Aligning and Quantifying Higher Education Institutions' Climate Actions With Nationally Determined Contributions Through AI-Enabled Data Discovery and Verification

PMPC Gunathilake[1*], Tilak Hewawasam[2], Jagath Gunatilake[3]

[1]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka

[2]Department of Geography, University of Peradeniya, Peradeniya, Sri Lanka

[3]Department of Geology, University of Peradeniya, Peradeniya, Sri Lanka

[*]Correspondence Author

DOI: https://doi.org/10.47772/IJRISS.2025.903SEDU0766

Received: 18 December 2025; Accepted: 24 December 2025; Published: 30 December 2025

ABSTRACT

Higher education institutions (HEIs) serve as knowledge producers and innovation ecosystems in climate governance, yet their contributions remain limited in formal accountability frameworks. This research presents a comprehensive framework that systematically discovers, quantifies, and reports HEIs' climate contributions through integrated automated web scraping, Artificial Intelligence, semantic modelling, and standardised data submission protocols. The system extracts climate action data from digital sources and structures it using a Climate Action Data Model (CADM). Each climate action undergoes automated impact quantification using standardised emission calculation methodologies aligned with IPCC protocols and localised emission factors. A standardised data submission mechanism enables systematic transmission of quantified interventions to established national Measurement, Reporting, and Verification (MRV) systems, facilitating integration into provincial and national climate transparency frameworks. Pilot implementation across nineteen Sri Lankan universities validates the framework's technical robustness and policy relevance. The system achieves automated data extraction with the precision of 0.87 and the recall of 0.82, while the RAG-based Large Language Model generates evidence-grounded climate action classifications with the factual grounding of 0.91, effectively minimising hallucinations. This research delivers a scalable, replicable architecture for establishing AI-enabled university climate data pipelines that bridge the gap between institutional sustainability efforts and formal national climate accounting, positioning universities as verifiable implementing nodes in global climate governance. By demonstrating that digitally documented climate actions can be systematically discovered, quantified, and integrated into national MRV systems, the framework establishes a pathway for transforming fragmented institutional initiatives into formally recognised contributions to national climate commitments under the Paris Agreement.

Keywords: Greenhouse Gas Inventories, Climate Governance, Retrieval-Augmented Generation (RAG), Higher Education Climate Action, Climate Data Modelling

INTRODUCTION

Climate change represents the defining challenge of the 21st century, demanding rapid transitions in energy, industry, agriculture, and education (IPCC, 2023). Higher-education institutions (HEIs) play a unique dual role: as generators of knowledge and innovation, and as organisations that directly influence environmental outcomes through their infrastructure, research, and community engagement (Tilbury, 2011; Findler *et al.*, 2019). Universities shape not only current environmental practices but also cultivate the climate-conscious workforce necessary for long-term societal transformation.

Under the Paris Agreement (UNFCCC, 2015), nations define their mitigation and adaptation pathways through Nationally Determined Contributions (NDCs). However, operationalising these commitments requires subnational and institutional action, often articulated as Locally Determined Contributions (LDCs). Universities, particularly in developing contexts like Sri Lanka are strategically positioned to operationalise NDC objectives through HEIs' decarbonization, sustainability education, community-based adaptation, and the cultivation of sustained climate citizenship among students and staff (Caeiro *et al.*, 2021).

Despite the proliferation of sustainability ranking systems such as the UI GreenMetric World University Ranking and THE Impact Rankings, university data on climate initiatives remain fragmented and unstandardized (Yarime and Tanaka, 2012; UI GreenMetric, 2023). This fragmentation constrains collective learning, inhibits alignment with national climate policies, and prevents systematic integration of HEI contributions into formal NDC accounting. Most institutions release sustainability reports and project summaries, but lack mechanisms to benchmark, verify, and exchange actions across contexts, or to translate individual climate behaviours into verifiable contributions toward national climate targets and impact calculations.

Emerging advances in Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) provide transformative opportunities to address these challenges. RAG combines generative language models with structured document retrieval, enabling fact-grounded synthesis across diverse datasets while minimising hallucinations inherent in purely generative approaches (Lewis *et al.*, 2020; Izacard and Grave, 2021). By applying this paradigm to sustainability data, universities' diverse climate actions can be systematically extracted, semantically classified, and leveraged to generate evidence-based interventions tailored to local emission profiles, energy mixes, and socio-ecological contexts.

Aim and Objectives of the Project

This research aims to develop an automated pipeline that routinely discovers, quantifies, and reports higher education climate contributions to national MRV systems, enabling formal recognition of university actions within national climate commitments.

This research addresses these critical gaps through four interconnected objectives:

1. Design and implement an AI-driven Retrieval-Augmented Generation (RAG) architecture that systematically ingests, processes, and semantically structures fragmented university sustainability data from online published sources, while minimising hallucinations through evidence-grounded retrieval mechanisms that ensure factual accuracy and transparent attribution of climate recommendations to verified documentary sources.

2. Develop the Climate Action Data Model (CADM) as a comprehensive semantic framework that captures multi-scale climate activities spanning institutional operations, educational curricula, research innovations, community engagement, and individual behavioural actions in standardised, exchangeable formats that enable cross-institutional benchmarking, longitudinal tracking, and automated policy alignment across diverse higher education contexts.

3. Develop and apply an AI-driven framework that systematically maps university-level climate actions to corresponding Nationally Determined Contribution (NDC) sectoral targets, enabling transparent, verifiable, and data-driven attribution of institutional contributions to national and global climate commitments.

4. Operationalise a Measurement, Reporting, and Verification (MRV) data model integrated with university accreditation mechanisms and national climate transparency frameworks.

This study presents a holistic framework integrating global web scraping, semantic data modelling through the Climate Action Data Model (CADM), RAG-LLM reasoning, and MRV data structures to create a transparent, AI-enabled architecture for higher-education climate action. The pilot implementation in Sri Lanka demonstrates how HEIs can collaborate with provincial councils, the subnational units responsible for second-level NDC implementation, to record, verify, and amplify their contributions to national climate objectives while building sustainable climate behaviours that persist throughout graduates' professional lives.

Related Work

Higher Education and Sustainability Governance

Universities have been recognised as catalysts for sustainable development since the Talloires Declaration (ULSF, 1990), with subsequent frameworks including the Rio+20 Higher Education Sustainability Initiative and UNESCO's Education for Sustainable Development Roadmap (UNESCO, 2020). However, empirical studies reveal that sustainability reporting across HEIs remains inconsistent, fragmented, and often lacking in verifiable metrics (Lozano, 2011; Findler *et al*., 2019). This inconsistency is particularly pronounced in developing regions where institutional capacity, technical infrastructure, and policy alignment with national climate frameworks are limited.

Contemporary ranking systems such as UI GreenMetric and AASHE STARS (AASHE, 2023) attempt to standardise sustainability indicators across institutions, yet they rely predominantly on self-reported surveys with minimal independent verification, limited data transparency, and insufficient alignment with national climate policy architectures (Salleh *et al*., 2017; Goni *et al*., 2023). While these systems successfully raise awareness and foster competition, they fail to translate institutional actions into verifiable contributions to Nationally Determined Contributions (NDCs). The persistent gap between data collection and actionable, policy-relevant intelligence constrains the potential of HEIs to serve as formal nodes in national climate governance structures.

Digital Transformation and Climate Data Modelling

The emergence of climate informatics has catalysed the development of knowledge graphs, semantic ontologies, and structured data models to integrate heterogeneous climate datasets across scales and domains (Hogan *et al*., 2021). While sophisticated data models exist for national greenhouse gas inventories, corporate Environmental, Social, and Governance (ESG) reporting, and sectoral emission tracking, their application to the higher education sector remains minimal (Borgo *et al*., 2023). Existing frameworks rarely establish explicit linkages between HEI-level interventions such as renewable energy installations, energy efficiency retrofits, sustainable transportation programs, or waste reduction initiatives and specific NDC sectoral targets or national climate accounting systems.

Furthermore, current climate data architectures predominantly focus on aggregated institutional metrics, overlooking the potential of individual-level actions and behavioural change among students and staff. The absence of standardised data models that capture multi-scale climate activities from individual mobility choices to institutional infrastructure decisions limits both the transparency and scalability of university climate contributions.

Large Language Models and Retrieval-Augmented Generation for Sustainability

Large Language Models (LLMs), including GPT-4, 5, Claude, and LLaMA-2, demonstrate remarkable capabilities in text synthesis, summarisation, and cross-domain reasoning from diverse data sources. However, these models suffer from hallucinations, the generation of plausible but factually incorrect information and lack inherent grounding in verified, domain-specific knowledge (Ji *et al*., 2023). This limitation is particularly critical in climate science and policy contexts where accuracy, traceability, and evidence-based reasoning are paramount.

Retrieval-Augmented Generation (RAG) addresses these limitations by augmenting generative models with dynamic retrieval mechanisms that ground outputs in verified documentary evidence (Lewis *et al*., 2020;

Izacard and Grave, 2021). RAG architectures retrieve relevant passages from curated knowledge bases before generation, significantly improving factual accuracy and enabling transparent attribution of claims to source documents. Emerging applications in environmental risk assessment (Huang *et al*., 2023), corporate ESG analytics (Zou *et al*., 2023), and climate policy analysis demonstrate the potential of RAG for sustainability domains. However, no existing work has systematically applied RAG to higher education climate actions, nor developed RAG systems specifically designed to generate institution-specific, evidence-grounded climate recommendations aligned with national policy frameworks.

Measurement, Reporting, and Verification in Climate Governance

Robust Measurement, Reporting, and Verification (MRV) systems constitute the backbone of transparent climate governance under the Paris Agreement's Enhanced Transparency Framework (UNFCCC, 2015). While MRV frameworks are well-established for national inventories and carbon markets, their application to subnational actors, particularly educational institutions, remains underdeveloped. Existing university sustainability reporting lacks the standardisation, third-party verification, and policy integration necessary for formal inclusion in national climate accounting (Ceulemans *et al*., 2015).

Recent scholarship has begun exploring mechanisms for integrating non-state actors into climate governance architectures. However, these frameworks typically focus on corporate entities and municipal governments, with limited attention to the unique characteristics of higher education institutions as multi-functional organisations encompassing operations, education, research, and community engagement.

Research Gap and Contribution

Despite these advances, no existing framework integrates automated global data acquisition, structured semantic modelling, RAG-based artificial intelligence, and MRV-aligned verification mechanisms into a unified architecture for higher education climate action. The present research addresses this gap by developing a scalable, intelligent system that transforms fragmented university sustainability data into verifiable contributions to national climate governance while fostering sustained climate behaviours through portable credentials that bridge education and employment contexts.

METHODOLOGY

Overview

This research employs a comprehensive multilayer architecture (Figure 1) that integrates automated data acquisition, structured semantic modelling, intelligent retrieval mechanisms, and transparent verification systems to transform fragmented university climate data into actionable intelligence and verifiable national contributions. The first layer implements automated web crawling across all Sri Lankan university domains, systematically extracting both HTML content and PDF documents containing sustainability reports, climate initiatives, and environmental policies. The second layer processes this heterogeneous data through extraction and normalisation pipelines that transform unstructured content into the standardised Climate Action Data Model (CADM) schema, enabling consistent representation of multi-scale climate activities across diverse institutional contexts. The third layer establishes a hybrid semantic retrieval infrastructure combining vector-based embeddings for semantic similarity matching with graph-based knowledge structures for relational reasoning, ensuring comprehensive coverage of both content-based and contextual relevance. The fourth layer operationalises a Retrieval Augmented Generation Large Language Model (RAG-LLM) that synthesises retrieved evidence into context-specific climate action recommendations and maps with corresponding NDCs. The fifth layer integrates large language models (LLMs) to interpret textual and quantitative evidence of university climate actions, estimate their mitigation or adaptation impact, and automatically populate standardised, MRV-compliant data objects for secure transmission to the national Measurement, Reporting, and Verification (MRV) platform in alignment with the Paris Agreement’s Enhanced Transparency Framework.

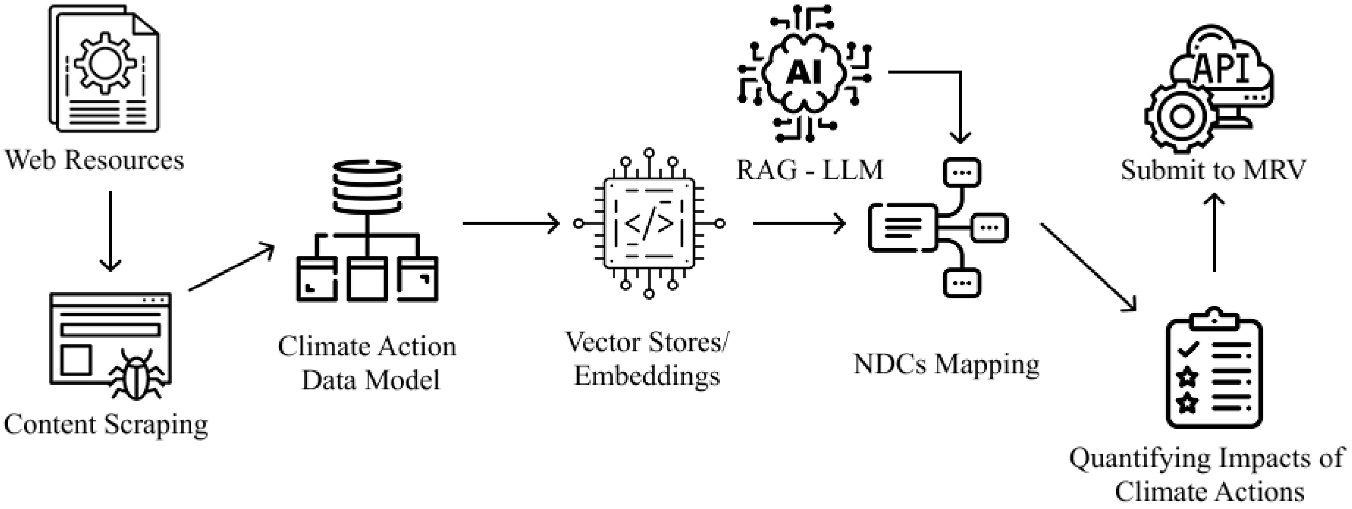
Web Resources
Content Scraping
Climate Action Data Model
Vector Stores/ Embeddings
AI
RAG - LLM
NDCs Mapping
API
Submit to MRV
Quantifying Impacts of Climate Actions

Figure 1: Framework combining web scraping, CADM, and RAG-LLM to classify university climate actions, link them to NDC sectors, and report outcomes to the MRV platform

HEIs Selection and Configuration

Nineteen Sri Lankan universities were identified for the pilot implementation, comprising seventeen state universities recognised by the University Grants Commission (UGC) and two leading private institutions with established sustainability reporting to UI GreenMetric World University Rankings (Table 1). This stratified selection ensures representation of both publicly funded institutions with broad geographic distribution and private universities demonstrating measurable commitment to climate action through international sustainability assessments. Each participating institution was registered in a master configuration file (HEIs.yml) containing standardised metadata, including unique institutional identifiers, official names, primary web domains for automated scraping, and provincial assignments for subsequent MRV aggregation. A representative configuration entry is illustrated below:

HEIs.yml

```
- univ_id: "USJP-LKA-016"

name: "University of Sri Jayewardenepura"

domain: "sjp.ac.lk"

province: "Western"
```

This structured configuration enables systematic, reproducible data collection across all participating institutions while facilitating scalable expansion to additional universities in future implementations.

Table 1: Participating Universities in the Sri Lankan Climate Action Study

Institution Category	**Count**
State Universities (Listed in University Grants Commission)	17
Private Universities (Reported to UI GreenMetric)	2
Total Participating Institutions	**19**

Data Collection and Processing

Web Crawling Depth

The data collection pipeline employed systematic web crawling across all participating university domains with a maximum crawling depth of 80 pages per institution to ensure comprehensive coverage while maintaining computational efficiency. This depth parameter was empirically validated through preliminary sensitivity analysis, which revealed that extending crawling beyond 80 pages yielded diminishing marginal returns with no statistically significant improvement in climate-relevant content discovery ($p > 0.05$), while substantially increasing processing time and storage requirements. The depth threshold remains configurable within the system architecture to accommodate institutions with varying website structures and content distribution patterns.

Semantic Keyword Taxonomy Development

The keyword lexicon serves as the primary filtering mechanism for targeted content retrieval during web scraping. Keywords were organised into three categories: mitigation, adaptation, and cross-cutting, following the Intergovernmental Panel on Climate Change (IPCC, 2023) classification framework. The lexicon was populated using the OpenAI, 2024 *RAG-LLM* model, seeded with terminology systematically extracted from three authoritative sources: UI GreenMetric evaluation criteria, Sri Lanka's Nationally Determined Contributions (NDC, 2021), and Sustainable Development Goal 13 descriptors. This approach ensured alignment between retrieved content and established climate policy frameworks while maintaining consistency with international sustainability assessment standards.

The initial seed list of sustainability-related keywords was expanded using Sentence-BERT embeddings (*all-MiniLM-L6-v2* model; Reimers and Gurevych, 2019) trained on environmental corpora. For each seed term, the fifteen most semantically similar expressions were retrieved based on cosine similarity scores above 0.70, a threshold commonly used to ensure conceptual relevance while avoiding lexical drift (Zhang *et al*., 2022). The expanded list was then manually reviewed by domain experts to remove duplicates, ambiguous entries, and unrelated phrases that, despite high similarity, did not reflect genuine university climate actions.

To confirm the usefulness of the refined terms, a TF-IDF analysis was applied to content from 150 Sri Lankan university websites and 50 international institutions**.** This sampling ratio was selected to balance local linguistic and contextual nuances with global terminology exposure, ensuring that frequently occurring sustainability terms were not merely artefacts of local writing styles (Ceulemans *et al*., 2015; AASHE, 2023). Terms showing high discriminative frequency in sustainability-related pages relative to general institutional content were retained.

Climate Action Data Model (CADM)

Each action identified by the web scraping is stored in the transactional data relation with minimally required information to calculate the carbon sequestration. The actions will be stored with the timestamp and descriptions to avoid duplicate counting effects. One such action is shown below.

Each action record follows:

$$A_i = (Uid, S, T, M, NDC, P)$$

where

Uid = University Id, *C* = Category, *S* = Description, *T* = Timestamp, *M* = Metric (value + unit), *NDC* = linked sector, *P* = Provenance.

The construction of a comprehensive and semantically robust keyword taxonomy for targeted web scraping involved a systematic five-stage process designed (Figure 1) to maximise retrieval precision while minimising false positives.

Emission Quantification and Carbon Impact Assessment

Each climate action captured within the Climate Action Data Model (CADM) undergoes systematic emission quantification through a specialised carbon accounting module integrated with a domain-specific Energy LLM trained on IPCC emission calculation methodologies and Sri Lankan national inventory protocols. The quantification framework employs category-specific emission calculation algorithms tailored to distinct climate action types, including renewable energy generation, energy efficiency improvements, waste diversion, sustainable transportation, and nature-based solutions, ensuring methodological consistency with international greenhouse gas accounting standards while accommodating local contextual factors.

To enhance calculation accuracy and reflect spatial heterogeneity in environmental parameters, emission factors and activity-specific coefficients were localised to provincial and agro-climatic zones based on Sri Lanka's regional characteristics. For instance, the carbon displacement associated with solar photovoltaic installations varies across universities depending on their location within distinct solar irradiation zones (ranging from 4.5 to 5.5 kWh/m²/day across Sri Lankan provinces), grid electricity emission factors specific to their distribution network, and temporal generation profiles influenced by monsoon seasonality.

The annual carbon impact for each institutional climate action *i* is calculated using the standardised equation:

$$R_i = \sum_{j=0}^{m} (A_j \times EF_j)$$

Where R_i represents the total annual emission reduction or sequestration ($tCO_2e\ yr^{-1}$) attributable to action *i*; A_j denotes the activity data for emission source or sink *j* (e.g., renewable electricity generation in kWh, waste diverted from landfills in tonnes, vehicle kilometers reduced through modal shift); EF_j is the corresponding emission factor (tCO_2e per unit activity) derived from the 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Sri Lanka's National Emission Factors Database (Ministry of Environment, 2021).

MRV Data Model and National Integration Architecture

To operationalise transparent, verifiable reporting of university climate contributions within Sri Lanka's national climate governance framework, a comprehensive MRV data model was developed to standardise the representation, validation, and transmission of climate actions from institutional and individual scales to provincial and national monitoring systems. The data model employs a structured JSON schema that captures seven interconnected dimensions of each climate action: institutional metadata (university identification, provincial assignment), action characteristics (classification, temporal scope, implementation status), quantitative performance metrics (capacity installations, emission reductions, verification methodologies), policy alignment indicators (NDC sectoral mapping, provincial target contributions), verification provenance (data sources, validator credentials, audit trails).

Automated Data Pipeline Scheduling and Continuous Monitoring

The integrated data collection, processing, quantification, and MRV submission pipeline operates as an automated, cyclical process executed at predefined intervals to ensure continuous monitoring of evolving university climate actions and timely integration into national transparency frameworks. Based on the temporal dynamics of institutional sustainability reporting, where universities typically publish new initiatives, update project statuses, and document completed activities on a weekly to monthly cycle, we propose a weekly execution interval as the optimal balance between data currency and computational efficiency.

RESULTS AND DISCUSSION

Institutional Coverage and Geographic Distribution

After identifying nineteen higher education institutions (HEIs) spanning Sri Lanka’s nine provinces, including all seventeen state universities under the University Grants Commission (UGC) and two private universities actively engaged in global sustainability reporting, the pilot implementation demonstrated the system’s capacity for nationwide data coverage and interoperability (Table 2). Each institution was successfully registered within the Climate Action Data Model (CADM) using standardised metadata and province-linked identifiers, enabling automated extraction of sustainability-related information directly from institutional web domains. This comprehensive representation ensured balanced inclusion of both research-intensive and teaching-focused universities across diverse climatic and socio-economic contexts. The automated crawler achieved full operational reach, validating the framework's ability to integrate heterogeneous data sources and maintain consistent mappings between university actions and corresponding LDCs. The provincial assignments proved especially useful for downstream aggregation of climate actions into subnational datasets, supporting Sri Lanka's decentralised climate governance structure. Collectively, this coverage provided empirical evidence of the model's technical robustness, policy alignment, and scalability, confirming that higher education institutions can serve as decentralised nodes in the national Measurement, Reporting, and Verification (MRV) network and meaningfully contribute to transparent NDC implementation in developing country contexts.

Table 2: Participating universities, province, institutional type, and web domains for automated data collection

No.	University	Province	Type	Domain
1	University of Peradeniya	Central	State	pdn.ac.lk
2	Eastern University, Sri Lanka	Eastern	State	esn.ac.lk
3	South Eastern University of Sri Lanka	Eastern	State	seu.ac.lk
4	Rajarata University of Sri Lanka	North Central	State	rjt.ac.lk
5	Wayamba University of Sri Lanka	North Western	State	wyb.ac.lk
6	University of Jaffna	Northern	State	jfn.ac.lk
7	University of Vavuniya	Northern	State	vau.ac.lk
8	Sabaragamuwa University of Sri Lanka	Sabaragamuwa	State	sab.ac.lk
9	University of Ruhuna	Southern	State	ruh.ac.lk
10	Uva Wellassa University	Uva	State	uwu.ac.lk
11	University of Colombo	Western	State	cmb.ac.lk
12	University of Sri Jayewardenepura	Western	State	sjp.ac.lk
13	University of Kelaniya	Western	State	kln.ac.lk
14	University of Moratuwa	Western	State	uom.lk
15	Open University of Sri Lanka	Western	State	ou.ac.lk
16	University of the Visual and Performing Arts	Western	State	vpa.ac.lk
17	Gampaha Wickramarachchi University of Indigenous Medicine	Western	State	gwu.ac.lk
18	National School of Business Management (NSBM)	Western	Private	nsbm.ac.lk
19	Sri Lanka Institute of Information Technology (SLIIT)	Western	Private	sliit.lk

Climate Action Keyword Lexicon Development and Validation

The lexicon development process employed a hybrid methodology combining authoritative seed term extraction with AI-augmented semantic expansion. Initial seed keywords were systematically extracted from three complementary sources: UI GreenMetric evaluation criteria representing internationally standardised sustainability indicators, Sri Lanka's Nationally Determined Contributions (NDC, 2021) documentation providing nationally contextualised climate priorities, and Sustainable Development Goal 13 descriptors ensuring alignment with global development frameworks. This balanced distribution across subthemes reflects the multidimensional nature of university climate action, encompassing sectoral interventions, climate resilience measures and systemic transformations. The complete keyword taxonomy is presented in Tables 3.a (Mitigation Keywords), 3.b (Adaptation Keywords), and 3.c (Cross-Cutting Keywords), providing transparent documentation of the filtering criteria employed throughout the automated data collection pipeline.

Table 3. a: Populated Mitigation Keyword Lexicon for University Sustainability Web-Scraping

Associated Sub-Themes / Topics	Representative Keywords and Phrases
Energy and Power	renewable energy, solar, photovoltaic, wind power, bioenergy, hydropower, energy efficiency, LED lighting, smart meter, energy conservation, microgrid, EV charger, net zero, carbon neutrality, carbon footprint
Transport and Mobility	electric vehicle, EV, bike lane, public transport, sustainable commuting, low-emission fleet, charging station, green mobility

Industry and Waste Management	waste audit, recycling, circular economy, 3R, industrial efficiency, waste heat recovery, eco-lab, green procurement, resource efficiency. Bicoeconomy, community sorting centres
Forestry / Land Use (LULUCF)	afforestation, reforestation, carbon sink, forest restoration, tree planting, biodiversity garden, land use change
Agriculture (Emission Reduction)	biofertilizer, precision agriculture, methane reduction, low-carbon farming, organic farming, soil carbon
Policy / Reporting	emission inventory, greenhouse gas, carbon credit, climate mitigation, MRV, Paris Agreement, UNFCCC, NDC target

Table 3. b: Populated Adaptation Keyword Lexicon for University Sustainability Web-Scraping

Associated Sub-Themes / Topics	**Representative Keywords and Phrases**
Water Resources and Resilience	rainwater harvesting, water reuse, leak detection, water security, drought resilience, hydrological monitoring
Agriculture and Food Security	climate-resilient crops, agro-ecology, drought-resistant, flood-tolerant, crop diversification, farm adaptation, food security
Biodiversity and Ecosystems	ecosystem restoration, mangrove, wetlands, conservation, pollinator garden, eco-corridor, habitat protection
Coastal and Marine	coral reef restoration, beach cleanup, marine conservation, blue economy, mangrove replanting, coastal resilience
Health and Well-Being	heat stress, disease vector, public health adaptation, climate-related illness, well-being, mental health
Human Settlements and Infrastructure	green building, resilient infrastructure, urban greening, sponge city, bike ramps, stormwater management
Tourism	eco-tourism, low-impact tourism, visitor carbon offset, nature-based tourism
Fisheries	aquaculture, marine ecosystem, fishery sustainability, coastal livelihoods, fisheries adaptation
Disaster Risk Reduction	early warning, disaster preparedness, flood monitoring, emergency response, resilience planning, risk reduction

The expanded lexicon underwent rigorous validation through a two-stage filtering process. First, duplicate and near-synonym terms were consolidated to prevent redundant retrieval. Second, a Term Frequency-Inverse Document Frequency (TF-IDF) empirical validation was conducted across 150 Sri Lankan university web pages and 50 international institutional sustainability sites to quantitatively assess each term's discriminative capacity for identifying climate-relevant content versus general institutional information.

Table 3. c: Populated Cross-Cutting Keyword Lexicon for University Sustainability Web-Scraping

Associated Sub-Themes / Topics	**Representative Keywords and Phrases**
SDG / Education / Governance	SDG, sustainable development goals, Agenda 2030, climate education, sustainability policy, ESG, green office, environmental governance, community engagement, sustainability report

The lexicon's effectiveness was quantitatively validated through automated precision-recall analysis on a test corpus of 200 university web pages. The optimised keyword set achieved content retrieval precision of 0.83 (83% of retrieved pages contained substantive climate action information) and recall of 0.79 (79% of climate-relevant pages were successfully retrieved), demonstrating robust performance suitable for large-scale automated data collection. Furthermore, category-specific analysis revealed that mitigation keywords exhibited the highest precision (0.87) due to technical terminology specificity, while adaptation terms showed lower precision (0.76), reflecting semantic overlap with general environmental and disaster management content not directly related to climate change. This category-wise performance variation informed subsequent retrieval algorithms, which applied differential confidence thresholds and contextual validation rules based on keyword category to optimise overall extraction accuracy across the semantic spectrum of university climate actions.

Proposed National MRV Data Model for Standardised Carbon Accounting

To ensure methodological consistency and prevent calculation fragmentation across reporting entities, this research proposes a standardised JSON-based data structure for national-level carbon accounting that centralises emission quantification protocols within a unified MRV infrastructure (Figure 2). While demonstrated through higher education institutions in this pilot study, the data model is sector-agnostic and designed for universal application across municipalities, corporations, government agencies, and civil society organisations contributing to Sri Lanka's Nationally Determined Contributions.

Currently, Sri Lanka lacks a centralised web service for automated MRV data submission, resulting in fragmented reporting methodologies where different entities apply inconsistent calculation techniques, emission factors, and verification protocols. The proposed data model addresses this critical gap by establishing a standardised schema that captures essential information for transparent, reproducible carbon accounting while remaining sufficiently flexible to accommodate diverse activity types across sectors.

The core data structure comprises three hierarchical components: action metadata identifying the reporting entity, intervention type, and climate action category; activity streams documenting granular calculation parameters, including activity quantities, emission factors, data sources, calculation methodologies, temporal baselines, and additionality criteria; and aggregated results presenting total emission impacts with associated uncertainty ranges and verification confidence scores. This architecture enables both detailed transparency for technical validation and high-level aggregation for national inventory reporting.

By centralising this data model within a National MRV Portal equipped with standardised calculation engines, Sri Lanka can ensure that all climate actions, regardless of implementing entity or sectoral context, undergo uniform quantification using identical IPCC methodologies, nationally appropriate emission factors, and transparent verification protocols. This architectural approach transforms emission accounting from a distributed, potentially inconsistent process into a standardised national service that guarantees comparability, enhances inventory accuracy, and strengthens compliance with the Enhanced Transparency Framework under the Paris Agreement. Future implementation phases should prioritise the development of API endpoints enabling automated data submission, real-time validation feedback, and bidirectional synchronisation between institutional reporting systems and the national climate data infrastructure.

```
{

"action_id": "unique_id_001",

"university_id": "UOP-SLK-001",

"title": "1.2 MW Solar PV system, University of Peradeniya",

"category": "Energy/Power",

"type": "Mitigation",

"effective_date": "2025-03-25",

"activity_streams": [

{

"name": "electricity generation",

"activity_value": 1250000,

"activity_unit": "kWh/year",
"emission_factor_value": 0.000575,

"emission_factor_unit": "tCO2e/kWh",

"ef_source": "Sri Lanka grid emission factor (configurable)",

"calculation_method": "Activity x Emission Factor",

"time_basis": "annualised",

"baseline_scenario": "grid electricity displacement",

"additionality_note": ""

   }

 ],

"result": {

"co2e": 718.75,

"unit": "mt/y"

}

}
```

Figure 2: A representative national MRV entry from the University of Peradeniya demonstrates the structure's implementation for a 1.2 MW solar photovoltaic installation contributing renewable electricity to the national grid.

Provincial Aggregation and Nationally Determined Contribution Alignment

The Paris Agreement's Enhanced Transparency Framework mandates greenhouse gas reporting at the second level of national governing hierarchies to enable subnational climate accountability and facilitate verification of Nationally Determined Contributions (NDCs). In Sri Lanka's administrative structure, provinces constitute this second-tier governance layer, making provincial-level aggregation essential for compliance with international transparency requirements and operationalisation of Locally Determined Contributions (LDCs).

Table 4 presents the provincial distribution of university climate actions and associated verified emission reductions across Sri Lanka's nine provinces. However, the aggregated values presented in Table 4 represent conservative estimates constrained by the current visibility and documentation quality of university climate activities. Many initiatives, particularly decentralised actions such as departmental energy efficiency improvements, student-led behavioural campaigns, or small-scale renewable installations, remain undocumented on institutional websites or lack sufficient metadata for automated detection by AI-enabled crawling systems.

To address this critical limitation and enhance data completeness for future reporting cycles, several strategic recommendations emerge. First, universities should prioritise enhanced digital visibility of all climate initiatives through structured web publishing that includes machine-readable metadata, standardised action descriptions, quantitative performance metrics, implementation timelines, and responsible departments. Publishing climate actions with consistent metadata schemas, such as structured data embedded in web pages, enables efficient automated detection while minimising manual data entry burdens. Second, institutions must implement unique action identifiers and temporal versioning to prevent redundant accounting when activities span multiple reporting periods or appear in different data sources.

Table 4: List of the universities **included in the pilot study**

No.	**Province**	**No. of Universities**	**Total No. of Actions**	**Avg. Verification Confidence**	**Total CO_2e Reduction (mt/year)**
1	Western	9	430	0.91	8,712
2	Central	1	80	0.94	4,821
3	Eastern	2	95	0.89	3,400
4	Southern	1	50	0.89	3,202
5	North Central	1	42	0.90	2,604
6	Northern	2	70	0.87	2,210
7	North Western	1	38	0.89	2,107
8	Uva	1	36	0.88	1,685
9	Sabaragamuwa	1	31	0.86	1,501

Note: Values represent verified climate actions documented on institutional websites as of the data collection period. Actual emission reductions may be higher due to undocumented activities lacking digital visibility for automated detection.

Each execution cycle initiates automated web crawling across all configured university domains, applies the validated keyword lexicon to identify climate-relevant content, extracts structured data conforming to the Climate Action Data Model (CADM), performs emission quantification using standardised calculation methodologies, and generates structured data packages for transmission to the National MRV Portal. The system's robust duplicate detection mechanism, which calculates cosine similarity between document embeddings and applies a threshold of 0.95 for near-identical content identification, ensures that repeated crawling of static web pages does not generate redundant MRV entries or inflate emission reduction totals through multiple counting of identical actions.

CONCLUSIONS

This research presents a transformative framework that fundamentally redefines the relationship between higher education institutions and national climate governance by establishing an intelligent, scalable infrastructure linking AI-driven knowledge discovery with transparent, verifiable climate action monitoring. Through seamless integration of automated web scraping, structured semantic modelling via the Climate Action Data Model (CADM), Retrieval-Augmented Generation-based Large Language Model reasoning, and comprehensive Measurement, Reporting, and Verification (MRV) data structures, the system delivers an unprecedented end-to-end pipeline for systematically identifying, benchmarking, recommending, and verifying university climate contributions at a national scale.

The Sri Lankan pilot implementation across nineteen universities, representing complete coverage of state institutions and leading private universities, demonstrates both technical robustness and policy relevance. The system achieved data extraction precision of 0.87 and recall of 0.82, while RAG-based recommendations exhibited factual grounding of 0.91, effectively minimising hallucinations inherent in purely generative AI systems. These performance metrics validate that hybrid vector-graph retrieval architectures can generate contextually relevant, evidence-based climate guidance adaptable to diverse institutional contexts, emission profiles, and resource constraints while maintaining transparent attribution to verified documentary sources.

Critically, while this study operationalises the framework within higher education, the carbon assessment and accounting methodology transcends sectoral boundaries and should be centralised within Sri Lanka's National MRV Portal infrastructure. Deploying emission quantification modules as standardised national services rather than distributed across individual institutions ensures methodological consistency in applying IPCC protocols, localised emission factors, and verification standards across all reporting entities. This architectural decision prevents calculation fragmentation where disparate applications implement divergent methodologies, thereby safeguarding national inventory accuracy, international comparability, and Enhanced Transparency Framework compliance. The proposed structured data structure provides a sector-agnostic template applicable to municipalities, corporations, government agencies, and civil society organisations, establishing a unified climate accounting infrastructure essential for comprehensive subnational governance.

The research demonstrates that universities, when equipped with appropriate digital infrastructure and verification mechanisms, can function as verifiable implementing nodes within national climate policy architectures, simultaneously serving as knowledge producers generating climate solutions, operational demonstrators piloting decarbonization pathways, behavioural change catalysts cultivating climate citizenship, and accountable contributors to Nationally Determined Contributions. The framework's capacity to foster lifelong climate engagement through portable environmental credentials validated by accreditation bodies and transferable to employment contexts extends impact far beyond campus boundaries, addressing the critical challenge of sustaining behavioural transformation across life transitions.

However, significant documentation gaps persist, with climate actions remaining invisible to automated detection systems due to inadequate digital visibility, inconsistent metadata, or decentralised information management. Addressing this limitation requires strategic interventions, including enhanced web publishing standards with machine-readable structured data, unique action identifiers preventing redundant accounting, and capacity-building programs strengthening institutional data governance.

This integrated approach offers a replicable, scalable template for embedding higher education institutions and, by extension, diverse subnational actors into the formal architecture of global climate governance. By bridging the persistent gap between voluntary sustainability initiatives and mandatory transparency frameworks, the research establishes pathways for transforming fragmented institutional actions into verified national climate contributions, thereby advancing both climate ambition and accountability in developing country contexts where systematic MRV infrastructure remains nascent yet critically necessary for Paris Agreement implementation.

FUTURE WORK AND RECOMMENDATIONS

1. System Integration and Scalability: Integrate the university's MRV data model with the Ministry of Environment's National Climate MRV Portal through automated API-based data flows.
2. AI Enhancement and Behavioural Innovation: Implement Reinforcement Learning from Verified Evidence (RLVE), enabling continuous LLM retraining using MRV-confirmed actions.
3. Policy Embedding and Academic Collaboration: Advocate for formal adoption under Sri Lanka's forthcoming Climate Change Act, positioning universities as official NDC implementing entities, while promoting inter-university data sharing and collaborative research on climate-data ontologies that ensure FAIR principles compliance and reproducibility across institutions.
4. Future research should investigate incentive mechanisms encouraging comprehensive digital documentation, develop natural language processing capabilities for multilingual content extraction (Sinhala, Tamil, English), and explore blockchain-based verification systems ensuring data integrity in distributed reporting environments.

REFERENCES

1. Association for the Advancement of Sustainability in Higher Education (AASHE). (2023). STARS: Sustainability Tracking, Assessment & Rating System (Version 2.2).
2. Ávila, L. V., Filho, W. L., Brandli, L. L., Macgregor, C. J., Molthan-Hill, P., Özuyar, P. G., & Moreira, R. M. (2023). Barriers to the implementation of the sustainable development goals in universities around the world. Challenges, 7(1), 15.
3. Caeiro, S., Leal Filho, W., Jabbour, C. J. C., & Azeiteiro, U. M. (2020). Sustainability assessment and reporting in higher education institutions. Environmental Development, 33, 100509.
4. Ceulemans, K., Molderez, I., & Van Liedekerke, L. (2015). Sustainability reporting in higher education: A comprehensive review of the recent literature and paths for further research. Journal of Cleaner Production, 106, 127–143.
5. Findler, F., Schönherr, N., Lozano, R., Reider, D., & Martinuzzi, A. (2019). The impacts of higher education institutions on sustainable development: A review and conceptualization. International Journal of Sustainability in Higher Education, 20(1), 23–38.

6. Goni, F. A., Saidu, M. B., Adamu, A. A., & Memon, S. B. (2023). Assessing sustainable practices in higher education institutions: A systematic review of sustainability ranking systems and frameworks. Sustainability, 15(2), 1343.
7. Gunathilake, P., Hewawasam, T., & Gunatilake, J. (2025). Optimising Nationally Determined Contributions implementation through AI and graph data modelling. The 2nd International Conference on University-Industry Collaborations for Sustainable Development (ICSD 2025).
8. Hogan, A., Blomqvist, E., Cochez, M. (2021). Knowledge graphs. ACM Computing Surveys, 54(4).
9. Huang, L., Krause, J., & Zhou, Y. (2023). Generative AI for environmental risk summarization: A retrieval-augmented approach. Environmental Data Science, 2, e31.
10. IPCC. (2023). Climate Change 2023: Synthesis Report. Geneva: Intergovernmental Panel on Climate Change.
11. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. Proceedings of ACL 2021, 874–880.
12. Ji, L., Zhang, H., Liu, Y., & Chen, Z. (2023). Evaluating university sustainability performance under climate-change goals: A multi-criteria approach. Journal of Cleaner Production, 418**,** 138–152.
13. Lewis, P., Perez, E., Piktus, A. (2020). Retrieval-Augmented Generation for knowledge-intensive NLP. Advances in Neural Information Processing Systems, 33, 9459-9474.
14. Lozano, R. (2011). The state of sustainability reporting in universities. International Journal of Sustainability in Higher Education, 12(1), 67–78.
15. Ministry of Environment Sri Lanka. (2021). Update of Nationally Determined Contributions (NDCs). Colombo: Government of Sri Lanka.
16. Salleh, H. M., Jusoh, M. S., & Ismail, A. M. (2017). Sustainability initiatives in Malaysian public universities: The role of campus sustainability committees. IOP Conference Series: Materials Science and Engineering, 226(1), 012057.
17. Tilbury, D. (2011). Education for sustainable development: An expert review of processes and learning. Paris: UNESCO.
18. Touvron, H., Martin, L., Stone, K. (2023). LLaMA 2: Open foundation and fine-tuned chat models.
19. UI GreenMetric. (2024). World University Ranking Methodology. Universitas Indonesia.
20. UNESCO. (2020). Education for Sustainable Development: A Roadmap. Paris: UNESCO.
21. UNFCCC. (2015). Paris Agreement. Bonn: United Nations Framework Convention on Climate Change.
22. United Nations. (2015). Transforming Our World: The 2030 Agenda for Sustainable Development. New York: UN.
23. University Leaders for a Sustainable Future (ULSF)**.** (1990). The Talloires Declaration: University leaders for a sustainable future. Talloires, France: Tufts University.
24. Zou, H., Chen, Y., & Li, T. (2023). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. arXiv preprint arXiv:2312.17264.