



OPEN ACCESS

Volume: 5

Issue: 4

Month: October

Year: 2026

ISSN: 2583-7117

Published: 13.10.2026

Citation:

Bhakti Pawar, Amritanshu Mishra, Janeekaa Kedar, Mrunal Chatur, Sudhanshu Bhoyar, Vijaya P. Balpande "TrueGreen AI: Intelligent Verification of Sustainability Claims for Smarter and Greener Consumer Decisions" International Journal of Innovations in Science Engineering and Management, vol.5, no.4, 2026, pp. 129-139

DOI:

10.69968/ijisem.2026v5i4129-139



This work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

TrueGreen AI: Intelligent Verification of Sustainability Claims for Smarter and Greener Consumer Decisions

Bhakti Pawar¹, Amritanshu Mishra¹, Janeekaa Kedar¹, Mrunal Chatur¹, Sudhanshu Bhoyar¹, Vijaya P. Balpande¹

¹Dept. of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, India

Abstract

Greenwashing, or the creation of environmentally misleading claims about products, is becoming a more prevalent phenomenon among products labelled as "eco-friendly," "natural," or "sustainable" without adequate evidence to prove such claims in the context of e-commerce. Prior research in greenwashing detection focuses on manual content analysis or single-modality text processing and is confined to food and beverages. Multimodal verification based on evidence in multiple modalities (text, image, speech), public benchmark datasets with annotated greenwashing instances, and explanation in machine learning or natural language processing approaches are relatively under-explored.

We introduce True Green AI, a multimodal framework for auditing environmental sustainability claims based on textual, visual, and verbal evidence. The system considers the representation of a sustainability claim as an aggregation of evidence in text, image, and speech, independent analysis of each modality, and aggregation of outcomes into an interpretable risk profile. In addition to the greenwashing classification based on semantic entailment, we introduce a metric called the Cross-Modal Divergence Index (CMDI) that quantifies cross-modal divergence in a brand's communication, presentation, emphasis, and evidence.

The framework is developed as a decoupled architecture, composed of a React 19 client and Node.js/Express service layer with Firebase persistence. Browser-based technologies are used for capturing camera and microphone input, AI-assisted analysis of claims and evidence is performed by our system, and generation of justification based on evidence retrieval is constrained. Our system generates an interpretable A-E assessment of a sustainability claim, along with the evidence used for the evaluation and component scores, as opposed to just numbers or classifications. Such an approach is inspired by research on consumer perception that indicates that interpretative, graded labels can enhance consumer understanding as compared to numeric or unadjudicated environmental information.

Our approach addresses four research gaps, including: multimodal verification, cross-modal consistency evaluation, retrieval-based generation, and interpretable output generation. However, quantitative evaluation is contingent upon the development of an annotated multimodal dataset of greenwashing instances, and the current study is presented as a framework for decision-making as opposed to an autonomous adjudication tool.

Keywords; Greenwashing, multimodal artificial intelligence, sustainability communication, natural language processing, explainable artificial intelligence, green computing.

INTRODUCTION

Environmental sustainability is a key concern in purchasing and investment decisions, and organizations are using environmental terms and visual cues to signal the associated attributes. By 2023, 456 environmental labels were identified in 199 countries and 25 industries [4]. There are no corresponding developments in the field of verification: according to the analysis carried out by the European Commission, more than 53% of green claims were found to be misleading, while 40% of claims lack evidence, and about half of the analysed environmental labels in EU markets lack any form of verification [4].

The issue becomes complicated by the nature of environmental impact as a credence attribute: unlike product attributes that are observable and testable in product usage, the environmental performance cannot usually be verified at the time of purchase. The knowledge of environmental terms among the consumers is also low: only 39% of Germans understand that “climate-neutral” labels may refer to the carbon offsetting, and only 41% of the surveyed U.S. adults can explain the concept of carbon neutrality [3].

The greenwashing is usually understood as the interaction of poor environmental performance with aggressive environmental communication [2], [4]. According to Wang et al., there are three main forms of operationalizing greenwashing [2]: false disclosure – the misrepresentation of the environmental practices; selective disclosure – the communication of positive information in conjunction with the exclusion of relevant negative information; and expressive manipulation – the communication of high environmental commitment in the absence of substantial environmental investments.

The effects are not limited to individual organizations, as the misleading environmental claims affect the trust in environmental communication, undermine the competitors’ environmental investments, misallocate resources, and interfere with the change in the consumption patterns needed to achieve climate goals [2]–[4]. Moreover, the conventional methods of greenwashing detection are vulnerable to vocabulary drift since organizations are increasingly using sophisticated sustainability terminology [5].

There are two approaches based on model-driven methods currently used in the academic literature. The first of these utilizes transformer-based zero-shot classification to match the previously unseen claim to the greenwashing taxonomy: EcoLens GRIE classifies semantic entailment of the scraped claims in relation to the Seven Sins of Greenwashing using the facebook/bart-large-mnli model, industry-weighting, and returns credibility, deception probability, and transparency scores [1]. The second one treats the greenwashing as a regression task over corporate-disclosure texts: Wang et al. extract the topic-distribution and topic-sentiment features from the dual-channel architecture, use the transformer-based modeling, and estimate the degree of greenwashing [2].

Both approaches operate with textual information predominantly, which is a serious limitation since the environmental communication does not have to include only text: the graphics, the symbols, the colors, the labels, and

other visual information may lead to greenwashing [4], and thus the product-level auditing has to incorporate computer vision capabilities along with the natural language processing [5]. The environmental communication is becoming increasingly visual and audio on the social media as well [6].

General-purpose multimodal models cannot solve this problem as they lack the context needed to assure the environmental, social, and governance (ESG) information and confuse the communication complexity with the intention to deceive [5]. Therefore, the multimodal greenwashing detection needs a structured approach that connects different evidences with the greenwashing taxonomy and the verification process.

Another problem is connected with how the results of the process are communicated. In the experiment conducted with 2,109 German consumers, the offset-based climate labels led to the misunderstanding of the climate impacts, and even the substantiated variants like “climate-compensated” and “climate-neutral, product CO₂-compensated” did not eliminate the problem, while the five-level traffic-light label allowed to make more realistic assessments [3]. Similarly, the certification information provided as raw evidence was not shown to have any significant effect on reducing the greenwashing perception [6]. Thus, the greenwashing auditing system should not present the raw evidence or binary verdict but interpretable assessment connecting the evidence and the classification result.

Based on the above discussion, True Green AI framework was developed that builds a sustainability claim from the combination of textual, visual, and verbal evidence; separately analyses each of the modalities; does cross-modal consistency analysis; retrieves and validates evidence; and at the end presents the interpretative risk profile. The main contributions of True Green AI framework are: (1) Multimodal Claim Representation and Ingestion, where the sustainability claim is created through a combination of textual, visual and spoken evidence with the help of browser-native multimodal capturing capability; (2) Cross-Modal Divergence Index to measure the inconsistency in environmental communication across modalities and the existing textual evidence; (3) Fusion-based risk assessment, where multiple analytical components are fused to generate a risk profile based on the greenwashing taxonomy; (4) Retrieval-based justification, where generation of explanation is constrained to structured and evidence-based outputs rather than being generated by an unrestricted

generative model to make the final classification; and (5) Interpretability and Exportability, where apart from A-E assessment, component-wise evidences and an auditable report can be generated.

The following sections contain Section II, Related Work, Section III, Methodology & Architecture, Section IV, Multimodal Framework, Section V, Outcomes and Limitations, and Section VI, Conclusion.

RELATED WORK

Conceptual Foundations

The concept of greenwashing was coined by Jay Westerveld in 1986 [2]. Delmas and Burbano later defined greenwashing in terms of bad environmental performance coupled with high environmental communication [2], [4]. For computerised analysis, the three-fold classification developed by Wang et al., identifying false disclosure, selective disclosure, and expressive manipulation is especially helpful [2].

The Seven Sins of Greenwashing constitute a more detailed taxonomy at the claim level for phenomena like vagueness, absence of proof, hidden trade-offs, irrelevance, and false labelling [1]. The True Green AI project has taken the Seven Sins as its main claim-level classification target but employs the three-fold classification as an interpretive framework for brand-level results. This is important since selective disclosure and expressive manipulation are all about inconsistencies between environmental communication and the facts behind it, something a single-claim assessment system is unable to address.

Poorly managed corporate responsibility rating systems do not provide protection against possible greenwashing, and even poorly managed certifications might lead to misleading environmental communications [6]. As a result, certification marks alone cannot be used as evidence of the truthfulness of claims but should be viewed in context as one piece of a wider verification process.

Detection Approaches and Their Limitations

Content analysis methods can detect greenwashing but might not give much information about its extent, and their effectiveness could decrease with an increase in the amount of disclosed information [2]. Indicators can estimate the extent of greenwashing but typically demand significant manual annotation and feature construction [2].

The results of the analysis by Wang et al. indicate that for six companies included in the 2022 China Greenwashing

List, indicator-system values varied from 0.31 to 0.50, while predictions made by a model were between 0.75 and 0.82 [2]. This gap was especially noticeable in case of a pharmaceutical company, which had a lot of disclosures, but received an underestimation of risk according to the indicator-based assessment despite greenwashing, thereby illustrating that disclosure volume alone cannot be considered proof of environmental responsibility.

Manual product-level auditing delivers a detailed analysis but does not allow scaling to the number of products available on digital markets [1], [5]; hence, automation solutions provide scalability along with structured verification. EcoLens GRIE classifies industries and detects Seven Sins based on sequence-to-sequence semantic entailment with facebook/bart-large-mnli; that is, it identifies semantic consistency with greenwashing indicators but not mere sentiment polarity [1]. Wang et al. utilize MacBERT-LDA for extracting explicit features of topic distributions, Bi-LSTM for implicit topic-sentiment features, channel fusing with the help of Transformer, and prediction of greenwashing degree with the help of interval type-2 fuzzy broad learning system [2].

Table I. Fused vs. Single-Channel Feature Performance [2]

Input Features	RMSE	MAE	MAPE
Fused topic + sentiment	0.1530	0.1244	19.09%
Topic distribution only	0.1657	0.1414	25.46%
Topic sentiment only	0.2507	0.1845	31.29%

The merged representation achieved higher performance than either representation by itself, and IT2F-BLS was better than BLS, CNN, and RNN baselines when applied to the same merged representation [2]. This supports the need to combine complementary sources of evidence in True Green AI.

Some recent surveys further emphasize retrieval enhancement, knowledge graphs, cross-source validation, and product-level computer vision as critical areas [5]. However, comparative studies between multimodal systems and environment-based scoring system (ESRS)-aligned environmental performance have failed to establish any negative correlation between the two [5].

Consumer Evidence

Labeling in food industry has been a subject of many studies on greenwashing. Systematic review of 711 Scopus indexed peer-reviewed articles revealed only 16 works

which fulfilled the authors' inclusion criteria despite of high exposure of food industry to environmental labeling [4]. Labelling and consumer trust have been recognized as key factors connected to the perception of greenwashing. Mechanisms include self-developed sustainability labels, combination of several labels, and green colour of packing; according to one study among 5,296 household products 95.5% used at least one greenwashing method such as false labeling or absence of evidence [4].

Dreist et al. conducted experiment with 2,109 participants and proved that interpretive labeling could help consumer assessment as compared to numeric or substantiated but not interpreted information [3]. In their survey of 145 Slovenian adults Vukasović et al. revealed positive correlation between environmental awareness and perception of ability to recognize greenwashing and between environmental exposure and attentiveness. Negative trust reaction was linked with greenwashing, whereas verifiable claims such as certification did not show statistically significant link to reduction of perceived greenwashing [6]. Authors of the latter study called their investigation exploratory; hence the results are regarded as indicative in the present work.

Research Challenges and Gaps

The literature review reveals four critical issues that have remained persistent to this date. The lack of an established dataset comprising instances of greenwashing, heterogeneity of ESG disclosure that affects the generalization capability, shift of language in sustainability communication, need for interpretability in legal and regulatory frameworks, asymmetrical costs of false positives and false negatives, and lack of sufficient context grounding in general-purpose multimodal models [5].

- **G1 – Modality:** Current systems lack integration of both visual and auditory claim evidence in taxonomy-based verification pipeline.
- **G2 – Cross-modal consistency:** Selective disclosure and expressive manipulation include inconsistencies between the communication and substance [2], while current detectors focus on measuring intra-textual properties of data.
- **G3 – Context grounding:** Multimodal generators in unconstrained systems can provide inconsistent ESG assessment in the absence of sufficient context grounding [5].
- **G4 – Interpretation:** Numerical and substantiated, but uninterpreted outputs might not be able to help consumers correct their misperceptions, while

interpretive graded labels have shown better results [3], [6].

True Green AI framework addresses all four gaps.

METHODOLOGY AND PROPOSED SYSTEM

System Overview

True Green AI is an integrated, evidence-based product verification system, which assesses the environmental and sustainable claims by applying a predefined R1-R6 verification framework, artificial intelligence (AI), and an additional machine learning (ML) component. True Green AI is implemented as follows: React.js Frontend, Node.js & Express.js Backend, Firebase Cloud Firestore Database, AI component for claim and evidence evaluation, Supervised Machine Learning Component on labelled product dataset; API tests are performed with the use of Postman. The system has a client-server architecture: React.js frontend is connected with the Node.js & Express.js backend via HTTP/REST APIs, and the backend interacts with the database, AI component, ML component, and R1-R6 Rule Engine. Every user is identified by a unique User ID.

Product Dataset Development

A well-structured product data set acts as the base for the verification process. Every product has a unique Product ID and is associated with one of the five categories that have been defined already: Staples, Biscuits, Beverages, Snacks, and Confectionery. The data set consists of the following columns: Product ID, brand name, product name, category, ingredients, quantity of ingredients when applicable, nutrition data, product claim, packaging details, evidence, evidence source, and verification.

Product Categorisation

Product classification offers a framework through which products can be verified. The staples include products such as atta, flour, cereals, pulses, rice, oil, and salt. Biscuits comprise Marie biscuits, digestive biscuits, cream biscuits, and cookies. Beverages comprise fruit beverages, fruit juices, coconut water, and other beverages. Snacks comprise chips, nachos, extruded snacks, and other savory products. Confectionery comprises candies, toffees, eclairs, and chocolates. Classification of products helps in determining the criteria used for verification and is also used in the machine learning part.

R1–R6 Rule-Based Verification Framework

The R1–R6 structure represents the main method of verifying claims, since it outlines clear criteria for

assessment of product claims in comparison with available evidence.

- **R1 – Claim-Ingredient Consistency:** assesses consistency between the product claim and its ingredients or formulation, by comparing the claim and available information about ingredients to determine if the product characteristic mentioned in the claim is justified.
- **R2 – Ingredient Quantities:** assesses consistency between the quantities or percentages of ingredients mentioned in the claim and the claim itself.
- **R3 – Natural/Organic Claims:** assesses product claims that refer to natural, organic, chemical-free, pesticide-free or naturally sourced products by the available evidence or certification if any.
- **R4 – Sustainable Ingredient Sourcing:** assesses claims about sustainable, responsible, ethical or traceable sourcing of the ingredients by available information or certification.
- **R5 – Claim Scope Consistency:** assesses the consistency of the available evidence in relation to the entire scope of the claim rather than part of it.
- **R6 – Packaging Material and Recyclability:** assesses environmental claims related to the material of packaging, recyclability, biodegradability, compostability, eco-friendliness, and plastic-free claims based on available information.

Rule-Based Verification Process

For every product, the back end system will analyze and assign rules that apply to R1 – R6 according to the product's category and claim type in the order Product Claim > Applicable Rule > Evidence > Rule Evaluation > Verification Score. Every applicable rule is given one of four statuses, including: 2 – Supported; 1 – Partially supported; 0 – Unsupported/Contradicted; and – Evidence Pending. The system allows the differentiation between claims that lack sufficient evidence and unsupported claims, therefore evidence absence doesn't equate to claim falseness.

AI-Assisted Claim Analysis

AI functions as an analysis layer in support of the processing of product information in natural language. The AI module helps to identify the product claims, classification of claims into types, identification of attributes, identification of potentially relevant R1-R6 rules, interpretation of text evidence, mapping of claims with evidence, and identification of any missing evidence. The statement "Made with natural ingredients and recyclable packaging" can be divided into two claims: one is natural-ingredients claim linked with R3, while the other one is the recyclable-packaging claim linked with R6.

Evidence Analysis Assisted by AI

The product evidence is available in both structured and unstructured forms, for example, ingredients, product specification, certifications, manufacturer information, and packaging descriptions. AI aids in interpreting and connecting evidence through Claim -> Evidence -> Applicable Rule format. The last criterion of the verification will be based on the pre-defined R1-R6 framework, and therefore, AI will assist in the interpretation and analysis but not in replacing the rules-based verification.

Machine Learning Integration

Supervised machine learning module enables data-driven learning in True Green AI. As opposed to the AI module, which mainly assists in interpreting natural language texts, the ML module learns patterns in a labelled product dataset. The selected features include product category, claim type, claim characteristics, ingredient-related data, availability of evidence, availability of certifications, packaging characteristics, and R1-R6 rule results. The target label is the verification outcome, which is generated using the R1-R6 verification framework: 2 – Fully Supported, 1 – Partially Supported, and 0 – Unsupported.

Preparing the Dataset and Training the Machine Learning Component

Machine learning dataset preparation includes data cleaning, feature selection, data transformation, labelling, dataset splitting, model training, and evaluation. Relevant product, claim, ingredient, evidence, and rule-related features are selected, categorical and text features are transformed into machine-readable forms, and verification labels are generated based on the R1-R6 framework. The labelled dataset is split into training and evaluation datasets, and depending on the final form of the dataset, a simple supervised classification algorithm such as Logistic Regression, Decision Tree, or Random Forest can be used. The model is then evaluated by accuracy, precision, recall, F1-score, and confusion matrix. Due to the small number of products in the dataset, the ML component is developed as a proof of concept and not as an autonomous prediction model.

Hybrid AI-ML-Rule Framework

True Green AI uses a hybrid AI-ML-Rule framework where each analytical component has its own task: R1-R6 rule engine provides primary and clear verification, AI layer performs natural language claim analysis and evidence interpretation, while the ML layer analyses the patterns in the labelled product dataset and provides secondary classification support. The overall flow is the following:

Product Data → AI Claim Analysis → R1-R6 Rule Mapping → Evidence Analysis → Rule Evaluation → ML Classification Support → Verification Result. The R1-R6 framework is still used as the basis for the final verification result.

User Identification and Verification History

In the backend, the User ID is stored along with the details of verification, which includes Verification ID, User ID, Product ID, product type, claim, Results R1–R6, evidence details, final result of verification, and timestamp. This relation can be depicted as User ID → Verification Request → Product → R1–R6 Evaluation → Final Verification Result. This enables retrieval of past verifications for an individual user.

Database Design

The Firebase Cloud Firestore database serves as the main database. The conceptual schema consists of five collections, which include Users (User ID, user info, creation time); Products (Product ID, brand, product name, category, ingredients, claims, packaging info); Rules (Rule ID, rule name, rule description, categories that apply); Evidence (Evidence ID, Product ID, Rule ID, evidence info, source, evidence status); and Verification Results (Verification ID, User ID, Product ID, R1 to R6 result, overall score, final status, reason, time).

Backend Development

The back-end is created by using Node.js to provide the server-side JavaScript runtime, along with Express.js for routing, middleware, and HTTP API capabilities. It acts as the orchestrating layer between React front-end, Firebase database, AI module, ML module, and rule engine (rules R1–R6), and performs tasks including receiving request from the front-end; checking the validity of the User ID and request parameters; fetching the selected product; fetching the rules and evidence; processing product claim using the AI module; performing the verification based on rules R1–R6; executing ML module if necessary; creating the verification result; storing the result along with the User ID; and sending the result to the front-end.

REST API Layer

The communication from the React client to the backend server is done using REST API over HTTP protocol. Typical operations are GET for fetching the products, product details, R1-R6 rules, and user verification history, POST for sending a verification request and storing a verification outcome, and PUT for updating the product and/or evidence details. The main format for data exchange is JSON, and a

sample of a verification request would be {"userId": "USER_ID", "productId": "PRODUCT_ID", "claim": "Product claim"}.

React.js Frontend

The frontend has been developed using React.js framework and handles all interaction with the users, including user registration and authentication; searching for products; selection of products; display of product information; submission of claims; trigger of verification process; display of results R1 to R6; explanations provided by AI model; verification status; and the verification history of a specific user.

Firebase Integration

Firestore stores the user data, product data, claims, rules, evidence, verification outcome, verification history, and labeled machine learning data. The backend manages all operations on the database with the User ID attached to each verification for associating the outcome with the respective user.

Postman-Based API Testing

The Postman test and validates the backend APIs independently. The User API testing includes valid User ID, invalid User ID, no User ID, and verification history of the particular user. The Product API testing includes product fetching, invalid Product ID, and fetching of product categories. The Verification API testing includes verification request, AI integration, verification of R1–R6 rules, ML integration, and storing the result of verification. Error testing includes no parameter, invalid IDs, empty requests, invalid request format, and database errors. The integration testing involves communication routes React → Express.js → Firebase, Express.js → AI, Express.js → ML, Express.js → R1–R6 Rule Engine, and Express.js → Firebase User Verification History.

End-to-End Verification Workflow

The entire process of verification includes: User Authentication → User ID Identification → Product Selection → Product Categorization → AI-Assisted Claim Extraction → Applicable R1-R6 Rule Identification → Evidence Gathering → AI-Assisted Evidence Analysis → R1-R6 Rule Validation → Machine Learning Assisted Classification → Verification Score & Status → Store in Firebase Firestore along with User ID → React.js Displaying of Results.

System Validation

Validation occurs on several fronts. Rule level validation involves testing of individual rules R1 to R6 on products and

evidence supporting the rules. AI level validation involves the claim extraction, claim classification, evidence interpretation, and rules identification. ML level validation involves the validation of the trained model by checking its accuracy, precision, recall, F1-score, and confusion matrix. API level validation involves the testing of endpoint functionality, requests, responses, and error handling through Postman tool. Validation of the database includes validation of creation, reading, and updating of users, products, evidence, and verification results. User level validation includes validation of each verification result by checking whether they match User IDs.

Technology Stack

React.js is used for front-end development, Node.js for back-end run-time environment, Express.js for REST API development, Firebase Cloud Firestore for database management, AI/LLM for claims and evidence interpretation, supervised machine learning for classification support, R1-R6 for primary verification, and Postman for API integration testing. The dataset preprocessing, training, and evaluation can be performed using Python along with machine learning framework. JSON facilitates the communication between front-end and back-end, whereas HTTP/REST is responsible for the communication itself.

Overall Methodological Framework

The full methodology process can thus be described as follows: Product Dataset → Product Categorisation → AI Claim Analysis → R1-R6 Rule Mapping → Evidence Retrieval → AI Evidence Interpretation → R1-R6 Verification → ML-Based Supporting Classification → Verification Result → Firebase Storage with User ID → React.js Visualisation. The methodology makes use of the R1-R6 evidence-based rule system as the central aspect of the system, where artificial intelligence is used for evidence interpretation, while the machine-learning aspect provides a limited form of measurable proof of classification based on the data set.

MULTIMODAL FRAMEWORK AND ANALYSIS

Problem Formulation

Claim of sustainability is expressed by a triple $C = \langle T, V, S \rangle$, where T stands for textual evidence consisting of textual descriptions on product label, packaging, ads, and reports; V stands for visual evidence such as images used in packaging, ads, and graphical representations of the environment; and S stands for spoken evidence such as audio track of the ad and verbal claims. Not all of them need

to be present; however, if there is one modality only, then the system works in reduced mode corresponding to single-modality approach used in prior work [1], [2]; full cross-modal analysis including CMDI is performed if at least two modalities are present. Analysis is done separately for each modality, and afterwards the framework carries out sector classification, semantic mapping of the Seven Sins, cross-modal consistency check, and retrieval-based justification.

Modality-Specific Ingestion

1) Textual Channel: Normalisation and tokenisation are applied on the texts to strip out stop words [1]. The TF-IDF weighting algorithm picks out terms which contribute to the claim risk profile, and thus, along with the transformer based classification, generates an inspectable signal [1]. The spoken transcripts are joined with the text channel.

2) Visual Channel: Images can be captured through the Media Streams API or uploaded directly. Visual analysis focuses on three areas. OCR extraction: text appearing on packaging, environmental symbols, certification marks, and claim statements is extracted and incorporated into the textual evidence pipeline. Colour-semantic analysis: green-dominant colour schemes are examined where environmental signalling is not supported by corresponding certification or evidence [4]. Certification-mark verification: certification marks are compared against a maintained reference set, and unrecognised or potentially imitative marks are flagged rather than automatically trusted, addressing documented concerns regarding certification reliability [4], [6].

3) Spoken Channel: Audio is captured through the Web Audio API and transcribed through the Web Speech API. The resulting timestamped transcript is routed into the textual analysis channel, while prosodic metadata is retained as an auxiliary signal. Because browser-native speech recognition has not been independently validated in this study, spoken output is conservatively treated as text-equivalent evidence within the fusion model.

Analysis Stages

1) Sector Classification: Merged textual information is classified against an industry-label set using zero-shot entailment for sector-specific weighting [1]; ingredient specificity may receive greater relevance in food-related claims, whereas material lifecycle may be more important for fashion-related claims.

2) Seven Sins Semantic Mapping: Text is evaluated for entailment against the Seven Sins definitions [1], with topic-

sentiment estimation performed in parallel, following the dual-channel principle described by Wang et al. [2].

3) Cross-Modal Consistency Assessment: The system records a divergence when visual evidence communicates an environmental claim lacking corresponding textual substantiation, such as a certifying body, measurable metric, or dated audit, and likewise when spoken environmental emphasis is disproportionate to available textual substantiation. This operationalises selective disclosure and expressive manipulation [2] as cross-modal phenomena.

4) Retrieval-Grounded Justification: A constrained call to @google/genai generates a natural-language justification. The model is not permitted to assign the final score; it must instead explain which evidence contributed to each classification and identify questions requiring further verification. Responses are constrained to structured JSON and validated server-side, limiting unsupported generative conclusions and addressing concerns regarding unconstrained multimodal generation in ESG applications [5].

Scoring Model

The framework computes four analytical components. Credibility Score (CS) measures the presence of verified environmental markers and the absence of deceptive indicators [1]. Deception Probability (DP) represents cumulative entailment scores associated with the Seven Sins [1]. Transparency Index (TI) evaluates the specificity of environmental language, penalising vague terminology and rewarding concrete, verifiable information [1]. Cross-Modal Divergence Index (CMDI) measures the number and severity of detected cross-modal inconsistencies, weighted according to the rhetorical prominence of the unsupported modality.

These are combined into an A–E output rather than presented only as raw numerical scores, motivated by evidence that interpretative multi-level labels improve assessment compared with purely numeric information [3]. Raw component scores remain in the evidence panel to preserve transparency, which matters because certification information presented without interpretation has not consistently reduced perceived greenwashing [6].

System Architecture

The system adopts a decoupled client-service architecture based on frontend-backend separation, extending existing approaches with browser-native multimodal capture and cross-modal analysis. Five layers

are used — presentation, capture, application/service, inference-integration, and persistence — and the architecture supports four primary visual outputs: the system architecture, multimodal fusion pipeline, end-to-end audit sequence, and verification dashboard.

Presentation Layer: a React 19/Vite 6/Tailwind CSS 4 client whose components correspond to the analytical pipeline, including a capture panel, modality tray, CS, DP, TI, and CMDI score indicators, evidence accordion, semantic-entailment results, justification panel, and side-by-side claim comparison. The system can use jsPDF and html2canvas to export completed audits as portable reports suitable for consumer, journalistic, or regulatory use.

Capture Layer: uses the Media Streams API for camera access, the Web Audio API for microphone capture, and the Web Speech API for transcription. Client-side preprocessing, including image compression and transcript-only transmission, reduces unnecessary data transfer and supports the green-computing objectives of the architecture.

Service Layer: a Node.js/Express REST API with representative routes POST /ingest, POST /analyze/text, POST /analyze/visual, POST /analyze/speech, POST /fuse, and POST /report. Middleware provides payload validation, rate limiting, and structured error handling, and modality analysers can operate asynchronously to reduce end-to-end latency.

Inference Integration Layer: an adapter isolates @google/genai calls behind a schema-validated interface. If a generated response fails validation, the system performs a bounded retry; if validation continues to fail, it falls back to a deterministic evidence-only summary, ensuring that failure of the generative component results in reduced explanation capability rather than an unverified explanation.

Persistence Layer: Firebase Authentication manages user identity, while Firestore stores claims, analyses, evidence, and audit records. An append-only audit trail ensures a previously generated grade is not silently replaced. Postman collections correspond to the API routes and include regression tests for documented failure cases such as vague-language-only claims, unrecognised certification marks, and visual claims based only on colour.

Evaluation Design

Evaluation has two tracks: Seven Sins classification, assessed by precision, recall, and F1-score against a manually annotated claim corpus [1]; and graded risk

assessment, assessed by MAE and RMSE against annotator consensus, following the regression approach of [2]. Both depend on an annotated multimodal corpus, which does not currently exist for the proposed application [5], so corpus construction is a necessary precursor to quantitative validation. Inter-annotator agreement, measured using Cohen's kappa (κ), should also be reported to establish annotation reliability. The system is intended as a decision-support tool rather than an autonomous adjudication mechanism: each graded output is accompanied by its supporting evidence and justification, while thresholds remain subject to human review.

RESULTS AND DISCUSSION

Functional Outcomes

The prototype demonstrates end-to-end operation involving multimodal ingestion, concurrent modality analysis, cross-modal divergence computation, constrained justification, graded A–E output, and exportable audit generation. At the systems level, this addresses the multimodal detection gap identified in prior research, where image-based analysis was identified as future work [1].

Expected Detection Performance

Quantitative validation depends on the corpus described in Section IV-F, so this work does not claim independently measured classification accuracy. The expected value of

multimodal fusion is instead grounded in transferable findings: Wang et al. showed that combining complementary channels reduced error across RMSE, MAE, and MAPE compared with either individual channel [2]. True Green AI extends this principle across modalities rather than combining only multiple feature types within text, assuming that complementary and partially independent evidence sources reduce uncertainty in the joint assessment. CMDI is accordingly expected to add information beyond text-only classification, particularly for selective disclosure or expressive manipulation; this remains a research hypothesis requiring validation through an annotated multimodal dataset.

Interpretability and User-Side Outcomes

The A-E system has been developed in order to make interpretation easier compared to purely numeric or dichotomous systems, driven by the experiment on 2,109 participants where numeric disclosure and claims did not help consumers perceive reality as long as an interpretive five-level label helped them see through greenwash [3]. This is the logic of the design, not an independently verified effect of True Green AI. The evidence panel explains the reasons for a particular rating, and since being environmentally aware means being attentive to communication about greenwashing [6], the certification-verification phase puts certification labels into context.

Comparative Positioning

Table 2. Comparative Positioning of True Green AI

Property	EcoLens GRIE [1]	Dual-Channel ESG Regression [2]	General Multimodal LLM Audit [5]	True Green AI
Modality coverage	Text	Text with dual internal channels	Text + image	Text + visual + spoken
Cross-modal consistency	No	No; intratextual	No	Yes; CMDI
Generative output constraint	Partial	N/A	No	Yes; schema validation and fallback
Interpretative graded output	Partial; numerical metrics	No; regression score	No	Yes; evidence-supported A–E assessment
Full-stack implementation	Yes	No	N/A	Yes; Node.js/Express/Firebase/React

The comparison illustrates the intended architectural distinction of True Green AI: the system combines multimodal evidence with explicit rule-based verification, cross-modal analysis, constrained generation, and interpretable output.

Implications

1. Regulatory: The EU Green Claims Directive suggests that the focus should be placed on substantiation as a way of preventing consumer misunderstanding. Nevertheless, studies have found that even substantiated “climate-neutral” versions of such claims can yield results that are as bad as

those yielded by the original statements [3]. Hence, a system that only confirms the presence of substantiation is not guaranteed to detect any misleading communication practices, which is why True Green AI sees substantiation as one of many inputs into a cross-modal assessment.

2. Consumer and Investor: Labeling and trust were named as the two key variables involved in the perception of greenwashing in the food industry [4], while a connection has been established between the perception of greenwashing and unfavorable reactions related to trust [6]. For consumers, a preliminary assessment could provide additional information prior to a purchase decision. For investors, there are studies showing that an indicator-based score could be different from the greenwashing degree predicted by a model, especially in case of high levels of disclosure that do not guarantee environmental honesty [2]. This is why the system avoids using the volume of disclosure as a proxy for environmental honesty.

3. Green-Computing: Issues related to overreliance on AI in environmental reporting have already been discussed in existing literature [5]. True Green AI uses several design approaches to reduce unnecessary computing power and data transfers – client-side preprocessing minimizes the amount of transferred data, the architecture guarantees that the generative inference is limited to the necessary stage and there is a non-generative alternative, and rate limiting prevents too frequent inference calls. The system does not make a claim of carbon neutrality of its functioning and quantified audit of energy and carbon usage is left for future research.

Threats to Validity and Limitations

Three primary constraints exist. The first is that due to the lack of an annotated multimodal corpus, independent results on classification accuracy cannot be reported, thus limiting the expected gains in Sections V-B and V-C to transfer results based on other validated work. The second constraint is that CMDI is an innovative measure without any prior validation, and the rule-based weights have to be validated empirically using ground truth. Thirdly, browser-based speech transcription has not been benchmarked independently in this work, and thus spoken information is considered text-equivalent input and not a separate linguistic modality. Other constraints are the relatively small product corpus and the supporting nature of the machine learning component.

CONCLUSION AND FUTURE WORK

True Green AI, which is a multimodal approach to environmental sustainability claim auditing using text,

image, and voice data, has been introduced in this paper based on four major gaps recognized through literature review. These gaps include the lack of visual and auditory modalities [1], [5], failure of intratextual methodologies in detecting communication-substance inconsistencies [2], unreliability of unbounded multimodal generative auditing [5], and numeric or substantiated-but-uninterpreted information may not adequately correct consumer misperception [3].

The system combines multimodal ingestion, AI-assisted claim and evidence interpretation, R1–R6 rule-based verification, supporting machine learning, cross-modal consistency assessment, and retrieval-grounded explanation. Its reference implementation uses a React 19/Vite 6/Tailwind CSS 4 client, browser-native multimodal capture, a Node.js/Express service layer, a schema-constrained @google/genai integration, and Firebase-backed persistence with an auditable verification history. A central feature is the proposed Cross-Modal Divergence Index, quantifying inconsistencies between environmental claims communicated through different modalities, together with an interpretable A–E output supported by component-level evidence rather than a binary decision or isolated numerical score. The system is positioned as a decision-support framework rather than an autonomous adjudicator, and quantitative validation remains necessary before it can be considered empirically established.

Four directions are proposed for future work. First, an annotated multimodal greenwashing corpus should be constructed and evaluated using inter-annotator agreement, providing the basis for quantitative evaluation of the classification and risk-assessment components. Second, the framework can extend to video and other temporal modalities, capturing environmental claims that develop across an advertisement. Third, the rule-based weighting of CMDI can be replaced or supplemented with a learned model once sufficient ground-truth data become available, building on the fusion methodology of [2]. Fourth, the retrieval-grounded justification layer can be extended through knowledge-graph integration using verified third-party environmental, emissions, and certification data [5], supporting longitudinal monitoring of sustainability claims and comparison between communicated claims and documented environmental performance over time.

Overall, True Green AI provides a structured foundation for multimodal environmental-claim auditing by combining transparent rules, AI-assisted interpretation, supporting

machine learning, cross-modal analysis, and evidence-grounded explanations.

REFERENCES

- [1] U. Mogili, M. Sahitya, V. Leelakrishna, K. Chandini, K. Mohini, and K. Venkataramana, "EcoLens GRIE: AI-driven detection of greenwashing claims using transformer-based NLP," *International Journal of Research in Applied Science and Engineering Technology*, vol. 14, no. III, pp. 1868–1873, Mar. 2026, doi: 10.22214/ijraset.2026.78204.
- [2] X. Wang, Y. Chen, and H. Yao, "Revealing the veil of greenwashing in ESG reports: Predicting the degree of corporate greenwashing based on thematic and sentiment features of text," *IEEE Transactions on Sustainable Computing*, vol. 10, no. 6, pp. 1173–1188, Nov./Dec. 2025, doi: 10.1109/TSUSC.2025.3594401.
- [3] D. Dreist, A. Zühlsdorf, A. Spiller, and S. Kühl, "Greenwashing in food labelling: Consumer deception by claims of climate neutrality and the importance of an interpretative labelling approach," *Food Quality and Preference*, vol. 122, Art. no. 105294, 2025, doi: 10.1016/j.foodqual.2024.105294.
- [4] W. S. Nugraha, D. Szakos, M. Süth, and G. Kasza, "Greenwashing in the food industry: A systematic review exploring the current situation and possible countermeasures," *Cleaner Responsible Consumption*, vol. 15, Art. no. 100227, 2024, doi: 10.1016/j.clrc.2024.100227.
- [5] A. Vishnoi, A. S. Chauhan, and M. I. Khan, "Detecting greenwashing through artificial intelligence: Methods and challenges," *International Journal of Innovative Science and Engineering Management*, vol. 5, no. 1, pp. 209–216, Mar. 2026, doi: 10.69968/ijisem.2026v5i1209-216.
- [6] T. Vukasović, M. Trglavčnik, and A. Faganel, "Greenwashing and green marketing on social media: Implications for trust-related reactions in the food sector," *Foods*, vol. 15, no. 10, Art. no. 1626, 2026, doi: 10.3390/foods15101626.