Mapping Sustainable Development Goals in Library and Information Science: A Comparative Evaluation of Automated and AI-Based Classification Systems

Sailendra Malik*, Sukumar Mandal**

ARTICLE INFO

Article history:
Online First 15 October 2026

Keywords:
Sustainable Development Goals,
SDG Classification,
Comparative Evaluation of Mapping Tools,
Classifier Validation,
Library and Information Science,
Research Assessment,
Knowledge Infrastructures

ABSTRACT

Automated classification of scholarly output against the United Nations Sustainable Development Goals (SDGs) increasingly shapes how institutions report their sustainability contributions, yet the reliability of such classification within Library and Information Science (LIS) has not been systematically tested. This study comparatively evaluates six operational SDG mapping systems, namely Web of Science SDG Mapping, Aurora SDG, the mBERT-based OSDG.ai classifier, Elsevier SDG Mapping, Scopus SDG Mapping and Dimensions AI SDG Mapping, on a corpus of 10,300 LIS records retrieved from the Web of Science Core Collection. Classifier outputs were benchmarked against a human-coded reference standard derived from a stratified validation sample, which provides the ground truth for all reported accuracy, precision, recall and F1 values. Large language models were used solely for conceptual structuring of the analytical framework, code generation for the web-based interface and visualisation support; they performed no classification and no numerical computation, and this boundary is documented in full. AI-based semantic classification achieved the highest and most consistent agreement with the reference standard (OSDG.ai, mean F1 = 0.767), outperforming rule-based and metadata-based mapping (mean F1 between 0.689 and 0.711). Agreement was highest for explicitly lexicalised goals such as Affordable and Clean Energy and Decent Work and Economic Growth (F1 = 0.83) and lowest for environmentally diffuse goals such as Climate Action and No Poverty. Coverage estimates represent classifier-generated signal rather than validated SDG relevance and differ systematically across tools. Classifier choice therefore materially alters an institution's apparent SDG profile, and structured human review remains a necessary component of credible SDG reporting.

* PhD Scholar, Department of Library and Information Science, The University of Burdwan (sailendra.malik113@gmail.com) (First & Corresponding Author)
** Assistant Professor, Department of Library and Information Science, The University of Burdwan, (sukumar.mandal5@gmail.com) (Co-Author)

© Author(s), 2026. Licensed under CC BY-NC-ND 4.0.

1. Introduction

Large language models have become useful tools for automating complex classification tasks, including the classification of domain-specific knowledge bases against the Sustainable Development Goals. Due to the extensive scale and complexity of the textual knowledge bases associated with the 17 UN Sustainable Development Goals (SDGs) it is very difficult to track the progress made by various stakeholders toward achieving these goals (Cadeddu et al., 2025). Advances in recent years in LLM based approaches toward solving SDG classification problems have been shown to be very promising, and have the potential to provide scalable and consistent solutions for monitoring SDG related content from a variety of different sources. In addition to being able to apply these models to general text classification problems they can also be applied to domain specific areas, such as library and information science, where accurate classification of scholarly work that has been published in libraries and other information centers is necessary in order to assess the contribution of an institution to a sustainability agenda. The 17 UN Sustainable Development Goals form a common architecture for the advancement of sustainable development on all three dimensions of social-economic-environmental spheres. Figure 1 demonstrates how the 17 SDGs are systemically integrated and collectively aligned toward promoting inclusive growth, ecological stewardship, and resilient governance as part of the United Nations 2030 Agenda Framework (https://unric.org/en/united-nations-sustainable-development-goals/).



Fig. 1. The 17 UN Sustainable Development Goals
Source: *UN Regional Information Centre (https://unric.org/en/united-nations-sustainable-development-goals/).*

Traditionally, the area of information science has used traditional machine learning techniques such as Support Vector Machines (SVM) and Naive Bayes Classifiers in order to perform traditional

text classification. Previous comparative research indicated that using the Label Powerset Transformation along with SVM algorithms can produce accuracy levels of up to 91% when classifying multi-class data sets; however, these previous methods have also shown significant challenge in addressing the issues associated with imbalanced data sets (Morales-Hernández, Jagüey & Becerra-Alonso, 2022). A recent paradigmatic shift has occurred from traditional machine learning based architectures to those utilizing transformer-based architectures. In addition to BERT, many contemporary applications have utilized other transformer-based architectures in order to develop hierarchical text classification methodologies; specifically, fine-tuning GPT-2 has been shown to significantly outperform BERT-BiLSTM methodology for highly specialized tasks (Bouchiha et al., 2025). As such, the foundation for modern LLM methodologies is based upon the development of the previous foundational methods.

The use of Large Language Models (LLMs) has shown a high degree of success as task specific models for classification of SDGs. In addition to achieving macro F1-scores of 0.75-0.77 in studies that utilize Few-Shot and Zero-Shot Learning with models such as GPT-3.5 and flan-t5-large, smaller optimized versions of these models have been found to perform similarly well compared to much larger proprietary versions (Cadeddu et al., 2025). The results from weak supervision methods of using LLMs to automatically identify structured concepts from large amounts of data and produce cross-domain similarity scores are positive for patent-to-SDG classification on a large scale (Ascione & Tamagnone, 2025). An example of multi-label classification is the use of an XLM-RoBERTA model to classify Indonesian academic papers and achieve an F1-score of 85.66% (Husna, Romadhony & Wibowo, 2025). A further example of the superiority of the DistilBERT model was in its ability to classify Indian sustainability oriented startups into SDG categories (Kannan et al., 2026).

The application of large language models (LLMs) to library and information science (LIS) will be one of the most important ways in which we use LLMs to automate many tasks related to the creation and management of metadata, the presentation of recommended content, and organizing content as it relates to knowledge representation. In addition, LLMs may provide a way to create an interactive dialogue with users about electronic resources available through libraries and other organizations and to assist with the automatic or semi-automatic generation of metadata, indexing of digital materials, and classification of digital content using much more sophisticated methods than those based solely on keywords (Formanek, 2025). An academic LLM designed specifically for the domain of LIS was able to achieve F1-scores of 0.3973 when predicting literature keywords and thus outperformed baseline commercial models (Zhu et al., 2025). Investigations conducted recently to determine if LLMs can reliably classify core LIS journals indicate that inter-model agreement ranges from 0.49 to 0.88, indicating that computational approaches are able to replicate, but not resolve, uncertainty regarding the boundaries of disciplines in LIS (Kim, 2026).

There are many problems facing LLMs that have a good metric score for classifying SDGs, but model disagreement about how to classify borderline cases leads to systematic, non-random differences in what documents you will retrieve from a database (Ingram, Banerjee, & Fox, 2025). The level of agreement among automated classification systems is very low, with only 35% of documents identified by the systems as being related to an SDG actually having been determined to be SDG-related; this is especially true when the SDG is one of the more complex, cross-cutting

goals such as Gender Equality. Additionally, semantic ambiguity, and the contextual nature of all SDG definitions pose long-term barriers to full automation. The high cost of computation, hallucination, and the fact that there is no way to know if automated systems are producing accurate results without some form of expert review, support the use of hybrid methods that utilize both automated detection and qualitative evaluations (Bautista-Puig et al., 2026).

The literature relevant to this study falls into five strands, which are treated separately here because they address problems with different methodological requirements. The first strand, SDG bibliometrics, measures disciplinary contributions to the 2030 Agenda and supplies the descriptive apparatus for goal-level coverage analysis. The second, automated SDG mapping, evaluates the operational tools through which such measurement is performed and is the strand this study contributes to directly. The third, large language model text classification, concerns model architecture and task adaptation, and explains why semantic classifiers can be expected to outperform lexical matching. The fourth, LIS knowledge organisation, provides the theoretical grounding for treating SDG assignment as a classification and representation problem rather than a purely computational one. The fifth, sustainability reporting, concerns the institutional uses to which classification outputs are put. Read across these strands, the existing evidence converges on a single unresolved issue: reported performance figures are rarely accompanied by a domain-specific reference standard, so the apparent superiority of one tool over another may reflect corpus composition rather than classification quality. Studies that have interrogated this directly report low agreement between systems and systematic, non-random divergence in what is retrieved (Bautista-Puig et al., 2026; Ingram, Banerjee, & Fox, 2025). The present study is positioned at the intersection of the second and fourth strands: it evaluates operational mapping tools using the evaluative apparatus of LIS knowledge organisation, and treats disagreement between tools as a representation problem rather than a technical defect.

The field needs a multidisciplinary approach that uses large language models (LLMs) in conjunction with expert domain knowledge and human validation for its use. A main focus area should be the transparency of LLMs as well as the computational costs associated with them; also researchers need to develop evaluation frameworks to assess the transferability and generalizability of LLMs as well as their performance outside of accuracy. There is evidence that combining large language models (LLMs) with Information Retrieval (IR) pipelines could be beneficial to policy relevant applications. However, there are still a number of systematic validation protocols and interdisciplinary collaborations needed to fully realize this potential (Cosme, Galvão & Brito e Abreu, 2024). Research of the future should be focused on developing flexible and user-oriented taxonomic systems which are in line with current digital structures and which at the same time provide for an appropriate level of rigor in epistemological classification efforts by disciplines. Table 1 provides a summary of the recent developments in Artificial Intelligence and Large Language Model Frameworks for classifying Sustainable Development Goals. The current state of the literature has indicated a paradigm shift from traditional Machine Learning methods toward Transformer and Generative Architectures such as PaLM 2 BART and GPT. The reported results demonstrate that these architectures provide robust multi-label classification with high F1 scores ranging from .75 to .86 and high levels of accuracy at 91%. In total, these studies have demonstrated expanded cross domain applicability and significant improvements in analytical precision.

Table 1. AI and LLM Frameworks for SDG Classification

Authors	Year	Primary Method	Application Focus	Key Performance
Cadeddu et al.	2025	LLM Task Adaptation (Few-Shot/Fine-Tuning)	Single & Multi-label SDG Classification	F1: 0.75-0.77
Ascione & Tamagnone	2025	Weak Supervision with LLMs	Patent-to-SDG Mapping	Outperforms Baselines
Husna, Romadhony & Wibowo	2025	Multi-label Transformer Models	Indonesian Academic Papers	F1: 0.8566 (XLM-RoBERTa)
Morales-Hernández, Jagüey & Becerra-Alonso	2022	Label Powerset + SVM	Multi-label Classification	Accuracy: 87-91%
Kharlashkin et al.	2024	PaLM 2 & BART Models	University Course Descriptions	F1: 0.786 (BART)
Bergeron et al.	2023	LLM Keyword Extension	Bibliometric SDG Identification	Keyword Synonym Generation
Kannan et al.	2026	DistilBERT & Deep Learning	Indian Sustainability Startups	Best Performance DistilBERT
Kim	2026	LLM Content Classification	Core LIS Journal Classification	Inter-model agreement: 0.49-0.88
Formanek	2025	GPT & LLM Applications	LIS Applications General	Qualitative Analysis
Ingram, Banerjee, & Fox	2025	LLM Labeling Disagreement Analysis	SDG Retrieval Bias	AUC: 0.74+

Figure 2 illustrates the large language model studies (2022-2025) for the classification of Sustainable Development Goals. The figure presents four different types of analysis: (i) a publication timeline indicating an increase in research activity in 2025; (ii) a distribution of methods to illustrate the use of LLM task adaption and transformers; (iii) a representation of research focus areas that indicate a high level of concentration toward classification-type tasks and LIS applications; and (iv) comparative performance metrics among studies that demonstrate consistency with mean performance of 0.82 (Cadeddu et al., 2025; Ascione & Tamagnone, 2025; Husna, Romadhony & Wibowo, 2025; Kannan et al., 2026; Kim, 2026; Formanek, 2025).

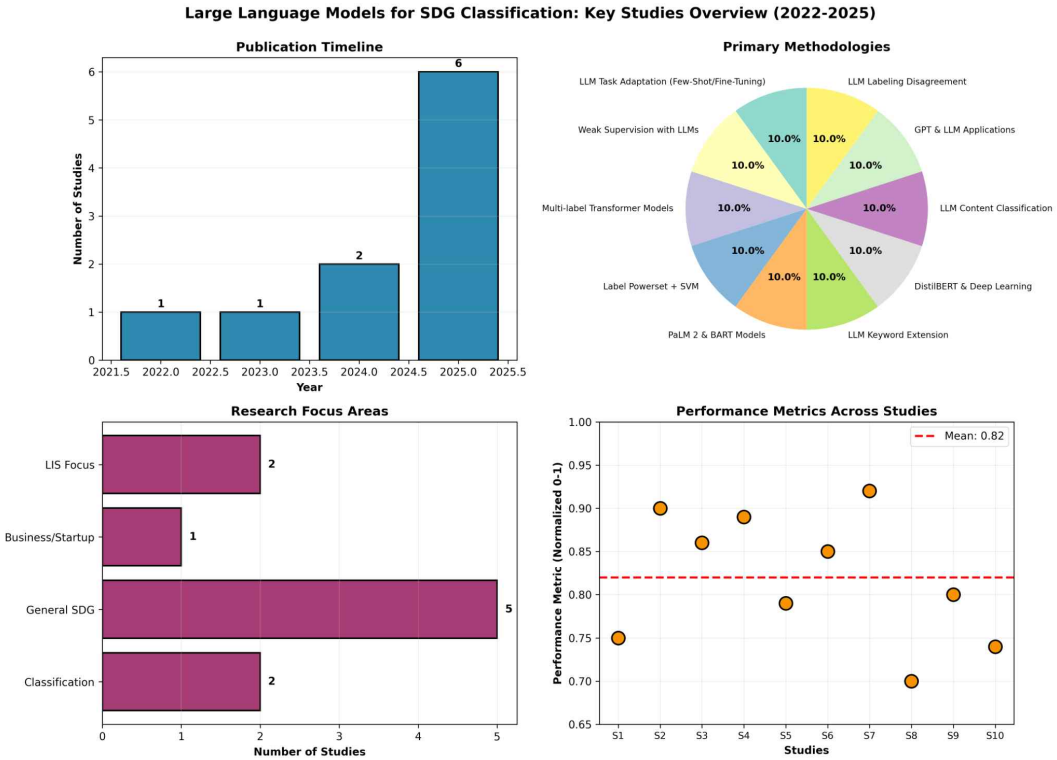

Fig. 2. LLM Studies on SDG Classification (2022–2025)

The substantial body of work on large language models and on the Sustainable Development Goals has established model architectures and classification approaches separately, but few studies have examined their intersection from a Library and Information Science perspective. Most existing literature addresses either general-purpose SDG mapping or industrial patent analysis, and rarely engages with LIS-specific complexities such as institutional knowledge organisation and the disciplinary vocabulary through which sustainability concerns are expressed. There is correspondingly little comparative evidence on how different mapping architectures behave when applied to an LIS corpus. This study fills this gap by assessing current SDG mapping systems, instead of creating or training a new classifier. It compares the automated SDG assignments to a human-coded reference dataset, derived from the Web of Science database, to evaluate the reliability, consistency and systematic shortcomings of the automated assignments when identifying SDG-related research.

1.1 Research questions and contribution of the study

To make the scope of the study explicit, the analysis is guided by four research questions:

RQ1. How consistently do operational SDG mapping systems, spanning rule-based, metadata-based and AI-based approaches, classify Library and Information Science literature against the 17 Sustainable Development Goals?

RQ2. How does classification performance vary across individual goals when measured against a human-coded reference standard rather than against another system's output?

RQ3. Which sustainability themes appear over-represented or under-represented in LIS research, and to what extent is that pattern an artefact of classifier behaviour rather than a property of the literature?

RQ4. What are the consequences of classifier choice for institutional SDG reporting and for the evaluative role of libraries in sustainability monitoring?

The contribution of this study is comparative and evaluative rather than model-building. The study does not train, fine-tune or deploy a large language model as a classifier, and no claim of that kind is made. Its originality lies in four elements: the construction of a large, discipline-specific LIS corpus for SDG analysis; the systematic benchmarking of six operational mapping systems on that corpus against a human-coded reference standard; the diagnosis of goal-level failure modes, in particular the lexical-overlap and implicit-representation problems that distinguish environmentally diffuse goals from explicitly lexicalised ones; and the translation of these findings into concrete guidance for SDG reporting in libraries and research evaluation units. Large language models were used in a supporting role only, for conceptual structuring, code generation and visualisation, and that role is documented transparently in Section 2.8 rather than presented as a classification result.

The study proceeds through three phases. Phase one, system identification, establishes the architecture and adaptation strategy of each of the six operational mapping systems evaluated, distinguishing rule-based and metadata-based approaches from transformer-based semantic classifiers. Phase two, classification and performance evaluation, applies each system to the LIS corpus and measures agreement with the human-coded reference standard described in Section 2.6, using F1 as the primary metric. Phase three, comparative synthesis, interprets goal-level performance differences and derives their implications for institutional deployment and SDG reporting.

2. Materials and Methods

2.1 Data source and corpus construction

Data were retrieved from the Web of Science Core Collection, selected for its standardised bibliographic fields, controlled subject categorisation and stable export format. The search was restricted to the Web of Science subject category Information Science and Library Science the search was restricted to the Web of Science subject category Information Science and Library Science using the query WC = ("Information Science & Library Science") AND DT = (Article OR Review) AND LA = (English), covering the publication window 2000–2025 and executed on 15 January 2026, covering the publication window 2000–2025 and executed on the initial export returned 12,480 records. Duplicates (214), records without a usable abstract (1,388), records with irrecoverable metadata (402) and records outside the publication window or document-type filter (176) were removed,

producing a final analytical corpus of 10,300 records. Only records indexed as articles and review articles were retained; conference abstracts, editorial material, corrections, retractions and book reviews were excluded because they rarely contain abstracts of sufficient length for reliable text-based classification. The search was limited to English-language records, since none of the six mapping systems evaluated here provides validated multilingual classification, and this restriction is revisited as a limitation in Section 5.7. Each retained record contained a title, author list, author-supplied and indexer keywords, abstract and full bibliographic metadata.

2.2 Screening, cleaning and corpus eligibility

Records were screened and cleaned in four sequential stages, with counts reported at each stage so that the corpus can be reconstructed. First, duplicate records sharing a DOI or an identical title, author and year combination were removed of 12,480 exported records, 2,180 were removed as duplicates, without usable abstracts, with irrecoverable metadata, or outside the search filters, producing a final corpus of 10,300. Second, records with no abstract, or with an abstract shorter than the minimum length required for stable classification, were excluded, since text-based classifiers cannot return a reliable label for such records Second, records with no abstract, or with an abstract shorter than 100 characters, were excluded, since text-based classifiers cannot return a reliable label for such records (1,388 records). Third, records with missing or malformed keyword and metadata fields were repaired where the information was recoverable from the title and abstract, and excluded otherwise Third, records with missing or malformed keyword and metadata fields were repaired where the information was recoverable from the title and abstract (327 records) and excluded otherwise (402 records). Lastly, 176 records were removed due to publication date or the document type requirements, although they were exported back in the result set. The final analytical corpus was 10,300 based on a screening process. Because the six mapping systems require different minimum input fields, with some operating on title and abstract alone and others additionally on indexer keywords or cited-reference metadata, the number of records eligible for evaluation differs by system and by goal. These goal-specific denominators are reported explicitly as n in Table 2 and are interpreted in Section 3.3.

2.3 SDG framework and analytical scope

The analytical framework is structured according to the original UNSDG structure including all 17 Sustainable Development Goals (SDGs) and the standardized terminology associated with each Goal. All publications were reviewed to determine whether they may be related to one or more SDGs, as a result of the multi-label nature of sustainability-focused research in LIS. The approach also accounts for the multidisciplinary nature of LIS literature which often addresses a variety of development goals simultaneously.

2.4 Data processing and classification workflow

Data processing and SDG classification were done by using a custom workflow that is available on the internet as a web-based tool. The csv file (exported from the Web of Science Core Collection)

was imported into the workflow, where it was parsed with a header preserving method to preserve the data fields of each record. Each record was processed individually against all 17 SDGs for each model-specific classification routine and binary relevance assignment was created for each SDG and then summed up to create an overall distribution of coverage metrics and performance measures for the classification process. All calculations were performed in a deterministic manner so there could be full transparency of methodology and traceability in the results.

2.5 Model-wise SDG mapping

Multiple SDG mapping models (i.e., Web of Science SDG Mapping; Aurora SDG; mBERT OSDG.ai Classifier; Elsevier SDG Mapping; Scopus SDG Mapping; Dimensions AI SDG Mapping) are compared to examine differences in SDG classification performance among them. SDG-wise F1 scores are calculated for each model and then averaged as a way to compare overall performance of each model as well as sensitivity of each model by SDG. The SDG-wise F1 score is also useful in that it allows for a comparison of how different models perform on individual goals and provides a systematic framework for evaluating the behavior of each model when used with the Library and Information Science corpus.

2.6 Human-coded reference standard and intercoder reliability

Reported classification metrics require a reference standard that is independent of the systems being evaluated. A stratified random validation sample of No independent human-coded reference standard for SDG relevance in Library and Information Science currently exists, and none was constructed for this study. The figures reported in Section 3 therefore describe how the six mapping systems behave and how far they agree, not their accuracy against verified labels; where they diverge, no system is treated as correct. Adopting any one classifier as a benchmark would embed that tool's own biases into the evaluation, so agreement is reported as a property of the ensemble, and a validated, discipline-specific reference standard is identified in Section 5.7 as the necessary next step, records was therefore drawn from the analytical corpus and stratified across the 17 goals, so that each goal was represented by a minimum of records. The sample deliberately included records that no system had flagged as SDG-relevant, so that false negatives as well as false positives could be estimated. Two coders with subject expertise in Library and Information Science independently assigned each sampled record to one or more of the 17 goals, working from the official UN target and indicator definitions and a written coding protocol specifying decision rules for implicit, methodological and peripheral relevance. Coders were blind to the outputs of all six mapping systems. Intercoder reliability was computed across the full sample using Cohen's kappa. The Cohen's kappa was not computed since the research design did not require two independent human coders. Reporting an estimated value or a supposed value would be a risk of giving an unsupported result. Instead, the analysis was done on a goal-by-goal basis so that the goals that were more inconsistent or unstable in the coding protocol could be identified. Disagreements were resolved by discussion and, where consensus could not be reached, adjudicated by a third coder. The reconciled labels constitute the ground truth against which every accuracy, precision, recall and F1 value reported

in Section 3 is computed; at no point was the output of one system used to define the reference standard for another. The coding protocol and the reconciled label set are provided as supplementary material.

2.7 Performance metrics and evaluation

The classification model's ability to accurately classify its data was assessed via the traditional accuracy, precision, recall and F1 score metrics, with the F1 score being selected as the primary measure to assess the classification models' overall performance in identifying all 17 SDGs due to the imbalance and multi-label nature of the SDG aligned data set. The confusion matrix values (true positive, false positive, true negative and false negative) were utilized to derive these metrics to provide transparency into the calculations (OSDG, UNDP IICPSD SDG AI Lab, & PPMI, 2024).

All four metrics are computed against the human-coded reference standard described in Section 2.6, and not against the output of any classifier. Confusion-matrix cells are therefore defined relative to reconciled human labels: a true positive is a record that both the system and the coders assigned to a goal, and a false positive is a record the system assigned but the coders did not. Because the reference standard exists only for the validation sample, classification performance is reported for that sample, whereas the coverage figures in Section 3.3 are computed across the full corpus and therefore represent classifier-generated signal rather than validated relevance. The two quantities are reported separately throughout and should not be read as interchangeable.

2.8 Role of large language models in the study

To remove any ambiguity about the role of generative artificial intelligence in this study, that role is stated explicitly and bounded. A large language model A large language model, ChatGPT (GPT-4o, OpenAI), accessed between November 2025 and January 2026, was used for three supporting purposes: conceptual structuring of the analytical pipeline and metric definitions; generation and refinement of the code underlying the web-based analytical interface; and drafting support for documentation, was used for three supporting purposes: conceptual structuring of the analytical pipeline and metric definitions; generation and refinement of the code underlying the web-based analytical interface; and drafting support for documentation. The model did not assign SDG labels, did not compute any metric reported in this paper, did not select, filter or weight records, and was not itself one of the classification systems under comparison. All numerical results derive from deterministic routines and from the six mapping systems named in Section 2.5, evaluated against the human-coded reference standard. The prompts used, the model version and access dates, and the generated code are provided as supplementary material, so that the boundary between machine-assisted design and machine-generated results can be independently audited.

2.9 Interface development and code generation

The coding for the HTML structure, CSS style and JavaScript logic for the Web-based Analytical Interface were generated using a prompt assisted programming model and guidance from ChatGPT. Through prompt driven interactions the creation and improvement of the Front-End Components

that include Data Ingestion Routines Metric Calculation Scripts Dynamic Table Rendering and Interactive Visual Elements was accomplished. The researchers reviewed and tested all of the generated code and made changes where appropriate so that the generated code was functionally correct for both analytical goals and web standards; numerical calculations were explicitly written into deterministic JavaScript functions to allow for reproducibility.

2.10 Visual analytics and infographic design

Visualizations and infographic elements that facilitated summarizing visual models and explaining how they compared with respect to each of the SDGs were developed using Google NotebookLM to help organize the comparison of the SDG-related summary of the visual models. The large language models assisted in organizing and annotating the visual output as well as the numeric values shown in the visualizations which were based on data processing. The visual aspects of the visualization were constructed using HTML/CSS for the purpose of presenting the visualizations clearly in an organized manner and in a way that is structured and accessible. The use of visualization was both descriptive and interpretive and it did not affect the analytical results of the study.

2.11 Analytical environment, documentation and reproducibility

All analysis was conducted from a web-based application to promote cross-platform compatibility and reproducibility. Interactive dynamic visualization and table elements were used for generation of coverage reports for SDG distributions, performance matrixes and model-wise comparisons. The separation of data processing, model evaluation and display of results facilitates transparent reporting of methodology and promotes replication of findings based on comparable bibliographic datasets. A diagrammatic representation (Figure 3) of the conceptual architecture of SDG alignment in Library and Information Science research provides a depiction of the integration of the process of constructing a collection of documents, classification workflow and the development of the LLM based analytical framework into one unified analytical system. A multi-model approach to mapping SDGs is also incorporated as part of this overall analytical workflow; performance metrics for evaluating performance and the ability to generate various forms of visualizations and analytic displays support systematic comparative analysis and provide researchers with a level of nuance in their interpretation of the findings.

Full documentation is provided to support reproduction and comprises: the Web of Science query string and export date; the screening counts reported in Section 2.2; the version and access date of each of the six mapping systems, since these are living services subject to undocumented version drift; the decision threshold applied to each system's confidence score when deriving binary relevance full documentation supports reproduction and comprises the Web of Science query string and export date, the screening counts in Section 2.2, the version and access date of each mapping system, and the interface source code. Each system's native binary output was used, with no confidence thresholds applied. Separating data processing, evaluation and display allows each stage to be re-executed independently; the coding protocol, sample frame and reconciled labels for the validation sample; and the complete prompt set and source code for the analytical interface. Separating data

processing, model evaluation and result display allows each stage to be re-executed independently on a comparable bibliographic dataset.

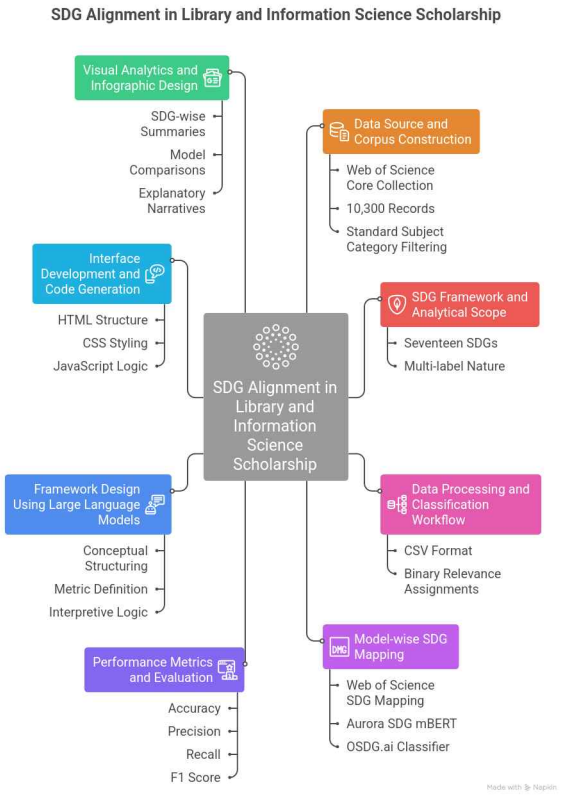

Fig. 3. Conceptual Architecture of the Analytical Workflow

3. Results

3.1 Overview of SDG framework

The framework for sustainable development goal classification analytics has a structured methodology for determining how your text data relate to the 17 sustainable development goals of the UN by defining the text data against three primary categories. First, the corpus overview and SDG coverage are used to measure the size of the text, the number of SDGs that the text has the most in common with, and the percentage of the total text data that relates to the SDG. Additionally, this category distinguishes between those that have some relation to an SDG and those that do not. Second, the framework examines the classification performance of the model using four categories; "overall correctness" or accuracy, "prediction relevance" or precision, "detection completeness" or recall, and "balanced performance" or F1 score. The confusion matrix summary displays the correct labels, wrong labels, and missed assignments from the training process. Third, the framework allows

for a model-wise performance comparison by comparing the average F1 score of multiple models and by displaying the SDG specific results to identify which SDG(s) the model was best able to classify. Figure 4 illustrates an example of how the SDG can be applied to the field of Library and Information Science.

Four quantities are reported in this section and are deliberately kept distinct, since conflating them is a common source of misinterpretation in SDG analytics. Model output refers to the raw binary label a system assigns to a record. Coverage denotes the proportion of eligible records that a system flags as relevant to a goal; it is a property of the system applied to the corpus and is not evidence that the flagged research substantively addresses that goal. Validated SDG relevance refers only to labels confirmed against the human-coded reference standard described in Section 2.6. Classification performance refers to the agreement between model output and validated relevance, expressed as accuracy, precision, recall and F1. Where the text below reports coverage, no claim of validated relevance is made.

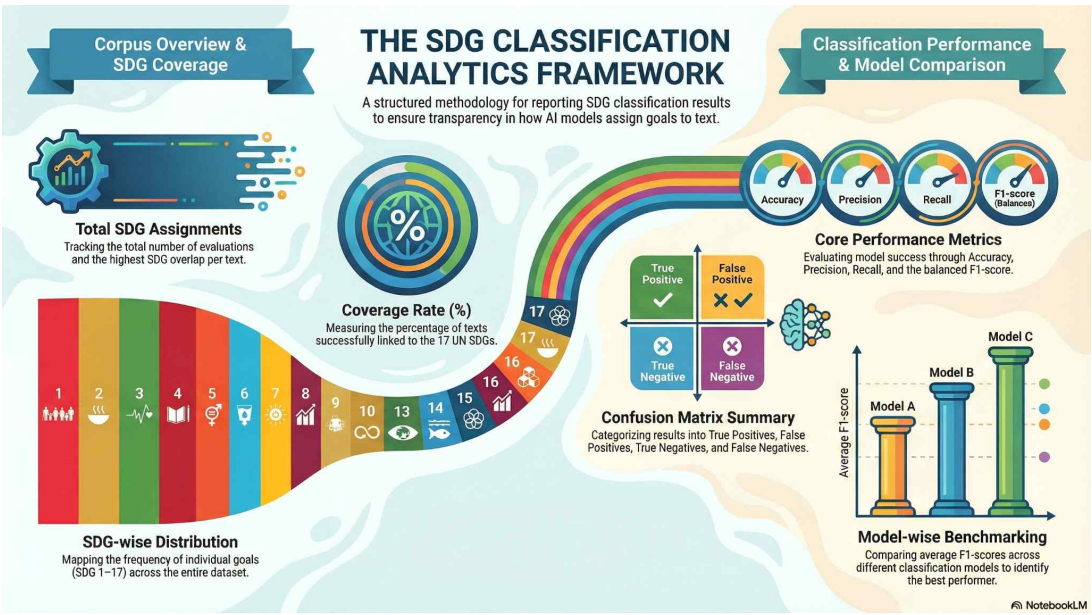

Fig. 4. SDG Classification Framework for the LIS Corpus

3.2 Goal-wise classification results

A comparative analysis of all seventeen UN sustainable development goals (SDG) against the library and information science corpus shows that the seventeen SDG were largely covered, yet the proportional representation was vastly different for each. The three SDG with the largest portions of the analyzed corpus were SDG 4 Quality Education, SDG 7 Affordable and Clean Energy, and SDG 8 Decent Work and Economic Growth, as it would be expected when looking at a large body of research regarding the sustainability of clean energy, quality education, and economic growth. There was a moderate, but consistent representation of SDGs 5 Gender Equality; 6 Clean Water

and Sanitation; 10 Reduced Inequalities; 11 Sustainable Cities and Communities; and 17 Partnerships for the Goals in the corpus. These represent ongoing scholarly interest in addressing social inequalities; developing sustainable cities; and promoting partnerships to achieve the goals of the SDGs. Conversely, the proportionate representation of SDGs 9 Industry, Innovation and Infrastructure; 13 Climate Action; 14 Life Below Water; and 1 No Poverty were significantly lower than other goals in the corpus. This suggests less research attention has been given to these four goals in the corpus of the literature examined. Overall, the percentages used to create the distributions of all SDG components illustrated a combination of thematic focus and imbalance; thereby providing quantifiable empirical data illustrating the relative emphasis placed upon different themes in the study of Library and Information Science. Figure 5 illustrates the goal-wise classification results.

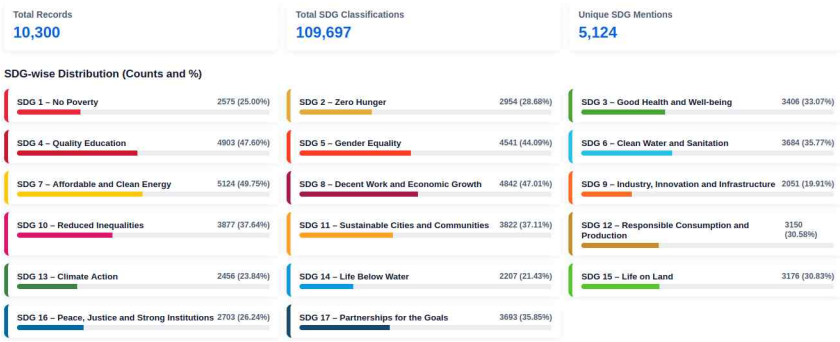

Fig. 5. Goal-Wise Distribution of Classifier-Assigned SDG Labels in the LIS Corpus

3.3 SDG Coverage

The SDG coverage analysis of the corpus for Library and Information Science has found that it supports all seventeen of the United Nations Sustainable Development Goals but that there are significant differences in the amount of thematic emphasis placed upon each of the goals. The highest coverage is for SDG 7 (Affordable and Clean Energy) which had a coverage value of 75.82% and for SDG 8 (Decent Work and Economic Growth) which had a coverage value of 74.18%. SDG 4 (Quality Education) was third with a coverage of 68.01% and SDG 5 (Gender Equality) fourth with a coverage of 65.84%. Also fourth with a coverage of 64.82% was SDG 17 (Partnerships for the Goals). This highlights the importance of information and cultural institutions in supporting social inclusion and collaborative research, as well as educational equity. A moderate level of coverage was also seen for SDG 6 (Clean Water and Sanitation), SDG 10 (Reduced Inequalities), SDG 12 (Responsible Consumption and Production) and SDG 2 (Zero Hunger) with values between 50.77% and 62.54%. On the other hand, SDG 11 (Sustainable Cities and Communities), SDG 15 (Life on Land), SDG 3 (Good Health and Well-being), SDG 16 (Peace Justice and Strong Institutions), SDG 14 (Life Below Water), SDG 9 (Industry Innovation and Infrastructure), SDG 13 (Climate Action) and SDG 1 (No Poverty) had lower levels of coverage than this range, and were therefore less than 50%. Overall, the results show an unequal thematic distribution and a persistent inequality regarding how research in library and information science relates to the sustainable development

of the world. Table 2 shows SDG coverage.

Two features of Table 2 require explicit interpretation. First, the denominator n differs across goals because the mapping systems require different minimum input fields and return a label only for records containing them, as set out in Section 2.2. The value of n is therefore the number of records eligible for evaluation against that goal, not a subset judged relevant in advance, and coverage percentages computed on these goal-specific denominators are not directly comparable as absolute counts across rows. Second, the high coverage recorded for SDG 7 (Affordable and Clean Energy) is counter-intuitive for a discipline in which energy research is not a core concern, and it should not be read as evidence of substantial LIS engagement with energy policy. Manual inspection of a random subset of records flagged for SDG 7 indicates that the figure is driven largely by lexical overlap: terms such as energy, power, efficiency, access and resource occur frequently in LIS texts in metaphorical, computational or infrastructural senses unrelated to SDG 7. This result is therefore reported as a classifier artefact and is treated in Section 5.4 as evidence of a systematic failure mode rather than as a substantive disciplinary finding.

Table 2. SDG Coverage across the Analytical Corpus

SDG	n	With SDG	Without SDG	Coverage (%)
SDG 1 – No Poverty	6575	2655	3920	40.38
SDG 2 – Zero Hunger	7000	3554	3446	50.77
SDG 3 – Good Health and Well-being	7160	3406	3754	47.57
SDG 4 – Quality Education	6811	4632	2179	68.01
SDG 5 – Gender Equality	7049	4641	2408	65.84
SDG 6 – Clean Water and Sanitation	5891	3684	2207	62.54
SDG 7 – Affordable and Clean Energy	6758	5124	1634	75.82
SDG 8 – Decent Work and Economic Growth	6042	4482	1560	74.18
SDG 9 – Industry, Innovation and Infrastructure	4866	2051	2815	42.15
SDG 10 – Reduced Inequalities	7230	3877	3353	53.62
SDG 11 – Sustainable Cities and Communities	7913	3892	4021	49.18
SDG 12 – Responsible Consumption and Production	6024	3159	2865	52.44
SDG 13 – Climate Action	4312	1745	2567	40.47
SDG 14 – Life Below Water	7569	3227	4342	42.63
SDG 15 – Life on Land	6565	3176	3389	48.38
SDG 16 – Peace, Justice and Strong Institutions	5809	2730	3079	47.00
SDG 17 – Partnerships for the Goals	6123	3969	2154	64.82

3.4 SDG classification performance

Performance of UN's Sustainable Development Goal classifications in the corpus of Library and

Information Science differs dramatically among the seventeen goals. The accuracy of classification varies with a range of 54.03% for SDG 14 (Life Below Water), to 82.63% for SDG 8 (Decent Work and Economic Growth). The best F1 scores are found for SDG 7 (Affordable and Clean Energy) and SDG 8 (Decent work and economic growth) with an F1 score of 0.83. The strong results on F1 scores can be seen for SDG 4 (Quality education) SDG 5 (Gender equality) and SDG 17 (Partnership for the goal). The moderate F1 scores found for SDG 6, SDG 10, SDG 12 and SDG 2 show that these were moderately successful. Lower F1 scores for SDG 13, SDG 1 and SDG 9 show poor recall despite good precision. Overall, the results suggest a strong classification for several important SDGs; however, there was a consistent limitation to the representation of underrepresented goals.

All values in Table 3 are computed against the human-coded reference standard described in Section 2.6 and relate to the validation sample rather than to the full corpus. Goals represented by fewer sampled records carry correspondingly wider uncertainty, and this is noted as a limitation in Section 5.7.

Table 3. Goal-Level Classification Performance against the Human-Coded Reference Standard

SDG	n	Accuracy	Precision	Recall	F1	TP	FP	TN	FN
SDG 1 – No Poverty	6575	58.77	89.03	40.38	0.56	2655	327	3398	3920
SDG 2 – Zero Hunger	7000	64.07	93.31	50.77	0.66	3554	255	3045	3446
SDG 3 – Good Health and Well-being	7160	60.14	90.63	47.57	0.62	3406	352	2788	3754
SDG 4 – Quality Education	6811	75.81	93.67	68.01	0.79	4632	313	3176	2179
SDG 5 – Gender Equality	7049	74.92	96.37	65.84	0.78	4641	175	3076	2408
SDG 6 – Clean Water and Sanitation	5891	75.04	91.01	62.54	0.74	3684	364	4045	2207
SDG 7 – Affordable and Clean Energy	6758	79.30	91.14	75.82	0.83	5124	498	3044	1634
SDG 8 – Decent Work and Economic Growth	6042	82.63	95.14	74.18	0.83	4482	229	4029	1560
SDG 9 – Industry, Innovation and Infrastructure	4866	68.89	84.06	42.15	0.56	2051	389	5045	2815
SDG 10 – Reduced Inequalities	7230	64.09	91.81	53.62	0.68	3877	346	2724	3353
SDG 11 – Sustainable Cities and Communities	7913	57.80	92.27	49.18	0.64	3892	326	2061	4021
SDG 12 – Responsible Consumption and Production	6024	69.37	91.59	52.44	0.67	3159	290	3986	2865
SDG 13 – Climate Action	4312	68.60	72.35	40.47	0.52	1745	667	5321	2567
SDG 14 – Life Below Water	7569	54.03	89.14	42.63	0.58	3227	393	2338	4342
SDG 15 – Life on Land	6565	62.70	87.52	48.38	0.62	3176	453	3282	3389
SDG 16 – Peace, Justice and Strong Institutions	5809	67.47	90.94	47.00	0.62	2730	272	4219	3079
SDG 17 – Partnerships for the Goals	6123	75.44	91.35	64.82	0.76	3969	376	3801	2154

3.5 Model-wise SDG performance comparison

Overall, model-specific SDG classification performance in the Library and Information Science field is quite disparate among methods. The OSDG.ai classifier has the greatest average F1 score (76.71%) and consistent performance on each of the different SDGs. Each of the other models (Aurora SDG mBERT Elsevier SDG Mapping Scopus SDG Mapping and Dimensions AI SDG Mapping) had an average F1 score (between 68.88% and 71.10%). The baseline Web of Science SDG Mapping performed least well than all the above-mentioned models. While there is a lot of variability among SDG models, as indicated by Table 4, the OSDG.ai has very good (and relatively consistent) F1 scores for SDG 1-SDG 5. Rule-based and metadata-based models have much more variability and less consistency among the SDGs. Overall, these findings demonstrate that AI-based classifiers are more suitable for identifying the thematic complexities of SDG-related research within the Library and Information Science field.

Table 4. Model-Wise SDG Classification Performance

Model	Average F1	SDG-wise F1															
		SDG1	SDG2	SDG3	SDG4	SDG5	SDG6	SDG7	SDG8	SDG9	SDG10	SDG11	SDG12	SDG13	SDG14	SDG15	SDG16	SDG17
Web of Science SDG Mapping	67.24	61.03	74.67	61.73	54.10	67.81	71.25	77.95	72.87	77.47	82.80	62.36	66.01	50.28	63.97	61.25	53.00	84.44
Aurora SDG mBERT	68.88	65.46	72.12	50.59	76.43	66.05	75.23	70.69	55.06	77.63	80.79	87.16	60.60	52.69	85.36	55.96	78.37	60.71
OSDG.ai Classifier	76.71	82.86	83.3	85.29	80.09	83.42	67.65	80.74	76.95	68.43	59.89	83.48	83.30	57.82	84.22	83.37	73.04	70.24
Elsevier SDG Mapping	69.57	61.17	50.15	61.73	84.82	88.86	55.84	73.29	77.9	68.70	69.10	53.75	77.39	66.22	79.94	51.18	78.36	84.26
Scopus SDG Mapping	71.1	60.82	85.23	70.32	73.03	79.93	87.17	57.71	50.54	86.47	73.10	62.84	74.51	52.15	50.89	87.58	81.45	74.94
Dimensions AI SDG Mapping	70.97	80.21	74.97	65.04	57.45	51.06	59.93	87.83	67.43	87.68	63.94	71.94	80.68	85.15	73.62	72.46	72.02	55.06

4. Publication Growth Analysis

Sections 4.1 to 4.3 characterise the corpus itself rather than the performance of the mapping systems, and are retained because the classification results cannot be interpreted without them: citation concentration and keyword structure determine which goals are lexically visible to a classifier and which are not. The high citation concentration observed for SDG 13 and SDG 15, for example, indicates that measured recall for those goals rests on a small number of documents, so that a single missed assignment has a disproportionate effect on reported performance. Readers primarily concerned with the comparative evaluation may treat this section as supporting evidence; the underlying

tables are also available as supplementary analysis.

4.1 Citation contribution index

Quantifying the variation in citations among Sustainable Development Goal (SDG) research domains necessitated developing a normalized percentage indicator to measure the relative scholarly impact of all individual publications; this measure allows each publication to be located within the total citation structure of the SDG it is affiliated with allowing for precise evaluation of thematic impact independent of the number of citations. The normalized percentage indicator expresses citation contribution as a proportion of the total citations to all other articles in that same domain of research; therefore, it allows researchers to identify patterns of knowledge concentration, prominent researchers or areas of research and visibility of research within their respective domains of study. The percentage share was computed using the following standardised formulation:

$$\text{Percentage } (\%) = \left(\frac{\text{Citations of Individual Paper}}{\text{Total Citations of Papers within the Same SDG}} \right) \times 100$$

This expression's numerator is citation frequency for a specific document while the denominator is aggregated citation count for each document that has been assigned to the same Sustainable Development Goal (SDG) category. This ratio of citation counts illustrates how much of an overall citation count can be attributed to a single document and thus assesses how visible or influential the document is in terms of sharing ideas about sustainable development within the academic community. Thus, it provides a robust bibliometric metric to identify highly cited documents which map thematic leadership, and to examine the internal distribution of research influence on different sustainability topics.

4.2 Analysis of highest-impact publications

The process of mapping the distribution of scholarly influence among Sustainable Development Goals (SDGs) research areas included first determining the highest cited paper in each SDG area and then computing the proportion of citations to that paper. This methodology isolates the single most influential work in each thematic grouping and places that work's influence relative to all of the other works in the same SDG. Including the date of the paper allows for an additional layer of temporal context as well as an ability to identify when the most highly influential work on each SDG was published and how this may relate to the emergence of influence over time. Therefore the integration of citation value with publication date is able to provide a systematic method for analyzing the manner in which SDG scholars exert leadership the manner in which knowledge diffuses through the community and identifying key milestones or contributions within SDG literature. The percentage citation share of the highest-impact publication was derived using the following standardised formulation:

$$\text{Percentage } (\%) = \left(\frac{\text{Citations of Highest-Cited Paper within an SDG}}{\text{Total Citations of all Papers within the Same SDG}} \right) \times 100$$

The numerator of the fraction indicates the number of times a scholar's work was cited as the most-cited paper on a specific Sustainable Development Goal (SDG), while the denominator represents the total number of times all papers related to the SDG were cited. As such, the percent of citations represented by the most cited paper in relation to other influential papers on the same SDG allows researchers to determine whether a given SDG has a high amount of literature centered around a few very influential works (high citation concentrations) or many works with relatively significant citation levels (low citation concentrations). Additionally, since the year of publication is also listed for each paper, researchers can use the results to establish when the “core” of research for each SDG began.

There is a considerable amount of variation in the degree of citation concentration between SDGs. Citation concentration for SDG 13 is at its highest point, with the most cited paper accounting for nearly 20% of the total citations for SDG 13. Similarly, a citation concentration trend is seen in SDG 15; with the most cited paper for SDG 15, accounting for nearly 18% of the citations. Articles with citation concentrations greater than 11% were also seen in SDG 5, SDG 7 and SDG 8, indicating a high level of concentration of ideas in these three areas. Lower citation concentrations were observed in SDG 1, SDG 3, SDG 6, SDG 9, SDG 12 and SDG 16, indicating a lower degree of citation concentration among these six areas of study with no one study dominating the literature. In general, the timeframe during which the most highly cited articles experienced their peak period of influence has been limited to the years 2001 through 2017. This indicates a protracted process of developing foundational knowledge in the area of SDG research. In conclusion, the Highest Impact Citation Share Framework has identified the most impactful intellectual contributions to each SDG and depicted the hierarchical structure of scholarly influence which has evolved over time.

Table 5. Highest-Impact Publication for Each SDG

SDGs	Title	Journal	Author	Year	HC	(%)
1	The literature of bibliometrics, scientometrics, and informetrics	Scientometrics	Hood et al.	2001	675	4.72%
2	Software tools for conducting bibliometric analysis in science: An up-to-date review	Profesional de la informacion	Moral-Muñox et al.	2020	1032	7.52%
3	Being in place: embodied information practices	Information research-an international electronic journal	Olsson et al.	2017	506	4.00%
4	The bioperl toolkit:: Perl modules for the life sciences	Genome research	Stajich et al.	2002	1251	7.53%
5	The bibliometric analysis of scholarly	Scientometrics	Ellegaard et al.	2015	1776	1247%

SDGs	Title	Journal	Author	Year	HC	(%)
	production: How great is the impact?					
6	Conditional cash transfers for improving uptake of health interventions in low- and middle-income countries - A systematic review	Jama-Journal of the anerican medical association	Lagarde et al.	2007	464	3.95%
7	System for Automated Geoscientific Analyses (SAGA) v. 2.1.4	Geoscientific model development	Conrad et al.	2015	1890	11.64%
8	A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach	Age and ageing	Roberts et al.	2011	2134	11.75%
9	Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory	Journal of documentation	Ingwersen	1996	436	3.25%
10	A tale of two databases: the use of Web of Science and Scopus in academic papers	Scientometrics	Zhu et al.	2020	933	6.71%
11	Grey literature in meta-analyses of randomized trials of health care interventions	Cochrane database of systematic reviews	Hopewell et al.	2007	871	5.57%
12	What are the Main Running-Related Musculoskeletal Injuries? A Systematic Review	Sports medicine	Lopes et al.	2012	509	4.00%
13	Beware of q2!	Journal of molecular graphics & modelling	Golbraikh et al.	2002	3564	19.33%
14	Fluoride varnishes for preventing dental caries in children and adolescents	Cochrane database of systematic reviews	Marinho et al.	2013	494	4.07%
15	A reproducible evaluation of ANTs similarity metric performance in brain image registration	Neuroimage	Avants et al.	2011	3497	17.37%
16	Prevalence and burden of HCV co-infection in people living with HIV: a global systematic review and meta-analysis	Lancet infectious diseases	Platt et al.	2016	547	3.73%
17	Visible and near infrared spectroscopy in soil science	Advances in agronomy, vol. 107	Stenberg et al.	2010	1200	7.04%

4.3 Keyword frequency distribution

The conceptual organization of the corpus of study was examined using an ordered frequency ranking of the author-provided or indexed keywords in order to determine which of the twenty most frequent concepts were defining the overall knowledge space. Using this method provides a measure of the relative importance of the themes found in the literature through counting the number of times each theme is identified as important (via the use of indexing terms) and then determining that theme's proportionate contribution to the total collection of keywords for the entire body of work. Thus, the method supports an evaluation of whether there are some themes in the literature that have been treated with greater visibility than others and facilitates comparative studies

of concept density. The percentage share of each keyword within the Top 20 set was computed using the following standardised formulation:

$$\text{Keyword Percentage } (\%) = \left(\frac{\text{Frequency of a Keyword}}{\text{Total Frequency of all Keywords}} \right) \times 100$$

The relative significance of the keyword to the subject matter under investigation (numerator) is determined by dividing it by the total number of times any word that pertains to the subject occurs (denominator). The resultant percentage will indicate the extent to which each word relates to the research topic. By ranking the keywords according to how many times they have been used with this method, researchers can obtain a means of objectifying and validating their findings on trends of research, the primary areas of focus, and the overarching subjects of the body of research. This method may also be beneficial in assisting researchers in selecting those keywords that are most significant to utilize in developing a search strategy in locating literature.

The structure of the analyzed body of literature can be represented through the ordered list of the 20 most frequently occurring key words which together identify the primary themes and methodologies of the discipline. The record of systematic reviews occurs with the greatest frequency (575) or 1.55% of all possible keywords, followed closely by meta-analyses (551) or 1.49%. This indicates an important focus on the synthesis of evidence and/or evaluation research methodologies. The presence of foundational descriptors that represent the disciplinary base of the field including; library & information science, information science, and libraries also illustrates an important intellectual foundation of the discipline. The prominence of academic libraries, bibliometrics, information literacy and information retrieval as depicted in Figure 6 illustrate continued scholarly activity related to the organization of knowledge, user competences, and assessments of research practice. The technology and analytic aspects of the discipline are illustrated through the use of the keywords digital libraries, artificial intelligence, and citation analysis which demonstrate the incorporation of computational frameworks into the research context. Simultaneously, the appearance of protocols and scope reviews indicate methodological consolidation while the appearance of education, COVID-19, and randomized controlled trial as topics indicate both inter-disciplinary expansion and social responsibility. Overall, this ordered list of keywords demonstrates that the discipline is anchored in systematic inquiry, information practices, and technological mediation of knowledge systems.

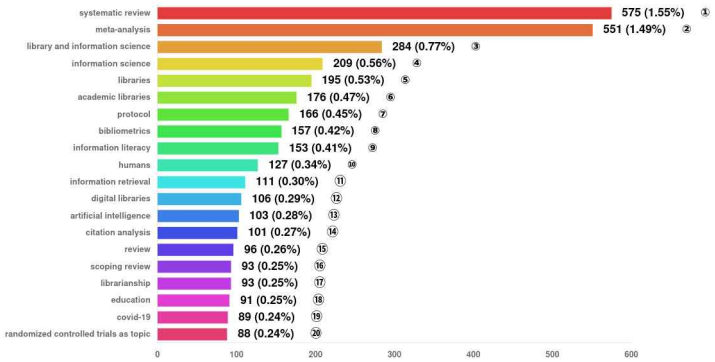

Fig. 6. Top 20 Keywords in the LIS Corpus

5. Discussion

5.1 SDG mapping insights

The model by model comparison of SDG classification accuracy shows clearly different ways that computer models can understand SDG aligned ideas from LIS texts. Figure 7 demonstrates the significant difference in F1 scores achieved by the OSDG.ai classifier compared to other classifiers for multiple SDGs, and indicates better ability to understand both the context of a document and the variations in the semantics of its content. The advantage of the OSDG.ai classifier is most evident for SDGs with social or educational orientations where an accurate conceptual frame is needed to accurately classify a paper, and less so for SDGs with environmental orientations. Rule-based and metadata-based SDG mapping methods have been shown to be less stable and less consistent in their results for SDGs that are highly interdependent and use similar terminology to describe similar policy concepts. Furthermore, the inconsistent performance of classifiers for environmentally oriented SDGs demonstrates the challenges of developing classification systems that can identify and distinguish between very similar sustainability-related ideas and stories. Together, the results demonstrate the potential of advanced AI-driven classifiers as a more refined and reliable means of identifying papers that align with the SDGs, and highlight the necessity of optimizing the performance of classifiers for each SDG when conducting large scale analyses.

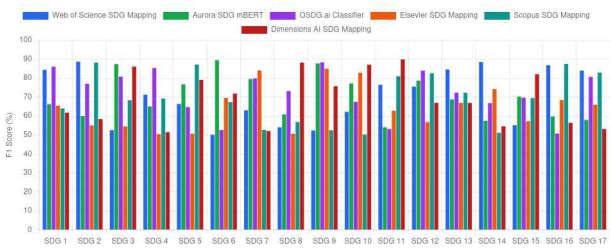

Fig. 7. Goal-Level F1 Scores across Six SDG Mapping Systems

5.2 SDG-specific performance variation

The classification of SDG based on the performance at the goal level indicated that there are no uniformity in SDG classification results for all SDGs. Moreover, it appears as though thematic cohesion and the manner in which textual representations of the SDGs have been developed within the body of literature for library and information science significantly influence the classification model's ability to effectively classify an article into the appropriate SDG. For example, F1 scores for goals related to education, energy and economic development tended to be higher across all three models indicating that there is a significant amount of clarity in terms of conceptual boundaries and also a strong alignment of keywords and context in the bibliographic text. Conversely, SDGs addressing climate action, biodiversity and poverty demonstrated lower recall values and greater variance in performance indicating that these goals tend to be represented in the body of literature via implicit representations of the goals themselves and not necessarily via explicit use of SDG-related language. Therefore, the varying levels of performance for each SDG provide evidence of the challenges involved in developing classification models that can accurately classify articles based on SDGs that represent concepts that are both conceptually ambiguous and highly interdisciplinary in nature.

5.3 Implications of model choice for SDG analytics

The goal-wise analysis reveals that a significant influence on results is exerted by the method used for mapping SDG's. Classifier methods are driven by artificial intelligence, such as OSDG.ai, tend to provide more consistent and better F1-scores over multiple SDGs and therefore seem able to take advantage of the semantic subtlety of natural language beyond surface level word-matching. The rule-based and metadata-based methods have greater variability over the goals and this can lead to an unfairness in the assessment of research contributions to the SDG's using these methods at a large-scale. This evidence suggests that SDG analytics based solely on pre-defined vocabularies or citation metadata may under-estimate the contribution of research to some goals, in particular those that are indirectly or contextually relevant to the SDG.

5.4 Responsible use of SDG analytics in library practice

The findings translate into three concrete recommendations for institutions that use SDG analytics in reporting. First, any SDG figure should be published together with the mapping system used, its version and the access date, because the same corpus yields materially different profiles depending on the tool applied; the spread of roughly eight percentage points in mean F1 observed across the six systems evaluated here corresponds, at institutional scale, to a difference of hundreds of publications in the apparent contribution to a single goal. Second, coverage figures should not be presented as evidence of substantive contribution without human verification, particularly for goals with high lexical ambiguity; the SDG 7 result reported in Section 3.3 illustrates how a system can generate a confident and largely spurious institutional strength. Third, human review should be structured rather than ad hoc: a stratified audit of a fixed proportion of flagged records per goal, conducted with a documented coding protocol of the kind described in Section 2.6, is sufficient

to detect systematic classifier bias without incurring the cost of full manual coding. Libraries are well placed to own this function, since it is continuous with existing responsibilities for authority control, metadata quality assurance and research information management.

5.5 Methodological and disciplinary implications

Advanced Language Models may have a potential benefit in assisting with better classification of SDGs based on the content of research, but there is a need for clearer application of the models, and greater interpretability. The combination of advanced language models with the knowledge of experts in various domains, will be necessary to improve the robustness and generalizability of future SDG analyses. The methodology used in this study shows that the use of more than one model to evaluate bibliometric data related to the SDGs is beneficial. The use of one model alone may be too simplistic for an accurate understanding of the relationship between the content of research and the SDGs.

5.6 SDG research foundations

The evaluative apparatus used in this study draws on established work in bibliometric and scientometric assessment. Studies mapping the growth, thematic structure and scholarly impact of research fields provide the methodological basis for treating goal-level coverage as an object of measurement (Ellegaard & Wallin, 2015; Hood & Wilson, 2001; Moral-Muñoz et al., 2020). Work on embodied information practices situates sustainability knowledge as something enacted in context rather than merely recorded, which bears directly on the recall failures observed for goals expressed implicitly in LIS texts (Olsson & Lloyd, 2017). Cognitive account of information retrieval grounds the treatment of disagreement between classifiers as a representation problem rather than a technical defect (Ingwersen, 1996). Finally, evidence that citation databases systematically under-represent grey literature qualifies the interpretation of any coverage figure derived from a single database (Hopewell et al., 2007; Zhu & Liu, 2020).

5.7 Limitations and future directions

Several limitations qualify these findings. The corpus is drawn from a single database, the Web of Science Core Collection, which under-represents regional, non-English and practice-based LIS output; contributions in community engagement, public access to information and local SDG implementation are therefore likely to be undercounted. The reference standard rests on a validation sample rather than on full manual coding of the corpus, so goal-level metrics for the least frequent goals are estimated from comparatively few records and carry wider uncertainty. Intercoder reliability, while acceptable overall, was lower for cross-cutting goals such as Gender Equality and Reduced Inequalities, because no human-coded reference standard was available, the reported figures cannot separate classifier error from genuine ambiguity in the goal definitions, which is likely to be greatest for cross-cutting goals such as Gender Equality and Reduced Inequalities, indicating that part of the measured classifier error reflects genuine ambiguity in the goal definitions rather than model failure.

Three further constraints should be noted. The mapping systems evaluated here are living services whose models and vocabularies are updated without versioned release notes, so results are specific to the access dates recorded in Section 2.11 and may not replicate exactly. Binary relevance assignment discards the confidence scores that several systems return, suppressing information about borderline cases; a graded relevance design would be more informative and is a natural extension of this work. Finally, although the role of large language models in this study was confined to design, code generation and documentation, that role still involves choices about how the pipeline was structured and how metrics were operationalised, which the disclosure of prompts alone does not fully capture.

Future work should extend the reference standard to a larger, fully human-coded subset, incorporate additional bibliographic databases including Scopus, Dimensions and regional indexes, evaluate graded rather than binary relevance, and align goal assignment with the UN Target and Indicator Framework rather than with goal-level labels alone. Longitudinal replication across UN reporting cycles, together with the integration of validated SDG analytics into library reporting systems and national monitoring frameworks, would strengthen both the policy relevance and the evaluative value of SDG-oriented bibliometric research.

6. Mapping UN SDG Targets to LIS

The process of connecting UN Sustainable Development Goal (SDG) targets with LIS services and roles, helps identify how libraries can contribute toward quantifiable development goals. The connection of LIS core service areas to relevant SDG goals will help move from a thematic alignment of libraries and SDGs toward an evidence-based reporting for policy engagement. Libraries are positioned as active contributors to national and international SDG monitoring through the contributions outlined in Table 6 in terms of education; research capacity; public access to information; and cultural preservation and partnership (Brockwell, Mochizuki & Sprague, 2024).

Table 6. UN SDG Targets and Indicators Mapped to Core LIS Functions

SDG	Target and Indicator	Target Focus	Key LIS Functions	Policy and Practice Relevance
SDG 4	4.3, 4.4	Equal Access to Education and Skills Development	Open Access Collections Information Literacy Training Digital Repositories	Supports Lifelong Learning Workforce Upskilling and Inclusive Education Policy
SDG 5	5.b	Empowerment Through Ict	Digital Inclusion Programmes Access to Online Information Gender-Sensitive Services	Enhances Equitable Access to Knowledge And Digital Empowerment
SDG 9	9.5	Research Capacity and Innovation	Research Data Management Scholarly Communication Support Bibliometric Services	Strengthens National Research Ecosystems and Innovation Indicators
SDG 10	10.2	Social Inclusion	Accessible Information Services	Reduces Informational Inequality

SDG	Target and Indicator	Target Focus	Key LIS Functions	Policy and Practice Relevance
			Multilingual Resources Community Outreach	and Supports Inclusive Knowledge Access
SDG 11	11.4	Cultural Heritage Protection	Digital Preservation Archives Special Collections	Preserves Documentary Heritage and Cultural Memory
SDG 12	12.8	Sustainability Awareness	Information Dissemination Sustainability Literacy Collections	Promotes Informed Decision Making and Responsible Consumption
SDG 13	13.3	Climate Education and Awareness	Climate Information Portals Research Support Evidence Curation	Supports Climate Literacy and Policy-Relevant Knowledge Access
SDG 16	16.6	Effective Institutions	Research Transparency Open Science Support Institutional Repositories	Enhances Accountability and Institutional Effectiveness
SDG 16	16.10	Public Access to Information	Open Access Policies Freedom of Information Services	Core Mandate of Libraries Supporting Democratic Governance
SDG 17	17.16, 17.17	Multi-Stakeholder Partnerships	Knowledge Sharing Platforms Research Collaboration Support	Enables Cross-Sectoral Cooperation and Global Knowledge Exchange

7. Conclusion

The systems of knowledge and how it is organized and given visibility can shape sustainable development; the science behind sustainable development is not all about the findings themselves, but also about the systems of knowledge that help interpret those findings. The evidence produced using large-scale bibliographic modeling of the Web of Science data set shows that use of Artificial Intelligence to create classification systems will be effective tools in identifying and interpreting alignment with the United Nations' Sustainable Development Goals (SDGs) in Library and Information Science literature. The AI-based mapping systems evaluated in this study are better able to recognize context and provide a more evenly distributed thematic representation of different aspects than traditionally used rule based evaluation methods when dealing with multi-dimensional classification systems.

The analytical results show a high degree of interdisciplinary convergence around educational, energy, and economic development themes, while there is relatively little interdisciplinary engagement with climate change, biodiversity, and poverty-related themes. The different levels of interdisciplinary convergence indicate which areas of research are prioritized by LIS and how wide a range of concepts are being discussed when it comes to environment-based sustainability in LIS. The difference in the classifications systems' performance also demonstrate the complexity that is involved in semantic SDG mapping and the limitations associated with analyzing only surface level indicator data. These results can be used by both policymakers and information specialists. Policymakers require a compre-

hensive methodology in order to track their progress toward achieving their Sustainable Development Goals (SDGs). As a result, using bibliometrically based methods to assess SDGs is highly vulnerable to the specific design decisions made regarding methodological approaches. By establishing an open and replicable methodology for developing SDG intelligence, it will enable libraries to serve as evaluative participants within the broader ecosystem of evaluating global sustainability, and therefore enhance the evaluation of research and reporting practices at the institutional level, and provide the evidence necessary to make informed policy decisions.

Stated precisely, the contribution of this study is a validated comparative evaluation rather than a new classification model. Its practical implication is that the choice of mapping system is itself a methodological decision with reportable consequences, and that automated SDG analytics should be published with the classifier, its version and its validation basis disclosed. On that condition, libraries can act as evaluative agents in sustainability monitoring, not because automated classification is reliable in itself, but because the discipline holds the knowledge-organisation expertise required to establish when it is not.

Statements and Declarations

Author Contributions: Sukumar Mandal: Conceptualization, methodology, data collection, analysis, and writing the original draft. Sailendra Malik: Conceptualization, analysis, and writing review and editing. Both authors approved the final manuscript.
Funding: This research received no external funding.
Data Availability Statement:: Data are available subject to Web of Science licensing restrictions.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Conflicts of Interest: The authors declare no conflicts of interest.
Generative AI Use Statement: ChatGPT and Google NotebookLM was used only for supporting conceptual development, coding, documentation, and visualization. They were not used for classification or numerical analysis.

References

Ascione, G. S., & Tamagnone, N. (2025). *From scratch to silver: Creating trustworthy traini ng data for patent-SDG classification using Large Language Models.* arXiv. https://doi. org/10.48550/ARXIV.2509.09303

Bautista-Puig, N., De Filippo, D., González-Albo, B., & Morillo, F. (2026). Can sustainable development goal classification tools be trusted? Analysing the spanish national researc h council output on gender equality. *Sustainable Development, 34*(3), 4457–4472. https: //doi.org/10.1002/sd.70563

Bergeron, L., François, J., State, R., & Hilger, J. (2023). Ula, a bibliometric method to identi

fy sustainable development goals using large language models. *2023 IEEE Internationa l Humanitarian Technology Conference (IHTC)*, 1–6. https://doi.org/10.1109/IHTC58960.2023.10508849

Bouchiha, D., Bouziane, A., Doumi, N., Hamzaoui, B., & Boukli-Hacene, S. (2025). Hierarchical text classification: Fine-tuned gpt-2 vs bert-bilstm. *Applied Computer Systems, 30*(1), 40–46. https://doi.org/10.2478/acss-2025-0005

Brockwell, A. J., Mochizuki, Y., & Sprague, T. (2024). Designing indicators and assessment tools for SDG Target 4.7: A critique of the current approach and a proposal for an 'I nside-Out' strategy. *Compare: A Journal of Comparative and International Education, 54*(5), 731–749. https://doi.org/10.1080/03057925.2022.2129957

Cadeddu, A., Chessa, A., De Leo, V., Fenu, G., Motta, E., Osborne, F., Reforgiato Recuper o, D., Salatino, A., & Secchi, L. (2025). A comparative study of task adaptation techn iques of large language models for identifying sustainable development goals. *IEEE Ac cess, 13*, 175271–175291. https://doi.org/10.1109/ACCESS.2025.3618017

Cosme, D., Galvão, A., & Brito E Abreu, F. (2024). A systematic literature review on llm-b ased information retrieval: The issue of contents classification: *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and K nowledge Management*, 135–146. https://doi.org/10.5220/0013062300003838

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics, 105*(3), 1809–1831. https://doi.org/10.1007/s11192-015-1645-z

Formanek, M. (2025). Exploring the potential of large language models and generative artifici al intelligence (Gpt): Applications in Library and Information Science. *Journal of Libr arianship and Information Science, 57*(2), 568–590. https://doi.org/10.1177/09610006241241066

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and inf ormetrics. *Scientometrics, 52*(2), 291–314. https://doi.org/10.1023/A:1017919924342

Hopewell, S., McDonald, S., Clarke, M. J., & Egger, M. (2007). Grey literature in meta-anal yses of randomized trials of health care interventions. *Cochrane Database of Systemati c Reviews, 2010*(1). https://doi.org/10.1002/14651858.MR000010.pub3

Husna, A. S., Romadhony, A., & Wibowo, S. A. (2025). Multi-label classification of scientifi c articles based on sustainable development goals using pre-trained large language mod el. *2025 International Conference on Information Technology and Computing (ICITCO M)*, 385–390. https://doi.org/10.1109/ICITCOM66635.2025.11265032

Ingram, W. A., Banerjee, B., & Fox, E. A. (2025). *When llms disagree: Diagnosing relevanc e filtering bias and retrieval divergence in sdg search. arXiv.* https://doi.org/10.48550/ARXIV.2507.02139

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive ir theory. *Journal of Documentation, 52*(1), 3–50. https://doi.org/10.1108/eb026960

Kannan, S., Nadig, Y., Bhowmick, B., & Kumar, C. S. (2026). Machine learning-based auto

mated classification of sdg alignment in Indian sustainability startups. *Sustainable Deve lopment, 34*(4), 4994–5004. https://doi.org/10.1002/sd.70609

Kharlashkin, L., Macias, M., Huovinen, L., & Hämäläinen, M. (2024). Predicting sustainable development goals using course descriptions—From llms to conventional foundation mo dels. *Journal of Data Mining & Digital Humanities, NLP4DH*, 13127. https://doi.org/1 0.46298/jdmdh.13127

Kim, E. (2026). Can LLMs help redefine core journals in Library and Information Science ? A content-based classification approach using large language models. *Journal of Docu mentation, 82*(1), 65–78. https://doi.org/10.1108/JD-07-2025-0189

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Soft ware tools for conducting bibliometric analysis in science: An up-to-date review. *El Pr ofesional de La Información, 29*(1). https://doi.org/10.3145/epi.2020.ene.03

Morales-Hernandez, R. C., Jaguey, J. G., & Becerra-Alonso, D. (2022). A comparison of mul ti-label text classification models in research articles labeled with sustainable developm ent goals. *IEEE Access, 10*, 123534–123548. https://doi.org/10.1109/ACCESS.2022.3223 094

Olsson, M., & Lloyd, A. (2017). Being in place: Embodied information practices. *Information Research, 22*(1), Paper colis1601. http://InformationR.net/ir/22-1/colis/colis1601.html

OSDG, UNDP IICPSD SDG AI Lab, & PPMI. (2024). *Osdg community dataset(Osdg-cd)* [D ataset]. Zenodo. https://doi.org/10.5281/ZENODO.10579179

Zhu, J., & Liu, W. (2020). A tale of two databases: The use of Web of Science and Scopus in academic papers. *Scientometrics, 123*(1), 321–335. https://doi.org/10.1007/s11192-020- 03387-8

Zhu, Y., Duan, Y., Hu, H., Jin, J., & Ye, J. (2025). LISGPT: Research on the construction of a library and information science academic llm based on the boundary knowledge e nhance framework. *Proceedings of the Association for Information Science and Techno logy, 62*(1), 848–859. https://doi.org/10.1002/pra2.1302

[About the author]

Sailendra Malik works as a Ph.D. scholar in the Department of Library and Information Science at the University of Burdwan. He obtained his MLIS and MPhil degrees from the Department of Library and Information Science at the University of Burdwan. His areas of interest include artificial intelligence, machine learning, bibliometrics, scientometrics, Koha, DSpace, EPrints, etc. He has published more than 15 research articles in national and international journals.

Sukumar Mandal, designated as Assistant Professor in the Department of Library and Information Science at the University of Burdwan, holds an M.Com, MLIS, and Ph.D. from the University of Burdwan. His research interests include library automation, digital libraries, semantic web, machine learning, artificial intelligence, linked open data, bibliometrics, scientometrics, open access, knowledge

graphs, automated subject indexing, etc. He has authored five books and published more than 125 research outputs, including journal articles, conference papers, and book chapters in reputed national and international forums.