ORIGINAL RESEARCH **Open access**

Scientific Accountability for Materials AI: A Conceptual Standard for Reporting Claims and Limitations

Fatima Zahra Amrani[1*], Youssef Benali[1]

Abstract

Artificial intelligence (AI) has rapidly expanded the scale and ambition of materials research, enabling property prediction, candidate screening, and data-driven optimization across large chemical and structural spaces. However, the field still lacks a discipline-specific standard for scientific accountability: a structured way to report what an AI output legitimately warrants, under which assumptions, and with what limitations. This gap is not cosmetic; it is epistemic. Materials AI often converts heterogeneous proxies (composition features, crystal graphs, microstructure descriptors) into numerical predictions. Yet, manuscripts frequently present these outputs as claims of generality, mechanism, or design readiness without specifying the scope conditions that would make such claims defensible. Recent progress in graph neural networks, benchmark suites, and large community datasets improves comparability. Still, it also amplifies risks of leakage, distribution shift, and proxy instability, which can inflate conclusions while remaining underreported. Meanwhile, uncertainty quantification and explainable AI are increasingly used as trust signals, even though both can be misunderstood when their semantics are not clearly stated, and their limitations are not operationalized for decision-making. We propose a novel conceptual standard—the Scientific Accountability Sheet (SAS)—which binds reported claims to explicit claim types, scope boundaries, evidence anchors, uncertainty semantics, and decision admissibility. SAS reframes “responsible reporting” as a scientific warrant structure rather than an optional best-practice appendix.

Keywords Materials informatics, Distribution shift, Scientific accountability, Reporting standard, Uncertainty semantics, Scientific claims

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Mohammed V University, Rabat, Morocco

Introduction

Materials science is increasingly shaped by computational systems that do not merely assist scientific reasoning but actively produce candidate rankings, predicted properties, and optimized design proposals. Within a single research cycle, AI can propose thousands of candidate materials, estimate their stability or functional performance, and prioritize which candidates merit expensive validation. This growth is enabled by advances in representation learning, benchmark-driven development, and rapidly expanding materials databases and community datasets [1–9]. Yet as materials AI becomes more capable, the field confronts a foundational weakness: performance has scaled faster than accountability.

Accountability in scientific writing is not synonymous with transparency, and it is not guaranteed by reproducibility. A study can be reproducible—meaning the reported numbers can be regenerated—while still being scientifically overclaimed, meaning the conclusions exceed what the evidence warrants. In materials AI, this risk is amplified because typical model inputs do not fully encode the scientific object of interest (a material and its properties). Properties depend on coupled

variables, including composition, processing history, microstructure, defect content, measurement protocol, and operating environment. These dependencies are often weakly represented or entirely latent in the data used to train models [5–7]. Consequently, a model can achieve high predictive accuracy on a curated dataset while remaining epistemically fragile under plausible shifts in processing or measurement context.

The literature already acknowledges that machine learning has become central in solid-state and molecular materials research, with applications ranging from screening and surrogate modeling to discovery acceleration [1–3]. It also acknowledges that "materials data" are not neutral objects: they are shaped by what is measured, what is published, and what is historically valued. However, many AI papers on materials still follow an implicit narrative pattern: a model is trained, metrics are reported, and the paper concludes with strong claims about scientific understanding or design readiness. This pattern creates a gap between what the model computed and what the manuscript asserts.

A key reason is that materials AI manuscripts routinely blend multiple claim categories without separating their warrant requirements. At a minimum, four claim types are ubiquitous:

1. Predictive claims ("the model predicts property Y from input X"),

2. Comparative claims ("this architecture outperforms prior approaches"),

3. Explanatory claims ("the model reveals governing features"),

4. Prescriptive claims ("candidate C is optimal/promising for application").

Table 1 maps each material's AI claim type (predictive, comparative, explanatory, prescriptive) to its minimum required scope boundaries, evidence anchors, uncertainty semantics, and decision admissibility, making claim legitimacy reviewable rather than rhetorical.

Table 1. SAS claim-to-warrant map: required scope, evidence, uncertainty meaning, and decision limits by claim type

Claim type (CT)	Minimum scope boundary required (SB)	Minimum evidence anchor required (EA)	Required uncertainty semantics (US): what must be stated	Decision admissibility (DA): what the claim can and cannot license	Red-flag patter
Predictive ("predict Y from X")	Dataset regime + material class + measurement protocol; what's not encoded (processing/defects/et)	In-regime evaluation + split rationale; leakage controls if relevant	Whether uncertainty is epistemic vs noise vs coverage, what high/low uncertainty implies for reliability	Can license: screening/ranking within scope. Cannot license: transfer to new regimes without stated evidence	"Acc benchma for r discove transfe
Comparative ("outperforms prior")	Same task definition, same label semantics, same splits, same leakage risk profile	Fair baselines + controlled comparison (ablation, same data access);	State whether uncertainty changes across models and whether	Can license: method ranking for that task. Cannot license: superiority in other regimes/tasks	"SOTA best fo predicti regime s

		report sensitivity to split choice	calibration differs (comparisons can invert under miscalibration)		
Explanatory (“reveals governing features/mechanisms”)	Representation scope must show the model actually encodes the purported explanatory variables (or admits proxies/latents)	At minimum: robustness of explanation under perturbations + alternative plausible explanations; ideally: causal/identifiability argument	Must state what the explanation output is (sensitivity, association, hypothesis support) and what it is not (mechanism)	Can license: hypothesis generation/model debugging. Cannot license: mechanistic claims unless a causal warrant is argued	Tr saliency “mecha ig confoundi h
Prescriptive (“candidate C is promising/optimal”)	Context scope must include deployment-relevant constraints (processing feasibility, lifecycle assumptions if claimed)	Evidence must include shift-aware robustness + failure regime disclosure; external/realistic validation pathway stated conceptually	Must map uncertainty to decision risk (what failure means, acceptable risk, what uncertainty bounds)	Can license: prioritization for validation under defined stakes. Cannot license: high-stakes adoption/certification	“Model r candidate usec risk/stak
Mixed claims (e.g., predictive + explanatory)	SB must satisfy the strongest claim’s requirements (usually explanatory/prescriptive)	EA must be typed per claim component; do not let predictive evidence “carry” explanatory parts	US must not be presented as a universal trust score; semantics must be claim-specific		

In practice, each type requires different evidence and different reporting obligations. Predictive claims require clear scope definition and evaluation. Comparative claims require baseline fairness and leakage control. Explanatory claims require restraint against mistaking correlational feature sensitivity for mechanism. Prescriptive claims require decision context: what actions are warranted, and under what failure consequences. Yet the field often treats these claims as interchangeable outputs of “a strong model.”

Benchmarking initiatives have helped standardize evaluation and make comparisons more systematic. Matbench, for example, provides a curated benchmark suite and reference procedures for supervised property prediction tasks [10]. Such infrastructure is valuable: it encourages repeatability and provides a shared language of progress. However, benchmark success can create a comparability illusion, where scores on a benchmark are implicitly treated as proof of general scientific validity. In reality, benchmark tasks are still constrained by their dataset regimes, label semantics, and split designs, which may not reflect deployment conditions in discovery or manufacturing contexts [10–12].

The same structural tension appears even more strongly in large community datasets designed to accelerate catalytic and atomistic modeling. The Open Catalyst 2020 dataset (OC20) explicitly frames generalization as a central challenge and provides tasks intended to test performance beyond memorization of training structures [8]. OC20’s scale and design

have been transformative, but they also illustrate that high performance does not automatically translate into learning transferable physical representations; even in large datasets, generalization across chemistry and geometry remains hard [8]. If this challenge persists at scale, accountability requirements become stricter, not weaker: claims must clearly state which kinds of transfer are supported and which are not.

This brings us to the problem of scientific accountability for claims, not only for code. Evidence from ML-for-science has shown that evaluation can be inflated by leakage and methodological shortcuts, producing confident-looking results that fail under realistic deployment conditions [12]. Materials AI is not immune: the combination of small datasets, correlated samples, compositional families, and shared provenance increases the risk that model success reflects hidden overlap rather than generalizable learning. Without an explicit reporting structure, manuscripts may overstate the reliability and portability of their conclusions.

Uncertainty quantification (UQ) is often proposed as a remedy. Indeed, uncertainty is essential when models are used to prioritize experiments and allocate costly resources. Yet UQ improves accountability only when authors report uncertainty semantics—what it represents, what it does not, and how it bounds claims. Benchmark studies of uncertainty-aware materials property prediction demonstrate that different UQ techniques behave differently and can be misinterpreted if treated as generic confidence [13]. In many papers, uncertainty is presented as a numerical supplement while claim language remains expansive. This reverses the logic of accountability: uncertainty should constrain claims, not decorate them.

Explainable AI introduces a parallel issue. Interpretability tools and explanation narratives can be highly valuable in materials contexts, especially for hypothesis generation and model debugging [14–16]. However, there is a known temptation to treat interpretability outputs (importance scores, saliency, learned attention) as a mechanistic justification. This is a categorical mistake when causal identifiability is absent, confounding is present, or representation semantics are unstable across contexts [14–16]. If explanation tools are used to support mechanistic conclusions, then the report must state the kind of explanatory claim being made and the limits that prevent causal overreach.

Therefore, the central issue is not whether materials AI is useful—it clearly is—but whether the field has a consistent way to prevent overclaiming and ensure that readers can independently judge what conclusions are warranted. We argue that materials AI needs a discipline-native reporting standard that focuses on claim legitimacy rather than only metrics. This standard must declare what claims are made, what scope makes them valid, what evidence anchors them, what uncertainty means, and what decisions the claims permit.

To address this need, we introduce a new conceptual framework: the Scientific Accountability Sheet (SAS). SAS is not a generic checklist and not a superficial “transparency” addendum. It is a structured reporting contract that binds manuscript claims to scope conditions and failure regimes, explicitly limiting how far model outputs can be promoted into scientific or prescriptive statements. The goal is not to slow progress, but to ensure that progress remains scientifically interpretable, reusable, and trustworthy within the regimes where materials decisions are made.

Theoretical Background & Literature Synthesis

Why “Materials AI outputs” are Not automatically “Scientific claims”

A materials AI system produces outputs: predicted values, rankings, uncertainty estimates, and sometimes explanations or generated structures. But a scientific claim is not an output. A claim asserts legitimacy: it implies that a statement remains warranted under specific conditions and can be treated as knowledge rather than mere computation. In materials AI, this conversion from output to claim is especially risky because representation proxies frequently mediate target mappings.

Materials ML has achieved strong predictive performance across many tasks, motivating broad optimism in AI-driven discovery [1–3]. Yet a prediction is epistemically conditional: it depends on the data's hidden assumptions and the evaluation regime used to justify performance. Small-data materials ML methods explicitly acknowledge that sparsity and heterogeneity are defining constraints, shaping not only performance but also robustness and interpretability [5]. This matters because in sparse regimes, model behavior can be dominated by dataset artifacts rather than stable structure–property regularities.

Representation choices further shape claim legitimacy. Structure-based learning has enabled predictive gains by encoding atomic neighborhoods and graph connectivity, providing a more physically meaningful input than hand-crafted features [4, 6]. At the same time, the same representation can be semantically incomplete: it may encode crystal structure but omit processing route; encode composition but omit defect states; encode microscopy texture but omit thermal history. This creates a structural risk: a model can appear scientifically strong while actually learning context-specific correlations.

Benchmarking progress and the risk of "Comparability inflation"

Benchmark suites provide major scientific value: they enable shared tasks, reproducible splits, and clearer progress tracking. Matbench is a prominent example, curating multiple materials property prediction tasks and providing evaluation protocols and baseline references [10]. Benchmarks like this reduce ambiguity in what "good performance" means.

However, comparability does not equal generality. Benchmarks often reflect historical dataset regimes and do not fully simulate real discovery conditions, in which the model is asked to extrapolate to new chemistries, microstructures, or process windows. Moreover, benchmark performance can encourage a rhetorical move: "this model performs best on Matbench," which can be read—implicitly or explicitly—as "this model is scientifically reliable." The latter does not follow unless scope, failure regimes, and decision admissibility are clearly reported.

Community datasets also illustrate this tension. OC20 scales data and focuses explicitly on generalization tasks in catalysis and DFT approximation [8]. Yet even with scale, OC20 emphasizes that generalization is difficult and that existing models may not yet capture transferable physical representations [8]. This is a critical lesson for reporting: large datasets do not eliminate the need for limitation disclosure; they magnify it, because models trained at scale are more likely to be reused across contexts.

Leakage, hidden overlap, and why "High Accuracy" Can be misleading

A persistent threat to accountable reporting is leakage—when evaluation accidentally reuses information that should be out-of-sample, producing inflated performance. In ML-for-science, leakage has been argued to create misleading "progress" by letting models exploit artifacts rather than learn generalizable regularities [12]. Materials AI is particularly exposed because many datasets contain correlated entries (e.g., related compositions, structure families, repeated prototypes), and split strategies may not adequately separate families or regimes.

Accountable reporting requires that authors disclose not just the final scores, but the meaning of the evaluation regime: what was held out, what remained shared, what type of shift was tested, and what type was not. Otherwise, readers cannot infer whether the model is appropriate for screening in nearby regimes or whether it is likely to collapse under modest extrapolation.

Uncertainty quantification: Necessary, but not sufficient

Uncertainty-aware prediction is often considered the gold standard for trustworthy materials AI. Indeed, uncertainty is essential for screening pipelines, active learning, and resource allocation decisions. However, uncertainty becomes scientifically meaningful only when its semantics are reported.

Benchmark work in materials property prediction with UQ shows that multiple UQ methods behave differently and that uncertainty estimates can be inconsistent or misleading under certain conditions [13]. This indicates that uncertainty should not be reported as a single number without interpretive commitments. If a paper reports uncertainty but does not state whether it represents noise, model uncertainty, coverage limitation, or label ambiguity, then uncertainty does not constrain claims; it merely accompanies them.

Therefore, accountability requires a shift from "report uncertainty" to "report uncertainty semantics and decision implications." A model with uncertainty estimates may still be unaccountable if the paper's claims ignore what uncertainty implies about failure probability and regime dependence.

Explainability and the mechanism temptation in materials AI

Interpretability tools are increasingly used to "explain" structure–property relations, select features, and generate scientific narratives. Reviews in explainable AI for materials emphasize both the promise and the pitfalls of interpretability: explanations may help debug models and highlight patterns, but they do not automatically provide mechanistic truth [14–16]. This becomes critical when explanations are promoted into causal or mechanistic claims without justification.

In materials science, mechanistic claims are high-value statements: they imply intervention guidance (changing microstructure to change property) and transferability (the mechanism persists across settings). But many XAI tools only reveal model sensitivity, not causal structure. Without explicit accountability constraints, manuscripts can elevate sensitivity into mechanism, producing scientific narratives that appear rigorous but are unwarranted.

Hence, accountability must include a mechanism for claim typing: explanatory claims require explicit caveats unless causal identifiability is established, confounding risks are addressed, and the representation captures the relevant physical variables.

The core gap: The field lacks a standard that binds claims to limits

Taken together, the materials AI literature provides rich tools: benchmarks [10], large datasets [8], structural representations [4, 6], small-data insights [5], and interpretability/UQ frameworks [13–16]. Yet one organizing structure remains missing: a claim-centered reporting standard that forces papers to say:

· What is the claim type?

· What is the scope boundary?

· What evidence anchors the claim?

· What does uncertainty mean?

· What decisions are admissible and what are not?

Without such binding, materials AI papers can remain technically correct but scientifically misleading through omission. This is precisely the accountability problem: the absence of a disciplined reporting grammar for claims and limitations.

Proposed conceptual framework

The Scientific Accountability Sheet (SAS): A standard for claim legitimacy

We propose the Scientific Accountability Sheet (SAS) as a conceptual standard for reporting scientific claims and limitations in materials AI manuscripts. SAS is designed to resolve a structural mismatch in current practice: models routinely generate predictions, rankings, uncertainty estimates, and explanation signals, while manuscripts often translate

these outputs into conclusions without explicitly stating the warrant conditions that make those conclusions valid. The result is a recurrent gap between computational success and scientific legitimacy, especially when claims migrate from dataset-bound prediction into broader interpretations of generality, mechanism, or design readiness [10, 12].

SAS is not a reproduction or reformatting of existing reporting guidelines. Its novelty lies in a different organizing principle: accountability is defined by claim–limit binding, not by the volume of reported detail. SAS, therefore, functions as a standardized epistemic contract between author and reader. Under this contract, each major conclusion must be accompanied by a structured disclosure of the conditions that make it legitimate, including where it applies, what evidence anchors it, what uncertainty means, and what decisions it can responsibly support.

SAS contains five modules, each designed to block a distinct and common pathway of overclaiming in materials AI. The first module is claim typing (CT), which requires that every headline conclusion be explicitly typed as predictive, comparative, explanatory, prescriptive, or as a clearly stated combination. This separation prevents benchmark-level predictive success from being implicitly elevated to a mechanistic explanation or decision-ready design recommendations without appropriate evidential support [10, 12, 14].

The second module is the scope boundary (SB), which requires an explicit scope declaration across the materials, context, and representation scopes. The materials scope specifies the relevant chemical family or structural class to which the claim applies. Context scope specifies the processing and measurement regimes under which the target is defined and meaningful. Representation scope specifies what the model inputs encode and what they omit. This module is necessary because representation fragility and context dependence are central limitations in materials AI, particularly under sparse and heterogeneous data regimes [5–7].

The third module is the evidence anchor (EA), which requires authors to state what class of evidence warrants each claim. Benchmark metrics provide one piece of evidence and may support narrowly defined predictive conclusions within the benchmark regime [10]. Robustness or generalization evaluation provides stronger evidence when the intended use involves transfer across regimes, which remains a nontrivial problem even in large dataset ecosystems [8]. Interpretability outputs can contribute evidence only when treated as hypothesis support rather than mechanistic proof, since explanation signals often reflect correlational sensitivity rather than causal structure [14–16]. By separating evidence classes, SAS constrains the common practice of treating “good numbers” as generalized scientific warrant [12].

The fourth module is uncertainty semantics (US), which requires authors to label what their uncertainty estimates represent: measurement noise, epistemic model limitation, sparse coverage, proxy mismatch, or label ambiguity. This module is motivated by the fact that uncertainty methods differ substantially in behavior and interpretation across tasks, and uncertainty cannot constrain claim strength unless its meaning is explicitly stated [13].

The fifth module is decision admissibility (DA), which requires mapping outputs and claims to permissible decision stakes. These stakes can include screening, prioritization for validation, optimization within a declared scope, or decisions that carry deployment relevance. This mapping is necessary because materials AI outputs frequently drive costly experimental actions and strategic design trajectories, and scientific legitimacy depends not only on predictive performance but on whether limitations are disclosed in a way that prevents misuse under high-stakes conditions [8, 12]. **Table 2** summarizes SAS as a claim–limit binding contract and specifies the minimum disclosures required per module. **Figure 1** shows the SAS from materials AI outputs to warranted claims.

Table 2. The scientific accountability sheet (SAS): modules, required disclosures, and the overclaiming pathway, each module block

SAS module	What the manuscript must explicitly report (minimum fields)	The specific overclaiming pathway it blocks	Example “accountable” phrasing (journal-style)

A. Claim typing (CT)	For each headline conclusion: claim type(s) Predictive / Comparative / Explanatory / Prescriptive; identify which statements are not being made	Semantic drift: upgrading prediction → mechanism or design advice without warrant	"We make a predictive claim within the stated dataset regime; we do not claim mechanistic causality or deployment readiness."
B. Scope boundary (SB)	Materials scope (family/class, chemistry/structure range); Context scope (processing, measurement protocol, environment); Representation scope (what inputs encode vs omit)	Scope laundering: implying "general materials law" while trained on narrow, proxy-defined conditions	"Valid for oxide perovskites within the sampled chemistry under the reported measurement protocol; processing history is not represented and is outside scope."
C. Evidence anchor (EA)	Evidence class supporting each claim: benchmark metrics, robustness/shift tests, ablation/probing, external validation, domain argument; state what evidence is missing	Numbers-as-warrant: treating strong metrics as proof of general scientific validity	"Benchmark results anchor in-regime predictive validity only; no evidence is provided for transfer across processing regimes."
D. Uncertainty semantics (US)	What uncertainty represents (epistemic, aleatoric, coverage gaps, label ambiguity, proxy mismatch); calibration/interpretation statement; failure meaning	Uncertainty-as-decoration: uncertainty reported, but claim language unchanged (still expansive)	"Uncertainty here primarily reflects coverage limitation; low uncertainty should not be interpreted as mechanistic certainty."
E. Decision admissibility (DA)	Allowed decision uses: screening, prioritization, optimization within scope, deployment-relevant decisions; state inadmissible uses and stakes alignment	Decision overreach: using a dataset-bound model to justify high-stakes selection/deployment claims	"Admissible for ranking candidates for validation; inadmissible for certification-level selection without regime-specific validation."

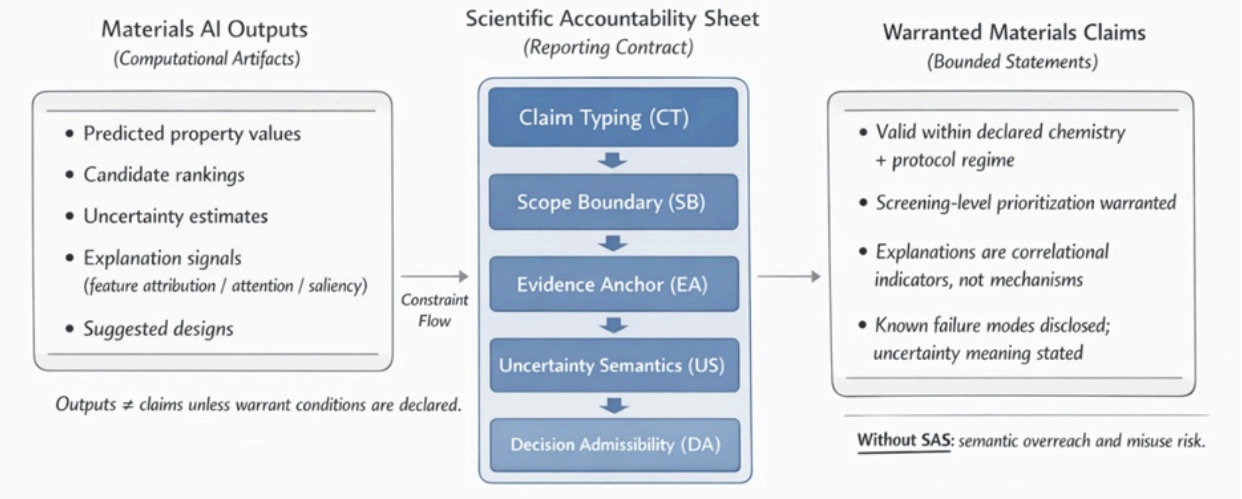

Figure 1. Scientific accountability sheet (SAS): from materials AI outputs to warranted claims

Propositions

This section formalizes the Scientific Accountability Sheet (SAS) into manuscript-grade propositions. These propositions are not implementation rules; they operate as theory-level statements about how accountability functions as an epistemic constraint on materials AI reporting. Each proposition is framed to be contestable, falsifiable in principle, and aligned with the distinctive characteristics of materials data regimes, representation choices, and decision stakes in materials informatics [1–5].

Proposition 1 — Accountability is a reporting property, not a model property

Scientific accountability in materials AI is determined by how scientific claims are written, bounded, and warranted in the manuscript, rather than by intrinsic model attributes such as architecture, accuracy, or interpretability. High-performing systems can still generate unaccountable science when generality, mechanistic validity, or design readiness are implied beyond the evaluation regime. At the same time, comparatively modest models can be fully accountable when scope limits and the meaning of uncertainty are disclosed in a binding manner [10, 12].

Proposition 2 — Most overclaiming arises from semantic drift across claim types

The dominant pathway to overclaiming in materials AI manuscripts is semantic upgrading, in which conclusions shift from predictive statements to explanatory or prescriptive language without a corresponding strengthening of evidential warrant. This drift is amplified when interpretability outputs are presented, because feature attributions and explanation maps can appear mechanistic while remaining correlational, especially under confounding and regime dependence that are common in materials datasets [14–16].

Proposition 3 — Scope-bounded validity is the primary unit of legitimacy in materials AI

In materials AI, legitimacy is most meaningfully defined at the level of scope-bounded conclusions rather than at the level of the model, dataset, or benchmark score. Since material properties depend on latent contextual variables such as processing history, defect states, microstructural distributions, and measurement protocols, validity cannot be assumed to transfer across regimes without explicit boundary conditions [5–7]. Benchmark suites improve comparability but do not eliminate regime dependence because benchmark tasks remain constrained by label semantics, sampling bias, and split design [10]. Large-scale datasets similarly demonstrate that scale does not remove generalization difficulty across chemistry and geometry [8].

Proposition 4 — Uncertainty improves accountability only when its meaning constrains conclusions

Uncertainty estimates enhance scientific accountability only when authors specify what uncertainty represents and allow that interpretation to limit the strength of reported conclusions. In materials AI, uncertainty may reflect noise, model limitations, sparse coverage, proxy mismatch, or misspecification, and different uncertainty methods behave differently across tasks. Without semantic interpretation, uncertainty becomes an accessory rather than a constraint, allowing expansive claim language to remain unchanged despite unresolved failure risk [13].

Proposition 5 — Interpretability increases accountability only when it restricts explanatory freedom

Explainability and interpretability can increase scientific accountability only when they narrow what the manuscript is permitted to conclude, rather than expand it. Interpretability tools are valuable for debugging and hypothesis formation, but explanations are often unstable under shift and reflect internal model sensitivities rather than causal structure. In materials settings with processing-driven confounding, interpretability outputs can encourage mechanistic narratives that are not warranted unless their epistemic status is explicitly bounded [14–16].

Proposition 6 — Benchmark performance is an insufficient warrant for prescriptive design conclusions

Performance improvements on benchmark prediction tasks are insufficient to justify prescriptive claims such as design recommendations or optimization decisions unless decision admissibility is explicitly mapped to scope and risk. Benchmarks such as Matbench standardize predictive evaluation [10], but prescriptive decisions require additional legitimacy conditions, including shift-aware robustness, failure regime disclosure, and uncertainty semantics aligned with

the decision stakes. Large dataset initiatives also illustrate that strong reported performance can coexist with unresolved transfer gaps across regimes [8]. At the same time, shortcuts in leakage and evaluation can inflate apparent progress without improving real-world reliability [12].

Proposition 7 — Accountability is a minimum condition for scientific reuse, not an optional ethics layer

Scientific accountability in materials AI should be treated as a minimum requirement for legitimate reuse rather than a supplementary responsible-AI gesture. As models and datasets are increasingly transferred across problems and integrated into discovery pipelines, unclear scope boundaries and unbounded claims can cause systematic misuse: models are applied outside their valid regimes, explanations are overread, and uncertainty estimates are misinterpreted. FAIR data stewardship improves accessibility and interoperability [17], and materials data infrastructures such as NOMAD operationalize FAIR principles for reuse at scale [16], but neither enforces claim discipline. SAS, therefore, complements FAIR by governing the legitimacy of claims derived from reusable data and models.

Results and Discussion

What SAS changes: from “performance narratives” to “warrant narratives”

A key contribution of the Scientific Accountability Sheet (SAS) is that it shifts the center of gravity of materials AI reporting away from performance-first storytelling and toward warrant-centered scientific communication. Standard manuscript conventions in materials informatics often privilege architectural novelty, benchmark rankings, and metric improvements as the primary evidence of value. SAS instead privileges warrant structure: a model result becomes scientifically meaningful only when the paper explicitly states what the result licenses as a claim, under what scope conditions that claim is valid, and what it does not license. This reframing directly addresses broader concerns that evaluation practices in ML-for-science can yield inflated or misleading conclusions when reporting is incomplete or when methodological vulnerabilities are not made visible to readers [17-20].

SAS also creates a shared language for comparing papers beyond accuracy. Two studies may achieve similar predictive performance, yet differ radically in the legitimacy of their implied conclusions: one may support only screening-level prioritization under a narrow regime, while another may assert mechanistic explanations or design prescriptions. Without SAS, these differences are typically negotiated implicitly through rhetorical strength rather than through explicit, reviewable reporting commitments. By requiring claim typing, scope boundaries, uncertainty semantics, and decision admissibility to be stated, SAS makes claim strength legible and therefore contestable in peer review and reuse decisions.

Distinctive failure modes in materials AI that SAS targets

SAS is motivated by recurring reporting failure modes that are particularly severe in materials informatics, where properties depend on latent context variables and where models are routinely used to guide costly actions. One major failure mode is regime collapse under processing shift. Many AI workflows for materials ignore processing history or represent it only weakly, even though processing governs defect structures, microstructural distributions, and, therefore, functional outcomes. In such settings, a model may appear reliable on a dataset while becoming invalid once processing pathways differ, even slightly, from the training regime. SAS addresses this by requiring scope boundaries to explicitly name the processing assumptions and microstructural conditions required for the claim to remain warranted [5–7].

A second failure mode is protocol-driven label instability. Material properties are not universal constants; they depend on test standards, measurement settings, sample preparation, and reporting conventions. Datasets can silently mix regimes, and models can learn protocol-specific regularities that are later interpreted as general laws. SAS responds by requiring the context scope and evidence anchor to disclose when labels are protocol-specific, when comparability is limited, and which transfers are or are not supported by the evidence.

A third failure mode is the overinterpretation of explanations. Explainable AI outputs, including feature attributions and attention-based signals, are often read as a mechanistic justification even when they reflect model sensitivity rather than causal structure. This is especially risky in materials science, where confounding by processing history and multi-scale emergence can cause strong correlations to masquerade as mechanistic drivers. SAS constrains this tendency by requiring explanatory claims to be typed explicitly and by restricting mechanistic language unless causal warrant is available and scope limitations are stated [14–16].

A fourth failure mode is the misuse of uncertainty as a trust token. Uncertainty estimates are increasingly reported as markers of responsibility, but uncertainty without semantic interpretation can be misread as a general guarantee of reliability. SAS, therefore, requires uncertainty semantics to be declared, distinguishing whether uncertainty represents noise, model limitation, coverage gaps, proxy mismatch, or other failure-relevant unknowns, and it requires mapping uncertainty to decision admissibility so that uncertainty actually constrains claims and actions [13, 21-24].

A fifth failure mode is benchmark legitimacy inflation. Benchmark success is frequently rhetorically elevated to broad scientific reliability, even though benchmark regimes may not reflect the distribution shifts encountered in real discovery and validation contexts. SAS prevents this inflation by requiring explicit scope boundaries and decision admissibility mappings that are consistent with the evaluation regime, rather than allowing benchmark performance to substitute for generalization warrant [8, 10, 20].

Relationship to existing practices: Why SAS is not redundant

SAS does not replace established values such as data stewardship, reproducibility, benchmarking, interpretability, or uncertainty quantification. Instead, it binds these practices into a coherent grammar of claim accountability. FAIR principles strengthen dataset reusability and traceability, enabling computational reuse and improved interoperability across materials infrastructures [17]. Domain-specific ecosystem efforts such as NOMAD operationalize FAIR-oriented practices for big-data materials science, expanding machine-actionable reuse and standardization at scale [16]. Benchmarks improve methodological comparability and support systematic progress tracking, making performance claims easier to evaluate within a fixed task framing [10]. Uncertainty quantification enables risk-aware ranking and supports resource allocation under limited labeling regimes, but only when uncertainty is interpreted correctly and aligned with decision stakes [13]. Explainability supports debugging and hypothesis formation, yet also introduces risks when correlational model sensitivity is promoted into mechanistic narratives without warrant [14–16].

What these practices do not guarantee, however, is that authors will refrain from upgrading claim strength without evidence. A paper can be FAIR-compliant, reproducible, benchmark-competitive, uncertainty-aware, and explainable, while still implying mechanistic or prescriptive conclusions that exceed what its evidence supports. SAS is designed precisely to close this gap by operating at the level of claims rather than tools. It functions as the missing warrant-binding layer, forcing a manuscript to specify what its outputs can responsibly justify and what they cannot [25-29].

How SAS supports peer review and scientific accumulation

If adopted, SAS enables a peer-review style closer to scientific validation logic and less dependent on informal impression management. Instead of reviewers relying on subjective judgments about whether a paper “sounds overclaimed,” SAS provides a structured basis for critique: whether claim type matches evidence type, whether scope boundaries are declared in a regime-aware manner, whether uncertainty semantics are consistent with claim language, and whether decision admissibility is appropriate for the evaluation regime shown. This shifts the assessment from rhetorical policing to warrant verification, making it harder for manuscripts to pass review based solely on persuasive language [30-32].

Over time, SAS also supports scientific accumulation by making reuse safer and more interpretable. As AI increasingly relies on transferring models across systems and integrating them into discovery pipelines, an unclear claim scope becomes a direct barrier to reliable reuse. By standardizing how limitations and decision boundaries are reported, SAS

increases the likelihood that future researchers can correctly interpret what a model or conclusion applies to, thereby reducing downstream misuse and preventing the accumulation of weakly warranted “results” in future work [33-35].

Limitations of SAS as a conceptual standard

SAS is intentionally a conceptual framework, and its limitations must be stated clearly to preserve scientific realism. First, SAS is not a substitute for good science. It cannot compensate for weak datasets, biased sampling, poor curation, or flawed experimental logic; it can only force authors to disclose the consequences of such weaknesses for claim validity. Second, SAS does not guarantee generalization. It reports boundaries and legitimacy conditions, but does not create transferability where none exists. Third, SAS is vulnerable to performative compliance. Authors could complete SAS modules superficially, treating them as formalities rather than as constraints on claim strength. This risk applies to all standards and must be managed through review norms that treat SAS as an evaluative instrument rather than a box-ticking exercise. Fourth, SAS imposes reporting overhead. It requires additional narrative discipline and structured disclosure, which may be viewed as burdensome. However, that burden is justified when materials AI claims influence costly experiments, industrial decisions, and long-term scientific directions.

Broader impacts: From responsible AI to scientifically legitimate AI

Trustworthiness in materials AI is often framed as a matter of responsible AI ethics, but SAS reframes it primarily as a scientific necessity. Accountability is required for scientific legitimacy, not only for societal responsibility. As AI becomes increasingly integrated into materials discovery pipelines and robotics-enabled or autonomous laboratory visions become more realistic, the credibility and durability of the field will depend on whether the literature clearly states what models know, what they do not know, and when they should not be used [14, 20]. In this sense, SAS is not merely a reporting proposal; it is a mechanism for protecting scientific accumulation itself by ensuring that materials AI claims remain bounded, interpretable, and decision-consistent as the field scales.

Conclusion

Materials AI is increasingly capable of generating predictions, rankings, and design suggestions that shape scientific narratives and experimental priorities. Yet the field lacks a discipline-native standard for ensuring that these outputs are reported with appropriate scientific accountability. The absence of such a standard enables semantic overreach: predictive success is rhetorically upgraded into a mechanistic understanding or prescriptive design readiness without explicit warrant conditions.

This conceptual manuscript proposed the scientific accountability sheet (SAS) as a new reporting contract for materials AI. SAS binds claims to explicit claim typing, scope boundaries, evidence anchors, uncertainty semantics, and decision admissibility. By shifting manuscript reporting from performance narratives to warrant narratives, SAS makes scientific legitimacy reviewable, comparable, and safer for reuse across regimes where materials decisions carry high cost and irreversible consequences.

Scientific progress in materials AI should not be measured solely by improvements in accuracy, but by whether the field can reliably communicate what its models warrant—and where they fail. SAS offers a structured way to make that communication a formal part of materials AI science.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 21 Nov 2023 Revised: 02 Jan 2024 Accepted: 01 Mar 2024
Published online: 18 July 2024

References

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. 2020;6:83.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2020;559:547–55.

Himanen L, Jäger MOJ, Morooka EV, Federici Canova F, Ranawat YS, Gao DZ, et al. DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247:106949.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6:138.

Wang AYT, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32(12):4954–65.

Kauwe SK, Graser J, Murdock RJ, Sparks TD. Can machine learning find extraordinary materials? npj Comput Mater. 2020;6:1–6.

Janet JP, Liu F, Nandy A, Duan C, Kulik HJ. Designing reliable and transferable machine learning models for chemical discovery. Chem Rev. 2020;120(9):987–1037.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2020;6:25.

Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2020;571:95–8.

Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. ACS Catal. 2021;11(10):6059–72.

Musil F, De S, Yang J, Campbell JE, Ceriotti M. Physics-inspired structural representations for molecules and materials. Chem Rev. 2021;121(16):9759–815.

Sun W, Zheng Y, Yang K, Zhang Q, Shah AA, Wu Z, et al. Machine learning-assisted materials discovery using failed experiments. Nature. 2021;593:399–403.

Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2021;361(6400):360–5.

Zuo Y, Chen C, Li X, Deng Z, Ong SP. Accelerating materials discovery with machine learning and robotics. Nat Rev Mater. 2021;6:5–27.

Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater. 2022;3(1):93.
https://doi.org/10.1038/s43246-022-00315-6.

Draxl C, Scheffler M. NOMAD: The FAIR concept for big data-driven materials science. MRS Bull. 2020;45(9):758–65.

Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2020;3:160018.

Batatia I, Kovács DP, Simm GN, Ortner C, Csányi G. The design space of E(3)-equivariant atom-centered interatomic potentials. J Chem Phys. 2022;157(17):174801.

Baird SG, Sparks TD. A framework for automated materials discovery in the small data regime. Patterns. 2022;3(4):100479.

Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. Patterns. 2023;4(9):100857.

Tsesmelis T, Nassar M, Reiser S, Tohidi M, Vogiatzis KD. Explainable machine learning in materials science: concepts, tools, and challenges. Patterns. 2023;4(5):100743.

Laakso J, Himanen L, Pouillon Y, Jäger MOJ, Foster AS. Updates to the DScribe library: new descriptors and derivatives. J Chem Phys. 2023;158(23):234802.

Holm EA. In defense of the black box. Science. 2022;378(6616):26–7.

Stanev V, Oses C, Kusne AG, Rodriguez E, Paglione J, Curtarolo S, et al. Machine learning modeling of superconducting critical temperature. npj Comput Mater. 2020;4:29.

Rosen AS, Fung V, Huck P, O'Donnell CT, Horton MK, Tripp MW, et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter. 2021;4(5):1578–97.

Jablonka KM, Ongari D, Moosavi SM, Smit B. Big-data science in porous materials: ML and automated workflows. Chem Rev. 2020;120(16):8066–129.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.
https://doi.org/10.1038/s41586-023-06735-9.

Chen C, Ong SP. A universal graph neural network potential for materials. Nat Comput Sci. 2022;2:718-28.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2020;4:1–17.

Yang Q, Zhang P, Zhang H, Zhang J, Chen L. Explainable artificial intelligence for materials science: A review. J Mater Inform. 2022;2:1–22.

Janet JP, Kulik HJ. Resolving dataset shift in ML for chemical discovery. Nat Mach Intell. 2021;3.

Gasteiger J, Becker F, Günnemann S. Gemnet: Universal directional graph neural networks for molecules. Adv Neural Inf Process Syst. 2021;34:6790-802.

Unke OT. SE(3)-equivariant models for atomistic ML. Nat Commun. 2021;12:7273.

Westermayr J, Marquetand P. Machine learning for electronically excited states. Chem Rev. 2021;121:9873–926.

Jha D, Ward L, Paul A, Liao W, Choudhary A, Wolverton C, et al. ElemNet: deep learning the chemistry of materials from only elemental composition. Sci Rep. 2021;11:1–9.