Concept Paper

ESGlass: Glass-Box ESG and Sustainability Reports

Chaoyue He *, Xin Zhou, Di Wang, Hong Xu, Wei Liu, Chunyan Miao

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2187.v1

Keywords: ESG reporting; sustainability reporting; glass-box reporting; provenance; content credentials; multimodal foundation models; multimodal agents; document intelligence; multimedia retrieval

 doi:10.20944/preprints202603.2187.v1

Concept Paper

ESGlass: Glass-Box ESG and Sustainability Reports

Chaoyue He [1,*], Xin Zhou [1], Di Wang [1], Hong Xu [1], Wei Liu [2] and Chunyan Miao [1]

[1] Alibaba-NTU Global e-Sustainability CorpLab, Nanyang Technological University, 50 Nanyang Ave, 639798 Singapore
[2] Alibaba Group, Hangzhou, China
* Correspondence: cyhe@ntu.edu.sg

Abstract

We introduce ESGlass, a glass-box paradigm for ESG and sustainability reports. Instead of treating the report page, file, or tagged fact as the native unit, ESGlass treats each material disclosure claim as an interactive evidence object that binds multimodal observations, derived computations, provenance, uncertainty, and stakeholder-specific renderings. This shift matters because sustainability evidence is increasingly dispersed across invoices, tables, sensor streams, forms, maps, satellite imagery, facility video, and internal workpapers, while generative AI makes polished but weakly supported narrative cheap to produce. Building on progress in ESG benchmarks, sustainability knowledge infrastructure, document AI, multimodal retrieval, agents, geospatial foundation models, and provenance standards, we formalize the report as a policy-conditioned view over claim–evidence–provenance graphs and distinguish asset provenance from claim provenance. We argue that glass-box reporting demands stronger targets than citation-style grounding, including minimal sufficient evidence sets, challenge retrieval, replayable transformations, omission semantics, and selective abstention. We outline a one-company energy-disclosure prototype, define task families and evaluation criteria, and surface governance issues such as selective disclosure, privacy-preserving drill-down, and false completeness. ESGlass reframes ESG and sustainability reports from polished narrative artifacts into inspectable multimedia disclosure objects, offering a concrete multimedia research agenda for systems that must not only generate disclosure, but also defend it.

Keywords: ESG reporting; sustainability reporting; glass-box reporting; provenance; content credentials; multimodal foundation models; multimodal agents; document intelligence; multimedia retrieval

1. Introduction

ESG and sustainability reports are becoming more regulated, more assured, and more machine-readable and more machine-analyzed, but not more inspectable. Standards such as IFRS S1/S2, the EU CSRD/ESRS stack, and the new sustainability assurance regime represented by ISSA 5000 are raising expectations around comparability, internal control, and evidence quality [1–4]. Yet the dominant artifact remains a polished report that compresses heterogeneous observations into prose, tables, and a few curated figures. The final narrative may mention a facility retrofit, a supplier land-use policy, or a year-over-year energy reduction, but the underlying support usually lives elsewhere—in satellite tiles, smart-meter logs, invoices, forms, spreadsheets, shipment records, and assurance workpapers.

This mismatch matters more now for two reasons. First, the evidentiary base of sustainability disclosure is increasingly multimedia and multi-source. Second, generative AI can now produce fluent, persuasive, and visually grounded narrative at a pace that far exceeds institutional verification. Much current work still aims large language models or vision-language models at the finished report: parse it, summarize it, answer questions about it, benchmark hallucination on it, or extract structured facts after publication [5–9]. These are valuable directions, but they preserve the asymmetry that matters most: the narrative remains first-class, while evidence stays secondary and scattered.

 doi:10.20944/preprints202603.2187.v1

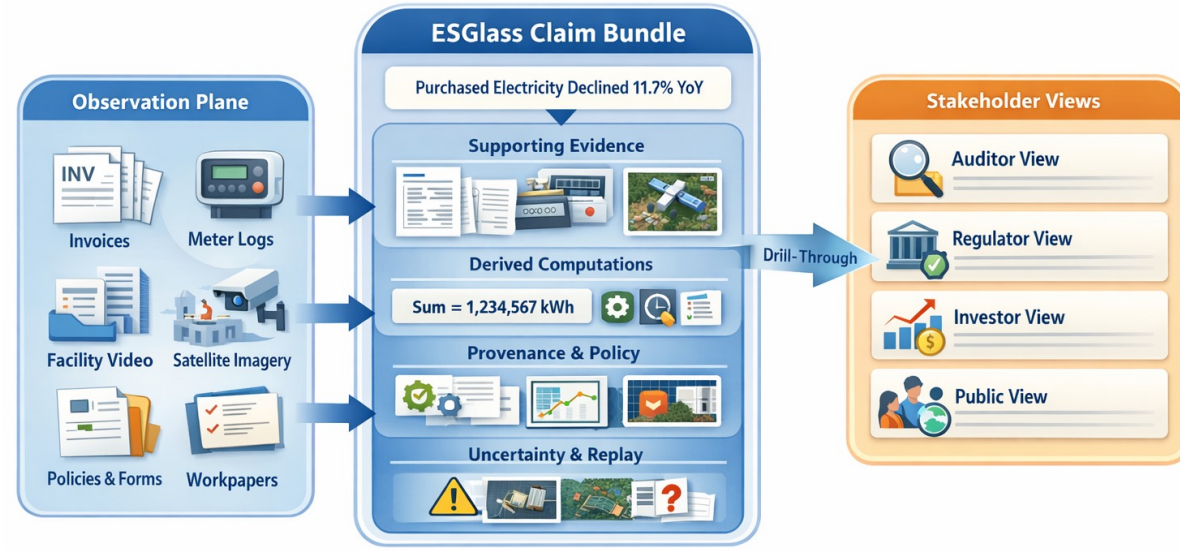

Figure 1. ESGlass turns ESG and sustainability reports into glass-box disclosure objects. Each material claim opens into a typed bundle that preserves multimodal observations, derived computations, provenance, uncertainty, challenge paths, and stakeholder-specific renderings. Overview diagram of ESGlass. On the left, an observation plane contains invoices, meter logs, facility video, satellite imagery, policies, and workpapers. In the middle, an ESGlass claim bundle contains a claim, support evidence, challenge evidence, derived computations, provenance and policy, and uncertainty with replay traces. On the right, an interaction plane renders auditor, regulator, investor, and public views. Arrows show drill-through from evidence to claim bundle to stakeholder-specific views.

At the same time, sustainability AI is rapidly expanding beyond basic report QA. EsGenius benchmarks whether LLMs understand ESG and sustainability concepts, MMESGBench probes multimodal reasoning over ESG materials, SSKG Hub and KG4ESG organize standards and domain relations into knowledge structures, and PCA-OS points toward system-level sustainability intelligence that spans data, policy, and decision support [8,10–13]. These advances are important, but they still leave one representational gap: the externally asserted report claim is not yet treated as the native multimedia object.

We argue for a sharper inversion. The report should not be the terminal product of evidence processing; it should be the interface to that process. In a glass-box setting, every material claim is born together with machine-readable support, derivation history, uncertainty, and challenge pathways. A sentence such as *"purchased electricity declined 11.7% year-over-year"* should not merely point to a page number or appendix. It should open into an inspectable bundle containing the relevant utility invoices, smart-meter traces, reporting-boundary rules, conversion factors, aggregation steps, counterevidence checks, freshness timestamps, and the model- or human-generated explanation used to render the final prose.

We call this paradigm *ESGlass*. The name is intentional: it combines ESG with the idea of glass-box disclosure. A glass report is transparent enough to inspect, layered enough to drill through, and structured enough to separate what was observed, what was derived, what was inferred, and what remains uncertain. It is not merely more readable; it is more contestable. The central representational unit is therefore no longer the page, the file, or the tagged fact. It is the *claim bundle*: a typed object that binds a disclosure statement to observed evidence, derived evidence, provenance, uncertainty, and explanations.

Contributions.

This paper makes four contributions. First, it introduces ESGlass as a claim-level glass-box paradigm for ESG and sustainability reports and explains why this is a more appropriate multimedia abstraction than the PDF page, tagged fact, or dashboard tile. Second, it formalizes the report as a policy-conditioned view over claim–evidence–provenance graphs, including support/challenge/insufficient semantics and the notion of minimal sufficient evidence sets. Third, it shows why current report analytics, tagging pipelines, knowledge graphs, RAG systems, and provenance standards provide useful ingredients but do not yet instantiate glass-box disclosure. Fourth, it outlines a sharp one-

 doi:10.20944/preprints202603.2187.v1

company prototype, a benchmark agenda, and evaluation criteria centered on sufficiency, contestability, replayability, and abstention.

2. Why Today's Reports Still Fail the Glass Test

The pressure to improve sustainability reporting is already producing better ingredients. External assurance is increasingly framed as a trust-restoring mechanism, while selective disclosure and greenwashing concerns remain active in the literature [14–16]. Inline XBRL and related digital-disclosure efforts make individual facts machine-readable [17,18]. Assurance standards raise expectations around traceability, controls, and reproducibility [4]. Internal dashboards and digital twins create richer operational visibility. Yet these layers are still fragmented. They improve *parts* of the disclosure pipeline without changing the unit around which the pipeline is organized.

A useful litmus test is the *sentence-click test*. If a reader clicks any material claim, can the system reveal all and only the evidence, transformations, caveats, and conflict checks needed to defend or challenge that specific statement? Static reports usually fail because they are narrative-first. Tagged reports usually fail because they make facts machine-readable but not multimedia support or derivation logic. Data rooms usually fail because they collect files without making them claim-centric or stakeholder-navigable. Dashboards usually fail because they expose internal metrics but do not externalize report-ready provenance, challenge logic, or policy-specific abstractions. The report, in other words, is still rendered after evidence assembly rather than being *defined by* it.

Table 1. Why today's reporting artifacts are useful yet still fail the glass test of claim-level multimedia assurance.

Artifact	Native unit	Primary strength	What is still missing
Static sustainability report	Page, paragraph, table cell	Human-readable storytelling, comparability of narrative sections, formal publication artifact	Narrative is first-class while evidence is externalized. Drill-down is shallow, transformations are rarely replayable, and cross-modal support is weakly linked.
Inline XBRL / digitally tagged report [17,18]	Tagged fact	Machine-readable numbers and selected textual facts; interoperability across consuming systems	Facts become extractable, but multimedia evidence, derivation semantics, counterevidence, and claim-level uncertainty remain largely outside the tagged object.
Assurance workpapers / evidence room	File, folder, workpaper trail	Rich raw support for internal assurance and audit processes	Strong for practitioner review, weak for claim-centric public navigation. File collections do not automatically tell a stakeholder why a specific sentence should be trusted, challenged, or withheld.
Operational dashboard / digital twin	Metric, stream, alert	Continuous monitoring and internal decision support	Excellent for operations, but not equivalent to external disclosure. Dashboards usually lack versioned narrative provenance, reporting-boundary logic, and stakeholder-specific evidence views.
ESGlass glass-box report	Claim bundle	Claim-level drill-down, replayable derivation, cross-modal evidence, challenge pathways, policy-aware rendering	Requires new methods for evidence sufficiency, counterevidence discovery, temporal/geospatial alignment, privacy-aware provenance, and human–AI assurance interaction.

 doi:10.20944/preprints202603.2187.v1

The deeper point is that the claim itself should become the organizing principle of disclosure. Claims are what management signs, what investors scrutinize, what auditors test, what regulators compare, and what affected communities challenge. A claim-centric representation aligns the computational object with the socio-technical object that actually matters. This is the core conceptual move of ESGlass.

A useful way to understand this move is through the lens of *boundary objects*. The claim bundle is shared across management, auditors, regulators, and public-interest users, but each actor inspects it at a different granularity and with different permissions. Today those groups often touch disconnected artifacts—the annual report, the audit workpapers, the dashboard, the supplier folder, the satellite portal. ESGlass proposes one typed object around which those views can be coordinated without forcing them to become identical.

3. Why Now: The Ingredients Exist, but Not the Glass-Box Abstraction

Several research streams have matured enough to make glass-box reporting plausible. On the perception side, multimodal foundation models such as Flamingo, BLIP-2, LLaVA, and Qwen3-VL have expanded practical capabilities for cross-modal reasoning, document parsing, visual grounding, long-video understanding, and structured extraction from invoices, forms, and tables [19–22]. In document AI, LayoutLMv3, Donut, and DocVQA show that visually rich documents can be treated as first-class multimodal objects rather than flattened OCR text [23–25].

Adjacent work on structured and visual fact verification is equally important. TabFact and FEVEROUS frame claims against tables and mixed text-table evidence [26,27]. ChartQA, MatCha, and DePlot show that charts can be parsed, derendered, and reasoned over as evidence-bearing objects rather than decorative figures [28–30]. ChartQA-X pushes further by treating explanation quality over charts as a modeling target [31]. These lines of work matter because ESG claims are often inseparable from tables, charts, and layout cues.

Retrieval and generation have also advanced. RAG-style systems demonstrate the general value of explicit external evidence for factual generation [32]. MuRAG extends that logic to image-text memory [33]. VDocRAG and RAG-Anything show that retrieval can now operate over visually rich documents, tables, formulas, and image-native corpora without flattening everything to brittle text chunks [34,35]. Self-RAG and RARR suggest that models can critique and revise their own outputs when evidence is weak or missing [36,37]. At the same time, hallucination diagnostics such as HallusionBench and multimodal verification benchmarks such as VERITE remind us that fluent cross-modal coherence is not the same as evidence-grounded truth [38,39].

The agentic turn is also relevant. ReAct-style frameworks and newer multimodal agent systems treat model outputs as plans with tool calls and intermediate traces instead of one-shot answers [40–42]. Disclosure assembly is naturally agentic: identify the claim, infer the evidence requirements, retrieve assets across modalities, compute or reconcile the KPI, search for contradictions, and compose an explanation together with an abstention or escalation decision when support is incomplete.

Geospatial AI is expanding what counts as observable evidence in sustainability contexts. GeoChat and SkySense bring remote-sensing reasoning closer to general multimodal pipelines [43,44]. SatlasPretrain and Prithvi-EO-2.0 show the rapid maturation of Earth-observation foundation models [45,46]. Domain literature has already argued that Earth observation can materially support ESG assurance and sustainable-finance workflows [47,48]. Provenance and authenticity standards such as PROV-DM, the broader data-provenance literature, and C2PA provide the beginning of a representational substrate for lineage and signed transformations [49–51].

There is also a fast-growing sustainability knowledge layer around reports and standards. EsGenius evaluates sustainability knowledge competence in LLMs; MMESGBench measures multimodal ESG reasoning; SSKG Hub structures sustainability standards as expert-guided knowledge graphs; KG4ESG organizes ESG entity and relation atlases; and PCA-OS sketches a systems view of climate adaptation intelligence [8,10–13]. These efforts reduce the semantic gap between raw evidence, stan-

 doi:10.20944/preprints202603.2187.v1

dards, and action, but they still stop short of redefining the report claim itself as a first-class multimedia evidence object.

Table 2. Adjacent research streams supply key ingredients, but they remain misassembled around the wrong object.

Research stream	Representative work	What it already enables	What is still missing for ESGlass
ESG report analytics	ChatReport [5], ESGReveal [6], ESG QA [7], MMESGBench [8], ESG-Bench [9]	Parsing, QA, extraction, and emerging multimodal reasoning benchmarks over existing reports	These systems begin from a finished report. They do not define each disclosure claim as a machine-readable evidence object with replayable lineage to raw assets.
Sustainability knowledge infrastructure	EsGenius [10], SSKG Hub [11], KG4ESG [12], PCA-OS [13]	Benchmarking ESG and sustainability knowledge; structuring standards, entities, and sustainability concepts; connecting disclosure logic to broader sustainability intelligence	These systems organize knowledge, but not the concrete claim-specific bundle that ties a sentence in a report to observed evidence, derived computations, omissions, and stakeholder-facing renderings.
Structured/visual claim verification	TabFact [26], FEVEROUS [27], ChartQA [28]	Verification over tables, mixed text-table evidence, and chart-centric reasoning	Benchmarks validate the importance of structured evidence, but they do not capture enterprise provenance, time-window policy, or stakeholder-facing report rendering.
Document intelligence	LayoutLMv3 [23], Donut [24], DocVQA [25], Qwen3-VL [22]	Parsing complex layouts, forms, invoice pages, and chart-rich documents	Parsing does not say whether the extracted assets are sufficient, challengeable, or policy-compliant support for a claim.
Multimodal retrieval and grounded generation	RAG [32], MuRAG [33], VDocRAG [34], RAG-Anything [35], Self-RAG [36], RARR [37]	Retrieval-aware generation, evidence injection, self-critique, and visually rich document retrieval	Disclosure needs more than citation-style grounding: it needs minimal sufficient evidence sets, typed transformations, counterevidence discovery, and versioned claim objects.
Multimodal agents	ReAct [40], T3-Agent [41], VisualAgentBench [42]	Tool use, decomposition, trajectory reasoning, and multi-step interaction	Agent traces must themselves be provenance-native, replayable, and policy-bounded; otherwise the reasoning chain becomes another opaque layer.
Geospatial and remote-sensing AI	GeoChat [43], SkySense [44], SatlasPretrain [45], Prithvi-EO-2.0 [46]	Region grounding, temporal earth-observation analysis, and foundation models for satellite time series	Earth observation rarely gets linked all the way to claim text, supplier identity, invoices, reporting boundary, and stakeholder-facing uncertainty notes.
Digital tagging and provenance	Inline XBRL [17], PROV-DM [49], Data Provenance [50], C2PA [51]	Machine-readable facts, asset lineage, and signed authenticity manifests	Asset history is not claim provenance. We still need representations for why certain evidence was chosen, how it was transformed, what policy was applied, and what counterevidence was ruled in or out.

The ingredients are therefore not absent; they are simply assembled around the wrong endpoint. Today, the usual computational question is: *given a report, what can a model recover from it?* The ESGlass question is different: *how should a report be constructed so that each claim is inspectable, replayable, and challengeable by design?* That is the missing conceptual leap.

4. ESGlass: Glass-Box Reports as Claim–Evidence–Provenance Graphs

4.1. Core Object Model

We define an ESGlass report $\mathcal{G}$ as a set of claim bundles:

$$\mathcal{G} = \{b_i\}_{i=1}^{n}, \qquad b_i = \langle c_i, E_i^{obs}, E_i^{der}, P_i, X_i, U_i \rangle. \tag{1}$$

Here, c_i is a narrative claim or KPI statement. E_i^{obs} contains observed evidence items such as invoices, sensor logs, satellite tiles, facility video, contracts, forms, or shipment records. E_i^{der} contains derived evidence, including joins, conversions, reconciliations, emissions-factor applications, and KPI calculations. P_i is a provenance graph over entities, activities, tools, models, and human actors. X_i stores explanatory artifacts such as natural-language rationales, chart captions, visual summaries, or

 doi:10.20944/preprints202603.2187.v1

challenge notes. U_i stores uncertainty, coverage gaps, freshness, unresolved conflicts, and abstention flags.

The report consumed by any stakeholder is then a *view* over this object:

$$D_{s,t} = V_{\pi_s,t}(\mathcal{G}), \tag{2}$$

where V renders the disclosure for stakeholder type s at time t under policy π_s. The policy includes reporting boundary, time window, materiality rules, conversion tables, access control, and presentation granularity. This simple view function captures a crucial shift: the report is no longer the underlying object. It is a policy-conditioned rendering of the object. The glass metaphor is operational here: different stakeholders look through the same underlying claim graph at different resolutions, with some layers transparent and some access-controlled.

A second key distinction is between *support* and mere *relevance*. Most retrieval systems optimize semantic relevance. Disclosure demands a stricter decision:

$$\sigma_\pi(c_i, \mathcal{E}_i) \in \{\texttt{support}, \texttt{challenge}, \texttt{insufficient}\}, \tag{3}$$

where $\mathcal{E}_i = E_i^{obs} \cup E_i^{der}$. The same asset may be relevant to a claim yet insufficient to support it under the applicable policy. A rooftop image of solar panels, for example, is relevant to the claim *"the solar array was operational in Q3"* but insufficient unless combined with temporally aligned inverter output, installation records, and maintenance status. This motivates a stronger target than ordinary retrieval: the *minimal sufficient evidence set*

$$S_i^\star \in \arg\min_{S \subseteq \mathcal{E}_i} \{|S| : \sigma_\pi(c_i, S) = \texttt{support}\}, \tag{4}$$

which need not be unique. In practice, systems may recover a family of approximately minimal support sets, but the conceptual goal is important: disclosure should expose not just relevant evidence, but the smallest typed bundle that would let a reasonable reviewer replay and defend the claim.

4.2. Three Conceptual Shifts

From relevance to sufficiency.

A disclosure system should search for the *minimal sufficient evidence set*, not just the top-k semantically similar assets. This turns retrieval into a typed, policy-conditioned reasoning problem. It also creates a new benchmark target: exact or approximate recovery of support sets and challenge sets.

From support-only retrieval to contestability.

Assurance is not complete when supporting evidence is found. It becomes meaningful when the system also asks what could weaken the claim: missing data windows, contradictory invoices, alternate explanations, or policy mismatches. Challenge edges should be first-class edges in the provenance graph, not hidden reviewer notes.

From one-shot generation to versioned claims.

A claim bundle is a living object. New observations can raise or lower freshness, attach late-arriving evidence, or trigger re-rendering of narrative sections. This is especially important for sustainability contexts in which data streams, satellite revisits, supplier boundaries, and emissions factors are updated over time.

4.3. Design Invariants

Evidence-first generation.

Narrative text should be rendered *after* evidence assembly, not before it. This reverses the increasingly common workflow in which a model drafts persuasive prose and a human later searches

 doi:10.20944/preprints202603.2187.v1

for supporting snippets. Evidence-first generation does not ban narrative assistance; it constrains narrative to be a downstream view over assembled support.

Strict layer separation.

The evidence object, the interpretation object, and the rendered narrative view should remain distinct. A model-generated explanation belongs in X_i and must carry its own provenance, but it should never silently upgrade itself into observed evidence. This separation is one of the clearest practical defenses against explanation leakage.

First-class omission semantics.

Most current interfaces say a lot about what is present and little about what is absent. A provenance-native system should represent missing intervals, stale modalities, inaccessible evidence, and unresolved scope changes explicitly. In disclosure, omission is not background noise; it is part of the epistemic state.

Policy-conditioned rendering.

Two stakeholders may need the same truth structure but different views. An auditor may inspect raw invoices and replay code; a regulator may inspect normalized cross-firm facts; a community user may inspect a privacy-preserving explanation card. A strong design therefore keeps one underlying claim graph while allowing multiple policy-aware renderings.

Continuous rather than annual truth maintenance.

Sustainability evidence changes between reporting cycles. A glass-box report should track freshness, supersession, and backfilled evidence so that a stakeholder can inspect both the current state and the state that existed when the claim was rendered. This encourages a move from annual "snapshot PDF" logic toward versioned disclosure maintenance and makes disclosure closer to a maintained knowledge object than a once-a-year publication event.

4.4. Why Asset Provenance Is Not Claim Provenance

ESGlass is not just "more provenance." It requires a different provenance target. Asset provenance answers questions such as where a file came from, who signed it, or which tool altered it. Claim provenance answers a harder set of questions: why a particular collection of assets was selected; which policy and time window governed the selection; how the KPI or sentence was computed; which conflicting assets were discovered; what remained unknown; and whether the system should have abstained.

This distinction matters because authenticity is necessary but not sufficient. A signed video can still mislead if it is temporally stale. A valid invoice can still support the wrong claim if the reporting boundary changed. A beautifully cited explanation can still overreach if it compresses multiple uncertain inferences into one smooth sentence. In practice, at least five threat surfaces recur:

1. **Selective evidence:** the shown evidence is genuine but incomplete.
2. **Temporal drift:** individually valid assets correspond to incompatible time windows.
3. **Policy misbinding:** the wrong reporting boundary, factor table, or supplier scope is applied.
4. **Explanation leakage:** a fluent rationale is mistaken for proof.
5. **False completeness:** the interface hides that important modalities or intervals are missing.

A mature system therefore needs not only lineage, but also explicit omission warnings, challenge links, freshness semantics, and abstention states.

4.5. Why This Is Not Just RAG with Citations

It is tempting to describe ESGlass as a domain-specific retrieval-augmented generation system with better references. That description is too weak. In ordinary RAG pipelines, citations often behave

 doi:10.20944/preprints202603.2187.v1

like *pointer semantics*: they show where some supporting material may reside, but they rarely specify whether the cited evidence is minimal, sufficient, policy-compliant, current, or contradicted elsewhere. The generated sentence remains primary and the evidence layer stays auxiliary. ESGlass reverses this dependency. The claim bundle is primary, and any narrative sentence is merely one rendering of the bundle. This is why ESGlass belongs naturally in multimedia research: the hard problem is not citation formatting, but cross-modal evidence assembly, alignment, and contestable rendering.

Three differences follow. First, disclosure requires *typed evidence requirements*. A sentence about energy consumption, land-use conversion, or supplier compliance is not satisfied by any semantically similar document. It demands a specific configuration of modalities, time windows, boundaries, and transformations. Second, disclosure requires *stable replayability*. If the same claim is regenerated a week later, the system should preserve the underlying support set, policy bindings, and unresolved conflicts unless the evidence state itself changed. Third, disclosure requires *negative evidence handling*. A strong system must search for what would overturn the claim, not merely what would let a language model complete it fluently.

This distinction also clarifies the role of agents. In a standard agentic RAG workflow, intermediate tool traces are often implementation details. In ESGlass, those traces become part of the evidentiary object because they document how the system interpreted the claim, which evidence classes it searched, which transformations it executed, and why it escalated or abstained. The agent is therefore not just an answerer with tools. It is closer to an automated junior assurance process whose decisions must remain inspectable.

5. Architecture and a Sharp Prototype

5.1. A Three-Plane Architecture

A first ESGlass implementation can be built as a three-plane stack. The *observation plane* stores the raw multimedia evidence: documents, images, maps, videos, tables, and time series together with credentials and basic metadata. The *assurance plane* decomposes claims into evidence requirements, retrieves candidate assets, aligns them across modality, recomputes derived values, and records support/challenge decisions. The *interaction plane* renders stakeholder-specific views, including narrative text, claim cards, conflict logs, and drill-down interfaces.

The most important architectural decision is where reasoning traces live. In ESGlass, they belong with the claim bundle rather than in an ephemeral agent log. This turns planner outputs, retrieval choices, reconciliation steps, and abstention decisions into inspectable parts of the disclosure object instead of invisible middleware.

This architecture is close enough to current AI practice to be buildable. Qwen3-VL- or document-model-based parsers can extract page-level structure from invoices and forms; VDocRAG-like retrievers can operate over visually rich document stores; RAG/self-critique modules can drive evidence-aware explanation; and agent controllers can orchestrate the multi-step workflow [22,34,36,41]. Standards knowledge graphs such as SSKG Hub and domain atlases such as KG4ESG can supply policy tables, factor mappings, terminology normalization, and entity resolution during claim bundle assembly [11,12]. The novelty is therefore not merely "using a VLM for ESG." It is changing the unit of multimedia computation from *answering questions about a report* to *constructing and defending a report claim as an evidence object*.

5.2. One Company, One Disclosure Chapter

A credible first prototype does not need to solve all of sustainability reporting. It can focus on one company and one disclosure chapter, such as the energy section of a manufacturing firm's annual report. Energy is strategically attractive because it already mixes multiple evidence regimes—invoices, meter data, maintenance logs, production context, facility imagery, and sometimes public geospatial evidence—without immediately requiring the full complexity of scope-3 accounting.

 doi:10.20944/preprints202603.2187.v1

Table 3 sketches three example claims. Each has a different evidentiary profile. C1 is primarily financial and time-series based. C2 requires visual and temporal grounding. C3 is the most ambitious because it is explanatory: it asks whether a stronger narrative conclusion is justified rather than whether a simple numeric difference exists. That mix is valuable because it immediately surfaces the distinction between *observed evidence* and *interpretive overreach*. A strong system should be able to support C1, potentially support or challenge C2 depending on time alignment, and often abstain on C3 unless sufficient contextual evidence exists.

Table 3. Illustrative ESGlass claim cards for a one-company energy-disclosure prototype.

ID	Claim	Core evidence stack and primary failure mode
C1	Purchased electricity declined 11.7% YoY	Invoices, smart-meter traces, facility boundary rules, unit normalization, and factor tables. *Failure mode:* reporting-boundary drift across years.
C2	Rooftop solar became operational in Q3	Work orders, installation records, rooftop imagery, inverter output, and maintenance events. *Failure mode:* installed does not necessarily mean operational.
C3	Demand reduction was not merely lower activity	Production volumes, operating hours, occupancy context, selected video segments, and anomaly logs. *Failure mode:* apparent efficiency may simply reflect lower utilization.

A prototype workflow can be decomposed into three stages. **Stage 1: asset ingestion and credentialing.** Collect utility invoices, meter exports, maintenance logs, work orders, selected imagery, and policy tables; assign stable identifiers, hashes, and access-control metadata. **Stage 2: claim bundle assembly.** For each draft claim, use an agent to infer evidence requirements, retrieve candidate support and challenge assets, compute or replay the KPI, and populate uncertainty fields. **Stage 3: policy-conditioned rendering.** Render the final disclosure as prose, KPI tables, and claim cards. Each sentence or KPI cell can be opened into its evidence bundle, challenge log, and replay trace.

The same prototype can later be extended to land-use or supply-chain chapters. Facility evidence would be replaced by supplier polygons, earth-observation time series, deforestation alerts, shipment records, customs forms, and contract metadata [47,48]. At larger scale, PCA-OS-like infrastructures could host cross-organization glass-box claims for adaptation and resilience reporting that link local evidence bundles to broader planetary decision layers [13]. The central concept—claim bundle as unit of disclosure—remains unchanged.

5.3. Private Assembly, Public Rendering

A practical deployment cannot assume that every evidentiary asset is publicly revealable. Utility invoices contain account identifiers, supplier files contain contractual terms, facility video may expose people, and supply-chain artifacts may reveal commercially sensitive relationships. A viable glass-box design therefore separates *internal evidentiary assembly* from *external stakeholder rendering*. The internal graph can carry raw assets, fine-grained provenance, and sensitive joins; the external view can expose hashed commitments, redacted snippets, normalized summaries, uncertainty statements, and signed verifier attestations about hidden steps.

This separation suggests an important research direction: *redaction-aware provenance*. The challenge is not simply hiding fields, but preserving contestability when some evidence cannot be shown in full. A public claim card may need to reveal that the system saw twelve monthly invoices, that all twelve matched a declared facility boundary, that anomalies were escalated, and that the replayed total equals the published KPI—without revealing account numbers or exact timestamps. Verifiability and full visibility are not the same thing. ESGlass therefore does not require universal radical transparency. It

 doi:10.20944/preprints202603.2187.v1

requires that every claim have a defensible public *evidence surface* whose limitations are explicit. In practice, that means the public object should disclose not only what is shown, but also what remains hidden, aggregated, or third-party verified.

6. Research Agenda

Claim-grounded multimodal indexing.

Traditional indexes organize corpora by page, chunk, or embedding. Provenance-native disclosure needs indexes organized around evidence needs: facility identity, supplier hierarchy, geolocation, time window, reporting boundary, factor table, and claim type. Retrieving semantically related assets is easier than retrieving *assurance-relevant* assets. Knowledge substrates such as SSKG Hub and KG4ESG can help normalize standards, entities, and factor tables, but the hard retrieval problem remains claim-specific sufficiency [11,12].

Minimal sufficient evidence sets.

A promising formal target is recovery of the minimal sufficient evidence set for a claim under a given policy. This differs from top-k retrieval because one additional invoice, calibration log, or change-of-boundary notice can flip a verdict. Learning minimality and sufficiency may require new supervision beyond answer strings or supporting spans.

Counterevidence and alternate explanations.

Current RAG pipelines are optimized to support the active prompt. Assurance systems must also find what weakens a claim. In energy reporting, a drop in electricity use could be explained by efficiency, reduced activity, metering changes, or reporting-boundary shifts. The system should actively search among these alternatives and expose which ones remain unresolved.

Temporal and geospatial joins.

Claims rarely live on a single timestamp or image. They span facilities, quarters, suppliers, and regions. Multimedia models must therefore align assets whose spatial resolution, revisit frequency, and semantics differ. A rooftop image from September, a meter series from August, and an invoice from October may all be individually relevant but jointly insufficient.

Policy-conditioned reasoning.

Support for a claim is not policy-free. Emissions factors, organizational boundaries, materiality thresholds, and jurisdiction-specific standards change the evidentiary logic. This suggests a research program in *policy-conditioned multimodal reasoning*, where the same evidence bundle may support one disclosure formulation but not another.

Provenance-aware explanation.

Explanation quality is not enough. We need explanation objects whose sentences can themselves be linked to evidence nodes, whose unsupported phrases can be flagged, and whose uncertainty is explicit. ChartQA-X and related work show that explanation generation can be modeled, but disclosure requires stronger faithfulness guarantees because explanations affect assurance outcomes [31].

Human–AI assurance interfaces.

The interface problem is substantial. Auditors need reproducibility and replay. Regulators need comparability and selective access. Public-interest users need accessible drill-down and challenge mechanisms. The same underlying object must therefore support multiple renderings without forking the truth structure.

 doi:10.20944/preprints202603.2187.v1

Continual reporting and version control.

Annual PDFs encourage a batch mindset, but many sustainability signals are continuous. A glass-box system should represent versioned claims, late-arriving evidence, stale-evidence warnings, and historical diffs. This creates a new multimedia time-travel problem: how should a stakeholder inspect what the system knew, and did not know, at the moment a claim was rendered?

7. Task Families and Benchmark Construction

The paradigm also suggests a cleaner benchmark agenda than generic long-document QA or pure sustainability-knowledge testing. EsGenius, MMESGBench, and ESG-Bench already show that sustainability knowledge and report reasoning are challenging for current models [8–10]. The next step is to construct datasets whose supervision target is not merely the answer, but the claim bundle itself.

A useful benchmark suite would include at least six task families. **Claim-to-evidence retrieval** asks the system to recover the minimal cross-modal support set for a sentence or KPI cell. **Challenge retrieval** asks it to find the evidence most likely to weaken that claim. **Provenance graph induction** asks it to reconstruct which assets, tools, actors, and transformations connect the raw corpus to the final disclosure. **Policy-conditioned verdicting** asks whether the available bundle supports, challenges, or is insufficient for the claim under an explicit reporting policy. **Provenance-aware report generation** asks for narrative that is faithful to the bundle and explicit about uncertainty. **Stakeholder interaction quality** evaluates whether different users can navigate from summary narrative to source evidence and detect conflicts efficiently.

These task families differ in an important way from existing fact-verification or document-QA benchmarks. TabFact, FEVEROUS, and ChartQA show how to supervise claims against structured or visual evidence [26–28]; DocVQA and VDocRAG show how to query visually rich documents [25,34]. But disclosure requires *typed, open-world supervision*. The benchmark must know not only which evidence supports a claim, but also when the bundle is incomplete, when policy assumptions change the verdict, and when evidence should be partially hidden for privacy reasons.

For that reason, future datasets should annotate more than support spans. They should include claim type, reporting boundary, relevant time window, geospatial scope, factor tables or policy rules, minimal support sets, challenge sets, freshness state, and access-control notes. They should also include claims that are intentionally unresolved. Training on only answerable claims would teach the wrong behavior. In real reporting, saying "insufficient evidence" is often the best possible output.

An especially promising construction strategy is *bundle-first annotation*. Instead of starting from a finished PDF and asking annotators to write questions, start from curated claim cards with attached evidence objects and then render stakeholder views from them. That yields supervision for retrieval, provenance, narrative rendering, uncertainty expression, and user interaction all at once. It also mirrors the conceptual proposal of the paper: the report becomes a view over annotated claim bundles rather than the source of truth itself.

8. Evaluation and Governance

The right evaluation target is not generic QA accuracy over long ESG PDFs. A stronger framework should assess whether a system can build, challenge, and communicate claim bundles under realistic incompleteness, privacy, and policy constraints.

Several benchmark design choices follow from Table 4. Datasets should annotate not only answers or support spans, but also claim types, minimal support sets, challenge evidence, temporal windows, geospatial scope, policy assumptions, and access-control constraints. They should include claims that are genuinely *unresolvable* from the available bundle, because abstention is an important competency rather than a failure case. And they should evaluate stakeholder interaction, since the system's value depends on whether users can navigate from summary narrative to source evidence and back again.

 doi:10.20944/preprints202603.2187.v1

Table 4. Evaluation dimensions for ESGlass systems.

Dimension	Example metrics or annotations	Why it matters
Claim coverage	Share of claims with populated bundles; modality and freshness coverage	Measures how much of the disclosure is inspectable rather than free-floating narrative.
Support-set quality	Match to minimal support sets; typed retrieval recall; support-set compression	Rewards systems that retrieve the right evidence rather than simply more evidence.
Challenge quality	Counterevidence recall; contradiction localization; alternate-explanation discovery	Assurance requires finding what weakens a claim, not only what supports it.
Replayability	KPI replay error; trace completeness; factor-table correctness	A published number should be reproducible from recorded inputs and transformations.
Selective abstention	Behavior under missing evidence; abstain precision/recall; calibration	A trustworthy system should prefer `insufficient` to a fluent but unsupported conclusion.
Temporal/geospatial consistency	Interval overlap; geospatial alignment; stale-evidence detection	Many sustainability claims fail because relevant assets do not line up in space or time.
Human utility	Time to verify; error-discovery rate; disagreement-resolution time	The interface should improve assurance work and public scrutiny, not merely generate citations.
Governance overhead	Privacy leakage; redaction fidelity; compute cost; workflow friction	A technically elegant system that is unusable or unsafe will not change practice.

Governance concerns are not peripheral. Selective disclosure is already a documented concern in sustainability reporting [16]. In a glass-box setting, the danger shifts from omitted sentences to omitted evidence surfaces. Privacy constraints may prevent raw evidence from being shown directly, which motivates redaction-aware proofs and role-specific views. Credentialed content can still mislead through omission or misleading aggregation. These issues suggest that ESGlass systems must be designed with explicit threat models rather than generic "trustworthy AI" aspirations.

One practical implication is that provenance-native reporting should separate three layers cleanly: **evidence objects**, **interpretation objects**, and **rendered narrative views**. Evidence objects carry raw or derived support plus lineage. Interpretation objects explain, summarize, and challenge those supports. Rendered narrative views turn the result into stakeholder-facing prose, tables, and figures. Conflating the layers is precisely what makes current AI-assisted reporting hard to audit.

8.1. Adversarial Stress Tests

Generic question-answering accuracy will miss the hardest failure modes of provenance-native disclosure. The more realistic adversary is not a model that invents a random number. It is a workflow that assembles *plausible but strategically incomplete* evidence. We therefore argue for stress tests built around five recurring cases: **authenticity laundering**, **temporal swap**, **boundary swap**, **proxy inflation**, and **redaction overreach**. Together they ask whether genuine assets can still be composed into a misleading claim through stale timing, wrong scope, weak proxies, or overzealous hiding.

These failure modes point to a deeper insight: the main object to defend against is not hallucination alone, but *epistemic overstatement*. A model, analyst, or company can overstate what the available evidence justifies even when every individual asset is real. This is precisely the territory in

 doi:10.20944/preprints202603.2187.v1

which greenwashing risks can migrate into AI-mediated workflows [15]. It is especially likely when explanations are smooth, charts are visually persuasive, and users see only the rendered narrative layer. Stress testing should therefore construct cases in which a compelling story is easier to generate than a well-supported one.

Human factors also matter. A polished claim card can create automation bias just as easily as a polished paragraph can. Review protocols should therefore test whether auditors, regulators, or public-interest users can *discover* hidden weaknesses rather than merely confirm visible support. In some settings, a slower interface that foregrounds unresolved conflicts may be more trustworthy than a faster interface that collapses everything into a single confidence score.

9. Conclusion

ESGlass does not claim that every ESG statement can be machine-verified or that provenance alone guarantees truth; judgment, privacy, and governance remain irreducible. Its claim is narrower and more actionable: as ESG and sustainability reporting becomes increasingly multimedia and AI-mediated, the report should no longer be a glossy terminal artifact but a glass-box interface to typed claim bundles that preserve evidence, derivations, uncertainty, and challenge paths. By recasting ESG and sustainability reports as policy-conditioned views over claim–evidence–provenance graphs, ESGlass opens a multimedia research agenda centered on sufficiency, contestability, replayability, and abstention, and offers a practical path from today's report-centric benchmarks and sustainability knowledge systems to genuinely inspectable disclosure.

Funding: This research is supported by the RIE2025 Industry Alignment Fund (Award I2301E0026) and the Alibaba–NTU Global e-Sustainability CorpLab.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. IFRS Foundation. IFRS S1 General Requirements for Disclosure of Sustainability-related Financial Information. International Sustainability Standards Board Standard, 2023.
2. IFRS Foundation. IFRS S2 Climate-related Disclosures. International Sustainability Standards Board Standard, 2023.
3. European Commission. Corporate Sustainability Reporting. European Commission, 2025.
4. International Auditing and Assurance Standards Board. International Standard on Sustainability Assurance 5000, General Requirements for Sustainability Assurance Engagements. International standard, 2024.
5. Ni, J.; Bingler, J.; Colesanti-Senni, C.; Kraus, M.; Gostlow, G.; Schimanski, T.; Stammbach, D.; Vaghefi, S.A.; Wang, Q.; Webersinke, N.; et al. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023, pp. 21–51.
6. Zou, Y.; Shi, M.; Chen, Z.; Deng, Z.; Lei, Z.; Zeng, Z.; Yang, S.; Tong, H.; Xiao, L.; Zhou, W. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *Journal of Cleaner Production* **2025**, *489*, 144572.
7. Parikh, P.; Penfield, J. Automatic Question Answering From Large ESG Reports. *International Journal of Data Warehousing and Mining (IJDWM)* **2024**, *20*, 1–21.
8. Zhang, L.; Zhou, X.; He, C.; Wang, D.; Wu, Y.; Xu, H.; Liu, W.; Miao, C. MMESGBench: Pioneering Multimodal Understanding and Complex Reasoning Benchmark for ESG Tasks. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 12829–12836.
9. Sun, S.; Wu, B.P.; Jin, M.; Bai, P.; Zhang, H.; Song, X. ESG-Bench: Benchmarking Long-Context ESG Reports for Hallucination Mitigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 39322–39330.
10. He, C.; Zhou, X.; Wu, Y.; Yu, X.; Zhang, Y.; Zhang, L.; Wang, D.; Lyu, S.; Xu, H.; Wang, X.; et al. EsGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 14623–14664.

 doi:10.20944/preprints202603.2187.v1

11. He, C.; Zhou, X.; Yu, X.; Zhang, L.; Zhang, Y.; Wu, Y.; Xiao, L.; Li, L.; Wang, D.; Xu, H.; et al. SSKG Hub: An Expert-Guided Platform for LLM-Empowered Sustainability Standards Knowledge Graphs. *arXiv preprint arXiv:2603.00669* **2026**.
12. He, C.; Zhou, X.; Wang, D.; Yu, X.; Xiao, L.; Li, L.; Xu, H.; Liu, W.; Miao, C. KG4ESG: The ESG Knowledge Graph Atlas. *Preprint* **2026**.
13. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. PCA-OS: A Planetary Climate Adaptation Operating System. *Preprint* **2026**.
14. Pizzi, S.; Venturelli, A.; Caputo, F. Restoring trust in sustainability reporting: the enabling role of the external assurance. *Current Opinion in Environmental Sustainability* **2024**, *68*, 101437.
15. Sneideriene, A.; Legenzova, R. Greenwashing prevention in environmental, social, and governance (ESG) disclosures: A bibliometric analysis. *Research in International Business and Finance* **2025**, *74*, 102720.
16. Roszkowska-Menkes, M.; Aluchna, M.; Kamiński, B. True transparency or mere decoupling? The study of selective disclosure in sustainability reporting. *Critical Perspectives on Accounting* **2024**, *98*, 102700.
17. XBRL International. Inline XBRL 1.1. Specification, 2013.
18. XBRL International. Digital Sustainability Disclosures with XBRL. XBRL International, 2025.
19. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **2022**, *35*, 23716–23736.
20. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International conference on machine learning. PMLR, 2023, pp. 19730–19742.
21. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892–34916.
22. Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* **2025**.
23. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. Layoutlmv3: Pre-training for document ai with unified text and image masking. In Proceedings of the Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 4083–4091.
24. Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; Park, S. Ocr-free document understanding transformer. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 498–517.
25. Mathew, M.; Karatzas, D.; Jawahar, C. Docvqa: A dataset for vqa on document images. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 2200–2209.
26. Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; Wang, W.Y. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164* **2019**.
27. Aly, R.; Guo, Z.; Schlichtkrull, M.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In Proceedings of the Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), 2021, pp. 1–13.
28. Masry, A.; Do, X.L.; Tan, J.Q.; Joty, S.; Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Proceedings of the Findings of the association for computational linguistics: ACL 2022, 2022, pp. 2263–2279.
29. Liu, F.; Piccinno, F.; Krichene, S.; Pang, C.; Lee, K.; Joshi, M.; Altun, Y.; Collier, N.; Eisenschlos, J. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 12756–12770.
30. Liu, F.; Eisenschlos, J.; Piccinno, F.; Krichene, S.; Pang, C.; Lee, K.; Joshi, M.; Chen, W.; Collier, N.; Altun, Y. DePlot: One-shot visual language reasoning by plot-to-table translation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 10381–10399.
31. Hegde, S.; Fazli, P.; Seifi, H. Chartqa-x: Generating explanations for charts. *arXiv e-prints* **2025**, pp. arXiv–2504.
32. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **2020**, *33*, 9459–9474.

 doi:10.20944/preprints202603.2187.v1

33. Chen, W.; Hu, H.; Chen, X.; Verga, P.; Cohen, W. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5558–5570.
34. Tanaka, R.; Iki, T.; Hasegawa, T.; Nishida, K.; Saito, K.; Suzuki, J. Vdocrag: Retrieval-augmented generation over visually-rich documents. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 24827–24837.
35. Guo, Z.; Ren, X.; Xu, L.; Zhang, J.; Huang, C. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323* **2025**.
36. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
37. Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A.T.; Fan, Y.; Zhao, V.; Lao, N.; Lee, H.; Juan, D.C.; et al. Rarr: Researching and revising what language models say, using language models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 16477–16508.
38. Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 14375–14385.
39. Papadopoulos, S.I.; Koutlis, C.; Papadopoulos, S.; Petrantonakis, P.C. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval* **2024**, *13*, 4.
40. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. React: Synergizing reasoning and acting in language models. In Proceedings of the The eleventh international conference on learning representations, 2022.
41. Gao, Z.; Zhang, B.; Li, P.; Ma, X.; Yuan, T.; Fan, Y.; Wu, Y.; Jia, Y.; Zhu, S.C.; Li, Q. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. *arXiv preprint arXiv:2412.15606* **2024**.
42. Liu, X.; Zhang, T.; Gu, Y.; Iong, I.L.; Xu, Y.; Song, X.; Zhang, S.; Lai, H.; Liu, X.; Zhao, H.; et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327* **2024**.
43. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. Geochat: Grounded large vision-language model for remote sensing. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 27831–27840.
44. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27672–27683.
45. Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; Kembhavi, A. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 16772–16782.
46. Szwarcman, D.; Roy, S.; Fraccaro, P.; Gíslason, O.E.; Blumenstiel, B.; Ghosal, R.; De Oliveira, P.H.; de Sousa Almeida, J.L.; Sedona, R.; Kang, Y.; et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing* **2025**.
47. Gu, Y.; Dai, J.; Vasarhelyi, M.A. Audit 4.0-based ESG assurance: An example of using satellite images on GHG emissions. *International Journal of Accounting Information Systems* **2023**, *50*, 100625.
48. Rapach, S.; Riccardi, A.; Liu, B.; Bowden, J. A taxonomy of earth observation data for sustainable finance. *Journal of Climate Finance* **2024**, *6*, 100029.
49. W3C Provenance Working Group. PROV-DM: The PROV Data Model. W3C Recommendation, 2013.
50. Glavic, B. Data provenance: origins, applications, algorithms, and models. *Foundations and Trends in Databases* **2021**, *9*, 209–441.
51. Coalition for Content Provenance and Authenticity. Content Credentials : C2PA Technical Specification. Version 2.3, 2025.