Do Not Compare Apples and Offsets: An ESG Alignment Gate for Trustworthy Report Analysis

Zexi Wang[1], Zewen Shen[2]

[1]Macquarie Business School, Macquarie University, Sydney, NSW, Australia
[2]School of Computer Science, University of Sydney, Sydney, NSW, Australia

Abstract

Corporate environmental, social, and governance (ESG) disclosures are increasingly processed with large language models (LLMs). Retrieval and source grounding improve traceability, yet they do not ensure that the evidence supports the judgement drawn from it. Comparisons may remain invalid when disclosures differ in metric identity, mathematical form, unit, scope, reporting period, baseline, organisational boundary, or evidential status. To address this decision-validity gap, we introduce the *ESG Alignment Skill*, an intermediate trust-control layer that converts disclosure passages into structured evidence records before authorising cross-company comparison or greenwashing-related claim-status classification. The comparability gate permits ranking only across aligned records, whereas the claim-status gate preserves the distinction among reported action, aspiration, and ambiguity.

Evaluation uses 200 comparability cases, 200 claim-status samples, and a 30-case adversarial pilot. The Skill reaches 90.0% comparability accuracy, compared with 87.0% for a direct LLM baseline, and achieves 100.0% claim-status accuracy, compared with 94.5%. Under the baseline, 15.7% of ambiguous claims are labelled as reported action and the unsupported-action rate is 8.1%; both rates are 0.0% for the Skill. These controlled results indicate that structured gating can reduce unsupported judgements while making missing or incompatible evidence visible. The system assesses textual support for a requested decision and does not independently verify corporate conduct, factual accuracy, or deceptive intent.

Keywords

Agent skills, ESG disclosure, large language models, comparability gating, evidence-bounded reasoning, claim-status classification

1. Introduction

Large language models (LLMs) are increasingly embedded in workflows that analyse environmental, social, and governance (ESG) reports for investment and compliance. Retrieval-augmented generation (RAG) and report-analysis systems can locate passages, extract metrics, and prepare comparison-oriented summaries [1, 2]. These capabilities improve access to disclosure evidence, yet the resulting judgement may still exceed what that evidence supports.

Cross-company comparison makes this weakness visible. Renewable-energy performance may be reported as a percentage share, absolute quantity, growth rate, emissions intensity, or future target. Although the passages concern the same topic, differences in metric form, unit, reporting period, scope, baseline, and organisational boundary may prevent a common ranking axis. A model can therefore reproduce genuine evidence and still generate a structurally unsupported comparison, consistent with findings on factuality and faithfulness in generated text [3, 4, 5].

The same problem appears when sustainability language is given more evidential weight than the passage warrants. Corporate reports often place measured outcomes beside future commitments and broad statements of progress. Once qualifications are compressed during summarisation, an aspirational target or vague claim may be presented as completed action. Greenwashing research likewise stresses the gap between environmental communication and substantiated organisational activity [6, 7, 8].

Research on RAG, schema-guided extraction, traceable report analysis, and ESG measurement provides important components for improving access to and representation of disclosure evidence [9, 2, 10]. Yet these approaches leave unresolved whether the extracted evidence permits the judgement

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.

zexi.wang@students.mq.edu.au (Z. Wang); zshe0057@uni.sydney.edu.au (Z. Shen)

requested by the user. We address this gap by defining an evidence-bounded decision space and operationalising it through the *ESG Alignment Skill*, which combines structured evidence records with comparability and claim-status gates. The evaluation examines whether this design supports valid comparisons, refuses structurally unsupported rankings, and preserves uncertainty when claim evidence remains incomplete.

2. Trustworthiness Problem in ESG Report Analysis

Retrieval and source grounding improve access to disclosure evidence, yet the validity of a final judgement depends on how that evidence is used. A passage may be correctly retrieved and accurately quoted while still supporting a comparison or claim classification that exceeds its evidential content. We define this outcome as an *unsupported ESG judgement*: a ranking, classification, or evaluative conclusion for which a required condition is missing, incompatible, or insufficiently specified.

This risk is pronounced in ESG reporting because firms describe similar sustainability topics through different metrics, scopes, periods, and forms of evidence. It emerges most clearly in two related tasks: cross-company comparison and greenwashing-related claim-status classification.

2.1. Unsupported Comparisons

Disclosures concerning the same ESG topic do not necessarily support the same comparison. Renewable-energy performance, for example, may be expressed as a percentage share, absolute quantity, annual growth rate, intensity measure, reduction from a baseline, or future target. Although these records concern related activities, they represent different mathematical objects.

Even where the metric form appears similar, comparability may depend on unit, reporting period, baseline, metric scope, greenhouse gas (GHG) scope, and organisational boundary. A company-wide value cannot automatically be compared with a facility-level measure, while renewable electricity as a share of electricity consumption differs from renewable energy as a share of total energy use. Related research on ESG ratings likewise shows that differences in scope and measurement contribute to divergent evaluations [10].

A ranking is supported only when the relevant records establish a common comparison axis. Missing or incompatible attributes instead require refusal, together with the condition that prevents the requested ordering. This structure also accommodates partial comparability: aligned records may form a bounded comparison group even when other disclosures must be excluded.

2.2. Unsupported Claim Upgrading

The same mismatch between evidence and conclusion appears when the status of a sustainability claim is strengthened during summarisation or classification. ESG reports often place measured outcomes beside future commitments and broad statements of progress. A target such as "we aim to reach net zero" may be presented as an apparent achievement once its future orientation is removed, while vague language may be treated as implementation evidence despite the absence of quantities, dates, or operational detail.

Such upgrading matters in greenwashing-related analysis because positive corporate language can appear more substantive when its evidential limitations are obscured [6, 7, 8]. The task examined here concerns the status supported by the supplied passage. Claim status reflects both evidential strength and temporal orientation: completed or measurable activity is classified as *reported action*, future targets or commitments as *aspirational*, and claims lacking sufficient support as *ambiguous*.

These labels describe the strength of the available textual evidence. They do not establish whether a company is truthful, deceptive, or compliant.

2.3. Evidence-Bounded Decision Space

Taken together, unsupported comparison and unsupported claim upgrading reflect the same underlying failure: the final judgement is stronger than the available evidence permits. We describe the set of conclusions supported by the disclosure as the *evidence-bounded decision space*.

Within this space, source evidence, extracted attributes, and decision permission remain distinct. Missing information is preserved rather than inferred, incompatible records remain outside the ranking set, and weak claims retain aspirational or ambiguous status. The ESG Alignment Skill translates this principle into a structured evidence record, a comparability gate, and a claim-status gate, providing the basis for the method that follows.

3. Method

3.1. Design Motivation and Overview

The evidence-bounded decision space developed in the preceding section provides the organising principle for the ESG Alignment Skill. Retrieval and source grounding improve access to evidence, yet they do not ensure that the evidence supports the judgement produced from it [3, 4, 5]. The Skill therefore operates between disclosure text and downstream judgement, constraining comparison and claim-status decisions to the attributes explicitly supported by the supplied passage.

Within this design, a large language model converts unstructured disclosure text into a schema-grounded evidence record. Structured intermediate representations have been shown to improve consistency in information extraction by making task-relevant fields explicit [9]. The resulting record is then evaluated by task-specific gates that determine whether the requested judgement should be allowed, refused, or flagged. Separating evidence representation from decision permission prevents missing or incompatible attributes from being absorbed into a fluent conclusion.

The output retains the source quotation, extracted attributes, warnings, and decision trace. An allowed judgement is accompanied by the fields supporting it, while a refusal identifies the failed alignment condition. Where the evidential status remains unresolved, the system preserves that uncertainty through a flag. Figure 1 presents the workflow from disclosure input to gated decision.

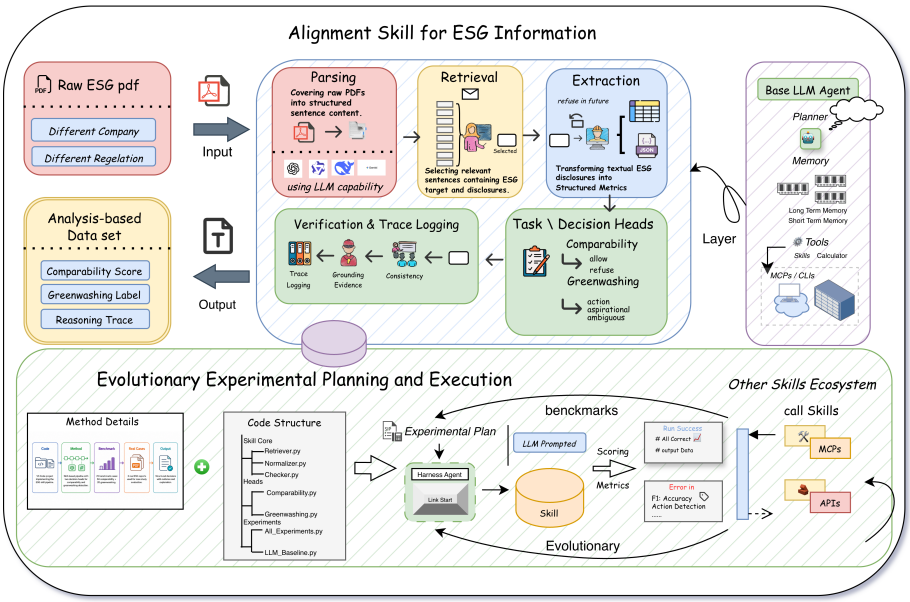

Figure 1: Environmental, social, and governance alignment workflow from disclosure text to structured evidence and allow, refuse, or flag decisions.

3.2. Evidence Record

Each metric or claim is represented through a fixed schema containing `company`, `metric_id`, `metric_type`, `reported_value`, `unit`, `reporting_year`, `baseline_year`, `target_year`, `metric_scope`, `ghg_scope`, `organisational_boundary`, `claim_status`, `source_quote`, `warnings`, and `alignment_note`. Consistent with structured extraction practices, these fields separate passage-level evidence from downstream inference [9].

Populated fields must be supported by the source passage. Missing information is stored as `null` and recorded as a warning when required for the requested judgement. `Metric_scope` identifies the relevant activity or denominator, while GHG scope and organisational boundary distinguish emissions coverage from whether a value applies to a reporting group, facility, product, or supply-chain segment.

The `metric_id` identifies the substantive quantity, whereas `metric_type` records its mathematical form, such as a percentage share, absolute quantity, intensity, growth rate, reduction, or target. Thematically related disclosures can therefore remain separate when their measurement forms are incompatible, reflecting evidence that differences in scope and measurement contribute to divergent ESG evaluations [10].

3.3. Comparability Gate

The comparability gate evaluates whether the extracted records define a common comparison axis. Metric identity and type must align, units must match or be explicitly converted, and reporting periods, metric scopes, GHG scopes, and organisational boundaries must be compatible. Baselines are required for change, reduction, or growth measures, but not for point-in-time shares.

Aligned records form a comparison group, while records with missing or incompatible attributes are excluded with the failed condition retained in the output. This permits partial comparability when only a subset of the retrieved disclosures supports ranking.

The worked application illustrates this outcome. Toyota reports a 42% renewable share of total electricity consumption in 2023, while Apple reports 38% for the same metric and year. Subject to compatible organisational boundaries, the two records can enter a shared comparison group. BP instead provides a 2050 carbon-neutrality commitment without a compatible quantitative renewable-share metric. The gate therefore permits a bounded Toyota–Apple comparison, excludes BP, and refuses a full three-company ranking. Figure 2 presents the extracted records and decision.

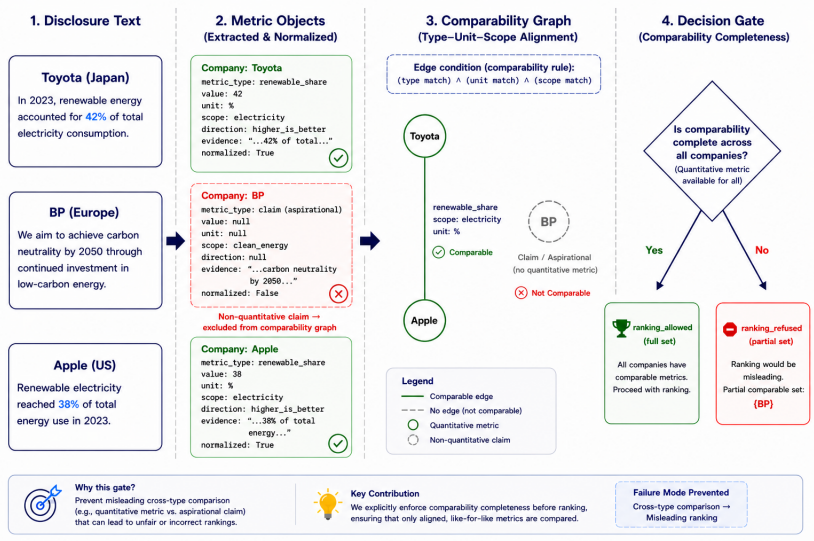

Figure 2: Worked application of the comparability gate using Toyota, Apple, and BP. The companies were selected to contrast two quantitative renewable-share disclosures that form an aligned subgroup with a future-oriented carbon-neutrality commitment lacking a compatible quantitative metric.

3.4. Claim-Status Gate

The second gate evaluates the evidential status of a sustainability claim. Research on greenwashing and environmental communication emphasises the importance of distinguishing substantiated organisational activity from forward-looking or weakly supported statements [6, 7, 8]. The gate operationalises this distinction through reported-action, aspirational, and ambiguous labels.

Claim status reflects both evidential strength and temporal orientation: completed or measurable activity is classified as *reported action*, future targets or commitments as *aspirational*, and claims lacking sufficient support as *ambiguous*. These labels describe the passage presented to the system. Reported-action status indicates that the disclosure reports implementation evidence, while ambiguous status preserves the absence of textual support for a stronger label. Neither category independently establishes whether the company is truthful, deceptive, or compliant. Table 1 summarises the operational distinctions.

Table 1
Claim-status labels and evidence signals.

Label	Example	Evidence	Rationale
Reported action	Installed 120 MW of solar capacity, reducing Scope 2 emissions by 18% from a 2020 baseline.	Completed activity, value, unit, scope, and baseline.	Implementation with a measured outcome.
Aspirational	We aim to reach net-zero Scope 1 and Scope 2 emissions by 2040.	Future orientation and target year.	Future commitment without a completed result.
Ambiguous	Our operations are environmentally responsible and show sustainability progress.	No metric, scope, or implementation evidence.	Insufficient support for action status.

3.5. Implementation and Output

The underlying model is Generative Pre-trained Transformer 4o (GPT-4o). Evidence extraction uses a temperature of 0.0, while preliminary claim-status classification uses a temperature of 0.3. The direct LLM baseline receives the raw disclosure passage and produces a final decision without an intermediate record. The Skill uses the same model family to populate the structured fields, after which the relevant gate constrains the final judgement.

Comparability decisions are derived from the extracted attributes and the alignment conditions described above. Claim-status decisions are checked against the evidence associated with each category, so future-oriented wording remains aspirational and passages lacking sufficient support remain ambiguous. The final output combines the decision with the source quotation, extracted fields, warnings, and alignment note.

Reproduction of the experiment requires the dated GPT-4o snapshot, prompts, parsing rules, application programming interface settings, retry policy, and repeated-run procedure. These details should be reported from the original experimental record.

4. Experiments and Results

4.1. Experimental Setup

Benchmarks. The evaluation follows the two decision problems addressed by the Skill. The comparability benchmark contains 200 cases testing whether aligned rankings are allowed and structurally invalid comparisons are refused. The greenwashing-related claim benchmark contains 200 samples labelled as reported action, aspirational, or ambiguous. A further 30 adversarial cases introduce scope ambiguity, implicit baselines, heterogeneous units, partial comparability, textual noise, and mixed organisational boundaries. These cases support qualitative error inspection and are kept separate from the headline accuracy results. Table 2 summarises the benchmark composition and evaluation targets.

Table 2
Benchmark composition and evaluation targets.

Benchmark	Size	Evaluation target
ESG comparability	200	Valid ranking and invalid-comparison refusal.
Claim-status classification	200	Reported action, aspiration, and ambiguity.
Adversarial pilot	30	Stress testing of irregular inputs and gate boundaries.

Models and metrics. The direct LLM baseline receives the passage and task instruction, then returns a final decision without an intermediate evidence record. The ESG Alignment Skill processes the same passage through extraction and gating. Holding the model family constant focuses the comparison on the structured representation and decision constraints, although the Skill introduces an additional extraction stage.

Accuracy is measured as the proportion of cases assigned the benchmark decision. In the comparability task, the over-refusal rate captures valid comparisons that are rejected, while the unsafe-allow rate captures invalid comparisons that are permitted. For claim status, the ambiguous-as-action rate measures ambiguous samples labelled as reported action. The dangerous-overclaim rate measures aspirational or ambiguous samples assigned the same label.

Because both systems are evaluated on the same cases, the comparison is paired. The available summary data do not contain the discordant-pair counts needed for a paired significance test, so the observed differences are interpreted descriptively.

4.2. Main Results

The Skill produces 180 correct comparability decisions among 200 cases, corresponding to 90.0% accuracy. The direct LLM baseline produces 174 correct decisions, corresponding to 87.0%. The observed difference is six cases. The over-refusal rate declines from 0.26 under the baseline to 0.20 under the Skill, while both systems record an unsafe-allow rate of 0.00. Within the benchmark, the remaining comparability errors therefore arise from rejecting valid cases rather than permitting invalid rankings.

The claim-status benchmark produces a larger difference. The Skill classifies all 200 samples according to the benchmark labels, compared with 189 correct classifications for the baseline. The corresponding accuracies are 100.0% and 94.5%. Baseline errors are concentrated in ambiguous claims: 15.7% receive reported-action status, and the dangerous-overclaim rate is 8.1%. Both observed rates are 0.0% for the Skill.

This concentration of errors indicates that the main behavioural difference emerges when favourable wording lacks implementation evidence. The structured record retains missing quantities, dates, scopes, and outcomes, allowing ambiguity to remain visible rather than being resolved in favour of action. Figure 3 presents the observed accuracy and risk rates.

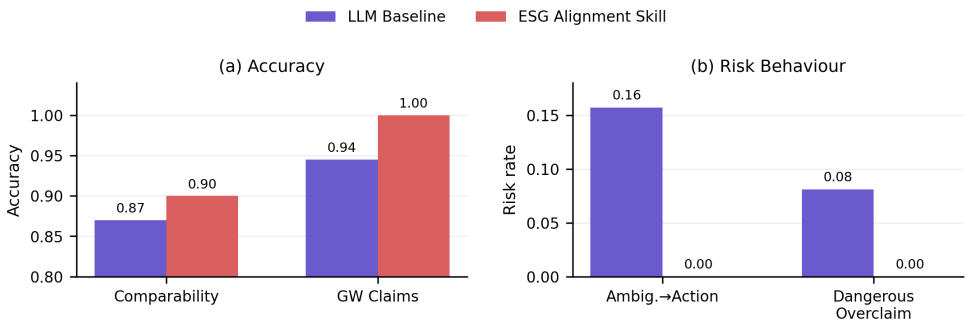

Figure 3: Observed benchmark accuracy and risk behaviour. Values are proportions. "Ambig.→Action" denotes the ambiguous-as-action rate, and "Dangerous Overclaim" denotes the share of aspirational or ambiguous samples labelled as reported action.

4.3. Error and Behaviour Analysis

Aggregate accuracy does not distinguish excessive refusal from unsafe allowance. In the comparability benchmark, both configurations avoid unsafe rankings under the assigned labels, while the Skill rejects fewer valid comparisons. The pattern is consistent with a clearer distinction between thematic similarity and metric-level comparability. At the same time, the remaining over-refusals indicate that metric normalisation and rule coverage still require refinement.

A related distinction appears in claim-status classification. Clear reported actions contain quantities, implementation verbs, reporting years, or measured outcomes, whereas clear aspirational claims contain future orientation and target dates. Ambiguous passages rely on favourable language without enough detail to support either implementation or a defined target. The baseline occasionally resolves this uncertainty in favour of action, while the Skill retains the missing evidence as a constraint.

The adversarial pilot was run with Kimi K2.6 rather than GPT-4o to examine whether the extraction and gating behaviour extends beyond the primary model configuration. Case-level results for the 30 adversarial inputs are reported in Tables 5 and 6. The Skill matches the reference decision in 22 of the 30 cases, corresponding to a 73.3% pass rate. Correct decisions include scope mismatches (`adv_001–002`, `adv_014`), unit mismatches (`adv_004`), type mismatches (`adv_007–009`), and mixed-boundary cases (`adv_021–024`).

The eight failures reflect two broader sources of error. *Rule-coverage errors* arise when the gate applies an incorrect blocking condition or fails to recognise a relevant incompatibility. These errors include embedding `baseline_year` in `metric_id` (`adv_010`), ignoring geographic scope (`adv_016`), and treating direction as a comparison precondition rather than a reported value (`adv_025`). *Extraction and normalisation fragility* appears in unit handling (`adv_005–006`) and synonym resolution. In particular, electricity, energy, and purchased power (`adv_003`); elec., green, and power (`adv_017`); and YoY, GHG, and CO2e (`adv_018`) are not consistently resolved to the required canonical representations.

These recurring behaviours are summarised in Table 3. Incompleteness alone does not trigger refusal; a judgement is refused or flagged only when the missing attribute is required for the decision at hand.

Table 3
Decision behaviour under common failure conditions.

Failure condition	Direct LLM baseline	ESG Alignment Skill
Missing required baseline	May rank from topical similarity.	Refuses a baseline-dependent comparison and records the missing year.
Different units	May compare within one ESG theme.	Blocks comparison until units align or are converted.
Different metric types	May combine shares, absolutes, intensities, growth, and targets.	Ranks only records with compatible metric types.
Future target	May treat commitment as completed action.	Retains aspirational status without implementation evidence.
Vague claim	May infer progress from positive wording.	Preserves ambiguity without measurable support.

Note. Entries summarise qualitative decision behaviour and do not represent additional performance estimates.

Given the pilot's limited size, these observations are treated as qualitative evidence rather than a general robustness estimate. The 8 failures indicate that gate quality depends jointly on accurate field extraction and adequate rule coverage: extraction may populate an incorrect field before the gate is applied, while an incomplete rule set may fail to recognise a relevant incompatibility. Repair targets are (i) unit normalisation must preserve system-level distinctions (e.g., volume/mass vs mass/mass) while collapsing scale-only differences (e.g., Mt vs kt), and (ii) extraction rules must expand synonym coverage for abbreviated and compound terms before the comparison gate is applied.

5. Discussion

The results are most informative when read against the central purpose of the ESG Alignment Skill: preventing a model from drawing a stronger conclusion than the disclosure evidence permits. On the comparability benchmark, accuracy rises from 87.0% to 90.0%, while the over-refusal rate declines from 26.0% to 20.0%. Both systems record an unsafe-allow rate of 0.0%. Within the current benchmark, the observed improvement therefore comes from retaining more valid comparisons without allowing additional invalid rankings.

This pattern reflects the distinction between topical similarity and metric comparability developed earlier in the paper. A direct LLM may recognise that several passages concern renewable energy and still overlook differences in metric form, unit, reporting period, scope, or organisational boundary. The Skill makes these attributes explicit before ranking is permitted. Its value lies partly in the final decision and partly in the record that explains how that decision was reached. Extracted fields, unresolved attributes, warnings, and alignment notes make it possible to identify whether an error arose during extraction, metric normalisation, or gate execution.

The claim-status results reveal a closely related effect. The baseline already performs well when disclosures clearly describe completed activity or a future commitment. Its errors are concentrated in passages where positive sustainability language is not accompanied by quantities, dates, implementation details, or measured outcomes. Under the direct baseline, 15.7% of ambiguous claims receive reported-action status, and 8.1% of the combined aspirational and ambiguous samples are upgraded in the same direction. The corresponding rates are 0.0% for the Skill. Preserving missing evidence therefore changes how uncertainty is handled: weakly supported language remains ambiguous instead of being resolved in favour of action.

The Toyota–Apple–BP application illustrates the same mechanism in a more concrete setting. Toyota and Apple can form a bounded comparison group when their renewable-electricity disclosures share a compatible metric form, unit, reporting period, scope, and organisational boundary. BP remains outside that group because its disclosure expresses a future carbon-neutrality commitment without a corresponding renewable-electricity share. The resulting decision retains the valid Toyota–Apple comparison while refusing a full three-company ranking. This outcome shows why comparability should be assessed at the level of specific records rather than inferred from a shared sustainability theme.

Refusal remains useful only when it is linked to the attribute required by the requested judgement. A missing baseline prevents comparison of a reduction or growth measure, yet may be irrelevant to a point-in-time percentage share. Likewise, different units may block a comparison until a valid conversion is available, while differences in wording alone should not. The remaining over-refusals suggest that the current rules do not yet capture every valid normalisation or equivalence. Further refinement should focus on unit conversion, denominator alignment, boundary reconciliation, and the recovery of supporting context from neighbouring report passages.

The findings should also be interpreted within the boundaries of the current evaluation. The benchmarks were constructed around the failure modes addressed by the two gates, so the reported results show behaviour under predefined alignment and claim-status conditions. The 30-case adversarial pilot broadens the range of examples through implicit baselines, mixed boundaries, textual noise, and heterogeneous units, although its scale does not support a general robustness estimate.

The evaluation combines evidence extraction, structured record construction, and gate execution. A schema-only baseline would help separate the effect of the intermediate representation from the effect of the decision rules. Reproducibility also depends on complete documentation of benchmark construction, annotation procedures, prompts, the dated model snapshot, runtime settings, and repeated-run procedures.

The Skill assesses whether the supplied disclosure supports a requested judgement. It does not independently establish whether the disclosure is factually accurate, materially complete, or consistent with external evidence. A reported-action label indicates that the passage reports implementation or a measured outcome; it does not confirm that the activity occurred. Extending the system to

broader greenwashing assessment would require external operational data, historical consistency checks, materiality analysis, and independent verification.

6. Conclusion

This study addresses a specific weakness in LLM-based ESG report analysis: accurately retrieved disclosure evidence can still be used to support an invalid comparison or an overstated claim classification. The ESG Alignment Skill responds to this problem by converting disclosure passages into structured evidence records and constraining downstream decisions through comparability and claim-status gates.

On the controlled benchmarks, comparability accuracy increases from 87.0% to 90.0%, accompanied by a reduction in over-refusal from 26.0% to 20.0% and no observed increase in unsafe allowance. Claim-status accuracy increases from 94.5% to 100.0%, while the observed ambiguous-as-action and unsupported-action rates fall to 0.0%. These results indicate that explicit evidence conditions can reduce unsupported judgements, particularly when positive sustainability language lacks implementation evidence.

The proposed Skill is a targeted control layer for report-based reasoning. Its role is to preserve the conditions under which a comparison or claim label is supported and to expose those conditions when they are absent. Broader evaluation with independently annotated disclosures, stronger baselines, and more diverse reporting settings is needed to determine how well this approach extends across sectors, jurisdictions, standards, and languages.

Acknowledgments

The authors thank the organisers and reviewers for their time and feedback.

Declaration on Generative AI

During preparation, the authors used generative artificial intelligence tools for language refinement, grammar checking, and editorial consistency. The authors reviewed the output and take responsibility for the final content.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. K"uttler, M. Lewis, W.-t. Yih, T. Rockt"aschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[2] J. Ni, J. Bingler, C. Colesanti-Senni, M. Kraus, G. Gostlow, T. Schimanski, D. Stammbach, S. Ashraf Vaghefi, Q. Wang, N. Webersinke, T. Wekhof, T. Yu, M. Leippold, 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, 2023, pp. 21–51. URL: https://aclanthology.org/2023.emnlp-demo.3/. doi:10.18653/v1/2023.emnlp-demo.3.

[3] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1906–1919. URL: https://aclanthology.org/2020.acl-main.173/. doi:10.18653/v1/2020.acl-main.173.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[5] J. Wallat, M. Heuss, M. de Rijke, A. Anand, Correctness is not faithfulness in retrieval augmented generation attributions, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, Association for Computing Machinery, 2025, pp. 22–32. doi:10.1145/3731120.3744592.

[6] M. A. Delmas, V. C. Burbano, The drivers of greenwashing, California Management Review 54 (2011) 64–87. doi:10.1525/cmr.2011.54.1.64.

[7] T. P. Lyon, A. W. Montgomery, The means and end of greenwash, Organization & Environment 28 (2015) 223–249. doi:10.1177/1086026615575332.

[8] Y. Kılınç, M. R. İnce, A. C. Badem, A multi-dimensional textual framework for detecting greenwashing in sustainability reporting, Discover Sustainability 7 (2026) 482. doi:10.1007/s43621-026-02890-x.

[9] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, H. Wu, Unified structure generation for universal information extraction, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5755–5772. URL: https://aclanthology.org/2022.acl-long.395/. doi:10.18653/v1/2022.acl-long.395.

[10] F. Berg, J. F. K"olbel, R. Rigobon, Aggregate confusion: The divergence of ESG ratings, Review of Finance 26 (2022) 1315–1344. doi:10.1093/rof/rfac033.

A. Detailed Benchmark Results

Table 4 reports the exact counts and diagnostic rates for the two primary benchmark tasks.

Table 4
Detailed benchmark results.

Metric	Direct LLM baseline	ESG Alignment Skill
ESG comparability		
Total cases	200	200
Correct decisions	174 / 200	180 / 200
Accuracy	87.0%	90.0%
Over-refusal rate	0.26	0.20
Unsafe-allow rate	0.00	0.00
Greenwashing-related claim-status classification		
Total samples	200	200
Correct classifications	189 / 200	200 / 200
Accuracy	94.5%	100.0%
Reported-action accuracy	1.00	1.00
Aspirational accuracy	1.00	1.00
Ambiguous-claim accuracy	0.843	1.00
Ambiguous-as-action rate	0.157	0.00
Dangerous-overclaim rate	0.081	0.00

Note. Over-refusal and unsafe-allow denote valid cases refused and invalid cases allowed, respectively. Ambiguous-as-action and dangerous-overclaim denote ambiguous claims, and aspirational or ambiguous claims, classified as reported action. Unmarked rates are reported as proportions.

B. Illustrative Benchmark Cases

B.1. Claim-Status Case

Input disclosure: “We continue to pursue circular-economy principles and recycled 48,000 tonnes of packaging material in financial year 2024 (FY2024).”

Expected label: Reported action.

Rationale: The sentence reports a quantified completed activity with a defined period despite its broader aspirational framing.

C. Example Structured Evidence Record


```
{
  "company": "Company A",
  "metric_id": "renewable_electricity_share",
  "metric_type": "percentage_share",
  "reported_value": "75",
  "unit": "%",
  "reporting_year": "2024",
  "baseline_year": null,
  "target_year": null,
  "ghg_scope": null,
  "organisational_boundary": "electricity used",
  "claim_status": "reported_action",
  "source_quote": "In 2024, 75% of the electricity used was obtained from
     renewable sources.",
  "warnings": ["GHG scope not applicable or not stated"],
  "alignment_note": "Compare only with aligned renewable-electricity shares."
}
```


D. Adversarial Pilot Record

Table 5
Adversarial pilot results for scope, unit, and type mismatches ($n = 12$).

Case ID	Type	Diff.	Gold	Decision	Res.	Gold rationale
adv_001	scope_mismatch	med.	False	False	PASS	A/B scope_1_2, C scope_1 mismatch
adv_002	scope_mismatch	hard	False	False	PASS	carbon footprint scope ambiguity blocked
adv_003	scope_mismatch	hard	False	True	FAIL	electricity vs energy vs purchased power scope drift
adv_004	unit_mismatch	med.	False	False	PASS	MWh vs GWh vs GJ unit mismatch
adv_005	unit_mismatch	hard	True	False	FAIL	2.5 million tonnes / 2.5 MtCO2e / 2,500 kt should normalise to same unit
adv_006	unit_mismatch	hard	False	True	FAIL	m3/tonne vs m3/ton vs L/kg different unit systems
adv_007	type_mismatch	med.	False	False	PASS	share vs growth_rate vs claim type mismatch
adv_008	type_mismatch	hard	False	False	PASS	absolute vs intensity vs claim type mismatch
adv_009	type_mismatch	med.	False	False	PASS	absolute vs intensity vs intensity_reduction type mismatch
adv_010	metric_id_mismatch	med.	True	False	FAIL	all fields identical; baseline_year should not block
adv_011	metric_id_mismatch	hard	True	True	PASS	both renewable_share, percentage points
adv_012	metric_id_mismatch	hard	False	False	PASS	baseline extraction difference causes metric_id divergence

Note. Gold is the reference decision and Decision is the Skill output. True permits comparison, False refuses it, and PASS indicates agreement. Diff. denotes difficulty.

Table 6
Adversarial pilot results for metric, comparability, extraction, and boundary cases ($n = 18$).

Case ID	Type	Diff.	Gold	Decision	Res.	Gold rationale
adv_013	partial_comparable	med.	False	False	PASS	C is claim type blocked; A–B locally comparable
adv_014	partial_comparable	hard	False	False	PASS	C scope_1_2_3 mismatches A/B scope_1_2
adv_015	partial_comparable	hard	False	False	PASS	C intensity_reduction vs A/B intensity type difference
adv_016	partial_comparable	med.	False	True	FAIL	A/B global, C European geographic mismatch
adv_017	extraction_robustness	med.	True	False	FAIL	synonym/abbreviation extraction failed (elec./green/power)
adv_018	extraction_robustness	med.	True	False	FAIL	synonym/abbreviation extraction failed (YoY/GHG/CO2e)
adv_019	extraction_robustness	hard	False	False	PASS	different metric granularities correctly distinguished
adv_020	extraction_robustness	hard	False	False	PASS	TRIR vs LTIFR correctly distinguished
adv_021	mixed_boundary	hard	False	False	PASS	target vs achieved vs commitment type difference
adv_022	mixed_boundary	hard	False	False	PASS	RECs vs on-site vs direct sourcing scope difference
adv_023	mixed_boundary	hard	False	False	PASS	market-based vs location-based methodology difference
adv_024	mixed_boundary	hard	False	False	PASS	supply chain vs operations vs palm oil scope difference
adv_025	direction_boundary	med.	True	False	FAIL	direction should not be a blocking gate (value, not condition)
adv_026	direction_boundary	hard	True	True	PASS	direction is part of value; does not affect comparability
adv_027	year_boundary	med.	True	True	PASS	reporting_year should not be a blocking gate
adv_028	year_boundary	hard	True	True	PASS	FY vs CY year normalisation correct
adv_029	taxonomy_boundary	hard	False	False	PASS	social metric not in taxonomy blocked
adv_030	taxonomy_boundary	hard	False	False	PASS	offsets vs reduction methodology difference blocked

Note. Gold is the reference decision and Decision is the Skill output. True permits comparison, False refuses it, and PASS indicates agreement. Diff. denotes difficulty.