Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Assisted Assessment of Sustainability Reports Using Large Language Models

Nasywa Alya Kharasani
Bina Nusantara University
Stephan Ardy
Bina Nusantara University
Surya Darma Wiputra
Bina Nusantara University
Marko Hermawan
Bina Nusantara University
Richard Wiputra
richard-slc@binus.edu

Bina Nusantara University

Research Article

Keywords: Sustainability Reporting, Sustainable Procurement Disclosure Index, AI-Assisted Assessment, Large Language Models, Design Science Research

Posted Date: May 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9250943/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI-Assisted Assessment of Sustainability Reports Using Large Language Models

Nasywa Alya Kharasani[1], Stephan Ardy[2], Surya Darma Wiputra[1], Marko Hermawan[1], Richard Wiputra[2]

Corresponding Author: Richard Wiputra[2]
Email: richard-slc@binus.edu

[1] *Management Department, BINUS Business School, Bina Nusantara University, Jakarta, Indonesia*
[2] *Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia*

Abstract

Sustainability Reporting (SR) has become a critical mechanism for corporate accountability. However, assessment approaches based on Disclosure Indices (DIs) are constrained by subjectivity, inconsistency, and time inefficiency because they rely on manual evaluation. This study addresses these limitations by developing an AI-assisted assessment system using a Design Science Research (DSR) approach. The proposed artifact integrates large language models (LLMs) with a two-tier web architecture and adaptive document processing. The system is evaluated using sustainability reports from 99 firms listed on the Indonesia Stock Exchange (IDX). Results demonstrate that the AI-assisted approach significantly improves efficiency, reducing the average assessment time per report from 92 minutes to 35 minutes, representing nearly a threefold improvement. However, accuracy analysis reveals limited agreement with human assessments, with a weighted kappa value of 0.307 and a mean absolute error (MAE) of 1.09, indicating that while the system captures general disclosure patterns, it lacks consistency in reproducing expert judgment. This study contributes to the literature by demonstrating the feasibility and limitations of AI-assisted sustainability disclosure assessment, positioning it as a complement to, rather than a replacement for, expert evaluation.

Keywords: Sustainability Reporting, Sustainable Procurement Disclosure Index, AI-Assisted Assessment, Large Language Models, Design Science Research

1. Introduction

Sustainability reporting (SR) has become a central mechanism of corporate accountability, driven by regulatory pressure and stakeholder demand for Environmental, Social, and Governance (ESG) transparency . This pressures companies to adopt Disclosure Indices (DIs) to structure and

evaluate sustainability disclosures. Among these, the Sustainable Procurement Disclosure Index (SPDI), proposed byLau et al. [19], provides a systematic framework for assessing sustainability and transparency in procurement processes.

However, prior studies indicate that the expanding scope of disclosure indices introduces several challenges, particularly in maintaining consistency and timely adherence to evaluation criteria. These processes remain highly dependent on manual content analysis, making them subjective, time-intensive, and difficult to scale [4, 11, 15]. Although manual evaluation of disclosure quality can reduce information asymmetry and improve market efficiency [1, 38], it is constrained by variability in human interpretation and limited processing capacity.

To address these limitations, this study proposes an AI-assisted approach for SPDI-based sustainability disclosure assessment. The proposed system leverages Large Language Models (LLMs) with structured prompt engineering to automate indicator-level evaluation of sustainability reports. Empirical evaluation on 99 firms listed on the Indonesia Stock Exchange (IDX) shows that the AI-assisted approach substantially improves efficiency, reducing assessment time by nearly threefold. However, the results also reveal limited agreement with expert judgment, indicating that while LLMs can capture general disclosure patterns, they struggle to reproduce nuanced human interpretation.

These findings highlight both the potential and limitations of AI-assisted sustainability disclosure assessment. Consistent with prior research on AI augmentation in evaluative tasks [27, 29, 39], the results demonstrate how AI can augment human evaluative work by improving scalability and reducing cognitive load, while human expertise remains essential for contextual interpretation and judgment under ambiguity. These results suggest that AI can serve as an efficiency-enhancing tool in sustainability assessment, though expert oversight remains necessary for reliable scoring[5].

The remainder of this paper is structured as follows. Section 2 reviews prior literature on sustainability disclosure, disclosure indices, and AI-based assessment. Section 3 presents the Design Science Research (DSR) approach adopted in this study. Section 4 describes the design of the AI-assisted assessment artifact, including the conceptual model, system architecture, workflow, and core algorithms. Section 5 details the implementation, covering ground truth construction, prompt engineering, and evaluation procedures. Section 6 discusses the implications of the findings for AI-assisted sustainability assessment. Finally, Section 7 concludes the paper and outlines limitations and directions for future research.

2. Literature Review

This section provides the theoretical foundation for the study by examining key concepts and frameworks related to sustainability reports (SRs), disclosure indices (DIs), and the implications of AI and LLMs for assessment.

2.1. Sustainability Disclosure Practices and Corporate Reporting Dynamics

A Sustainability Report (SR) is a critical disclosure document that companies must produce to remain relevant in the global business landscape. SR is an obligation because it enhances stakeholders' understanding of corporate operations and accountability. Environmental, Social, and Governance (ESG) considerations are fundamental to this document, as they demonstrate how companies achieve the Sustainable Development Goals (SDGs). Hence, the growth of SR is related to transparency, driven by increasing stakeholder expectations and evolving regulatory requirements that now mandate SRs as part of annual or integrated reporting [6, 9].

Sustainability reporting adoption is now near-universal among large firms, with increasing emphasis on disclosure quality and assurance. In Indonesia, climate-related risk disclosure increased from 77% in 2021 to 88% in 2022, reflecting growing recognition of the need for ESG transparency.

2.2. Measurement Frameworks for Sustainability Disclosure

The Sustainability Disclosure Index (DI) is a commonly used tool in SR. There is a difference between DI and Sustainability Standards based on their purposes. DI is a systematic and quantitative method for assessing how transparent companies are in disclosing their sustainability performance to stakeholders [19]. On the other hand, standards are rules or guidelines that companies are expected to follow when preparing SR or other ESG-related reports [32]. Driven from these differences, DI is constructed by identifying a list of disclosure items based on established standards, such as DJSI, GRI, SIGMA, UNGC, and even the National Voluntary Guidelines on Social by the Ministry of India [24].

Building on the general concept of DI, more specific instruments have been developed, such as the Sustainability Report Disclosure Index (SRDI) and the Sustainability Reporting Index (SRI). SRDI refers to the direct application of DI in line with established sustainability standards, such as GRI [24]. Prior research developed a coding scheme for measuring sustainability disclosure quality, where a value of 0 indicates no disclosure and a value of 1 indicates disclosure. In contrast, SRI has a distinct approach, combining more than one existing sustainability standard (other than GRI) and featuring six assessment scales, which range from 0 to 5. Regarding SP, Lau et al. [19] developed the Sustainable Procurement Disclosure Index (SPDI), which assesses businesses' SP performance using 33 GRI indicators.

2.3. Artificial Intelligence and Large Language Models for Automated Disclosure Assessment

Artificial Intelligence (AI) has emerged as a significant tool for sustainability assessment, particularly through its capacity for large-scale textual analysis. Recent literature demonstrates that AI technologies, including natural language processing and machine learning, can facilitate systematic evaluation of corporate disclosures [18, 34]. Furthermore, textual analysis has been employed to extract ESG initiative factors from sustainability reports, demonstrating the potential of computational methods for disclosure assessment [35]. However, an AI-assessor has not been implemented for SR, particularly for evaluating the accuracy of the company's reported information. AI has the potential to improve assessment efficiency, although its outputs remain constrained by prompt design and lack the interpretive flexibility of human judgment [16]. This study uses Large Language Models (LLMs), which perform tasks such as document summarisation and content classification [8, 33] to automate the assessment of sustainability disclosures against SPDI indicators [2, 22].

2.4. Related Works and Research Gaps

AI-driven automation has shown potential for reducing organisational burdens in sustainability accounting [7]. Recent studies have applied text-mining and dictionary-based approaches to assess ESG disclosure compliance across industries and regulatory frameworks [23, 31], positioning AI as a catalyst for sustainable accounting. In parallel, Minn [25] have explored related approaches in other assessment domains, demonstrating AI's broader capacity to process textual data and reduce manual evaluation workload.

Although AI has shown potential in ESG reporting and sustainable accounting, existing research does not adequately specify the role of procurement disclosures in these areas. Current approaches lack alignment with SPDI indicators, particularly in capturing supply chain transparency, supplier standards, and other procurement-related sustainability practices. Moreover, much of the existing literature remains conceptual, with limited empirical testing and practical application. This gap underscores the pressing need for AI-driven models that are closely tied to SPDI.

Existing disclosure studies often take a generalised approach to ESG or SR, rarely focusing on the unique characteristics of procurement-related data, which involve complex layers of contextual indicators. A practical solution to this problem is to leverage NLP and machine learning to systematically process procurement-specific disclosures and translate qualitative narrative data into quantifiable metrics. Such an approach is crucial for enhancing the reliability and comparability of SPDI assessments, supporting regulatory compliance, and informing decision-making for investors, regulators, and other stakeholders.

3. Research Approach

This study adopts Design Science Research (DSR) to develop and evaluate an AI-assisted assessment artifact. DSR is well-suited to Information Systems research where the goal is to create innovative artifacts that solve real-world challenges [3, 13, 14].

This study applies DSR by developing an Artificial Intelligence (AI)-based prototype system to automatically evaluate SR. The methodology is divided into two stages: prompting and AI creation. The prompting stage, which includes theoretical development, will be incorporated into the AI prompting; meanwhile, the AI-Creation stage will focus on the AI's technical development. Peffers et al. [28] 's six steps of Design Science (DS) are adopted in this research, which spans from problem identification to the final communication of results.

The process begins by defining the relevant problem and explaining why it matters to the research, followed by formulating specific goals that the researcher must meet. Once the groundwork is laid, the artifact design is created for the AI-based assessment system. The latter stages focus on testing the system, examining how effectively it assesses the SR, sharing the outcomes, innovations, and practical benefits of the developed artifact. A detailed overview of each stage, as applied in our research, is presented in Table 1.

Table 1 Design Science Research (DSR) Stages Applied in Developing the AI-Assisted SPDI Assessment System

DSR Stage	Related Details
Stage 1: Problem Identification	● Human assessment is prone to bias due to subjectivity, varying interpretations, and inconsistent evaluation results. ● Human assessment is more time-consuming because it requires manual review, interpretation, and judgment of complex information
Stage 2: Objectives	● To create an AI-based assessment system that improves efficiency in sustainability disclosure assessment. ● Creating an AI-based assessment system significantly reduces assessment time by automating data processing and analysis.
Stage 3: Design	● Adaptation of the SPDI scoring from the original binary (0–1) scheme used in prior studies to a finer-grained 0–4 scoring system to capture varying levels of disclosure quality. ● Retrieved and analysed SR contents using text extraction and Optical Character Recognition (OCR). ● Used LLM API to detect index disclosure based on SPDI.

	● Human assessment for the final check of AI-based assessment products.
Stage 4: Demonstration	● Run the AI-based assessment system on a web application using four different SRs. ● Display the differences of SPDI scores between the ground truth of human-assessment and AI-based assessment.
Stage 5: Evaluation	● How accurately does the AI result towards human assessment as the ground truth?
Stage 6: Communication	● Discussion of model results and the practical contributions. ● Analyse the key differences in SPDI scores between a human assessor and a developed AI. ● Presenting the final artifact and analysing the practical implications of AI-assisted SR assessment.

4. Artifact Design: AI-Assisted SPDI Assessment System

To operationalise the proposed framework, this study develops a dedicated artifact that integrates sustainability standards, disclosure indices, and AI technologies. The artifact is described across four subsections: the conceptual model, which outlines the theoretical foundation; the system architecture, which specifies the technical design; the system workflow, which details end-to-end interactions; and the algorithms that drive data extraction and indicator evaluation.

4.1. Conceptual Model of AI-Assisted SPDI Evaluation

The conceptual model proposed a simplified yet systematic framework for applying AI to evaluate SR through the SPDI. The model consists of three main elements: Data Input, SPDI Aim, and Output, which are interconnected to ensure a structured flow of information from reporting standards to measurable outcomes.

The data input stage draws from three primary sources. First, the GRI Standard serves as the normative foundation, providing globally recognised guidelines for sustainable disclosure. These standards are translated into the SPDI framework, which operationalises procurement-related disclosure into measurable indicators. Second, the SR of listed companies provides the empirical data that will be evaluated using the SPDI framework. Together, these inputs form the raw material for the subsequent AI-driven analysis.

In the SPDI AI stage, the process is mediated by Prompt Engineering, which functions as the critical bridge between the SPDI framework and the capabilities of Large Language Models (LLMs). Prompt engineering ensures that SPDI indicators are accurately represented in queries submitted to the AI system. This structured interaction allows the Large Language Model (AI) to extract, classify, and score relevant information from SR in a manner consistent with SPDI indicators.

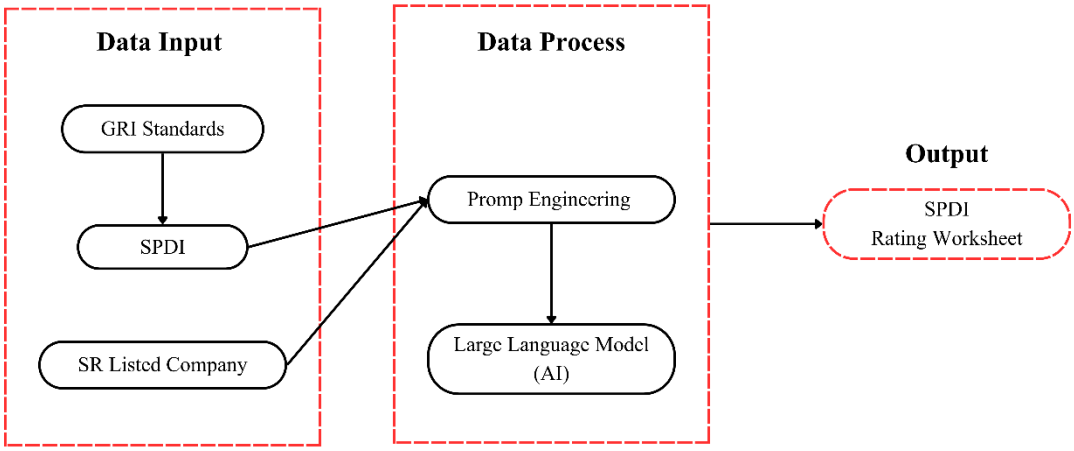


Figure 1 Conceptual Model of the AI-Assisted SPDI Assessment Framework Integrating GRI Standards, LLM Processing, and Structured Outputs

The final output of the process is the Sustainability Rating Worksheet, which consolidates the AI evaluation results into a structured, replicable format. This worksheet provides transparency and consistency, enabling cross-company comparison of sustainability performance. Overall, this conceptual model demonstrates how integrating international sustainability standards, a structured disclosure index, and AI technologies can streamline the assessment of SR. The conceptual model is visualised in Figure 1.

4.2. System Architecture of the AI-Assisted Assessment Platform

The application employs a two-tier web architecture that separates the presentation layer (frontend) from the application and data layers (backend). The frontend is implemented with Next.js for server-side rendering, while user authentication is managed through Clerk.

The backend is centred on a Python API that orchestrates data flow and business logic. Upon receiving authenticated requests, the API retrieves source documents, performs text extraction, and constructs structured instructions for Gemini LLM through a dedicated prompt-engineering module. The LLM performs document understanding, evidence extraction, and indicator mapping. Data persistence combines a relational database for structured results and object

storage for source documents and generated worksheets. Figure 2 illustrates the overall system architecture.

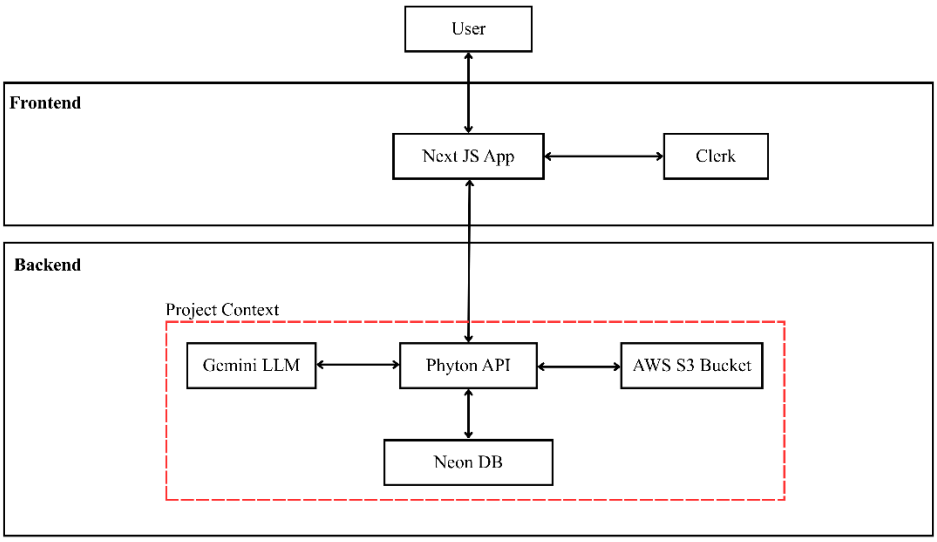


Figure 2 System Architecture of the AI-Assisted SPDI Platform Showing Frontend, Backend, LLM Integration, and Data Storage Layers

End-to-end execution proceeds from user authentication and document upload through text extraction, prompt construction, and model invocation, with results persisted for querying and traceability. The design ensures reproducibility by recording prompt hashes, model versions, and document checksums. Security is enforced with JWT verification, TLS in transit, encryption at rest for S3 and Neon, and per-project data isolation to support multi-tenancy. Scalability is achieved through Next.js rendering strategies, serverless Postgres autoscaling, and elastic object storage. Reproducibility and auditability are ensured by recording prompt hashes, model versions, parameters, evidence locations, and document checksums, enabling exact re-runs and systematic error analysis.

4.3. System Workflow for AI-Based SPDI Evaluation

Furthermore, the system workflow is visualised using a sequence diagram to illustrate the dynamic interactions among components over time. This diagram illustrates the entire process, from a user uploading a file to the system displaying the final analysis results, as shown in Figure 3 below.

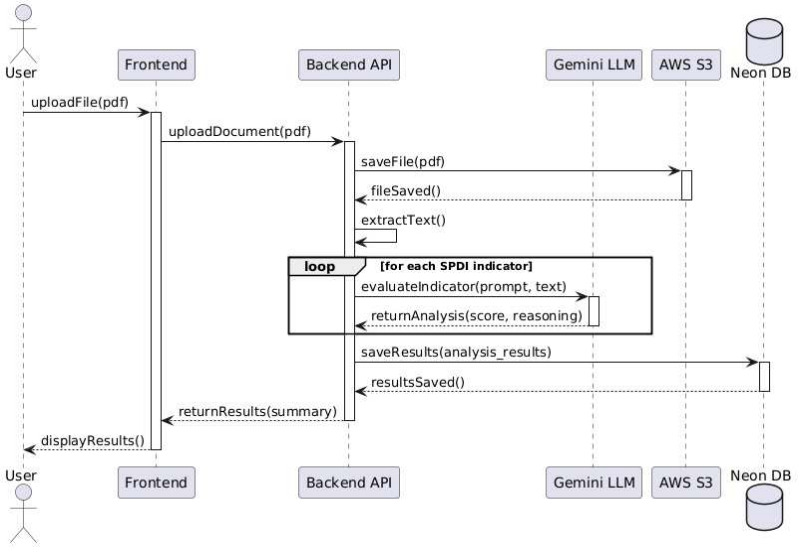


Figure 3 End-to-End Workflow of the AI-Assisted SPDI Assessment Process from Document Upload to Structured Evaluation Output

The system begins with a user-initiated upload of a PDF document via the web frontend. The backend persists the original artifact in object storage, then performs text extraction. For each SPDI indicator, the backend constructs a task-specific prompt that combines the indicator requirements with the extracted text and dispatches it to the LLM, which returns a structured score and rationale.

Results persisted to a relational store, while raw files remain in object storage, supporting reproducibility and traceability. The frontend renders SPDI scores and explanatory reasoning to the user.

Finally, the frontend renders the summary to the user, providing SPDI scores and explanatory reasoning, and may expose links to the underlying artifacts for verification purposes. The architecture thus enforces a clean division of responsibilities: the frontend handles interaction and visualisation; the backend orchestrates storage, extraction, and model calls; the LLM performs indicator-specific evaluation; and the storage tier guarantees durability and traceability. This design facilitates extensibility (by adding indicators or models), auditability (by retaining both inputs and outputs), and reproducibility (through deterministic extraction and structured result storage).

4.4. Core Algorithms for Document Processing and Indicator Evaluation

This section consists of two main algorithms. PDF Text Extraction with OCR Fallback, which ensures accurate text capture from both native and scanned sustainability reports. AI-Powered Indicator Evaluation, which uses a large language model with structured prompts to assess SPDI

indicators and generate scores, reasoning, and usage data. Together, these algorithms form a robust pipeline that spans from raw document ingestion to structured evaluation.

PDF Text Extraction with OCR Fallback

The algorithm in Figure 4 is crucial for handling the variety of PDF formats found in SR. It first attempts standard text extraction, then automatically switches to an AI-powered Optical Character Recognition (OCR) process if it detects that the PDF is a low-quality scan or an image-based PDF. This adaptive approach ensures a robust data input for subsequent AI analysis.

Algorithm 2: EXTRACT_PDF_TEXT_WITH_OCR_FALLBACK

```
1. FUNCTION EXTRACT_PDF_TEXT_WITH_OCR_FALLBACK(pdf_content):
       // Attempt standard text extraction with PyPDF2
2.     text_pypdf2 = PYPDF2_EXTRACT(pdf_content)
       // Check if the extracted text is too sparse (indicates a scanned PDF)
3.     IF average_characters_per_page(text_pypdf2) < threshold:
           // Fallback to image-based OCR using Gemini 1.5 Pro
4.         LOG("Detected scanned PDF. Starting Gemini OCR process."

           // Convert each PDF page to an image
5.         pdf_pages_as_images = CONVERT_PDF_TO_IMAGES(pdf_content)
6.         text_gemini_ocr = “”

           // Use Gemini to extract text from each image
7.         FOR each page_image IN pdf_pages_as_images:
8.             text_gemini_ocr += GEMINI_VISION_API_CALL(“Extract all text from
   image”, page_image)

           // Compare the quality and return the best result
9.         IF LENGTH(text_gemini_ocr) > LENGTH(text_pypdf2):
10.            RETURN text_gemini_ocr
11.        ELSE:
12.            RETURN text_pypdf2
13.
```

```
14.     ELSE:
          // Return the standard extraction if it was successful
15.       RETURN text_pypdf2
16.
17.   END FUNCTION
```

Figure 4 Adaptive PDF Text Extraction Algorithm with OCR Fallback for Robust Processing of Machine-Readable and Scanned Documents

AI-Powered Indicator Evaluation

The algorithm in Figure 5 leverages the power of a pre-trained LLM through precise prompt engineering. Instead of being trained on a labelled dataset, the LLM is given structured instructions to evaluate an SPDI indicator. This approach enables highly flexible, efficient assessment without the need for extensive manual data labelling or model training.

Algorithm 3: EVALUATE_INDICATOR_WITH_GEMINI

```
1. FUNCTION EVALUATE_INDICATOR_WITH_GEMINI(extracted_text,
   indicator_details):
       // Identify relevant sections of the report based on
2.   keywords = indicator_details.keywords
3.   relevant_sections = FIND_RELEVANT_SECTIONS(extracted_text, keywords)

       // Prepare a structured prompt for the LLM
4.   prompt = CREATE_PROMPT(
5.     "You are an expert ESG analyst. Assess the following report.",
6.     indicator_details.description,
7.     relevant_sections
8.   )
      // Send the prompt to the Gemini LLM
9.   response = GEMINI_API_CALL(prompt)

      // Parse the LLM's response to extract structured data
10.    score = PARSE_SCORE_FROM_RESPONSE(response)
11.    reasoning = PARSE_REASONING_FROM_RESPONSE(response)
```

```
    token_usage = PARSE_TOKEN_USAGE_FROM_RESPONSE(response)

    RETURN score, reasoning, token_usage

END FUNCTION
```

Figure 5 LLM-Based Indicator Evaluation Algorithm for SPDI Scoring with Structured Prompt Engineering

5. Implementation and Evaluation

This section explains the implementation of the proposed AI-assisted SPDI assessment system, with particular attention to data preparation, assessment design, and evaluation procedures. Subsection 5.1 describes the development of the ground truth and the preparation of SPDI-based assessment data. Subsection 5.2 explains the design of the AI-assisted assessment framework, including the use of scoring worksheets, guideline construction, keyword strategies, and reference excerpts from sustainability reports. Subsection 5.3 presents the evaluation procedures used to assess the system's efficiency and accuracy. The overall implementation structure, spanning from standards definition (L1) to evaluation (L5), is summarised in Figure 6.

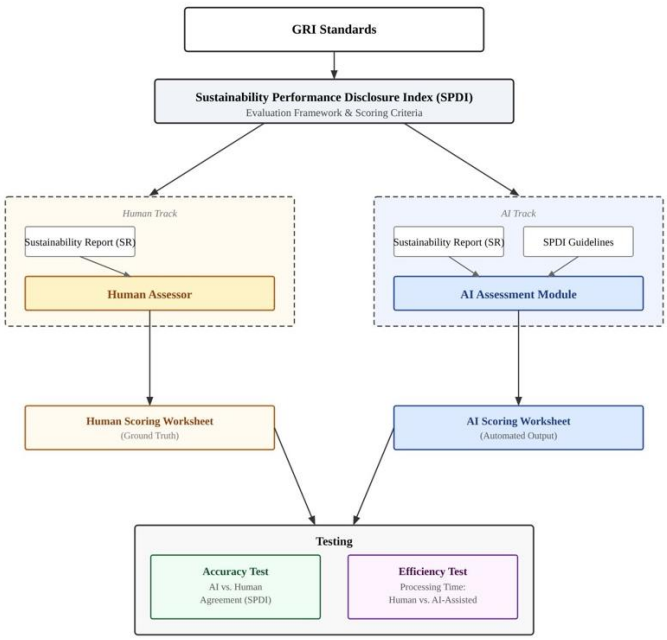


Figure 6 Implementation Framework of the AI-Assisted SPDI Assessment System

5.1. Ground Truth Development and SPDI-Based Data Preparation

This stage corresponds to Layers L1–L2 in Figure 6, where GRI standards are operationalised into the SPDI framework, and ground truth is established. The ground truth was established through

manual assessment by a trained human assessor with expertise in sustainability reporting and GRI standards. The primary assessor independently evaluated all 99 reports using a structured coding protocol based on the SPDI framework. To validate scoring consistency, a subset of reports was independently coded by a second assessor, and discrepancies were resolved through discussion and consensus. While a formal inter-rater reliability coefficient was not computed across the full dataset, the adjudication process ensured that the ground truth reflects considered expert judgment rather than a single unverified assessment. The authors collected the SRs from the top 99 companies listed on the Indonesia Stock Exchange (IDX) as ranked by Kompas100. The authors specifically selected the SRs from 2024 because this period provides the most recent and relevant data available for the analysis.

Once all sustainability reports are compiled, the disclosures are evaluated using the Sustainable Procurement Disclosure Index (SPDI), adapted from the Global Reporting Initiative (GRI) standards. The SPDI consists of 33 procurement-related indicators that provide a structured framework for assessing transparency and sustainability in corporate procurement activities Lau et al. [19]). A five-point scale coding scheme (0-4) outlined in the SPDI will be used in this study. A score of 4 indicates that the information related to a specific index is comprehensively reported, while lower scores, such as 0 or 1, reflect limited or absent disclosures. If a particular index frequently receives a core of 4 across companies, it suggests it is well-addressed in the SR. Conversely, a lower score implies insufficient information.

5.2. Prompt Engineering and AI-Assisted Assessment Framework Design

This stage operationalises the AI assessment track in Layer L3 and produces structured outputs in Layer L4. The implementation distinguishes between a human assessment track and an AI-assisted assessment track, enabling parallel evaluation and direct comparison. After all information elements and the scoring system are established, the next step is to create the Scoring Worksheet based on the human assessment. The worksheet contained the names of the 99 companies, their sectors, the 33 SPDI indicator results from the previous step, and the SPDI score for each company. In addition to the scoring worksheet, the guideline worksheet is also essential for supporting this step. The guidelines worksheet includes all GRI indicators listed in the SPDI framework, along with definitions and disclosure expectations outlined by the GRI. The arrangement of this worksheet is designed to allow AI to interpret the documents accurately, leveraging the knowledge of a human assessor.

Keyword matching strategies include broad, phrase, and exact matching [20, 26]; this study adopts an exact-match approach to ensure precision in identifying disclosed information.

Table 2 Examples of Indicator-Specific Keywords Used in Prompt Engineering for SPDI Evaluation

Index	Keywords Frequently Reported	Related Statement in SR
2-9-b	1. Sustainable Governance Structure 2. ESG Structure 3. Director of Sustainability 4. Sustainability Governance Structure 5. Governance Structure	Overall, responsibility for sustainability activities within Jasa Marga is carried out by the President Director together with other Directors, with full supervision conducted by the Board of Commissioners. In accordance with OJK Regulation No. 51/POJK.03/2017 on the Implementation of Sustainable Finance for Financial Services Institutions, Issuers, and Public Companies, the Company has established a governance structure by appointing employees, officials, and/or work units responsible for implementing sustainability within Jasa Marga. This addition is made to comply with the regulations. Additionally, it demonstrates the company's commitment to supporting sustainability programs and achieving the Sustainable Development Goals (SDGs).

As part of an example-based approach, the following process in this study incorporates reference excerpts from manually assessed SRs. This strategy has been shown by Brown et al. [8] and Minn [25] to enhance interpretability in prompting tasks. The reference-based approach ensures that the LLM follows the researcher's screening decisions rather than relying solely on synthetic or hypothetical examples. For each indicator, one or two real-world quotations are drawn from the listed companies, all of which have been evaluated by the researcher. In cases where the index lacks a reference for specific points (0–4), as no company has achieved those scores, the explanation will be described according to the researchers' understanding. Other reference

methods applied; this involves researchers directly referring to the SRs according to the required index, following careful consideration, particularly concerning the highest scores.

After this process, the AI was given the 99 sustainability reports (SRs) to process sequentially, with each report evaluated indicator by indicator against the provided guidelines. The assessment includes not only the scores themselves but also the reasoning behind each score, enabling transparency in how the AI evaluated the reports.

5.3. Evaluation of AI-Based Assessment: Accuracy and Efficiency

This stage corresponds to Layer L5, where both human and AI outputs are evaluated. To enhance the speed and validity of the AI system's scores, an evaluation comparing the results of a human assessor (ground truth) and the AI system is being conducted. There are two types of evaluation tests: efficiency and accuracy. Each test offers a distinct perspective on the system's effectiveness. The following subsections explain the procedures for each evaluation test.

A. Efficiency Evaluation: Assessment Time Analysis

The measurement assesses how quickly a human or AI can evaluate a single SR. A previous manual measurement by a human assessor took over 1 month to assess 99 SRs, with approximately 1 SR requiring 1/2 hour to 1/2 day to be fully assessed, indicating a time-consuming process. This also highlights the limitations of manual assessment, particularly given the large number of reports. Hence, this test demonstrates how AI can reduce the time required to assess SR. The procedures are listed in Table 3 below.

Table 3 Procedure for Efficiency Evaluation of Human and AI-Assisted Assessment

Speed Test Procedure	
1	The assessor prepares the SR for evaluation.
2	The assessor prepares the manual, AI evaluation sheets, and a timer (from a mobile phone or laptop).
3	The assessor conducts the assessment manually (human evaluation).
4	The assessor starts the timer at the beginning of each assessment category (Governance, Economic, Social, and Environmental) and stops it upon completion of that category.
5	The recorded time is entered into the evaluation sheet.
6	Steps 4-5 are repeated until all categories have been assessed.

This speed test was conducted by four independent human assessors who were not affiliated with the researchers of this study. The assessors were university students with

no prior knowledge or experience in sustainability reporting. These assessors were given the researchers' guidelines to assess the SR.

The findings indicated a consistent advantage of AI in reducing assessment time, though the magnitude of the efficiency gain varies across cases. AI-assisted assessment consistently required less time than manual evaluation, with efficiency gains ranging from 24.25% to 170.46% across the four tested reports. On average, the manual assessment takes 92 minutes and 27 seconds to complete. In comparison, the average completion time for AI assessments is 35 minutes and 1 second. The results represent a nearly threefold reduction in time efficiency. Specifically, the highest efficiency was recorded in Indofood, with a reduction rate of 170.46%, followed by Antam at 67.11%, BJB at 50.60%, and Semen Indonesia at 24.25%.

B. Accuracy Evaluation: Agreement with Human Assessment

The test aims to evaluate the accuracy of the AI assessment by comparing it to a human-assessed score. In this context, the human assessor is treated as the ground truth, reflecting expert-informed evaluation based on established SPDI guidelines. 99 SRs will be equipped for this accuracy test. The procedures are listed in Table 4 below.

Table 4 Procedure for Accuracy Evaluation Using Human Ground Truth and AI-Assisted Scores

Accuracy Test Procedure	
1	Human assessor manually assessed all 99 SRs.
2	Human assessor enters the assessed score in the scoring worksheet.
3	AI generates a score for 99 SRs.
4	AI-generated assessment score in the scoring worksheet.
5	Side-by-side comparison.

Weighted kappa is used to compare the ground truth and AI-assisted scores, accounting for the degree of disagreement [21]. The results indicate a weighted kappa value of 0.307, reflecting minimal agreement.

Mean Absolute Error (MAE) quantifies the average deviation between two sets of scores [36]. The resulting value of 1.09 indicates that the AI model deviates by approximately one point on average. This result is consistent with findings from Bland–Altman analysis, further illustrating the degree of agreement and the extent of variation between the two ratings by analysing the

average difference and the spread of discrepancies across observations [12]. Figure 7 presents the Bland-Altman graphical plot.

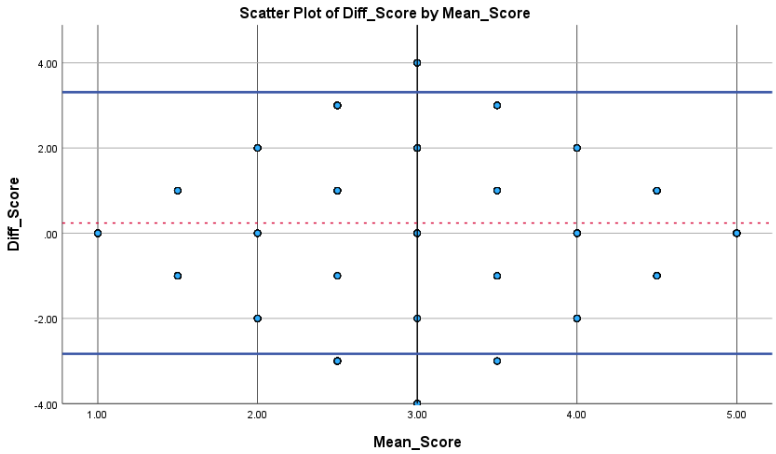


Figure 7 Bland–Altman Analysis of Agreement Between AI-Assisted and Human Assessment Scores

The Bland-Altman plot further supports the finding of minimal systematic bias between ground truth and AI-assisted analysis, as indicated by the mean difference of 0.24, which is close to zero. However, the wide limits of agreement, ranging from negative 2.83 to 3.31, highlight substantial variability in the differences.

Another finding can be observed at the firm level, where several distinct patterns emerge in the differences between the ground truth and AI-assisted scores. To quantify firm-level agreement, we computed a simple exact-match percentage: for each firm, the proportion of the 33 SPDI indicators where the AI score exactly matched the human score. A particularly notable contrast is observed between Perusahaan Gas Negara (PGN) and Zyrexindo Mandiri Buana. PGN demonstrates a relatively low exact-match rate of 15%, whereas Zyrexindo Mandiri Buana achieves a substantially higher rate of 92%.

This disparity can be further explained by an additional unweighted analysis at the individual rating level, in which matching scores were assigned a value of "1" and non-matching scores were assigned a value of "0". This approach was employed to capture the exact agreement between the two ratings, thereby providing a clearer comparison of how consistently the AI replicates human evaluations across different firms.

6. Discussion

This section interprets the results from the evaluation stage (Layer L5 in Figure 6), focusing on the efficiency and accuracy of the human and AI assessment tracks.

6.1. Efficiency Gains in AI-Assisted Sustainability Assessment

The results demonstrate a consistent advantage of AI-assisted assessment in terms of processing speed. Across all evaluated cases, AI reduced assessment time by 24.25% to 170.46%, with the average time decreasing from 92 minutes for manual assessment to 35 minutes for AI-assisted evaluation. This represents a nearly threefold improvement across governance, economic, social, and environmental disclosure categories.

These findings highlight the role of AI as an efficiency-enhancing component in sustainability assessment workflows. Rather than replacing human evaluators, AI enables rapid processing of large volumes of reports, thereby reducing cognitive and operational burdens. This aligns with prior literature on AI-augmented decision-making, which positions AI as a complementary tool that supports human capability in repetitive and data-intensive tasks [10, 17, 30, 37].

6.2. Accuracy Limitations and Variability in AI-Assisted Evaluation

Despite efficiency gains, the accuracy results indicate limited agreement between AI-assisted and human assessments. The weighted kappa value of 0.307 suggests minimal agreement, while the mean absolute error (MAE) of 1.09 indicates that AI scores deviate by approximately one point on average. Bland–Altman analysis further reveals minimal systematic bias but substantial variability, as reflected in wide limits of agreement.

These findings indicate that AI can approximate general disclosure patterns but struggles to reproduce nuanced human judgment. A key explanation lies in the difference between pattern-based and semantic evaluation. The AI system relies on explicit linguistic signals, particularly keyword matching, to identify relevant disclosures. In contrast, human assessors interpret meaning more flexibly, recognising semantically equivalent expressions even when exact keywords are absent.

This divergence explains why AI may assign low scores when predefined keywords are missing, while human assessors assign higher scores based on contextual understanding. Conversely, AI may assign higher scores when disclosures are explicitly structured, reflecting its sensitivity to clearly defined criteria.

The results also suggest that human assessment is not error-free. Instances of re-evaluation indicate that human assessors may overlook relevant disclosures or apply inconsistent judgment, particularly when processing large volumes of information. This highlights that variability exists in both human and AI assessment processes, albeit driven by different mechanisms.

Additional variability may arise from document characteristics. Reports containing non-machine-readable elements, such as low-resolution images, limit the AI system’s ability to accurately extract and interpret information, further contributing to discrepancies between AI and human scores.

Overall, these findings demonstrate that AI-assisted assessment can approximate broad disclosure patterns but cannot yet replicate expert scoring reliably. AI contributes scalability and consistency in applying explicit evaluation rules, while human assessors provide contextual interpretation and judgment in the face of ambiguity. These results suggest that a practical pathway forward involves using AI as an initial screening tool, with expert review for final scoring decisions.

This highlights the importance of designing AI-assisted evaluation workflows that combine computational efficiency with human interpretive flexibility, particularly in domains characterised by unstructured data and subjective judgment.

7. Conclusion and Limitation

The final section summarizes the study's key findings. It reflects on its practical and theoretical contributions and its boundaries. The conclusion highlights how the AI-assisted SPDI assessment advances SR evaluation by improving efficiency and consistency while recognising its current accuracy limitations. The subsequent subsection further discusses the main constraints and suggests potential directions for further refinement.

7.1. Conclusion

This study demonstrates the feasibility of AI-assisted sustainability disclosure assessment using the SPDI framework. The development of an AI-assisted system highlights the potential of NLP and information-extraction techniques to address key challenges in sustainability reporting, including time inefficiencies and a lack of standardisation in manual assessments.

Implementing a keyword-driven prompting approach is an initial step toward automating sustainability disclosure analysis, offering greater consistency, transparency, and replicability than traditional manual methods. The prototype demonstrates that automation can reduce analysts' cognitive workload by speeding up assessments. However, improvement is needed to better align the system with expert judgment. Using the Weighted Kappa Test, minimal agreement was found between the ground truth and the AI-based assessment. Alongside, the MAE and Bland-Altman Graphic also supported this finding. These indicate that the AI-based assessment demonstrates limited reliability and consistency in replicating the ground truth.

This study contributes by (1) operationalising an AI-assisted SPDI assessment artifact using Design Science Research, (2) empirically evaluating the agreement between AI-generated and expert-assigned disclosure scores, and (3) identifying the practical conditions under which AI-assisted assessment can complement expert evaluation in sustainability reporting.

7.2. Limitations and Further Improvement

Despite its promising outcomes, this project presents several limitations that warrant attention in future development. First, prompt optimisation: refine prompts to better guide LLMs' reasoning.

This process enhances the clarity, structure, and specificity of prompts, enabling the system to accurately capture the information disclosed by the SPDI framework through SR. The next step involves domain-specific fine-tuning, such as models like Gemini or a comparable open-source model with an established score. This fine-tuning enables the model to learn the specific language patterns, terminology, and unique contexts of SR, especially in Indonesia, thereby enhancing overall accuracy beyond what generalized LLMs can achieve. Lastly, improvements are needed in the AI-based system to enhance accuracy towards the ground truth. A key limitation lies in the reliance on keyword-driven prompting, which constrains the model's ability to capture semantically equivalent disclosures expressed in varied linguistic forms.

Ethics Statement

Ethical Approval and Accordance: The research was reviewed and approved by the Bina Nusantara University Research Ethics Committee under approval no: 002/EC/IV/2026. With this approval, the study met the research integrity guidelines of the School of Information Systems, Bina Nusantara University, which also reviewed it and granted ethical clearance.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare that they have no commercial, financial, or personal relationships that could be construed as potential conflicts of interest in the conduct and reporting of this research. The development and evaluation of the AI-assisted SPDI assessment system were conducted independently for academic purposes without external sponsorship or competing interests.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contribution Statement

Marko Hermawan: Conceptualisation; Writing – Review and Editing, Supervision: Project Administration.

Richard Wiputra: Conceptualisation; Methodology; Writing – Review and Editing; Supervision: Project Administration.

Nasywa Alya Kharasani: Methodology; Data Collection and Ground Truth Assessment; Formal Analysis; Writing – Original Draft Preparation.

Stephan Ardy: System Development and Technical Implementation.

Surya Darma Wiputra: Formal Analysis; Writing – Review and Editing.

Declarations

Ethical Approval

This study was reviewed and approved by the Ethics Committee of Bina Nusantara University (Indonesia). All research activities were conducted in accordance with international ethical standards, including the principles of Respect for Persons, Beneficence, and Justice as outlined in the Belmont Report, as well as the Declaration of Helsinki. The study also complies with national regulations, including Peraturan BRIN No. 22 Tahun 2022 and ethical guidelines issued by the Kementerian Kesehatan Republik Indonesia (2021).

No experimental interventions involving human participants were undertaken in this study. Data collection was limited to a minimal-risk survey designed to capture participants' perceptions. The study posed negligible risk, with full protection of participant rights ensured throughout the research process. Prior to participation, all individuals were informed of the study's objectives and provided informed consent voluntarily, with the option to withdraw at any stage without any consequences. All data were anonymized, securely stored, and utilized solely for academic and research purposes.

Content to Participate

All participants were provided with comprehensive information regarding the study's objectives, procedures, potential risks, and expected benefits. Informed consent was obtained from each participant prior to survey completion, in accordance with approval granted by Bina Nusantara University. Participation was entirely voluntary, and participants had the right to withdraw at any point without any adverse consequences. The study did not involve any participants under the age of 18.

Consent to Publish

Not Applicable

References

[1] Abed, S., Basil Al-Najjar, A.-J., & Roberts, C. (2016). How to Measure Annual Report Narratives Disclosure? Empirical Evidence from Forward-Looking Information in the UK Prior the Financial Crisis. *Managerial Auditing Journal*, *31*, 338–361.

[2] Abro, A. A., Talpur, M. S. H., & Jumani, A. K. (2023). Natural Language Processing Challenges and Issues: A Literature Review. *Gazi University Journal of Science*, *36*(4), 1522–1536. https://doi.org/10.35378/gujs.1032517

[3] Adam, M. T. P., Gregor, S., Hevner, A., & Morana, S. (2021). Design Science Research Modes in Human-Computer Interaction Projects. In *AIS Transactions on Human-Computer*

Interaction (Vol. 13, Issue 1, pp. 1–11). Association for Information Systems. https://doi.org/10.17705/1thci.00139

[4] Adams, C. A., Alhamoood, A., He, X., Tian, D. J., Wang, L., & Wang, Y. (2021). The double-materiality concept Application and issues. In *Global Reporting Intiative*.

[5] Arakawa, R., & Yakura, H. (2023, April 19). AI for human assessment: What do professional assessors need? *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3544549.3573849

[6] Boadi, M., Twum, S., & Boadi, J. A. (2025). Exploring The Multifaceted Determinants Of Sustainability Reporting: A Holistic Perspective. *The International Journal of Business Review*, *8*(1). https://doi.org/10.17509/tjr.v8vi1.70365

[7] Bose, I., Kar, A. K., Lee, G. M., Leung, A. C. M., & Kim, D. J. (2025). Artificial Intelligence and the Bright Internet: Transforming the Internet for a Better Future. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-025-10632-z

[8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. http://arxiv.org/abs/2005.14165

[9] Dasinapa, M. B. (2024). The Integration of Sustainability and ESG Accounting into Corporate

[10] Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. *52nd Hawaii International Conference on System Science*. https://doi.org/https://doi.org/10.24251/HICSS.2019.034

[11] Dunfjäll, M. (2025). Materiality in Transition: Challenges and Opportunities in Corporate Sustainability Reporting under the CSRD. *European Journal of Risk Regulation*. https://doi.org/10.1017/err.2025.10016

[12] Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

[13] Hevner, A. R., & Chatterjee, S. (2015). *Design Science Research in Information Systems*. http://eduglopedia.org/reference-syllabus/AIS_Reference_Syllabus_Design_Science_Research_in_IS.pdf

[14] Hevner, A., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1). https://www.researchgate.net/publication/201168946

[15] Huq, A. M., & Mohammadrezaei, M. (2024). A review of ex ante and ex post materiality measures, and consequences and determinants of material disclosures in sustainability reporting. In *Journal of Accounting Literature* (Vol. 47, Issue 5, pp. 71–98). Emerald Publishing. https://doi.org/10.1108/JAL-04-2024-0084

[16] Jungmin, C., & Chao, M. M. (2024). For Me or Against Me? Reactions to AI (vs. Human) Decisions That Are Favorable or Unfavorable to the Self and the Role of Fairness Perception. *Personality and Social Psychology Bulletin*, *1*(21). https://doi.org/10.1177/01461672241288338

[17] Kapoor, R., & Ghosal, I. (2022). Will Artificial Intelligence Compliment or Supplement Human Workforce in Organizations? A Shift to a Collaborative Human – Machine Environment. *International Journal on Recent Trends in Business and Tourism*, *06*(04), 19–28. https://doi.org/10.31674/ijrtbt.2022.v06i04.002

[18] Kulkov, I., Kulkova, J., Rohrbeck, R., Menvielle, L., Kaartemo, V., & Makkonen, H. (2024). Artificial intelligence - driven sustainable development: Examining organisational, technical, and processing approaches to achieving global goals. *Sustainable Development*, *32*(3), 2253–2267. https://doi.org/10.1002/sd.2773

[19] Lau, K. H., Yadlapalli, A., Abdulrahman, M. D. A., Chhetri, P., & Thai, V. (2023). Disclosure index development for sustainable procurement: An australian perspective. *Journal of Cleaner Production*, *420*. https://doi.org/10.1016/j.jclepro.2023.138357

[20] Li, H., & Yang, Y. (2022). Keyword targeting optimisation in sponsored search advertising: Combining selection and matching. *Electronic Commerce Research and Applications*, *56*. https://doi.org/10.1016/j.elerap.2022.101209

[21] Li, M., Gao, Q., & Yu, T. (2023). Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. In *BMC Cancer* (Vol. 23, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12885-023-11325-z

[22] Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science,* (2nd ed.). Marcel Decker, Inc. https://surface.syr.edu/istpub

[23] Lukács, B., Rickards, R. C., Molnár, P., Limbach, Z., & Suta, A. (2025). ESG disclosure topics and reporting frameworks: exploratory research across automotive, construction, and energy industries. Discover Sustainability, 6, 649. https://doi.org/10.1007/s43621-025-01533-x

[24] Meutia, I., Kartasari, S. F., & Daud, R. (2022). Voluntary Assurance of Sustainability Reports: Evidence from Indonesia. *Accounting Analysis Journal*, *11*(1), 44–53. https://doi.org/10.15294/aaj.v11i1.58810

[25] Minn, S. (2022). AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2022.100050ï

[26] Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, *42*(9), 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050

[27] Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors and Ergonomics Society*, *39*(2), 230–253.

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

[29] Shneiderman, B. (2020a). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4). https://doi.org/10.1145/3419764

[30] Spillias, S., Tuohy, P., Andreotta, M., Annand-Jones, R., Boschetti, F., Cvitanovic, C., Duggan, J., Fulton, E. A., Karcher, D. B., Paris, C., Shellock, R., & Trebilco, R. (2024). Human-AI collaboration to identify literature for evidence synthesis. *Cell Reports Sustainability*, *1*(7). https://doi.org/10.1016/j.crsus.2024.100132

[31] Suta, A., Molnár, P., Limbach, Z., Lukács, B., & Rickards, R. C. (2025). Dictionary-based assessment of European Sustainability Reporting Standard (ESRS) disclosure topics. Discover Sustainability, 6, 486. https://doi.org/10.1007/s43621-025-00930-6

[32] van Oorschot, K. E., Aas Johansen, V., Lynes Thorup, N., & Aspen, D. M. (2024). Standardisation cycles in sustainability reporting within the Global Reporting Initiative. *European Management Journal*, *42*(4), 492–502. https://doi.org/10.1016/j.emj.2024.04.001

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*. http://arxiv.org/abs/1706.03762

[34] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-019-14108-y\

[35] Wang, J. (2025). Textual analysis reveals ESG initiatives linked to idiosyncratic risk in China. Discover Sustainability, 6, 862. https://doi.org/10.1007/s43621-025-01780-y

[36] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79–82. www.int-res.com

[37] Xu, J. Q., Wu, T. J., Duan, W. Y., & Cui, X. X. (2025). How the Human–Artificial Intelligence (AI) Collaboration Affects Cyberloafing: An AI Identity Perspective. *Behavioral Sciences*, *15*(7). https://doi.org/10.3390/bs15070859

[38] Ying Qiu, X., Jiang, S., & Deng, K. (2013). Automatic Assessment of Information Disclosure Quality in Chinese Annual Reports. In Guodong Zhou et al. (Ed.), *NLPCC 2013: Natural Language Processing and Chinese Computing* (pp. 288–298). Springer. https://doi.org/10.1007/978-3-642-40319-3_31

[39] Zimmerman, A., Janhonen, J., & Beer, E. (2023). Human/AI relationships: challenges, downsides, and impacts on human/human relationships. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00348-8