EmeraldMind: A Knowledge Graph–Augmented Framework for Greenwashing Detection

Georgios Kaoukis*
g.kaoukis@athenarc.gr
Archimedes, Athena Research Center
Athens, Greece

Ioannis-Aris Koufopoulos*
j.koufopoulos@athenarc.gr
Archimedes, Athena Research Center
Athens, Greece

Eleni Psaroudaki
h.psaroudaki@athenarc.gr
Archimedes, Athena Research Center
Athens, Greece

Danae Pla Karidi†
danae@athenarc.gr
Archimedes, Athena Research Center
Athens, Greece

Evaggelia Pitoura
pitoura@uoi.gr
University of Ioannina
Ioannina, Greece
Archimedes, Athena Research Center
Athens, Greece

George Papastefanatos
gpapas@athenarc.gr
IMSI, Athena Research Center
Athens, Greece

Panayiotis Tsaparas
tsap@uoi.gr
University of Ioannina
Ioannina, Greece
Archimedes, Athena Research Center
Athens, Greece

Abstract

As AI and web agents become pervasive in decision-making, it is critical to design intelligent systems that not only support sustainability efforts but also guard against misinformation. Greenwashing, i.e., misleading corporate sustainability claims, poses a major challenge to environmental progress. To address this challenge, we introduce EmeraldMind, a fact-centric framework integrating a domain-specific knowledge graph with retrieval-augmented generation to automate greenwashing detection. EmeraldMind builds the EmeraldGraph from diverse corporate ESG (environmental, social, and governance) reports, surfacing verifiable evidence, often missing in generic knowledge bases, and supporting large language models in claim assessment. The framework delivers justification-centric classifications, presenting transparent, evidence-backed verdicts and abstaining responsibly when claims cannot be verified. Experiments on a new greenwashing claims dataset demonstrate that EmeraldMind achieves competitive accuracy, greater coverage, and superior explanation quality compared to generic LLMs, without the need for fine-tuning or retraining.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Information systems** → *Information retrieval*.

*Both authors contributed equally to this research.
†Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
WWW '26, Dubai, United Arab Emirates
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792997

Keywords

Greenwashing Detection, Retrieval Augmented Generation (RAG), Knowledge Graph, Responsible AI, Evidence-based Explanations

ACM Reference Format:
Georgios Kaoukis, Ioannis-Aris Koufopoulos, Eleni Psaroudaki, Danae Pla Karidi, Evaggelia Pitoura, George Papastefanatos, and Panayiotis Tsaparas. 2026. EmeraldMind: A Knowledge Graph–Augmented Framework for Greenwashing Detection. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774904.3792997

Resource Availability:
The source code and datasets are available at: https://doi.org/10.5281/zenodo.18338874 and https://github.com/ai4greenwashing/EmeraldMind.

1 Introduction

Greenwashing refers to the corporate practice of conveying a misleading impression of environmental responsibility through advertisements, statements across media channels, and traditional communication platforms. Common examples include presenting regulatory compliance as a product's environmental benefit, or advertising green claims to entire products when they pertain only to specific parts or aspects. Greenwashing misleads consumers and investors with false sustainability claims, undermining trust, hindering genuine environmental efforts, and allowing harmful practices to persist while delaying urgent action on climate change.

Journalists assess potential greenwashing by seeking evidence in regulatory records from authorities such as the Advertising Standards Authority (ASA) or corporate Environmental, Social, and Governance (ESG) reports. ESG reports are published annually by organizations to disclose their performance and practices via standardized, non-financial metrics, i.e., key performance indicators (KPIs). KPIs are defined by the European Union's Sustainable

Finance Disclosure Regulation and Corporate Sustainability Reporting Directive. Monitoring these metrics and cross-checking them against other corporate information, such as advertising campaigns and news coverage, can uncover inconsistencies, including environmental impacts that are inaccurately reported or overstated.

Greenwashing detection represents a specialized subset of fact-checking research, presenting some unique challenges when realized through automated retrieval-augmented generation (RAG) pipelines. The first refers to the extraction and modeling of domain-specific information that can be used in automated verification pipelines [1, 14, 18]. Traditional fact-checking pipelines relying on generic sources, such as scholarly literature or general web information, often lack the specialized evidence required for sustainability claims, like emissions KPIs. This data gap results in incomplete or misleading verification results [20, 35]. Thus, *greenwashing detection demands tailored retrieval and processing of multimodal content from specialized knowledge sources like ESG reports and regulatory records beyond general repositories.*

Second, the design space of RAG pipelines can be vast, with systems sourcing and combining information from diverse repositories (e.g., internal knowledge bases, web data, ESG reports). This leads to highly variable outputs that depend heavily on retrieval quality and prompting strategies, with sensitivities further amplified by ambiguous greenwashing definitions [3]. In high-stakes domains like sustainability, binary labels are insufficient; *models must provide evidence-backed justifications to ensure trust and accountability among stakeholders* [20, 29, 30, 38]. Accuracy-focused evaluation can be misleading, as high scores can hide excessive abstentions, i.e., cases where the system does not make a judgment, or ungrounded guesses, both of which obscure the system's reasoning and limit practical adoption. For real-world use, RAG-based systems must maintain high coverage, balancing accuracy with evidence-backed assessments. Consequently, the challenge lies in designing a suitable RAG pipeline that *maximizes the number of claims assessed with well-supported, evidence-based justifications, while minimizing unsupported predictions and abstentions.*

Third, a major obstacle for greenwashing detection remains the scarcity of annotated datasets [3]. The lack of high-quality annotations limits reliable model evaluation and constrains domain-specific fine-tuning shown to improve performance [22]. These tasks require substantial annotated data, which is costly, time-intensive to create, and demands significant expertise. *This limitation underscores the need for benchmarks that enable scalable evaluation and fine-tuning of greenwashing detection models through annotated sustainability claims.*

To address the above challenges, we propose `EmeraldMind`, a domain-specific RAG framework for greenwashing detection. Given a textual sustainability claim, `EmeraldMind` determines whether it constitutes greenwashing and provides a fact-based justification, or abstains when evidence is insufficient. Specifically, `EmeraldMind` analyzes ESG claims by integrating corporate ESG reports and KPI definitions to ground LLM reasoning in evidence-based verdicts. Our framework is built on two complementary custom sustainability knowledge stores: the *EmeraldGraph* that captures ESG-specific entities and relations, as extracted from ESG reports, KPI definitions [8], and widely known greenwashing claim examples, and the *EmeraldDB* that captures raw ESG text. The two stores are coupled with tailored retrieval mechanisms to gather relevant evidence for each claim. By grounding Large Language Model (LLM) reasoning in this domain-specific knowledge, our system produces a verdict plus an evidence-backed justification, or abstains when insufficient facts exist, enabling auditable outputs that address accuracy-only evaluation limitations. To evaluate our framework, we propose a semi-synthetic data benchmark named `EmeraldData`, which can also be used to fine-tune other models. To the best of our knowledge, this is the first work to propose an end-to-end *knowledge-based retrieval-augmented generation* framework for *greenwashing detection*. Our key contributions can be summarized as follows.

(1) **`EmeraldMind` framework**: We introduce a domain-specific, knowledge-based, and abstention-aware RAG pipeline. Given a sustainability claim, `EmeraldMind` produces a verdict (greenwashing, not greenwashing, or abstain) along with a natural-language justification grounded in retrieved evidence.

(2) ***EmeraldGraph* knowledge graph:** We develop a structured sustainability knowledge graph centered on company-specific ESG entities and relations. Our retrieval algorithm prioritizes critical entities such as companies and KPIs to ensure that relevant context is extracted for responsible reasoning and precise claim evaluation.

(3) **`EmeraldData` benchmark:** We release a novel evaluation dataset for greenwashing detection, comprising 620 semi-synthetic corporate sustainability claims, each paired with a ground-truth label. This benchmark facilitates transparent evaluation and comparative testing of greenwashing detection systems.

(4) **Experimental evaluation:** We conduct a thorough experimental evaluation comparing `EmeraldMind`'s variants, demonstrating superior performance and emphasizing the critical role of justification quality in greenwashing detection evaluation.

The paper structure is: Section 2 reviews related work; Section 3 provides the `EmeraldMind` framework overview; Section 4 details the evidence stores construction; Section 5 describes retrieval and reasoning variants; Section 6 introduces the `EmeraldData` benchmark, followed by experimental results and justification analysis in Section 7. Finally, Section 8 concludes and outlines future work.

2 Related Work

Fact-Checking and Greenwashing Detection with LLMs. Automated fact-checking systems often rely on structured resources to ground verification and produce explanations; yet, LLMs still trail human experts in fact-checking accuracy [4]. Specialized models fine-tuned on domain-specific data yield the best performance [22]; however, their success hinges on the availability of substantial annotated datasets, which are prohibitively expensive and time-consuming to produce, demanding significant domain expertise.

To mitigate this, synthetic data generation methods [23, 37], and frameworks like UNOWN [2] have emerged to automate dataset creation at scale across textual and tabular modalities. In the environmental domain, greenwashing detection presents additional challenges due to ambiguous definitions and data scarcity [3]. Recent work has explored the use of machine learning and LLMs to evaluate ESG-related statements [3]. For example, the `ESGenius` benchmark provides structured reasoning datasets for evaluating environmental question-answering performance of LLMs [15], while GLITTER [1] employs RAG to cross-reference corporate disclosures with

external data, evaluating the consistency between self-reported narratives and external ESG ratings. These efforts underline the potential of LLM-driven reasoning to identify deceptive sustainability narratives, but they lack explicit grounding in a structured ESG knowledge context.

To the best of our knowledge, `GreenClaims` [11], introduced by Fornasiero [10], is the only available dataset specifically for greenwashing detection. It contains third-party verified cases derived from reputable news sources and regulatory records like the ASA. However, its small size limits broader applicability, necessitating expanded and continuously updated datasets [3].

Knowledge Graphs (KGs) and RAG for ESG Fact-Checking. The hallucination problem in LLM outputs motivated the integration of retrieval-augmented generation (RAG) methods, where relevant external evidence is incorporated into LLM prompts to improve factual accuracy [30]. By retrieving relevant evidence before generation, RAG systems strengthen verification-oriented tasks. Environmental domain examples include ChatReport [25] and ChatClimate [33], which incorporate ESG-specific retrieval and summarization for sustainability news and corporate reports. These pipelines, however, primarily focus on open-ended question answering rather than greenwashing claim verification.

Building on these foundations, KG-based RAG solutions integrate entity-level reasoning into retrieval and generation, improving factual precision and interpretability over generic text retrieval approaches. [14] for instance, uses RAG for question answering from news articles. However, none of the ESG-specific methods currently provides an end-to-end pipeline for greenwashing detection in sustainability claims.

Automatic and semi-structured KG construction remains an active research direction [5, 41]. In the ESG domains, several structured ontologies and knowledge graphs have been developed. OntoSustain [32, 43] formalizes corporate sustainability concepts around Global Reporting Initiative (GRI) indicators, while KnowUREnvironment [16] and SustainGraph [12] integrate environmental and Sustainable Development Goals related entities into comprehensive graphs. Other efforts, such as RSOKG [42] and ESGOnt [34], unify standards across the GRI and the European Sustainability Reporting Standards (ESRS), reinforcing the value of structured sustainability indicators. Despite their value, these ontologies have not been incorporated into end-to-end greenwashing detection pipelines.

3 The EmeraldMind Framework

In this section, we present an overview of the `EmeraldMind` framework. `EmeraldMind` is a domain-specific knowledge-graph RAG-based framework for sustainability claim verification and greenwashing detection. Given an input claim, `EmeraldMind` assesses its veracity and decides whether it constitutes greenwashing. Sustainability claims are often vague or self-declarative and cannot be reliably validated using the language model's internal knowledge alone. Consequently, to produce verdicts that are both evidence-based and explainable, `EmeraldMind` grounds its decisions in an external ESG-based context rather than relying solely on the LLM's latent knowledge.

`EmeraldMind` adopts a two-stage architecture that explicitly separates the construction of evidence from its use in reasoning (see

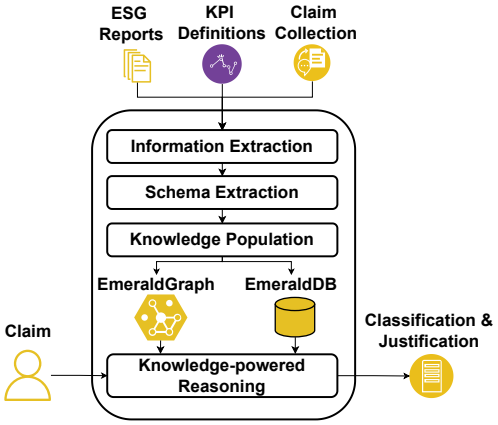


Figure 1: `EmeraldMind` Framework

Figure 1). The **Evidence Stores construction phase** (presented in Section 4) serves a critical role: it transforms unstructured multimodal ESG reports into both structured, queryable representations and vectorized natural language chunks. This phase normalizes heterogeneous content, induces a semi-automatically derived KPI sub-schema, and constructs two complementary stores: *EmeraldGraph*, a property graph encoding ESG entities and relations, and *EmeraldDB*, a document repository that preserves textual context and provenance metadata. Together, they constitute the verifiable `EmeraldMind` Evidence Stores.

During the **Knowledge-powered Reasoning phase** (presented in Section 5), the framework receives a user-provided claim and operates over the preconstructed evidence stores. By grounding the claim in *EmeraldGraph* and retrieving semantically related chunks from *EmeraldDB*, we create the context to support the reasoning. The final classification module delivers a verdict, classifying the claim as greenwashing or not, or abstains if the evidence is inconclusive. In all cases, the system generates a concise justification grounded in the retrieved facts.

4 EmeraldMind Evidence Stores

In this section, we detail the construction of the evidence stores, which comprise the *EmeraldDB* document store and the *EmeraldGraph* knowledge graph.

EmeraldDB. *EmeraldDB* is a vectorized document store that helps retrieve relevant information to support reasoning. Each ESG report is broken down into smaller sections, and key metadata for these sections, such as the report name, the referenced company, year, and page, are stored in *EmeraldDB* so they can be efficiently accessed when needed.

EmeraldGraph. *EmeraldGraph* is a domain-specific knowledge graph. It offers the following features: (1) it captures ESG data (e.g., facility-level emissions) that are typically absent from generic knowledge graphs, (2) it enforces structural clarity by distinguishing semantically similar but logically distinct entities (e.g., targets vs. actual performance), and (3) it improves auditability by grounding verdicts in explicit reasoning paths (e.g., `Company` → `reportsKPI`)

rather than opaque LLM justifications. *EmeraldGraph* is defined as a labeled property graph $G = (V, E, T, S, L, \mathcal{A}, \tau, p)$, where:

- V is the set of nodes, each representing a real-world entity mentioned in an ESG report (e.g., a company, facility, KPI observation, sustainability goal, etc.).
- T is a finite set of entity types (node labels) used in the graph schema (e.g., `Company`,`Facility`,`KPIObservation`, etc.).
- $\tau : V \to T$ is the function associating every node $v \in V$ with its entity type $\tau(v) \in T$. For example, $\tau(v)$ = Facility if v represents a plant.
- L is the set of relationship types (edge labels).
- $E \subseteq V \times L \times V$ is the set of directed labeled edges (u, ℓ, v) representing facts, e.g., (`ACME`, `reportsKPI`, `EmissionObservation`).
- $S \subseteq T \times L \times T$ is the domain schema, i.e., the set of allowable typed relationships. For example, $(Facility, locatedIn, Location) \in S$ constrains the graph construction.
- $\mathcal{A}$ is the set of key–value (properties) pairs. For example, `KPIObservation`: {`value`, `unit`, `year`}).
- $p : V \cup E \to \mathcal{A}$ maps each node/edge to its properties (e.g., $p(v) = \{value : 2300, unit : tCO2e, year : 2025\}$).

The construction of G requires (1) assigning each node v a type $\tau(v) \in T$ and (2) attaching its properties $p(v)$. A central challenge is ensuring schema consistency. Every extracted fact (u, ℓ, v) must satisfy three constraints: (i) the typed relationship must conform to S (i.e., $\tau(u) \xrightarrow{\ell} \tau(v)$), (ii) the resulting edge (u, ℓ, v) must belong to E, and (iii) all properties in p must satisfy type constraints (e.g., numeric fields for KPIObservation nodes).

To construct these stores from ESG reports, KPI definitions, and claim collections, we execute the following stages: (1) Information Extraction, which extracts parsed text from the reports, (2) Schema Extraction, which induces a domain schema T for ESG concepts, (3) Knowledge Population that populates the evidence stores into the complementary architectures of the *EmeraldDB* and the *EmeraldGraph*.

Information Extraction. It processes raw ESG reports into parsed documents used by *EmeraldGraph* and *EmeraldDB*. Corporate ESG reports are highly multimodal: narrative text, tables, charts, and figures all carry quantitative information. The main challenge is to extract the full semantic content without losing context [1, 44].

We employ a two-channel information extraction pipeline that processes textual and multimodal content separately, as shown in Figure 2. The text channel uses a PDF parser (PyMuPDF [28]) to extract paragraphs, headings, footnotes, metadata, and tables. The multimodal channel renders each page as an image and uses a vision-language model to recover semantic content from charts, figures, and tables, including structured fields (series, axes, units), captions, and descriptive context that would otherwise be lost. We reconcile outputs from both channels through alignment and deduplication, retaining both text and structured representations.

For each report, we produce a parsed representation consisting of text spans, normalized tables, and chart or figure descriptions with associated metadata. *EmeraldDB* stores these spans as retrieval passages, forming a document store for semantic retrieval. *EmeraldGraph* consumes the same parsed structures as input to downstream

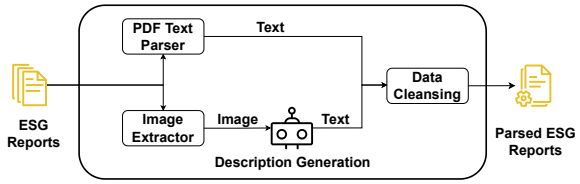


Figure 2: Information Extraction

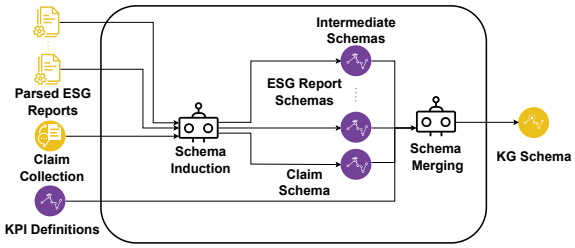


Figure 3: Schema Extraction

schema extraction and knowledge population, where typed entities and relations are constructed.

Schema Extraction. It synthesizes a domain-specific schema S that constrains all entities and relations in *EmeraldGraph*. S defines the allowed entity types, relation labels, and attribute domains; for instance, a `Facility` node can only be linked to a `Location` node via a `locatedIn` relation. The primary challenge in schema design lies in balancing expressivity and restriction. The schema must accommodate structurally heterogeneous and semantically inconsistent ESG reports (where, e.g., KPIs often vary in labels, units, and layouts), while preventing invalid entities or relations (arising, e.g., from confusing targets, baselines, and actual performance metrics).

The schema S is synthesized from three sources as depicted in Figure 3: (i) data-driven patterns observed in the parsed ESG corpus, yielding a merged schema S_{data} of frequent entity–relation configurations induced from each report, (ii) a regulatory subschema S_{reg} of key performance indicators derived from official standards, and (iii) a claim-driven schema S_{claim} constructed from known greenwashing examples by abstracting common claim patterns. These partial schemas are merged into a unified schema $S = S_{\text{data}} \cup S_{\text{reg}} \cup S_{\text{claim}}$ that specifies allowed entity types (e.g., `Company`, `Facility`, `KPIObservation`, `SustainabilityClaim`) and relation types (e.g., `reportsKPI`, `setsGoal`) along with their attribute domains. One key design choice is a company-centered schema, where a single `Organization` node anchors all entities and relations for each company, normalizing heterogeneous disclosures around a consistent corporate identity. For example, all emission reports by Company X attach to that `Organization` node, ensuring structural clarity. This schema-based modeling enables efficient context retrieval, constrains inference to valid schema patterns for company-centered claims, and supports explainable evidence trails.

Knowledge Population. Using the parsed reports and extracted schema, we instantiate the *EmeraldDB* and *EmeraldGraph* stores with actual data from reports as depicted in Figure 4. To populate the *EmeraldDB*, each report is segmented into chunks d_i, and

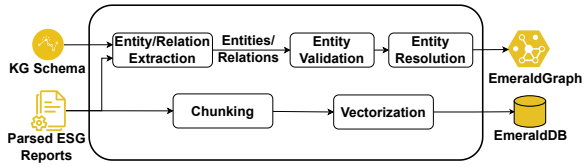


Figure 4: Knowledge Population

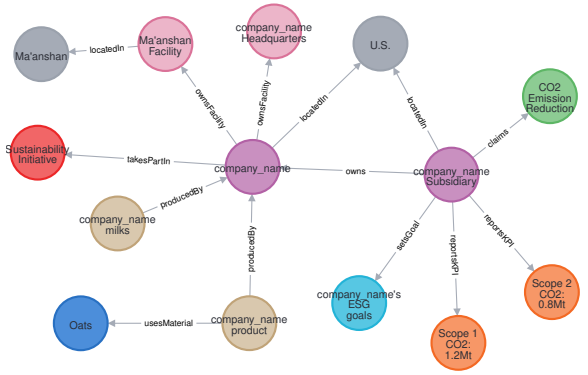


Figure 5: EmeraldGraph snippet centered on an anonymized company. Node types in the EmeraldGraph knowledge graph are color-coded as follows: Initiative, Location, Material, Product, Goal, Sustainability Claim, Organization, Facility, KPIObservation.

an embedding is computed for each. We store for each chunk: d_i=(embedding, reportID, company, year, chunkID, page_number). Chunks have a 250-token size, with a 50-token overlap to preserve contextual meaning, while enabling effective reasoning.

To populate the *EmeraldGraph*, report content is mapped to entities and relations defined by schema S. To address extraction accuracy and ensure entity uniqueness (e.g., resolving "ABC Corp." vs. "ABC Corporation"), we map noisy document mentions to schema-typed entities and relations. We further employ embedding-blocking [17, 21, 27, 36] to prevent duplicate or misaligned segments, ensuring each real-world entity corresponds to a single node. For each parsed ESG report, an LLM extracts candidate triples (u, ℓ, v). A candidate triple (u, ℓ, v) is only admitted if its type is valid against the schema S under $\tau(u) \xrightarrow{\ell} \tau(v)$. Figure 5 shows a real-world snippet from EmeraldGraph centered on an anonymized company.

5 Knowledge-powered Reasoning

In this section, we present the Knowledge-powered Reasoning phase (Figure 6), evaluating a textual sustainability claim c (e.g., "*Company X reduced its CO_2 emissions by 30% in 2023*") using evidence from *EmeraldGraph* and *EmeraldDB*. We consider three pipeline configurations: `EM-KGRAG`, which uses only graph evidence H from *EmeraldGraph*; `EM-RAG`, which uses only document chunks $\{d_i\}$ from *EmeraldDB*; and `EM-HYBRID`, which uses as context the justifications produced by `EM-KGRAG` and `EM-RAG`.

Claim Grounding. Given a claim c, claim grounding identifies the target company and parses c for other key elements: KPIs,

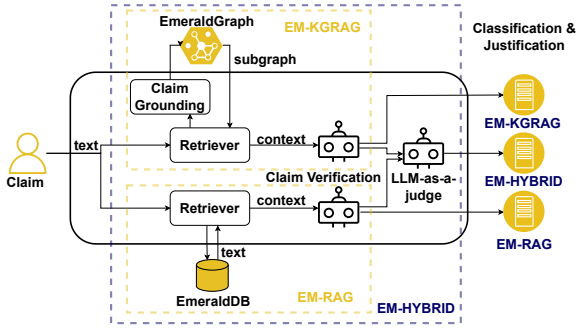


Figure 6: Knowledge-powered Reasoning

numeric values, policy mentions, and goals. Then, it retrieves the Organization node $v_{\text{company}} \in V$ from *EmeraldGraph* that represents the identified company, and grounds the other key elements by mapping them to nodes $v \in V$ of corresponding types $\tau(v) \in T$ in the graph. Each extracted element is mapped to a schema type in S (e.g. `KPIObservation`, `Policy`, `Goal`) and the node's attributes are populated using the property function p. For example, in the claim "*Company X reduced its CO_2 emissions by 30% in 2023*", we would link "Company X" to the $v_{company}$, and resolve the "30%" emission into a new node $v_{observation}$ connected to $v_{company}$, where $\tau(v_{obs}) = \texttt{KPIObservation}$, and $p(v_{observation}) = \{\text{value} : 30, \text{unit} : \%, \text{year} : 2023\}$. Claim grounding outputs the resolved $v_{company}$ node and a node set V_{claim} of the key-elements defined by their specific types τ and property maps p.

We identify two key challenges in claim grounding. First is its linguistic variability and ambiguity. Sustainability claims often use imprecise language (e.g., "net-zero" vs. a concrete target, or "our emissions" without a unit or year) and industry-specific terminology. For example, the term "flaring" must be recognized as "burning gas" and thus "CO_2 emissions"; otherwise, the system might store it as a SafetyIncident instead of an EmissionObservation. Mapping such phrases to precise schema types requires robust LLM-based parsing. We therefore employ a schema-based retrieval algorithm that restricts retrieval to the types explicitly referenced in the claim, focusing reasoning on concrete entities and avoiding unrelated graph regions. Second, the claims do not necessarily include KPI elements, which are imperative for the retrieval of quantifiable evidence. To mitigate this, we augment the claim-extracted node set V_{claim} by adding a KPIObservation node u with $\tau(u) = \texttt{KPIObservation}$ (initialized from the claim text) to ensure the retrieval algorithm always queries KPIObservation nodes, even for implicit KPI mentions.

Schema-Based Context Retrieval. Schema-based context retrieval extracts graph-based context from *EmeraldGraph*, which is used for reasoning. A key challenge is extracting a compact, claim-specific evidence subgraph. Naïve neighborhood expansion around v_{company} either floods the context with loosely related nodes or omits multi-hop facts critical for greenwashing detection (e.g., links to nodes topologically deeper in the graph, such as Material).

Schema-based context retrieval addresses this challenge through the following approach: Using the grounded company node v_{company} and the set of schema types $\{\tau_j\}$ identified in claim c, we construct

Algorithm 1 Schema-driven Context Retrieval

Input: $v_{company}$ (organization node ID), $\mathcal{V}_{claim}$ (nodes extracted from claim grounding), top_n (no. retrieved nodes per type), $threshold$ (similarity threshold), k (max path length in hops)
Output: H (subgraph with retrieved nodes and edges)

```
H ← ∅
for all v ∈ V_claim do
    V_r ← Retrieve(v_company, τ(v), k)        ▷ Collect all nodes of the specified type within company's k-hop neighborhood
    V_sim ← ∅                                  ▷ Nodes that pass the similarity threshold
    for all u in V_r do                        ▷ Compute similarities and select top-n nodes
        sim ← CosSim(Embedding(v), Embedding(u))
        if sim ≥ threshold then
            V_sim ← V_sim ∪ {(u, sim)}
        end if
    end for
    V_tn ← TopN(Sort(V_sim), top_n)            ▷ Top-n most similar nodes
    for all u ∈ V_tn do                        ▷ Compute the shortest paths for top-n nodes
        path ← ShortestPath(v_company, u)
        H ← H ∪ {path}
    end for
end for
return H
```

an evidence subgraph $H \subseteq G$. Algorithm 1 performs, for each type τ_j, a breadth-first search from v_{company} up to k hops, collecting all nodes of type τ_j. It then ranks these candidates by cosine similarity between their node embeddings and the corresponding key elements found in the claim and keeps the top-n most similar nodes per type above a threshold. For each selected node v, we compute the shortest path $P(v_{\text{company}}, v)$ and add all nodes and edges on these paths to H. The resulting context subgraph consists of reasoning paths rather than isolated facts.

Document Retrieval. Document retrieval retrieves textual evidence from *EmeraldDB*. Using the company name extracted from the claim, we use the identifier *company* to limit our search to relevant chunks within the document store. We compute the embedding of the claim c and return the top-m chunks based on cosine similarity. Restricting to the target company avoids irrelevant documents and ensures that retrieved chunks are contextually aligned with the claim, improving factual grounding.

Classification and Justification. In `EM-KGRAG` and `EM-RAG` variants, the claim c and its corresponding context are fed into a prompted LLM, which performs contextual reasoning to produce the final verdict (Greenwashing, Not Greenwashing, Abstain) and a factual justification for the classification label. The `EM-HYBRID` variant adopts an LLM-as-a-judge setup: given the claim and the two (label, justification) pairs from `EM-RAG` and `EM-KGRAG`, a verifier LLM selects the better-supported explanation and outputs its label and justification.

6 The EmeraldData Benchmark

A fundamental limitation of existing research on greenwashing is the absence of large-scale annotated real-world benchmarks containing verified instances of greenwashing [3]. Several factors contribute to this scarcity, including the vagueness of greenwashing definitions, context-sensitive claims, annotation complexity requiring domain expertise, and legal and reputational implications of labeling corporate claims as deceptive.

Table 1: Summary of the datasets used in the experimental evaluation with G as the number of Greenwashing and NG as the number of Not Greenwashing.

Dataset	No. Claims	G	NG
`GreenClaims`	51	24 (47%)	27 (53%)
`EmeraldData`	620	225 (36%)	395 (64%)

To the best of our knowledge, the only relevant publicly available greenwashing detection dataset is `GreenClaims` [11]. It contains only a limited number of 91 claim samples, from which only 51 were usable in our evaluation due to the availability of corresponding ESG reports. To overcome this limitation, we introduce `EmeraldData`, a larger semi-synthetic dataset (620 instances), constructed via a four-stage pipeline inspired by [2, 26, 37]. First, we used the smaller `GreenClaims` benchmark to extract 37 (company, year) unique pairs. These allow us to align claims with the ESG reports used for the creation of `EmeraldMind` Evidence Stores. Second, we collect relevant articles from reliable news sites that cover topics including, but not limited to, greenwashing, sustainability news, ESG news, company news related to various ESG goals, and any accusations or litigation that a company may face. We filter them by (company, year) pairs as defined in the first step, ensuring contextual relevance. Third, we prompt an LLM with article metadata to generate both truthful (non-greenwashing) and refuting (greenwashing) claims, yielding 620 candidate instances. Finally, the same model assigns a label to each claim and produces a brief textual justification anchored to the source article, enabling transparent, article-grounded evaluation. Table 1 summarizes the datasets.

7 Experimental Evaluation

7.1 Experimental Setup

We evaluated all `EmeraldMind` variants: `EM-RAG`, `EM-KGRAG`, and `EM-HYBRID`, benchmarked against a baseline LLM relying solely on internal knowledge under two prompt configurations. The first (zero-shot prompt) uses only claim text and greenwashing definitions from the EU Green Claims Directive [6]. The second (few-shot prompt) adds few-shot examples found in [3]. This comparison demonstrates how domain-specific evidence improves trustworthiness over generic learned patterns, aligning with responsible AI principles. Evaluation metrics include classification accuracy, coverage, and justification quality.

All code, prompts, and benchmarks are publicly available in: https://github.com/ai4greenwashing/EmeraldMind. Retrieval hyperparameters are in the Appendix Table 6. Experiments were conducted on a system equipped with an NVIDIA RTX A6000 GPU with dual AMD EPYC 9335 32-core processors and 128 GB RAM, using `gemma-27b-it` for inference, `prometheus-13b-v1.0` for ILORA evaluation, and for ranking / hybrid judging `prometheus-7b-v2.0`. Prometheus [19] is an open-source evaluator LLM designed for reproducible, fine-grained assessment as a practical alternative to human evaluation.

Table 2: EmeraldGraph Statistics

Graph Metrics		Centrality by Node Type	
Metric	**Value**	**Node Type**	**Avg. Degree**
No. of Entities	53,748	Organization	17.95
No. of Relationships	59,344	Country	6.40
Avg. Total Degree	2.21	Location	3.43
Avg. Shortest Path Length	4.58	Facility	2.63
Diameter	15	Material	1.77

Table 3: Top-5 Entity and Relationship Types by Frequency

Top-5 Entities		Top-5 Relationships	
Entity Type	**Count**	**Relation Type**	**Count**
KPIObservation	24,809	reportsKPI	24,832
Initiative	4,060	takesPartIn	4,388
SustainabilityClaim	3,458	setsGoal	3,475
Goal	3,414	claims	3,446
Organization	3,020	locatedIn	3,396

7.2 EmeraldGraph Knowledge Graph

We constructed *EmeraldGraph* using the pipeline described in Section 4, extracting structured facts from 37 publicly available ESG reports. Table 2 shows that *EmeraldGraph* is a sparse, company-centered graph. The 53,748 entities and 59,344 relationships indicate low edge density and predominantly small local neighborhoods. The centrality statistics by node type reveal that Organization nodes (companies) act as high-degree hubs in a core-periphery topology, inducing star-like networks centered on companies.

Table 3 highlights how the graph density is concentrated on KPI- and claim-related structure. KPIObservation accounts for 24,809 nodes, which is approximately 46% of all entities, and `reportsKPI` for 24,832 edges, which is approximately 42% of all relationships. This implies dense connectivity between the two entity types. The remaining frequent entity types define the dominant local motifs around each company, capturing performance reporting, goal setting, initiatives, and geographic scope. Overall, the degree distribution and type frequencies confirm that *EmeraldGraph* allocates most of its structural capacity to modeling company-centric KPI observations, claims, and goals, which are precisely the regions exploited by our schema-based retrieval algorithm. The entire *EmeraldGraph* schema is available in the Appendix Figure 8.

7.3 Experimental Results

Classification Performance. To evaluate `EmeraldMind`, we report accuracy (correct/answered), coverage (answered/total), and overall accuracy (accuracy × coverage), revealing both performance and selectivity. Overall accuracy penalizes excessive abstentions while rewarding correct predictions on verifiable claims for practical utility. Table 4 summarizes these performance metrics across both evaluation datasets.

`EmeraldMind` variants substantially outperform the baseline in coverage across both datasets, achieving 2-4 times higher rates (49-77% vs. 19-31%) while maintaining competitive accuracy (77-93% vs. 93-100%). Overall accuracy, accounting for abstentions, shows that `EM-HYBRID` achieves the highest effective performance (up to 70.59% on `GreenClaims` few-shot), as it leverages LLM-as-a-judge pairwise comparison to select superior justifications from `EM-RAG` and `EM-KGRAG`, suggesting that a good quality justification leads to more accurate classification. Few-shot prompting consistently boosts the coverage of the `EmeraldMind` pipelines compared to zero-shot, though baseline coverage remains stagnant or declines. This confirms RAG-based reasoning benefits from example-guided reasoning, reducing abstentions. On the smaller `GreenClaims` dataset (51 claims), `EM-KGRAG` excels with the highest zero-shot accuracy (93.55%) and few-shot overall accuracy (68.63%). Conversely, `EM-RAG` dominates the larger `EmeraldData` (620 claims) with superior coverage (62-77%) and overall accuracy (55-62%). `EM-HYBRID` maximizes performance in both by successfully combining the two variants' performance strengths.

Justification Quality. The metrics in Table 4 do not fully capture the quality of reasoning behind the justification generation. Prior research has shown that classification metrics do not assess hallucination, factual consistency, or the soundness of model reasoning [7]. To address this, we follow approaches that incorporate LLM-driven qualitative assessment [13]. We adopt an LLM-as-a-judge paradigm [39] with two complementary evaluation strategies. First, *single answer grading* (*absolute LLM judge* [31]), which scores each `EmeraldMind` variant and baseline justification independently based on established criteria. Second, *pairwise comparison* (*relative LLM judge* [31]), which compares justifications from `EM-RAG`, `EM-KGRAG`, and the baseline directly. Since `EM-HYBRID` selects the higher-ranked justification between `EM-RAG` and `EM-KGRAG` from pairwise comparisons, we exclude it from this evaluation.

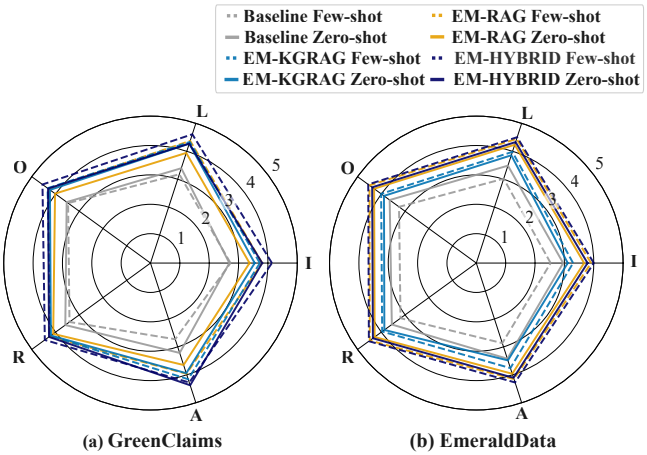
(1) Single Answer Grading. For the evaluation of the justifications, we use ILORA [40], an Explanation Quality Evaluation Method employing a 5-point scale across five criteria, where 1 indicates the lowest quality and 5 the highest: **Informativeness (I)** - provision of new information, such as background or additional context; **Logicality (L)** - coherent reasoning process and causal link to the outcome; **Objectivity (O)** - objectivity of the answer and bias-free analysis; **Readability (R)** - grammatical correctness, structural clarity, and ease of comprehension; **Accuracy (A)** - alignment with the true label and its accurate reflection of the result. Figure 7 shows radar charts comparing ILORA scores across prompt-pipeline combinations.

In both datasets, `EM-RAG` and `EM-KGRAG` consistently outperform the baseline across all ILORA metrics. Few-shot prompting improves explanation quality over zero-shot across all `EmeraldMind` variations, while the baseline's limited knowledge hampers its ability to provide factual explanations, leading to more abstentions and lower quality. `EM-HYBRID` achieves the highest ILORA scores across all criteria and datasets, under corresponding prompting conditions, demonstrating coherence between individual scoring and pairwise comparisons, as it selects the superior justification between `EM-RAG` and `EM-KGRAG`.

Specifically, in the `GreenClaim` benchmark (average scores depicted in Figure 7(a)), `EM-KGRAG` excels in Objectivity and Accuracy under the few-shot setting. Conversely, in the larger `EmeraldData` benchmark (Figure 7(b)), `EM-RAG` outperforms `EM-KGRAG` across all

Table 4: Classification performance across benchmarks, pipelines, and prompt variations. Overall Acc. measures correct predictions across all claims. Accuracy and Coverage measure conditional performance on non-abstained claims.

Prompt	Pipeline	GreenClaims				EmeraldData			
		Accuracy	Coverage	Overall Acc.	No. Abstains	Accuracy	Coverage	Overall Acc.	No. Abstains
Zero-shot	Baseline	93.33%	29.41%	27.45%	36	94.21%	25.97%	24.52%	459
	EM-RAG	82.14%	54.90%	45.1%	23	87.92%	62.74%	55.16%	231
	EM-KGRAG	93.55%	60.78%	56.86%	20	92.51%	49.52%	45.81%	313
	EM-HYBRID	89.47%	74.51%	66.67%	13	85.78%	68.06%	58.39%	198
Few-shot	Baseline	100.00%	31.37%	31.37%	35	94.21%	19.52%	18.39%	499
	EM-RAG	77.14%	76.47%	52.94%	12	85.19%	69.68%	59.35%	188
	EM-KGRAG	89.74%	76.47%	68.63%	12	88.03%	60.65%	53.39%	244
	EM-HYBRID	92.31%	76.47%	70.59%	12	83.80 %	74.68%	62.58%	157



(a) GreenClaims (b) EmeraldData

Figure 7: ILORA scores for the different prompt-pipeline combinations, highlighting performance on key metrics.

ILORA metrics, showcasing superior explanation quality under both prompting conditions. The baseline consistently exhibits the lowest performance, especially in Readability and Logicality. These results confirm that few-shot prompting enhances reasoning depth and explanation reliability across datasets and models.

(2) Pairwise Comparison. We conduct relative evaluation using a 3-way LLM-as-a-Judge instead of pairwise. Specifically, an LLM ranks all three justifications (`Baseline`, `EM-RAG`, `EM-KGRAG`) simultaneously under ILORA metrics. Table 5 presents the count of first, second, and third place rankings for each pipeline across dataset-prompt combinations using the relative judge paradigm outputs. To derive an overall ranking, the Borda count [9] method is applied, assigning 3 points for first place, 2 points for second, and 1 point for third place. For all the dataset-prompt variation, the resulting Borda scores (see Appendix Table 7) consistently yield the ranking:

$$\texttt{EM-RAG} \succ \texttt{EM-KGRAG} \succ \texttt{Baseline}.$$

To assess the significance of differences across pipeline justifications, we conducted a Friedman test followed by Nemenyi post-hoc analysis [24]. The Friedman test revealed a highly significant difference across methods ($x^2 = 1823.83$, $p < 0.0001$), rejecting

Table 5: Comparison of Baseline, RAG, and GraphRAG counts grouped by dataset-prompt.

Dataset	Prompt	Pipeline	1st	2nd	3rd
Greenclaims	Zero-shot	Baseline	1	11	39
		EM-RAG	38	13	0
		EM-KGRAG	12	27	12
	Few-shot	Baseline	0	1	50
		EM-RAG	39	12	0
		EM-KGRAG	12	38	1
EmeraldData	Zero-shot	Baseline	12	231	377
		EM-RAG	553	62	5
		EM-KGRAG	55	327	238
	Few-shot	Baseline	13	21	586
		EM-RAG	551	65	4
		EM-KGRAG	56	534	30

the null hypothesis that all pipelines perform equally. Nemenyi post-hoc tests confirmed all pairwise differences exceed the critical difference (CD=0.064), `EM-RAG` vs `EM-KGRAG` ($|\bar{R}| = 0.982 >$ CD), `EM-RAG` vs `Baseline` ($|\bar{R}| = 1.638 >$ CD), and `EM-KGRAG` vs `Baseline` ($|\bar{R}| = 0.656 >$ CD).

8 Conclusions and Future Work

We propose `EmeraldMind`, a domain-specific RAG-based greenwashing detection framework comprising three variants. `EM-HYBRID` achieves the highest overall accuracy, combining `EM-RAG`'s broad coverage with `EM-KGRAG`'s accuracy. Few-shot prompting improves coverage without sacrificing accuracy, and domain-specific retrieval substantially enhances selective prediction without accuracy trade-offs. Future work should investigate how graph retrieval strategies (e.g., hop limits, schema constraints, traversal heuristics) influence performance and explore alternative hybrid methods for integrating textual and graph evidence. Incorporating additional knowledge sources, such as regulatory data and sustainability taxonomies, could further enhance coverage and justification quality.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

[1] Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or gold? Deriving structured insights from sustainability reports via large language models. *EPJ Data Science* 13, 1 (2024), 41.

[2] Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown Claims: Generation of Fact-Checking Training Examples from Unstructured and Structured Data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. Association for Computational Linguistics, Miami, Florida, USA, 12105–12122. https://aclanthology.org/2024.emnlp-main.675

[3] Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M. Suchanek. 2025. Corporate Greenwashing Detection in Text – a Survey. arXiv:2502.07541 [cs.CL] https://arxiv.org/abs/2502.07541

[4] Kevin Matthe Caramancion. 2023. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. In *2023 IEEE Future Networks World Forum (FNWF)*. IEEE, Baltimore, MD, USA, 1–6. doi:10.1109/FNWF58287.2023.10520446

[5] Bohan Chen and Andrea L. Bertozzi. 2023. AutoKG: Efficient Automated Knowledge Graph Generation for Language Models. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, Sorrento, Italy, 3117–3126. doi:10.1109/BigData59044.2023.10386454

[6] European Commission. 2023. Proposal for a Directive of the European Parliament and of the Council on the substantiation and communication of explicit environmental claims (Green Claims Directive). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52023PC0166. COM(2023) 166 final.

[7] Hoang Anh Dang, Vu Tran, and Le-Minh Nguyen. 2025. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence* 8 (2025), 1622292.

[8] EFFAS and DVFA. 2009. Key Performance Indicators for Environmental, Social & Governance Issues: Financial Analysis and Corporate Valuation. https://ec.europa.eu/docsroom/documents/1547/attachments/1/translations/en/renditions/native Endorsed by EFFAS; Available via European Commission Docsroom.

[9] Peter C Fishburn and William V Gehrlein. 1976. Borda's rule, positional voting, and Condorcet's simple majority principle. *Public Choice* 28, 1 (1976), 79–88. doi:10.1007/BF01718459

[10] Mattia Fornasiero. 2024. *Exploring the Effectiveness of Large Language Models in Greenwashing Detection*. Master's Thesis. Erasmus University Rotterdam, Rotterdam, Netherlands. https://thesis.eur.nl/pub/72537/

[11] Mattia Fornasiero. 2024. Greenwashing Detection Dataset. https://github.com/DizzyPanda1/GreenwashingDetectionDataset. Accessed: 2025-10-22.

[12] Eleni Fotopoulou, Ioanna Mandilara, Anastasios Zafeiropoulos, Chrysi S. Laspidou, Giannis Adamos, Phoebe Koundouri, and Symeon Papavassiliou. 2022. SustainGraph: A Knowledge Graph for Tracking the Progress and the Interlinking among the Sustainable Development Goals' Targets. *Frontiers in Environmental Science* 10 (2022), 1003599. doi:10.3389/fenvs.2022.1003599

[13] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594 [cs.CL] https://arxiv.org/abs/2411.15594

[14] Tanay Kumar Gupta, Tushar Goel, Ishan Verma, Lipika Dey, and Sachit Bhardwaj. 2024. Knowledge Graph aided LLM based ESG Question-Answering from News. In *Proceedings of the 2nd International Workshop on Knowledge Graphs for Sustainability (KG4S 2024) (CEUR Workshop Proceedings, Vol. 3753)*. CEUR-WS.org, Hersonissos, Greece, 65–78. https://ceur-ws.org/Vol-3753/paper6.pdf

[15] Chaoyue He, Xin Zhou, Yi Wu, Xinjia Yu, Yan Zhang, Lei Zhang, Di Wang, Shengfei Lyu, Hong Xu, Wang Xiaoqiao, Wei Liu, and Chunyan Miao. 2025. ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. Association for Computational Linguistics, Suzhou, China, 14623–14664. https://aclanthology.org/2025.emnlp-main.739

[16] Md. Saiful Islam, Adiba Mahbub Proma, Yi-Shuan Zhou, Syeda Nahida Akter, Caleb Wohn, and Ehsan Hoque. 2022. KnowUREnvironment: An Automated Knowledge Graph for Climate Change and Environmental Issues. In *AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*. AAAI Publications, Arlington, VA, USA. https://www.climatechange.ai/papers/aaaifss2022/3

[17] Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Zixuan Li, Jian Guo, Huawei Shen, and Yuanzhuo Wang. 2024. Unlocking the Power of Large Language Models for Entity Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 7566–7583. doi:10.18653/v1/2024.acl-long.408

[18] Anna Kiepura and Jessica Lam. 2025. ClimateCheck2025: Multi-Stage Retrieval Meets LLMs for Automated Scientfic Fact-Checking. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, Tirthankar Ghosal, Philipp Mayr, Amanpreet Singh, Aakanksha Naik, Georg Rehm, Dayne Freitag, Dan Li, Sonja Schimmler, and Anita De Waard (Eds.). Association for Computational Linguistics, Vienna, Austria, 293–306. doi:10.18653/v1/2025.sdp-1.28

[19] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, Vienna, Austria, 36 pages. https://openreview.net/forum?id=8euJaTveKw

[20] Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2025. Automated fact-checking of climate claims with large language models. *npj Climate Action* 4, 1 (2025), 17.

[21] Huahang Li, Longyu Feng, Shuangyin Li, Fei Hao, Chen Jason Zhang, Yuanfeng Song, and Lei Chen. 2024. On Leveraging Large Language Models for Enhancing Entity Resolution: A Cost-efficient Approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*. Association for Computing Machinery, Boise, ID, USA, 1271–1281. doi:10.1145/3627673.3679576

[22] Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, 163–181. doi:10.18653/v1/2024.findings-naacl.12

[23] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 462–477.

[24] Peter Bjorn Nemenyi. 1963. *Distribution-free multiple comparisons*. Princeton University, Princeton, NJ, USA.

[25] Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Singapore, 21–31. doi:10.18653/v1/2023.emnlp-demo.3

[26] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot Fact Verification by Claim Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 476–483. doi:10.18653/v1/2021.acl-short.61

[27] Ralph Peeters, Aaron Steiner, and Christian Bizer. 2024. Entity Matching using Large Language Models. arXiv:2310.11244 [cs.CL] https://arxiv.org/abs/2310.11244

[28] pymupdf. 2025. PyMuPDF. https://github.com/pymupdf/PyMuPDF

[29] Ihsan A. Qazi, Zohaib Khan, Abdullah Ghani, Agha A. Raza, Zafar A. Qazi, Wassay Sajjad, Ayesha Ali, Asher Javaid, Muhammad Abdullah Sohail, and Abdul H. Azeemi. 2025. Scaling Truth: The Confidence Paradox in AI Fact-Checking. arXiv:2509.08803 [cs.SI] https://arxiv.org/abs/2509.08803

[30] Subhey Sadi Rahman, Md. Adnanul Islam, Md. Mahbub Alam, Musarrat Zeba, Md. Abdur Rahman, Sadia Sultana Chowa, Mohaimenul Azam Khan Raiaan, and Sami Azam. 2026. Hallucination to Truth: A Review of Fact-Checking and Factuality Evaluation in Large Language Models. *Artificial Intelligence Review* 59, 1 (2026), 23 pages. doi:10.1007/s10462-025-11454-w

[31] Aishwarya Sahoo, Jeevana Kruthi Karnuthala, Tushar Parmanand Budhwani, Pranchal Agarwal, Sankaran Vaidyanathan, Alexa Siu, Franck Dernoncourt, Jennifer Healey, Nedim Lipka, Ryan Rossi, Uttaran Bhattacharya, and Branislav Kveton. 2025. Quantitative LLM Judges. arXiv:2506.02945 [cs.CL] https://arxiv.org/abs/2506.02945

[32] Aida Usmanova and Ricardo Usbeck. 2024. Structuring Sustainability Reports for Environmental Standards with LLMs guided by Ontology. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Association for Computational Linguistics, Bangkok, Thailand, 168–177. https://aclanthology.org/2024.climatenlp-1.13

[33] Saeid Ashraf Vaghefi, Dominik Stammbach, Jingwei Ni, Tobias Schimanski, Marc Rivera, Julia Bingler, Chiara Colesanti Senni, Mathias Kraus, and Markus Leippold. 2023. CHATCLIMATE: Grounding Conversational AI in Climate Science. *Communications Earth & Environment* 4, 1 (Dec. 2023), 480.

[34] Annas Vijaya, Faris Dzaudan Qadri, Linda Salma Angreani, and Hendro Wicaksono. 2025. ESGOnt: An ontology-based framework for Enhancing Environmental, Social, and Governance (ESG) assessments and aligning with Sustainable Development Goals (SDG). *Resources, Environment and Sustainability* 22 (2025), 100262. doi:10.1016/j.resenv.2025.100262

[35] Gengyu Wang, Lawrence Chillrud, and Kathleen McKeown. 2021. Evidence based Automatic Fact-Checking for Climate Change Misinformation. In *Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM 2021) (SocialSens 2021: International Workshop on Social Sensing)*. AAAI Press, Palo Alto, California, USA, 7 pages. https://workshop-proceedings.icwsm.org/abstract.php?id=2021_39

[36] Tianshu Wang, Xiaoyang Chen, Hongyu Lin, Xuanang Chen, Xianpei Han, Le Sun, Hao Wang, and Zhenyu Zeng. 2025. Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics, Abu Dhabi, UAE, 96–109. https://aclanthology.org/2025.coling-main.8

[37] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2448–2460. doi:10.18653/v1/2022.acl-long.175

[38] Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics* 12 (2024), 334–354.

[39] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

[40] Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo. 2025. From Predictions to Analyses: Rationale-Augmented Fake News Detection with Large Vision-Language Models. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. Association for Computing Machinery, Sydney, NSW, Australia, 5364–5375. doi:10.1145/3696410.3714777

[41] Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *Comput. Surveys* 56, 4 (2023), 1–62.

[42] Yuchen Zhou, Yuan Cao, and Alexander Perzylo. 2024. Towards Digital Sustainability Reporting: An Ontology for Mapping of Indicators in GRI and ESRS. In *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI: Proceedings of the 20th International Conference on Semantic Systems (Studies on the Semantic Web, Vol. 60)*. IOS Press, Amsterdam, Netherlands, 191–207. doi:10.3233/SSW240016

[43] Yuchen Zhou and Alexander Perzylo. 2023. OntoSustain: Towards an Ontology for Corporate Sustainability Reporting. In *ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference (CEUR Workshop Proceedings, Vol. 3632)*. CEUR Workshop Proceedings, Athens, Greece, 181–184. https://ceur-ws.org/Vol-3632/ISWC2023_paper_462.pdf

[44] Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, Zongxiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *Journal of Cleaner Production* 489 (2025), 144572. doi:10.1016/j.jclepro.2024.144572

A Appendix

Schema Visualization. An interactive version of Figure 3 is available on our **GitHub repo** (Download file and open from browser).

Retrieval Hyperparameters. The default hyperparameters are summarized in Table 6.

Table 6: Default Hyperparameters for Retrieval

Pipeline	Symbol	Hyperparameter	Value
EM-KGRAG	top_n	No.nodes retrieved per type	3
	τ	Similarity Threshold	0.2
	k	Max path length (hops)	3
EM-RAG	top_m	No. chunks retrieved	8

Borda Scores. Table 7 presents the Borda scores derived from the *relative LLM judge* for all dataset-prompt variations.

Table 7: Borda Scores

Dataset	Prompt	Baseline	EM-RAG	EM-KGRAG
Greenclaims	Zero-shot	64	140	102
	Few-shot	52	141	113
EmeraldData	Zero-shot	875	1788	1057
	Few-shot	667	1787	1266
		1658	3856	2538

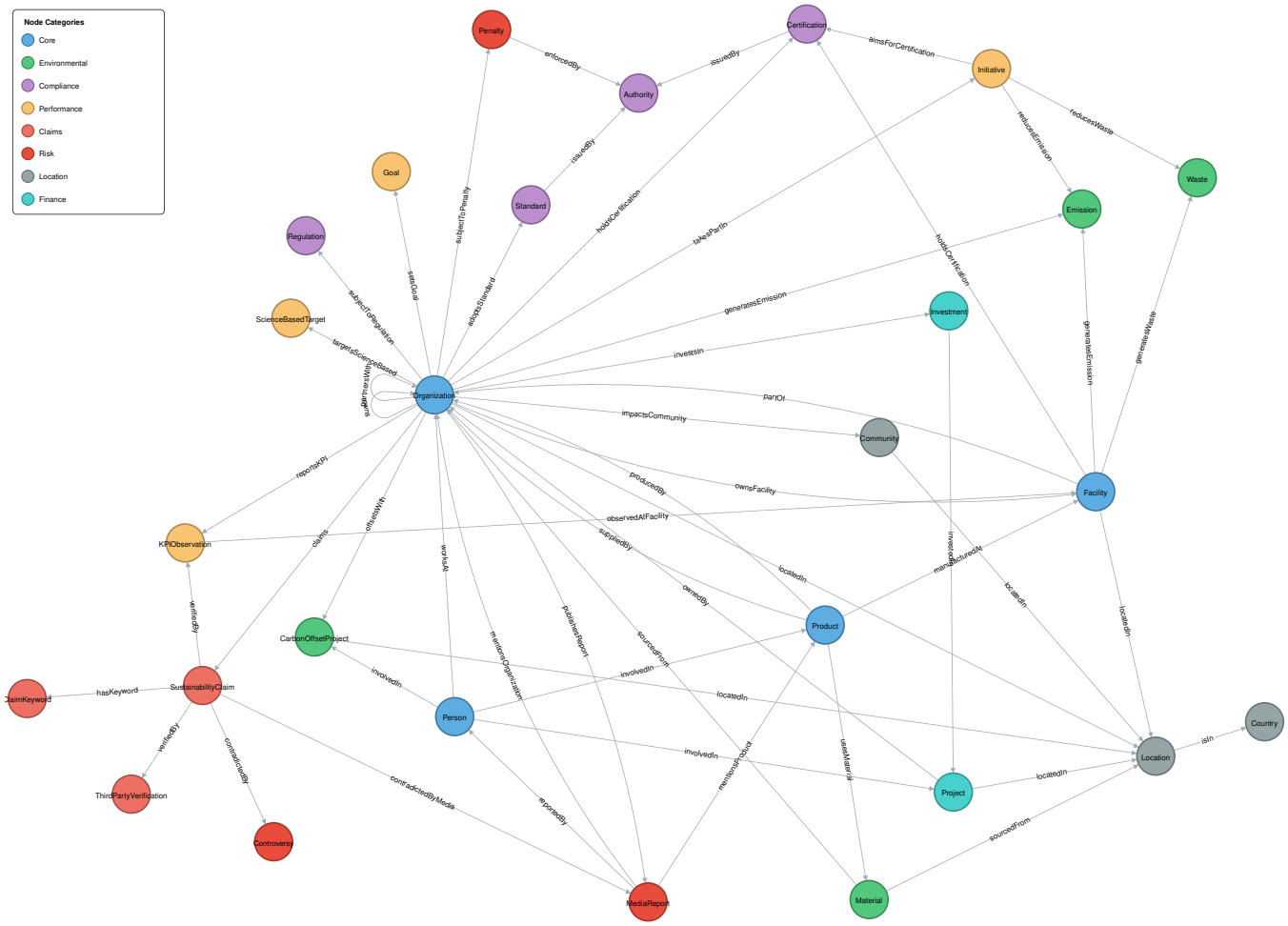


Figure 8: EmeraldGraph Schema