

Developing and Assessing Credible AI Workflows for Decarbonization Insights

Lessons for corporate users



**Data
Driven
EnviroLab**

October 2026
Silvia Landa, Asfay Raihan,
Damon Lim, and Angel Hsu
datadrivenlab.org

Executive Summary

Generative artificial intelligence (GenAI), which relies on large language models (LLMs), is a potentially game-changing technology for producing sustainability insights. Commercially available chatbots, like ChatGPT, Claude, and Gemini, provide powerful and simple interfaces for users to easily interact with technical materials for a range of corporate sustainability and decarbonization objectives. For example, these tools can answer questions like: Which emission hotspots within high-emitting commodity supply chains should a business prioritize, and how can they best navigate procurement processes and collaborate with suppliers to reduce emissions? GenAI chatbots and tools can now easily provide these answers, but are these responses accurate and reliable enough for a company to base its procurement decisions and strategy on?

This learning document outlines our lab's process for developing an AI-assisted workflow to create 12 low-carbon sourcing playbooks for procurement and sustainability teams. The playbooks detail where material-related emissions arise in these 12 commodities and what buyers can do to reduce emissions. We applied the same AI drafting instructions and review steps across commodities, while adapting the specific evidence base to generate recommendations for each commodity-specific playbook. Researchers, technical experts, and individuals with relevant supply-chain expertise reviewed the content for accuracy.

What we did

An AI-assisted workflow, utilizing two widely available commercial LLM-based chatbots, was developed to assist in the drafting of the playbooks. We compared the outputs of these two AI models (that utilized the same prompting structure) by asking technical experts and individuals with corporate sustainability experience to review them. We then selected one model to finalize the drafting process.

AI was utilized to generate the initial draft content, identify claims requiring further verification, locate supporting evidence, and identify potential errors. Researchers and expert reviewers reviewed the evidence, revised the playbook drafts, and flagged potential errors or inconsistencies to investigate further. We then corrected and/or refined the AI-generated content and assessed whether each claim was sufficiently supported with underlying evidence. Independent expert reviewers provided an additional quality-control step and made final editorial decisions. The same workflow was applied consistently across all playbooks, with an iterative review process that allowed drafts to be returned for additional research, verification, or revision before approval.

How quality was assessed

We assessed the AI-generated sourcing playbooks using a structured review process. Paragraphs and individual table rows were treated as separate review units, with enough surrounding context retained to evaluate each claim accurately. AI supported the review by helping locate relevant evidence and flag potential issues. Two researchers then independently evaluated each item, and a final reviewer resolved disagreements while also considering inputs from technical and business experts where relevant.

Each review item was assessed along three dimensions: **evidentiary support**, whether the statement was supported by evidence; **conceptual accuracy**, whether sustainability concepts were applied correctly; and **actionability**, whether the intended user could realistically act on the guidance. In this learning document, we use the term “**greenslop**” to describe AI-generated sustainability guidance that appears credible but fails one or more of these quality checks.

What we found

1. AI Added Value in Bounded, Repeatable Tasks

AI accelerated initial drafting, claim extraction, source retrieval, and preliminary assessments, particularly where the task was clearly defined and supported by accessible evidence.

2. Human Validation Remained Essential

For each review passage, AI provided an initial judgment of whether the content met our quality criteria. This initial AI judgment exactly matched final human judgment in 345 of 778 assessed instances (44.3%). In the remaining 55.7%, human evaluators modified or rejected at least part of the AI's assessment, reinforcing the need for expert validation rather than AI-only reviews.

3. Evidence Integrity Was a Greater Risk Than Fabrication

The most common errors were not outright fabrications. Instruction inconsistency, citation or source problems, and unsupported overstatements accounted for approximately 67.9% of all identified errors, while direct factual contradiction and fabrication were only identified four times each. For businesses, the greater risk is content that appears credible but fails to meet user specifications, lacks adequate sources, or overstates the available evidence. If left undetected, these issues can negatively affect downstream business decisions, including weakening supplier comparisons, undermining sustainability claims, or leading to poorly supported implementation decisions.

4. Technical Accuracy Did Not Guarantee Actionability

Even when guidance was technically credible, it was not always practically relevant. Human evaluators still had to assess whether a recommendation fits the buyer's actual decision-making authority, the user's ability to influence suppliers, and the operational realities of implementation. In addition, corporate usefulness depends on whether recommendations could realistically be acted upon in the user's decision context.

What we recommend

- **Near term: Strengthen the foundation for credible and repeatable use.**

Playbook developers and research teams should build a structured library of confirmed errors and corrections, establish clear ownership and a formal update process for each playbook, and validate revised guidance with intended procurement and sustainability users before release.

- **Longer term: Scale automation selectively, with human accountability built in.**

AI and workflow teams should use advanced or agentic AI only for well-defined tasks such as routine maintenance, evidence retrieval, and preliminary accuracy review; automatically escalate unresolved or high-risk issues to qualified experts; and periodically re-evaluate models using consistent test cases, evidence standards, and human-review criteria.

Bottom line: AI can accelerate defined parts of sustainability content development, but credible corporate use depends on traceable evidence, expert interpretation, operational review, and accountable human verification.

1. Introduction

Commodity-level decarbonization strategies require companies to account for different emissions sources, production processes, supplier capabilities, and levels of influence across the value chain. To support this process, we developed commodity-specific sourcing playbooks, designed to translate these complexities into practical guidance for procurement and sustainability teams. Each playbook provides evidence-based guidance tailored to the emissions profile, production context, and supply-chain characteristics of a specific commodity. The playbooks help users identify emissions hotspots, evaluate decarbonization levers, and prioritize actions aligned with their position and influence in the value chain.

Each playbook addresses three linked questions:

- How can a company assess material emissions across its value chain?
- What processes and factors drive those emissions?
- What practical steps can a business take, including the use of supplier-engagement tools and templates?

Twelve high-emitting commodities were examined: aluminum, cement and concrete, cotton, glass, glycerine, palm oil, paper and pulp, plastics, rubber, semiconductors, steel, and synthetic graphite.

GenAI was introduced to test whether a repeatable, AI-assisted workflow could reduce the effort required to produce and update technical guidance while helping accelerate climate action. The aim was not to replace subject-matter expertise, but to understand where AI could contribute meaningfully, where its limitations emerged, and what safeguards would be needed before its outputs could be reliably used in corporate-facing guidance.

We deliberately divided responsibilities between AI and human reviewers based on the type of task. AI supported drafting, claim extraction, evidence retrieval, and preliminary assessments, while a human reviewer remained responsible for technical interpretation, business practice, incorporating industry nuance, and final approval. This division of responsibilities reflects evidence that AI-human performance varies by task. For instance, [Vaccaro et al. \(2024\)](#) found that AI-human combinations performed on average worse than the better of humans or AI working alone. However, the outcomes varied substantially by task: performance losses were concentrated in decision-making tasks, while content-creation tasks showed significantly greater gains. This distinction informed our workflow: we used AI primarily for tasks where it could accelerate content development and evidence review, while retaining human judgment for decisions requiring technical expertise, contextual interpretation, and accountability.

2. AI-human Playbook Development Workflow

We developed the sourcing playbooks using a two-phased approach. In the trial and model-selection phase, we tested an AI-assisted workflow and evaluated different model configurations to determine how AI could best support the development of these playbooks. We then applied the selected approach consistently across the remaining commodities in a standardized development phase. Across both phases, generative AI supported structured drafting and initial evidence review, while human reviewers remained responsible for checking facts, interpreting technical information, assessing business relevance, and approving the final content.

2.1 Trial Phase and Model Selection

Before applying the process across all 12 commodities, we conducted a controlled trial to develop, test, and refine the AI-human workflow. The trial consisted of four steps:

Step 1. Finalize the playbook structure and scope.

We designed the decarbonization playbooks while considering the practical decision needs of procurement and sustainability teams. Potential users from The Climate Pledge's Science Club provided input on procurement decision-making, cross-functional coordination, and the guidance needed to move from planning to implementation. Their feedback informed several rounds of revision and resulted in a user-centered, action-first structure: users identify their position and influence in the value chain, prioritize actions available today, understand key emissions drivers and hotspots, select decarbonization levers aligned with organizational readiness and supplier influence, and apply practical tools for supplier engagement and decision-making. We also strengthened guidance for limited or inconsistent data, clarified the respective roles of procurement and sustainability teams, and incorporated examples addressing cost, trade-offs, and implementation constraints. The final playbook structure is provided in [Appendix A](#).

Step 2. Develop and iteratively refine the AI prompts.

We translated the playbook structure into a standardized prompt that could be applied across all commodities. The prompt specified a common set of requirements, including intended audience, scope, lifecycle boundaries, evidence, terminology, writing style, citations, and format, while allowing us to incorporate commodity-specific information where needed. We then tested and refined the prompt through several rounds to improve the relevance, consistency, and traceability of the generated content. Potential users also reviewed the evolving prompt structure and provided feedback. Section 3 describes the prompt-development process and key lessons in greater detail.

Step 3. Evaluation of AI models.

We used the standardized prompt and initial source materials to generate draft playbooks using two GenAI models and compared their outputs. The following sections explain how we assessed the models.

3.a. Factual assessment.

We evaluated whether factual and technical claims in the model outputs were supported by cited evidence. The initial AI-assisted factual assessment was intentionally narrow and mechanical, focusing on whether claims could be verified against the underlying sources.

Human evaluators then checked the AI assessment and classified potential issues into three main error types:

- **Factual error:** The claim is objectively incorrect based on reliable evidence.
- **Unfounded fabrication:** The claim, number, or citation cannot be verified or appears to have been invented.
- **Mathematical error:** The claim contains an incorrect calculation, comparison, conversion, or percentage.

3.b. Expert review and validation.

Technical experts and potential end-users reviewed the drafts after factual assessment. Their review addressed issues that factual checking alone could not reliably identify, including:

- **Reasoning integrity:** for example, whether the evidence on emissions hotspots was relevant to the proposed decarbonization lever and whether that lever supported the recommended action.
- **Causal discipline:** whether the draft made stronger causal claims than the evidence could support.
- **Real-world usefulness:** whether the recommendations reflected the actual level of control, sequencing, and decision points available to procurement and sustainability teams.

Experts reviewed outputs from both models and provided feedback on technical accuracy, clarity, relevance, and practical usefulness. These steps served two purposes: First, they helped us identify recurring weaknesses and determine whether additional error types or review stages were needed for rollout. Second, they informed the project-specific model-selection decision. Section 4 describes the expanded assessment method and the resulting error patterns identified.

Step 4. Select the final model and confirm the workflow.

Based on the comparative assessment and expert review, the project team selected one model for the rollout phase. During the trial, the selected model provided stronger factual grounding, more concise content, and more practical recommendations. This finding was specific to the tested models, prompts, and source materials evaluated in this project and should not be interpreted as evidence that the same model will perform best across tasks, industry context, or in the future.

The trial findings also helped us to establish the standardized development workflow for the playbooks: standardized inputs, AI draft generation, AI-assisted factual and error assessment, expert review for human validation and revision, corporate review, legal review, and final approval. Establishing these responsibilities and review points before rollout was consistent with National Institute of Standards and Technology (NIST) guidance, which emphasizes clearly defined roles, pre-deployment testing, and risk-management activities proportionate to identified risks ([NIST, 2024](#)). The trial was deliberately iterative: findings from factual assessment and expert review revised the prompt, source package, and workflow before we extended this to all 12 commodities. This process matches human-AI design guidance on testing system behavior, communicating limitations, and enabling users to identify and correct unsuitable outputs ([Amershi et al., 2019](#)).

2.2 Standardized development workflow

Following model selection, we applied a common seven-stage workflow across the commodity portfolio.

1. **Standardized inputs.** We assembled the finalized playbook structure and prompt template, as well as any commodity-specific material we identified to ground the AI-generated outputs (e.g., previous drafts or other source material).
2. **AI draft generation.** The selected model produced a structured first draft from the supplied inputs and instructions.
3. **Error assessment.** An AI-assisted process identified potential errors and claims requiring additional evidence. Human evaluators then examined that assessment and determined whether the evidence supported the draft (Section 4).
4. **Human revision.** Researchers revised the draft, corrected unsupported or inaccurate content, clarified boundaries, and improved organization and presentation.
5. **Technical and corporate review.** Technical experts assessed scientific validity, commodity-specific nuance, causal reasoning, and consistency among emissions drivers, levers, and recommended actions. Corporate reviewers assessed whether recommendations were realistic for procurement and sustainability professionals, accounted for the control available to different users, and could be implemented in real business settings.
6. **Legal review.** Legal reviewers examined the revised playbook for compliance issues, accuracy of external claims, and potential legal or communications concerns.
7. **Finalization and approval.** We addressed outstanding comments, completed editorial revisions, and prepared the approved playbook for deployment.

Although presented sequentially, the workflow was in reality iterative. Questions raised during human, technical, corporate, or legal review could return the draft to an earlier stage for source checking or revision. An AI-generated statement was not accepted simply because the model produced it or because an automated assessment classified it as supported.

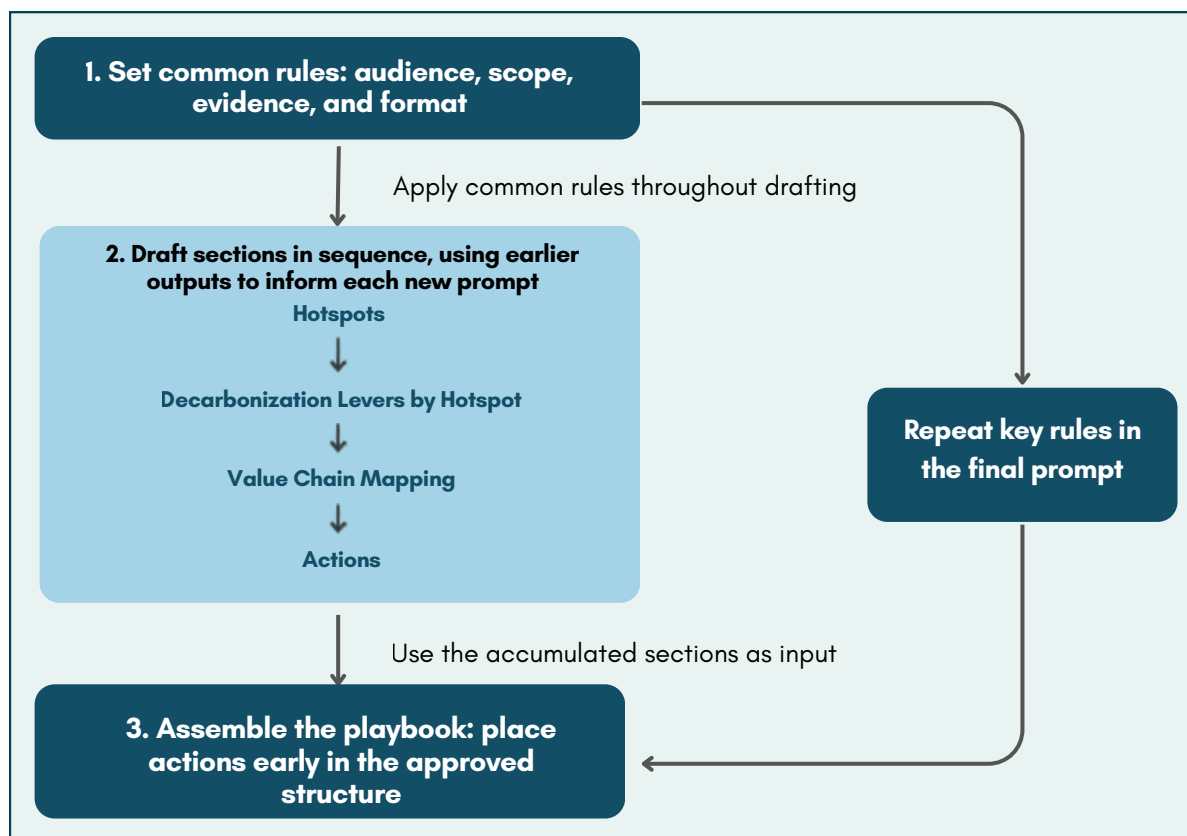
3. Prompt Development Lessons

Rather than asking an AI model to produce an entire playbook in a single, lengthy prompt, we divided the task into a sequence of smaller, connected prompts (e.g., a structured prompt chain). Each prompt focused on one section of the playbook in greater depth, and the output from that section was carried forward to inform the next step. For example, AI first identified a commodity's major emission sources, then used these findings to identify relevant decarbonization approaches and translate them into practical actions. This connected process produced more focused and reviewable outputs because each section could be assessed and revised independently without repeatedly regenerating the entire playbook. Once all sections were developed and reviewed, their outputs were brought together in a final synthesis step to produce the complete playbook draft.

We began by setting common instructions that applied across the playbook, including the intended audience, scope, evidence requirements, terminology, and writing style. We then developed and tested prompts for individual sections before combining them into the full sequence. Although the supporting technical analysis was developed first, the final playbooks were organized around the needs of a procurement or sustainability professional, with recommended actions presented first and the supporting evidence and explanation following them. Box 1 illustrates this step-by-step process.

Breaking a complex task into smaller, connected prompts through prompt chaining is consistent with existing practices, where complex tasks are divided into smaller prompt-driven steps and intermediate outputs are passed to subsequent stages ([Wu et al., 2021](#)). Our approach also followed broader prompt-design principles, including providing sufficient context, defining the required output, separating prompt components, using reusable templates, and refining prompts through repeated testing ([Geroimenko, 2025](#)). [Appendix B](#) records the resulting prompt architecture.

Box 1. How we connected the prompts



3.1 Prompt Design, Testing, and Iterative Refinement

During the trial, we tested six components of the prompt system, guided by prompt-design principles (Geroimenko, 2025).

1. Establishing Common Rules and Quality Standards.

We first established a common set of rules to guide the model through the prompt sequence. These rules defined the model's role, intended audience, writing style, terminology, tone, legal and communications boundaries, evidence requirements, citation format, units, and three user types. The rules also included consistency among emissions hotspots, decarbonization levers, and recommended actions. Testing showed that broad instructions such as "be concise" or "make the guidance actionable" were too open to interpretation and not sufficiently precise on their own. We therefore translated these general expectations into more specific, reusable global rules covering role and audience; plain, jargon-limited language; terminology, legal, evidence, citation, uncertainty, and unit standards; continuity among drivers, hotspots, levers, and actions; and recommendations tailored to three user types with different levels of control and supplier influence. Before drafting the individual sections, the model was instructed to review these general rules and confirm that it would apply them consistently in the subsequent sections. [Appendix B.2](#) lists these rules.

2. Structuring Prompts for Specific Playbook Sections.

We developed a separate prompt for each major section, such as emissions hotspots, value chain mapping, decarbonization levers, and recommended actions, rather than asking the model to generate the entire playbook in a single step. Each prompt defined the section's purpose, required headings, level of detail, evidence requirements, and output format.

For example, the emission hotspot prompt limited the output to two or three hotspots and required the model to explain their causal drivers, uncertainties, and comparison pitfalls. The lever prompt organized recommendations according to company maturity and influence over commodity-related decisions and applied them across the three user types. This modular design allowed us to identify and revise instructions for different contexts and needs without redesigning the entire workflow. It also reduced the risk of unfocused or incomplete outputs.

3. Testing continuity across the conversation.

We tested whether information generated in one section was carried forward accurately into later sections. The drafting sequence began with the technical foundation: identifying emissions hotspots and drivers, as well as decarbonization levers. Later prompts utilized these outputs to develop immediate actions and practical tools. This order was deliberate. Although the final playbook presents recommended actions near the beginning for ease of use, those actions were generated only after the underlying technical analysis had been first developed. This process helped ensure that recommendations were directly linked to the emissions hotspots and decarbonization levers identified earlier in the process.

Each prompt explicitly instructed the model to use previous responses as context and maintain consistent terminology. We reviewed whether hotspot names, emissions drivers, user types, and recommendations remained aligned as the conversation progressed.

4. Testing visual generation.

We initially tested whether the model could generate value-chain graphics directly. Most reviewers found the resulting visuals often too dense, difficult to interpret, or visibly AI-generated. We therefore removed AI-generated graphics from the workflow and playbook. The prompt instead asks the model to produce a detailed brief for a human designer, covering stage order, plain-language labels, hotspot locations, marker placement, figure explanation, footer, and legend. This retained the model's ability to organize technical information while delegating visual production to a human designer.

5. Consolidating and Validating the Final Playbook.

After drafting the individual sections, we tested a separate prompt that assembled them into the complete playbook. At this stage, web search was disabled, and the model was instructed to use only the working draft and approved structure. It could reorganize and tighten wording, but it could not add, remove, or materially change the content based on external sources.

To reduce the risk that important instructions would be overlooked in a longer prompt, the final prompt repeated the most important quality checks. It required the model to follow the approved structure, maintain consistency across sections, use clear language, provide complete citations, define abbreviations and technical terms, and remove generic or recognizably AI-like phrasing. It also specified the page limit and required the final output to serve both procurement and sustainability audiences.

6. Standardizing the Prompt Across Commodities.

We designed the prompts as reusable templates by separating stable rules from commodity-specific inputs. Placeholders such as [COMMODITY] could be replaced without changing the core workflow, audience definitions, section requirements, or final checks. This made a consistent process possible across all 12 commodities while allowing scientific evidence, hotspots, and recommended actions to vary.

3.2 How Expert and Corporate Feedback Shaped the Prompt Design

Across several rounds of expert review at the beginning of this process, before developing the playbooks, feedback on both the playbook structure and prompt design highlighted a central challenge: the playbooks needed to be both actionable and concise. We incorporated this feedback iteratively, refining the shared prompting template and playbook structure in four areas:

- First, we introduced a two-layer content structure. The main playbook became a concise, action-first guide, while technical explanations, manufacturing detail, caveats, and supporting context were moved to later sections, appendices, or companion resources. The prompt was revised so that immediate actions appear early in the final playbook even though the underlying technical evidence was generated first.
- Second, we aligned recommendations with the user's actual level of control and influence. Earlier drafts sometimes recommended actions more appropriate for manufacturers than for downstream buyers or designers. The revised prompt

distinguished among organizations that purchase a commodity directly, those that purchase products containing it, and those that have stronger supplier relationships. It also clarified the respective roles of procurement and sustainability teams, identified collaboration points, and indicated when suppliers or other stakeholders needed to be involved in a proposed recommendation.

- Third, we strengthened clarity and consistency requirements. Reviewers called for simpler headings, less jargon, clearer commodity boundaries, consistent units, and definitions for acronyms and technical terms. The prompt was also tightened to limit unsupported use cases and reduce generic or recognizably AI-generated language.

- Fourth, we narrowed the substantive scope to decision-relevant content. Some early drafts placed excessive emphasis on frameworks such as the Greenhouse Gas Protocol, Environmental Product Declarations, and the Science Based Targets initiative. The revised prompt included these frameworks only when they directly supported a procurement or decarbonization decision. The objective was to keep the playbooks focused on helping users make meaningful action, rather than serving as a general guide or reference.

3.3 What We Learned About Human-AI Communication

A central lesson from this project was that effective prompting for a complex business product requires more than clear instruction. The prompt must operate as a structured specification, defining the audience, purpose, evidence requirements, scope, terminology, decision context, responsibilities, and expected outputs. See Box 2 for an example of the initial prompt setup. For procurement and sustainability applications, these elements need to be explicit enough to guide the model consistently and provide human evaluators with a clear basis for assessing whether the resulting guidance is well supported, internally consistent, relevant to the user's role, and suitable for practical decision-making.

We also found that some design tensions require structural solutions rather than better wording alone. The need to be both comprehensive and concise was addressed by separating action-oriented guidance from supporting technical detail and by generating the technical evidence first, section by section, before assembling the final playbook. This helped preserve the connection between emissions drivers, decarbonization levers, and recommended actions while keeping the final product focused and practical for corporate users.

Finally, continuity and consistency could not be assumed across a long prompt sequence. Critical requirements had to be carried forward explicitly through stable templates, clearly defined commodity-specific inputs, and repeated quality checks in the final production prompt. These controls improved the relevance and consistency of AI-generated drafts but they did not remove the need for human review. Researchers, technical experts, and corporate reviewers remained essential for validating evidence, interpreting technical nuance, assessing feasibility and decision context, and ensuring that the guidance reflected how procurement and sustainability teams actually function. Section 4 sets out how that review was structured and what it found.

Box 2. Initial Prompt Configuration

AI-Human Workflow

Step 0. Rules

Human task

Do not ask content questions yet. Run the AI prompt below using extended thinking.

AI prompt

Global rules for all steps

Use these rules for every section unless a later step says otherwise.

Role: Write as a domain expert in [COMMODITY] manufacturing, climate science, and procurement-oriented decarbonization.

Audience: Procurement and Sustainability Officers.

Writing style: Use simple but accurate language. Keep the tone professional, natural, and human. Be succinct without leaving out important information. Avoid unnecessary repetition. Make the content actionable and approachable, not overly technical.

Terminology: Spell out each term in full the first time before using an abbreviation. Define specialized terms when they first appear. Avoid unexplained shorthand such as Tier 1 or Tier 2, or explain clearly what it means. Assume the reader may not know technical [COMMODITY], emissions, or supply chain jargon.

Tone: Avoid generic phrases such as “it is important to note” or “in today’s world.” Use direct, practical language and write as if advising a colleague, not writing a report.

Legal and communications guardrails: Avoid broad, absolute, or promotional sustainability claims. Use precise, evidence-bounded language. Do not advocate for expanded, stricter, or new regulatory mandates. Present regulations only as market, reporting, or procurement context. Do not provide legal advice, compliance instructions, or interpretations of regulatory requirements. Use neutral, descriptive language where wording

could sound like legal guidance.

Evidence quality: Prioritize recent, authoritative, and checkable sources. Use web search when content is requested. Explicitly acknowledge uncertainty where relevant. Avoid claims of environmental superiority unless supported by comparative peer-reviewed evidence. Keep units consistent throughout: kg CO₂e/kg material or t CO₂e/t material. Round carbon intensity values to 1 to 2 significant figures unless greater precision is necessary.

Citations: Every factual or technical claim must have a linked citation. Use inline links, not footnotes. Do not cite general statements. In the final document, keep inline citations in the body and include a full APA-style reference list at the end.

Consistency: Maintain consistent terminology across sections. Do not redefine the same concept differently. Ensure alignment between hotspots, levers, and actions.

Context continuity: Use previous answers in this conversation as context and ensure alignment with earlier sections.

User types:

- A. I buy this commodity directly: some control, limited influence.
- B. I buy products that include this commodity: low control, rely on suppliers.
- C. I have strong supplier relationships: higher influence, can influence.

Output: Provide one sentence confirming that you will wait for a specific section and will follow all rules for future responses.

Step 1. Structure

Human task

Upload "Structure Final" and "[COMMODITY]_PRINT." Do not ask content questions yet. Run the AI prompt below using extended thinking.

AI prompt

Study the uploaded file "Structure Final." Use "[COMMODITY]_PRINT" as the primary source for tone, framing, and relevant draft content unless later instructions require otherwise. Do not produce content, summaries, recommendations, or drafted sections yet. Only confirm that you understand the structure and will use it when I request specific sections. Do not cite anything yet and do not infer missing sections. Follow the structure exactly when asked later.

Output: Provide one sentence confirming that you understand "Structure Final" and will follow it later.

Step 2. Initial Setup**Human task**

State the commodity and audience below. Run the AI prompt below using extended thinking and web search.

AI prompt

I am developing a decarbonization playbook for:

Commodity: [COMMODITY]

Audience: Procurement and Sustainability Officers

Do not generate any section content yet. Wait until I provide a specific section. For all future sections, follow the legal and communications guardrails, evidence quality rules, terminology rules, writing style requirements, user types A, B, and C, and consistent use of units.

Output: Provide one sentence confirming that you will wait for a specific section and will follow all rules for future responses.

4. Evidence Quality and Error Patterns

We assessed the quality of evidence to evaluate whether claims in the AI-generated sourcing playbooks were supported and faithfully represented, whether sustainability concepts and boundaries were valid, and whether recommendations fit the intended decision context. AI produced a preliminary assessment record, while human reviewers retained responsibility for final coding and revision.

4.1 AI-assisted evidence assessment

We used a Python-based review pipeline to organize and support the evidence assessment. The pipeline treated each paragraph or table row as a separate review unit while preserving enough surrounding context to interpret the claim accurately. It identified claims requiring verification and compared them with the drafting instructions and available sources. When a cited source did not resolve a claim, the process searched for additional relevant evidence. AI then suggested whether the item contained an error and which type of error might apply. Appendix C provides the technical details.

The pipeline produced a structured preliminary record for human review, with each unit classified as Error, No Error, or Potential issue requiring evaluator judgment. It was instructed not to invent evidence, quotations, titles, locations, or links. Visual content required rendered review where text extraction could not preserve key information. All AI-generated assessments remained preliminary until validated by human evaluators.

4.2 Human review and revision process

Two researchers serving as independent human evaluators examined the original unit, supporting evidence, preliminary AI assessment, proposed error codes, and relevant expert comments. A final evaluator compared their judgments and adjudicated unresolved cases. Technical and corporate expert input was entered at three points: initial coding, final adjudication, and playbook revision. We used expert comments to guide the error assessment, but we still required and validated evidence before recording an error.

The retained record included original text, preliminary assessment, coder judgments, expert comments, supporting evidence, adjudicated codes, and the resulting revision. Figure 1 summarizes the process, and Appendix D provides the error coding rules and workflow.

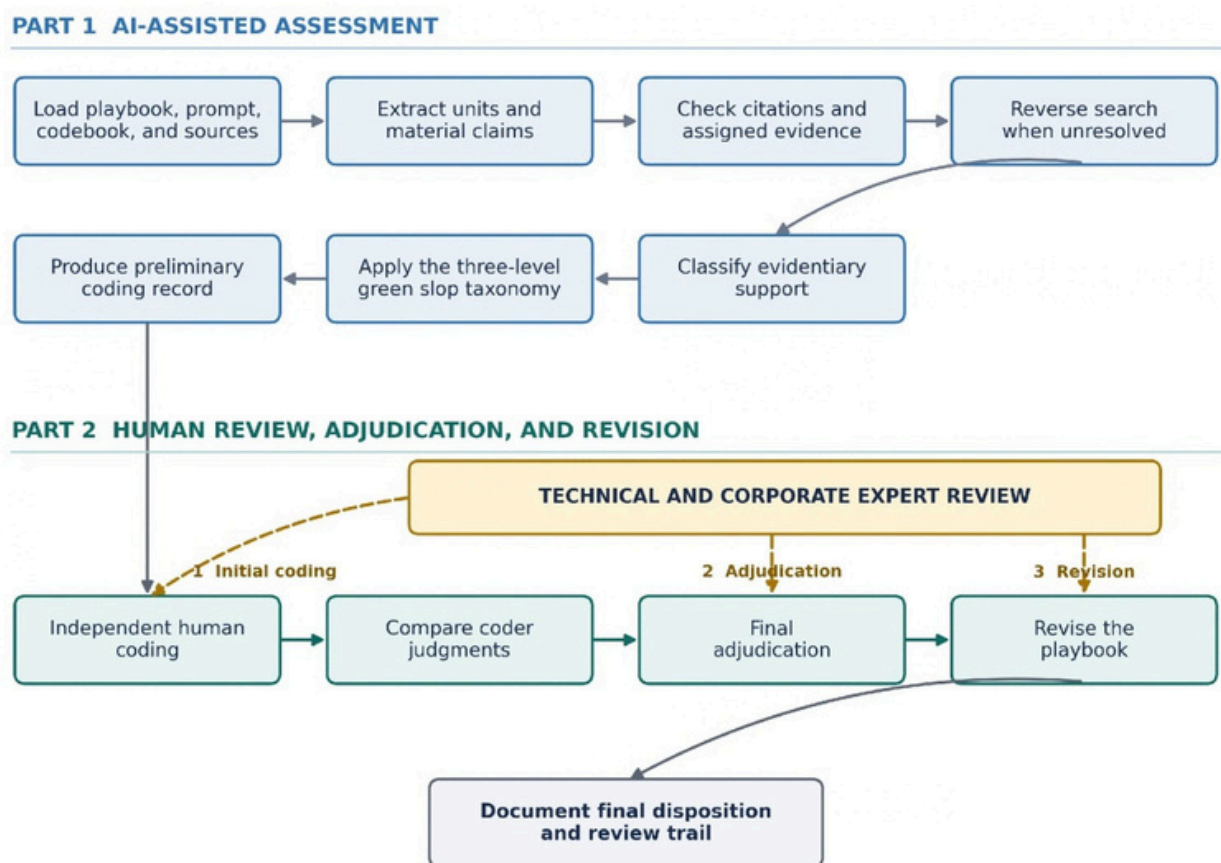


Figure 1. AI-assisted assessment, human review, adjudication, and the three points at which technical and corporate expert input informed the process.

4.3 A three-level greenslop taxonomy

Conventionally, human reviewers check for AI hallucination by looking for fabrication, factual contradiction, internal inconsistency, or unfaithful sourcing. These tests are necessary but insufficient for sustainability guidance. We use the term “greenslop” as a project-specific working term for AI-generated guidance that appears credible but fails one or more tests of evidence integrity, sustainability-domain validity, or decision-context utility.

Table 1. Greenslop taxonomy

Level	Assessment focus	Core question
1	Evidence, accuracy, and faithfulness	Is the claim correct, traceable, faithful to instructions and sources, and internally consistent?
2	Sustainability-domain validity	Are sustainability concepts, accounting boundaries, standards, protocols, and methods used correctly?
3	Decision context and implementation quality	Does guidance fit the user’s authority, prerequisites, risks, safeguards, sequence, and decision?

Review items could be classified as “No Error” when no problem was identified. When an item contained more than one distinct problem, evaluators could assign multiple error codes.

[Appendix E](#) provides the complete working codebook.

4.4 Observed error patterns

Distribution of final error assignments

We reviewed 778 paragraphs and table rows from AI-generated sourcing playbook drafts covering ten commodities. We found at least one error in 282 items and no errors under our review rules in 496 items. Some paragraphs had more than one type of error, producing 371 error-code assignments in total. Figure 2 displays these error assignments and the number of times they were recorded. [Appendix E](#) provides the complete adjudicated dataset.

Level 1: Evidence, Accuracy, and Faithfulness

Level 1 was the most prevalent error category, with 280 assignments, representing about 75.5% of all error-code assignments. The most common errors were instruction inconsistency (124), source inconsistency or citation misattribution (74), and embellishment or unsupported intensification (54). Together, these three categories accounted for 67.9% of all assignments.

Direct factual contradiction (4), factual fabrication (4), quantitative errors (2), temporal errors (5), and geographic errors (1) were much less common.

These results show that the main AI-generated errors were not fabricated facts. More often, AI failed to follow evidence requirements, made claims that were not fully supported by the cited source, or used stronger language than the evidence justified. Table 2 provides some examples.

Level 2: Sustainability-Domain Validity

Level 2 errors were recorded 36 times, accounting for 9.7% of all error-code assignments. Accounting or boundary logic errors were the most frequent category (20), followed by sector conceptual errors (15) and standards, protocol, and method compliance errors (1).

These errors differed from Level 1 because the issue was not primarily whether evidence was cited or utilized properly, but whether the output applied sustainability concepts and sector-specific logic correctly. As illustrated in Table 2, an output could therefore appear technically plausible while using incompatible units, inappropriate boundaries, or an incorrect interpretation of a sector standard.

Level 3: Decision Context and Implementation Quality

Level 3 errors were recorded 55 times, accounting for 14.8% of all error-code assignments. Generic or poorly contextualized answers (25) and risk and safeguard errors (24) accounted for most Level 3 assignments, followed by actor authority or control errors (5) and sequencing errors (1).

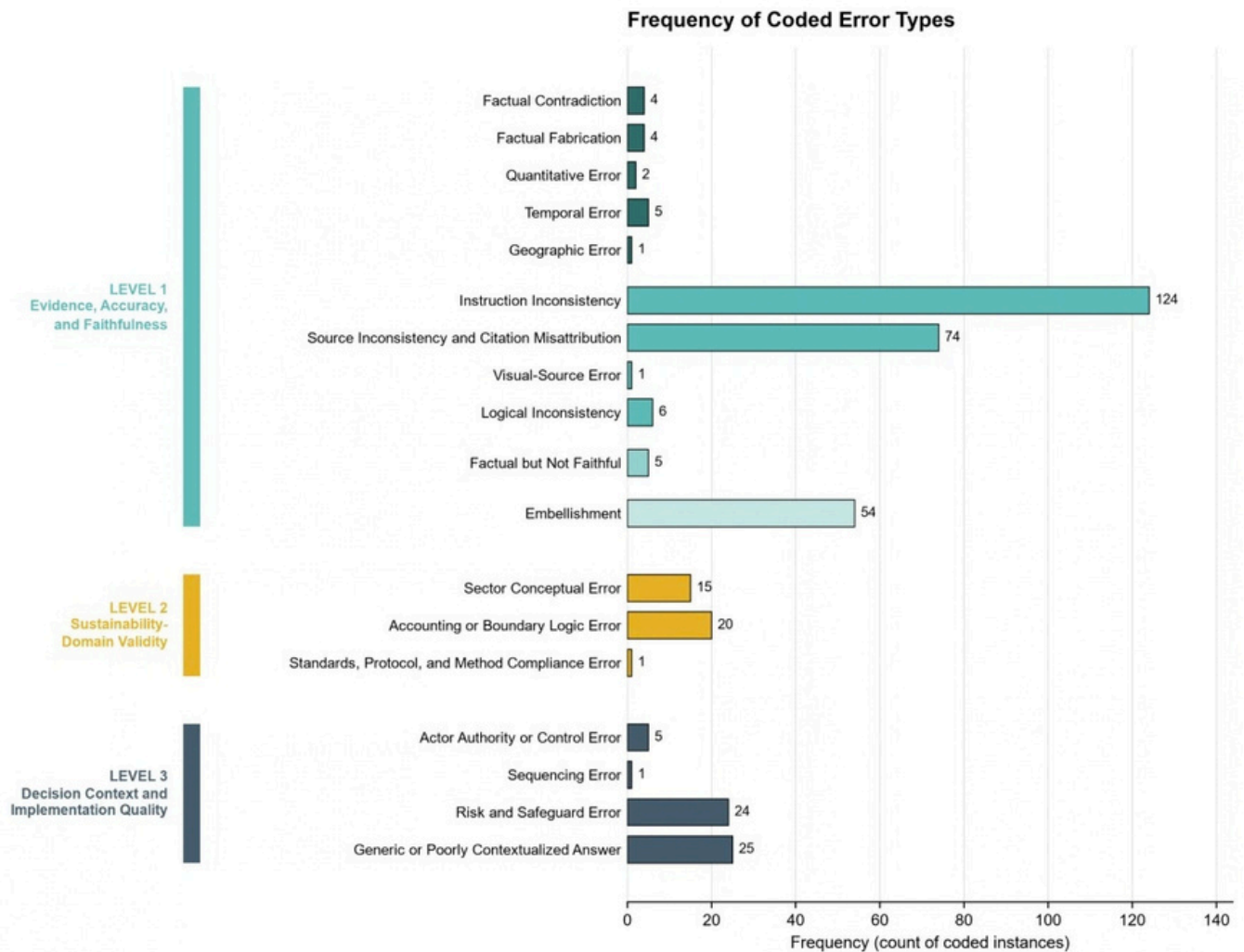


Figure 2. Frequency of final adjudicated error-code assignments. The chart reports 371 assignments across 778 assessed instances. A total of 282 instances received at least one error code, and an individual instance could receive multiple codes. The error categories are therefore not mutually exclusive at the instance level. Instances classified as “No Error” are not included in the frequency bars.

These errors generally involved recommendations that could appear reasonable in principle but are actually incomplete or impractical when considered in the context of actual implementation. Recurring problems included insufficient consideration of operating conditions, omitted safeguards or prerequisites, recommendations directed to actors without adequate leverage, and actions proposed before the information or approvals required for implementation were in place.

Table 2. Representative examples of error patterns identified in AI-generated commodity playbooks

Level and error type	What the error involved	Example
Level 1: Factual contradiction	AI-generated content includes a checkable claim that is contradicted by established facts or reliable external evidence, independent of the sources provided to or retrieved by the system.	Synthetic Graphite: The playbook referred to "lithium-ion phosphate (LFP)." LFP stands for lithium iron phosphate (LiFePO_4), not lithium-ion phosphate.
Level 1: Instruction inconsistency	AI-generated content fails to follow an explicit, clear, and permissible user instruction.	Cotton: The playbook stated that Better Cotton had "19% lower emissions intensity" without the required citation, making the claim difficult to trace to supporting evidence.
Level 1: Source inconsistency and citation misattribution	AI-generated content misrepresents, overstates, reverses, or wrongly attributes source evidence explicitly provided in the output.	Palm oil: The playbook attributed a 98% reduction in midstream emissions from biogas capture to Subramaniam et al. (2010), although the cited source did not support or include this figure.
Level 1: Logical inconsistency	AI-generated content contradicts itself or reaches a conclusion that does not follow from its own premises.	Glass: The playbook stated, "No external actors are required. This phase relies solely on direct communication between procurement teams and existing suppliers." The two statements were inconsistent about the actors required for implementation.
Level 1: Embellishment and unsupported intensification	AI-generated content that introduce exaggeration, sentiment, or bias. This content might also express a finding with greater certainty, magnitude, or generality than the evidence supports.	Aluminium: The playbook stated that "Procurement teams can drastically cut supply chain GHG emissions simply by shifting volume toward suppliers operating in renewable-powered grids..." The terms "drastically" and "simply" overstated both the expected magnitude of the emissions reduction and the ease of implementing the sourcing change.

Level and error type	What the error involved	Example
Level 2: Sector conceptual error	AI-generated content misuses or conflates sustainability, environmental, climate, procurement, or institutional concepts.	Glass: The playbook asked whether an individual facility had a Science Based Targets initiative (SBTi)-validated target, although SBTi validates targets submitted by companies and financial institutions, not individual facilities.
Level 3: Generic or poorly contextualized answer	AI-generated content fails to use decision-relevant context provided or clearly established by the prompt, assigned sources, or task context, resulting in advice that does not adequately address the stated decision need.	Glass: The playbook recommended progressively requiring minimum levels of verified post-consumer cullet without considering differences in cullet availability, quality requirements, contamination limits, manufacturing technology, or local collection systems.
Level 3: Risk and safeguard error	AI-generated content recommends an action while omitting or materially understating a risk, uncertainty, evidence limitation, or safeguard that was requested or is necessary for responsible implementation.	Paper and pulp: The playbook recommended reducing packaging basis weight by a fixed percentage each year without specifying a baseline, measurement method, functional-equivalence test, performance safeguard, or exception process.

Across the three error levels, our findings show that weaknesses in AI-generated sustainability guidance extend beyond factual correctness. Errors also arose from how evidence was used, how sustainability concepts were applied, and whether recommendations were sufficiently qualified and contextualized for practical decision-making. Low or zero frequencies should not be interpreted as evidence that an error type is unimportant or infrequent outside of this project. These results describe the composition of confirmed errors in this selected set of playbooks and do not measure their severity, business impact, or prevalence in other AI-generated sustainability content.

Human agreement with the preliminary AI assessment

Preliminary AI assessment and final human judgment were in full agreement for 345 of 778 instances (44.3%). Partial agreement occurred in 87 instances (11.2%), while no agreement occurred in 346 instances (44.5%). Human adjudicators modified or rejected at least part of the preliminary AI assessment in 433 instances, representing 56% of the dataset.

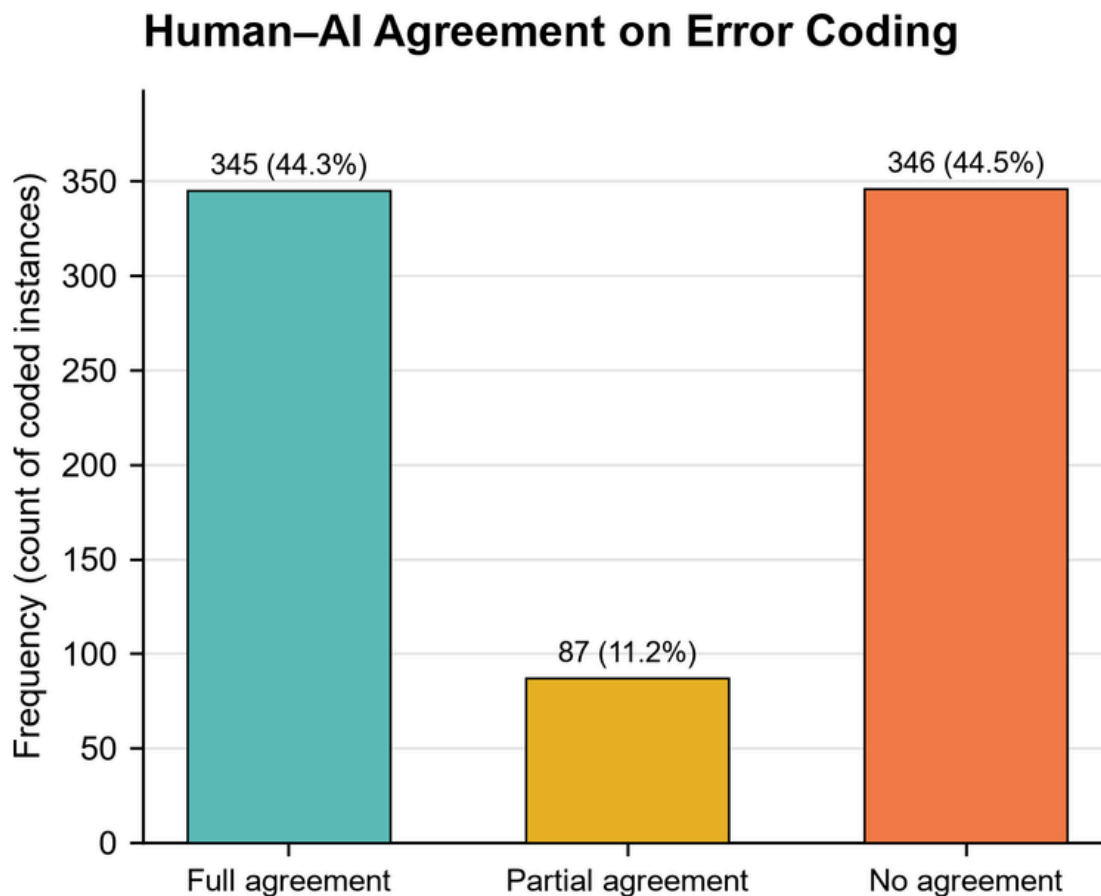


Figure 3. Agreement between the preliminary AI assessment and final human-adjudicated coding, $n = 778$ assessed instances.

Full agreement required the AI model and human evaluators to reach the same “Error” or “No Error” judgment, and when an error was detected, assign the same error code/type. “Partial agreement” indicated that the human reviewer agreed with at least one AI-suggested error code but disagreed with at least one other AI-identified classification and/or identified additional error types that the AI had missed. “No agreement” indicated that human reviewer reversed the “Error” or “No Error” judgment or retained none of the AI-proposed codes.

The 44.3% full-agreement rate should not be interpreted as a general measure of AI accuracy, especially for other domains and contexts. This figure is based on this specific dataset, coding protocol, model configuration, and exact-match definition. The findings suggest that AI can assist with preliminary screening and documentation, but human review remains necessary for verifying evidence, interpreting sector-specific concepts, assessing implementation risks, and determining whether recommendations are suitable for their intended business context.

5. Corporate Implications and Conditions for Credible AI Use

Key finding

AI added value in bounded, repeatable tasks such as drafting, evidence retrieval, claim screening, and preliminary error assessment, but it could not reliably perform end-to-end validation. Human evaluators modified or rejected the preliminary assessment in 433 instances (or in 56% of evaluated instances), particularly when decisions required judgment about source quality, evidence nuance, sustainability boundaries, business context, or implementation feasibility. The findings therefore support the use of AI as part of a structured human-review process, rather than as a substitute for human validation or final approval.

5.1 Strengths and limitations of source verification

In our project, AI was effective at extracting claims, locating supporting evidence, comparing claims with sources, flagging weak support, and recording preliminary findings consistently. Its limitations became more apparent when judging source quality, determining whether a claim extended beyond the evidence, or recognizing when differences in geography, technology, units, boundaries, or time periods made a comparison misleading. The results show that credible-sounding guidance can still misrepresent evidence, exceed the intended scope, or omit important context.

5.2 Speed, consistency, and scalability

AI accelerated the initial drafting, claim extraction, source retrieval, and standardized preliminary assessments for the development of the playbooks. A common prompt architecture also improved consistency across commodities. However, these efficiencies shifted substantial effort and time into verification, correction, expert interpretation, and editorial review. Based on the project team's estimate, each claim-level error assessment required, on average, about three minutes from each of two first-stage evaluators and one minute from the final evaluator. Across 778 assessed claims, this represents approximately 5,446 minutes, or 91 person-hours, of human review. This estimate excludes time spent setting up the review process, resolving complex cases, obtaining expert input, revising the playbooks, and completing editorial review. Because we did not systematically track these additional activities or compare the process of developing these playbooks with one conducted without AI, we cannot determine whether AI reduced the total development time.

5.3 Why technical accuracy did not ensure actionability

A technically credible decarbonization lever may still be impractical for a specific organization. Actionability depended on the company's position in the value chain, decision authority, supplier influence, organizational readiness, commercial feasibility, operating context, and planning horizon. Corporate reviewers therefore distinguished between actions companies could take directly and those they could influence through supplier engagement, purchasing requirements, contracts, collaboration, or longer-term market development. They also helped prioritize immediate actions and separate essential decision guidance from supporting technical detail.

5.4 Safeguards for Credible Corporate AI Use

Avoiding greenslop requires more than a generic human reviewer to verify content at the final step. [Hsu et al. \(2024\)](#) identified a human cognitive risk: expert reviewers may prefer more detailed AI-generated answers even when those answers are less factually accurate. [Green \(2022\)](#) argues that human-oversight policies can create a false sense of security when the people assigned to review algorithmic systems cannot perform that role effectively. Rather than relying on human oversight alone, Green proposes institutional oversight that requires organizations to justify both the use of an algorithm and the effectiveness of any proposed human-review mechanisms before any adoption of AI.

The seven safeguards below combine findings from this project, relevant literature, and an internal project report summarizing interviews conducted in the summer of 2024 with 17 corporate professionals whose work involved AI. The interviews were conducted by a Data-Driven EnviroLab partner at the Ackerman Center for Excellence in Sustainability (ACES), Kenan-Flagler Business School, University of North Carolina at Chapel Hill.

Table 3. Safeguards for credible corporate AI use

No	Safeguard	What credible use requires	What this project did
1	Documentation and traceability	<i>A recommendation should be traceable back to its source, reviewers' decision, and correction.</i> Interview participants emphasized the importance of source traceability and trustworthy input data so that AI-generated guidance can be linked to credible evidence and users can assess the reliability of its sources. This aligns with algorithmic-auditing approaches that emphasize documenting decisions and evidence throughout the AI development lifecycle to strengthen transparency and accountability (Raji et al., 2020).	We retained source links and excerpts, preliminary assessments, coder judgments, adjudicated codes, expert comments, and resulting revisions. Prompts and the codebook remained project records.
2	Pilot testing and ongoing monitoring	<i>Any AI workflow should be tested before rollout, then followed by the monitoring of workflow performance and collection of feedback.</i> Interview participants emphasized pilot and stress testing before wider implementation, including testing under realistic conditions, collecting user feedback, and identifying problems before broader use. NIST (2024) likewise recommends pre-deployment testing, evaluation under conditions similar to deployment, ongoing monitoring, and incorporation of post-deployment feedback.	We compared two models in a trial, refined section and end-to-end prompts, and piloted factual assessment. Continuous post-publication monitoring was not established and remains a gap.
3	Technical expert review	<i>Qualified reviewers are needed to provide domain-specific expertise and evaluation.</i> Interview participants emphasized technical expert review to question outputs, verify supporting evidence, assess technical accuracy, and provide quality assurance where automated checks are insufficient. Climate-domain research similarly supports combining evidence retrieval and automated evaluation with expert	We included technical experts to assess scientific validity, commodity evidence, boundaries, and causal reasoning. Their input informed initial coding, adjudication of unresolved error cases, and playbook revision.

No	Safeguard	What credible use requires	What this project did
		assessment of whether generated claims are supported and faithful to the underlying evidence (Hsu et al., 2024 ; Leippold et al., 2025 ; Thulke et al., 2025).	
4	Practical implementation review	<i>Check whether the intended buyer can take the action and identify the costs, supplier cooperation, and operational constraints involved.</i> Interview participants emphasized the need for business and implementation review before acting on AI-generated guidance, particularly to assess feasibility, cost, integration with existing systems, practical consequences, actionability. This aligns with research showing that fluent and technically plausible LLM outputs may still be unsuitable for real-world decisions when they fail to reflect professional standards, local context, implementation constraints, or workflow requirements. Evidence from legal, climate adaptation, and clinical contexts show that effective review must therefore go beyond factual accuracy to assess whether AI-generated guidance is contextually appropriate, practically usable, and safe to apply in decision-making (Dahl et al., 2024 ; Hager et al., 2024 ; Xie et al., 2025).	We worked with potential corporate users and asked them to review the structure and prompts as well as assess whether the actions proposed were appropriate given procurement authority, supplier influence, organizational maturity, feasibility, sequence, and planning horizon.
5	Legal and compliance review	<i>Refer legal, regulatory, compliance, privacy, cybersecurity, and reporting issues to the appropriate specialists for review and approval.</i> Interview participants emphasized the need for specialized review or approval when AI use may involve legal, regulatory, compliance, privacy, data-security, or ethical risks. They also highlighted	We included legal review as a formal approval step before finalization. We also adjusted rules in our prompts to prohibit legal advice and broad or promotional sustainability claims.

No	Safeguard	What credible use requires	What this project did
		safeguards such as approved-tool lists, restrictions on public AI systems, internal governance processes, and security requirements. This aligns with Raji et al. (2020) , who included a documented, multidisciplinary review body in their framework for internal algorithmic auditing. Their framework supports formal specialist review and recorded approval, although it focuses mainly on ethical and social-impact concerns.	
6	Clear decision ownership	<i>Identify who reviews, who approves, and who is responsible for updates.</i> Interview participants emphasized clear human ownership and accountability, with AI serving as a supporting tool while identifiable people remain responsible for questioning outputs, deciding how guidance should be used, and making final decisions. Wieringa's (2020) systematic review identifies explicit actors, roles, levels, and responsibility boundaries as central to algorithmic accountability across lifecycle stages.	We included researchers (as error evaluators), technical experts, corporate reviewers, legal reviewers, and final approvers. AI assessments remained preliminary and were not published without human review and oversight.
7	Risk-tiered review	<i>Apply stronger review where an error could materially affect spending, public claims, or other consequential decisions.</i> Interview participants emphasized that review intensity should match the risk and sensitivity of the use case. Routine applications may require less scrutiny, while uses involving financial, legal, public, or reputational consequences require stronger review	We escalated unresolved or sensitive claims through assigned evidence review, reverse searches, expert adjudication, and legal review. A complete enterprise risk-tiering framework was not implemented.

No	Safeguard	What credible use requires	What this project did
		and due diligence. NIST (2024) similarly recommends oversight and risk-management activities scaled to organizational risk tolerance and the priority of identified risks.	

6. Recommendations for Future Playbook Development

The recommendations below follow directly from the evidence-quality findings and corporate implications discussed in Sections 4 and 5. The immediate priority is to make the process of developing such playbooks more reliable, repeatable, and easier to maintain before expanding the role of automation. In the near term, organizations can use the lessons from this project to strengthen processes, clarify review standards, and define accountability. Over the longer term, they can expand automation selectively, focusing on tasks where the evidence, governance, and human-review systems are robust enough to support it.

6.1 Near-term system improvements

Build a structured correction and decision library

To benchmark and improve AI-assisted workflows over time, organizations could maintain a reusable library of confirmed errors and corrected statements. Rather than simply recording what went wrong, the library should preserve enough context to show why a correction was needed and how the issue was resolved. Early entries could focus on recurring problems such as source inconsistency, failure to follow instructions, or recommendations that lack sufficient contextualization.

The value of the library would grow as it becomes a reference for improving prompts, aligning reviewers, and testing model performance. However, corrections should not be transferred automatically from one issue domain to another. Before reusing an example, reviewers should confirm that the underlying evidence and decision context are sufficiently comparable. Where possible, each case should also indicate the stage of the workflow where the problem is best addressed.

Establish clear ownership and a formal update process

Any AI-assisted workflow needs clear ownership and a defined process to ensure that guidance is relevant given current the science. Each playbook should have an accountable owner, a documented evidence cutoff date, and a regular review cycle, with additional updates when important changes affect the underlying evidence or recommendations. These changes might include new research, shifts in technology or regulation, user feedback, or material changes to the AI workflow itself. Updates should also be documented in a way that preserves traceability, including what changed, why the change occurred, and who approved the revision.

Validate playbooks with intended users before release

Revised playbooks should also be tested with the people expected to use them in practice. Procurement and sustainability professionals can help assess whether the guidance is clear, relevant to their decision context, and realistic to act on. User testing should focus on whether readers can identify the actions that matter most to them, understand the evidence and uncertainty behind those actions, and distinguish between areas they can directly control and those they can only influence.

6.2 Longer-term opportunities

Explore agentic AI only for bounded maintenance and first-layer review tasks

As workflows mature, more advanced or agentic AI could take on a larger role in routine monitoring and first-layer quality review. For example, it may help identify outdated information, surface inconsistencies, and prepare issues for further review. Its role, however, should remain one of support rather than final judgment. Cases involving uncertain evidence, conflicting interpretations, or higher-risk decisions should still be routed to the relevant human experts. Responsibility for validation, interpretation, approval, and publication should remain clearly with human reviewers.

Re-evaluate tools as models change

Because AI models will continue to evolve, the tools used in any AI-assisted workflow should be re-evaluated periodically against a stable set of test cases. The focus should be on whether a model produces traceable, reliable, and usable outputs with a reasonable level of correction and expert review. While the ultimate goal may not be to automate as much as possible, re-evaluation can identify where newer models genuinely improve the quality or efficiency of the overall workflow.

7. Conclusion

This learning document traced the development of a trusted AI-assisted workflow for producing sustainability and decarbonization guidance from start to finish: generating initial content, checking claims against evidence, and incorporating expert review. In doing so, we highlight the promise and pitfalls for avoiding “greenslop” when using GenAI for sustainability.

We find that AI’s promise is clearest when tasks are well defined. GenAI adds value in initial drafting, claim extraction, evidence retrieval, and standardized preliminary reviews, while a common prompt architecture improves consistency across commodities. These benefits are real but task-specific.

Although generative AI saved time on some tasks, those gains were partly offset by the effort required to validate its outputs. This became especially clear during review: in 433 of 778 assessed instances (56%), human reviewers modified or rejected at least part of the preliminary AI assessment. Humans “in the loop” of the AI workflow contributed different forms of expertise, from interpreting technical evidence and system boundaries to assessing whether recommendations were feasible, appropriate, and clearly communicated. Legal review provided an additional check on sensitive claims. The AI-assisted workflow therefore did not replace expert judgment but supported specific tasks, while human reviewers retained responsibility for interpretation and decision-making.

The broader lesson is that trusted AI use depends not simply on the model but on the workflow around it. Traceable evidence, clear quality standards, domain-specific review, decision-context relevance, and accountable human approval all shape whether the resulting AI guidance would be credible and useful. AI can help reduce a team’s workload when its assigned tasks are clearly specified and evidence is sufficiently stable. Where evidence is fragmented or contested, or where the stakes are high, AI’s role should be to support qualified human judgment rather than replace it.

For sustainability guidance in particular, AI should not be utilized merely to generate content that sounds convincing. AI should be used to produce guidance that is well supported, technically sound, relevant to the user’s decision context, and practical to act on. A structured AI-human workflow can help achieve this while reducing the risk of credible-sounding but unreliable or impractical outputs.

References

- Amershi, S., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., & Bennett, P. N. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64–93. <https://doi.org/10.1093/jla/laae003>
- Geroimenko, V. (2025). *The essential guide to prompt engineering: key principles, techniques, challenges, and security risks*. Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-86206-9>
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681. <https://doi.org/10.1016/j.clsr.2022.105681>
- Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9), 2613–2622. <https://doi.org/10.1038/s41591-024-03097-1>
- Hsu, A., Laney, M., Zhang, J., Many, D., & Farczadi, L. (2024). Evaluating ChatNetZero, an LLM-Chatbot to Demystify Climate Pledges. *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, 82–92. <https://doi.org/10.18653/v1/2024.climatenlp-1.6>
- Leippold, M., Vaghefi, S. A., Stambach, D., Muccione, V., Bingler, J., Ni, J., Senni, C. C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Luterbacher, J., & Huggel, C. (2025). Automated fact-checking of climate claims with large language models. *Npj Climate Action*, 4(1), 17. <https://doi.org/10.1038/s44168-025-00215-8>
- NIST, G. M. (2024). *Artificial intelligence risk management framework: generative artificial intelligence profile*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
- Thulke, D., Kemmler, J., Dugast, C., & Ney, H. (2025). Listen to the Context: Towards Faithful Large Language Models for Retrieval Augmented Generation on Climate Questions. *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change*.
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. <https://doi.org/10.1145/3351095.3372833>
- Wu, T., Terry, M., & Cai, C. J. (2021). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *ArXiv*. <https://doi.org/10.48550/arxiv.2110.01691>
- Xie, Y., Jiang, B., Mallick, T., Bergerson, J., Hutchison, J. K., Verner, D. R., Branham, J., Alexander, M. R., Ross, R. B., Feng, Y., Levy, L.-A., Su, W. J., & Taylor, C. J. (2025). MARSHA: multi-agent RAG system for hazard adaptation. *Npj Climate Action*, 4(1), 70. <https://doi.org/10.1038/s44168-025-00254-1>

Appendices

[Appendix A. Final playbook structure](#)

[Appendix B. Prompts for playbook development](#)

[Appendix C. Code structure and technical specifications](#)

[Appendix D. Error coding rules and workflow](#)

[Appendix E. Complete working codebook](#)

[Appendix F. Complete error dataset](#)

Suggested Citaition

Landa, S., Raihan, A., Lim, D., & Hsu, A. (Oct 2026). Developing and Assessing Credible AI Workflows for Decarbonization Insights: Lessons for corporate users. Chapel Hill, NC: Data-Driven EnviroLab.



**Data
Driven
EnviroLab**

October 2026
Silvia Landa, Asfay Raihan,
Damon Lim, and Angel Hsu
Designed by Eunji Lee
datadrivenlab.org