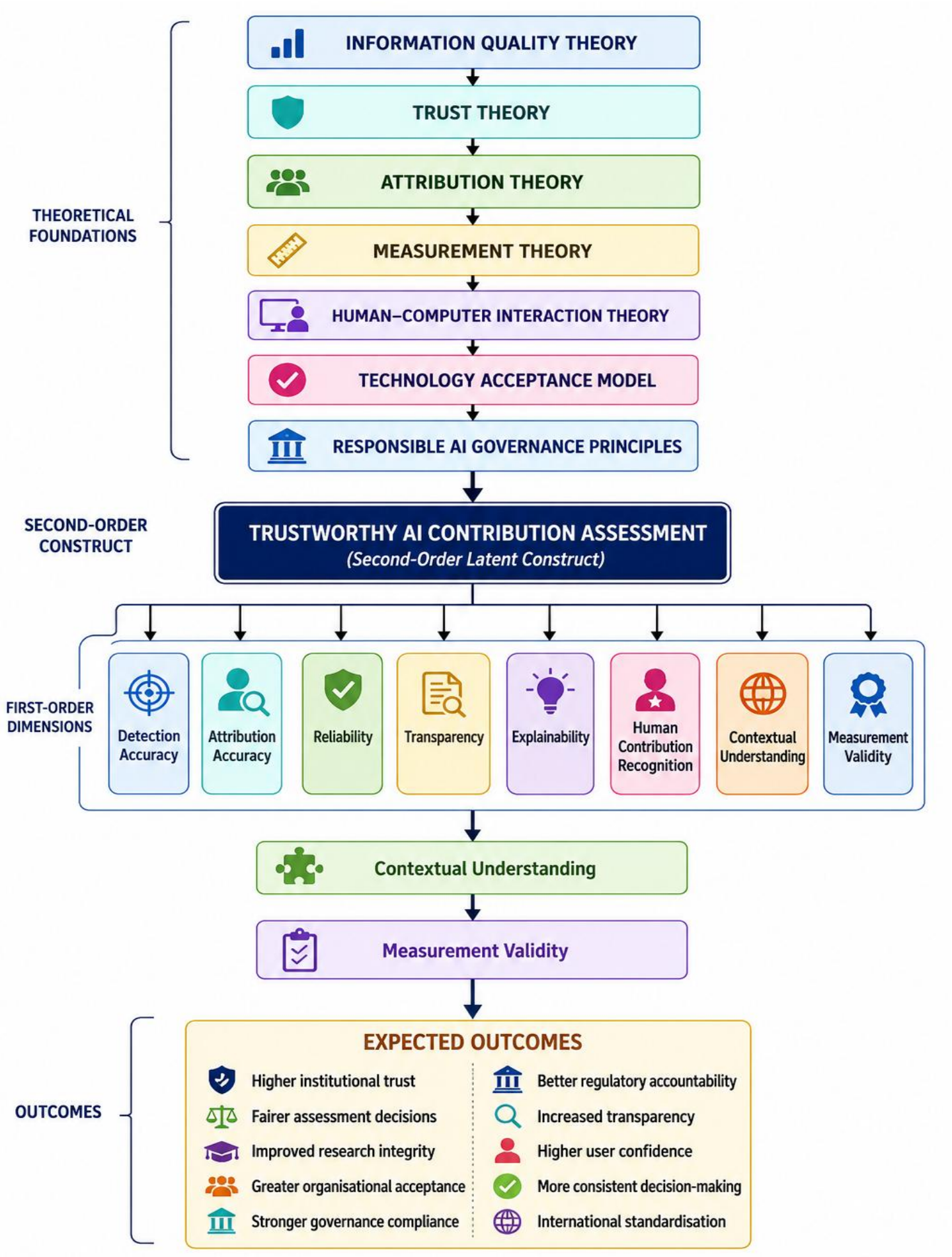
✗ = Dimension not directly addressed.

APPENDIX G

Nomological Network of the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI)

The proposed nomological network positions **Trustworthy AI Contribution Assessment** as a second-order reflective construct supported by seven complementary theoretical perspectives. These theories collectively explain the emergence of the eight first-order HAICATI dimensions. Together, the dimensions determine the overall trustworthiness of AI contribution assessment systems, which is expected to contribute to improved institutional trust, evidence-based decision-making, responsible AI governance, organisational adoption, research integrity, and international standardisation. The model therefore links theoretical foundations with anticipated practical outcomes and provides a basis for future hypothesis development and Structural Equation Modelling.

Figure G1: Proposed Nomological Network



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

Appendix H – Construct-to-Theory Mapping Matrix

HAICATI Dimension	Primary Theory	Supporting Theory	Governance Principle
Detection Accuracy	Information Quality Theory	Measurement Theory	Accuracy
Attribution Accuracy	Attribution Theory	Human–Computer Interaction Theory	Human Agency
Reliability	Measurement Theory	Trust Theory	Robustness
Transparency	Information Quality Theory	Responsible AI Governance	Transparency
Explainability	Trust Theory	Responsible AI Governance	Explainability
Human Contribution Recognition	Human–Computer Interaction Theory	Attribution Theory	Human Oversight
Contextual Understanding	Human–Computer Interaction Theory	Information Quality Theory	Fairness
Measurement Validity	Measurement Theory	Information Quality Theory	Accountability

Appendix I

Interpretation of the Human–Artificial Intelligence Contribution Assessment Trustworthiness Index (HAICATI)

Option 1: Latent Variable Interpretation (Journal Standard)

This is the approach used in most **Q1 information systems, management, psychology, and SEM-based journals**.

You **do not** calculate a simple total score.

Instead:

1. Estimate the eight first-order latent constructs using Confirmatory Factor Analysis (CFA).
2. Estimate the second-order latent construct (HAICATI) using Structural Equation Modelling (SEM).
3. The software (AMOS, LISREL, Mplus, SmartPLS, R-lavaan) estimates the HAICATI score.

In this case, the interpretation is based on **latent factor scores**, for example:

Standardised HAICATI Score	Interpretation
< -1.50 SD	Very Low Trustworthiness
-1.50 to -0.50 SD	Low Trustworthiness
-0.49 to +0.49 SD	Moderate Trustworthiness
+0.50 to +1.49 SD	High Trustworthiness
> +1.50 SD	Very High Trustworthiness

This is statistically the strongest approach because SEM accounts for measurement error and differential item loadings.

Option 2: Practical Index (Useful for Organisations)

Many reviewers like having an **interpretable index** that practitioners can calculate without SEM.

Because every item is scored **1–7**, you can calculate the mean.

Step 1

Calculate each dimension.

Example

Detection Accuracy

DA1 + DA2 + DA3 + DA4 + DA5

÷5

Suppose

6
6
5
7
6

Mean

=6.0

Repeat for all eight dimensions.

Step 2

Calculate HAICATI

Average the eight dimension scores.

Example

Dimension	**Mean**
Detection Accuracy	6.0
Attribution Accuracy	5.5
Reliability	5.8
Transparency	4.7
Explainability	5.1
Human Contribution Recognition	5.4
Contextual Understanding	5.0
Measurement Validity	5.6

Overall HAICATI

=(6.0+5.5+5.8+4.7+5.1+5.4+5.0+5.6)

÷8

=5.39

Interpretation

Because the scale ranges from **1 to 7**, the index also ranges from **1.00 to 7.00**.

I would recommend the following interpretation table.

HAICATI Score	Interpretation
6.50–7.00	Excellent Trustworthiness
5.50–6.49	High Trustworthiness
4.50–5.49	Moderate Trustworthiness
3.50–4.49	Marginal Trustworthiness
2.50–3.49	Low Trustworthiness
1.00–2.49	Very Low Trustworthiness

What each category means

Excellent Trustworthiness (6.50–7.00)

The assessment system demonstrates consistently high performance across all eight HAICATI dimensions. It provides accurate, reliable, transparent, explainable, contextually appropriate, and psychometrically valid evaluations while recognising meaningful human intellectual contribution. Such systems may be considered suitable for high-stakes decision-making, provided appropriate human oversight remains in place.

High Trustworthiness (5.50–6.49)

The assessment system performs well across most dimensions, with only minor limitations. Although improvements may still be desirable, the system generally produces dependable and scientifically credible assessments appropriate for most educational, organisational, and research applications.

Moderate Trustworthiness (4.50–5.49)

The system demonstrates acceptable performance but exhibits notable weaknesses in one or more dimensions. Results should be interpreted cautiously and supported by additional evidence, particularly when used in decisions affecting academic integrity, employment, or professional practice.

Marginal Trustworthiness (3.50–4.49)

The system shows inconsistent performance and should not be relied upon as the primary source of evidence. Human review, corroborating documentation, and contextual information are essential before consequential decisions are made.

Low Trustworthiness (2.50–3.49)

The assessment system demonstrates substantial weaknesses across several dimensions. It is unsuitable for high-stakes decision-making and should be used, if at all, only for exploratory or low-risk purposes while improvements are implemented.

Very Low Trustworthiness (1.00–2.49)

The system fails to demonstrate the minimum standards required for trustworthy AI contribution assessment. It should not be used for educational, organisational, regulatory, or legal decision-making because its outputs lack sufficient scientific credibility.