Online Resource 1: System Conformance with Credibility Criteria

Shaena Ulissi, Andrew Dumit, P. James Joyce, Krishna Rao, Steven Watson, Sangwon Suh

1 Introduction

To serve sustainability, AI has to be built for sustainability. We can't just force AI onto a climate problem or ask an off-the-shelf large language model (LLM) for GHG data and blindly trust the results. The AI system needs to be encoded with sustainability intelligence, data and methodologies. Its outputs and intermediate steps need to be benchmarked against environmental studies and for sensitivity and reproducibility; calculations need to be checked for correctness; and assumptions, citations, and quality measures need to be output for further review and data quality improvements.

That said, the guidance has not kept up with the technology. No existing studies that we are aware of propose consistent and feasible evaluation criteria that carbon footprinting providers can use when implementing and verifying AI-based systems. The Product Footprints AI systems (Auto-Mapper and Advanced AI Model) are built with these criteria as guiding principles, and we are encouraging other organizations, standards-setters, and industry associations to build on these criteria to establish a high bar for AI quality in product footprinting.

This section describes how our models meet each of the criteria. As there is very little ground truth data for cradle-to-gate GHG emissions, Watershed experts curate benchmarking datasets and evaluation data and make more deterministic "unit evaluations" in cases where they are feasible. While LLMs rarely achieve 100% accuracy and human experts often disagree on the correct approach for LCA, the goal is to achieve useful, defensible results that are at least as good as a typical human—and to build the structure in a way that we can continually monitor and improve the model performance by adding more data, guidance, human expertise, and upgrading as the underlying technology improves. As the saying goes, "All models are wrong, but some are useful."

2 Criteria for Credible AI-Assisted Footprinting Systems

The Auto-Mapper is an example of the mapping case (Case 1). As such, it targets the criteria in Tables 1 and 2 of the paper. The advanced AI model is an example of the lifecycle modeling case (Case 2) and additionally targets the criteria in Tables 3 and 4. For each criterion, we repeat the summary from the paper and respond with how Watershed's models align. As shown below, our AI-CF system is rigorously developed and evaluated, with clear benchmarks and paths to continue to improve and measure quality.

When developing test cases and running evaluations, Watershed climate and data scientists consider the environmental implications of excess AI use. We strive for the smallest useful dataset size and testing frequency that allows us to track improvements and notice regressions without wasting energy and resources. As we continue to improve the modeling performance or expand into untested areas, we augment the tests with more test cases or additional evaluations.

3 Case 1: Auto-Mapper Criteria

3.1 Criterion 1.1: Benchmark Performance

Criterion: AI-generated mappings should demonstrate their ability to map to defensible choices relative to expert mappings when tested against a validation set of products with established "ground truth" mappings.

Details from paper: The precise benchmark result will depend on the complexity of the products and available datasets to map to, so we do not propose a specific target for this metric here—but if a labeled dataset were to be made available, it could be used to set a common target. For example, mappings where there are exact matches should likely achieve near 100% benchmark performance because there is high agreement among experts and clear choices. If experts systematically map incorrectly, this benchmark may also be less than 100%. Complex plastic materials or chemicals mapped to a data source like ecoinvent should also achieve a high benchmark performance, but a result less than 100% is likely acceptable as there will be less agreement among experts as to the best and defensible mappings. A system with a lower performance may be acceptable if automated mappings are only used for small parts that are not material to a given company's purchased materials. To demonstrate this performance, the model provider should prepare documentation, visuals, or calculations that evaluate the performance at this mapping step. The performance evaluations should be repeated whenever the model undergoes a major update, to ensure there are no regressions in quality.

What we do: We maintain labeled validation sets and expert-reviewed mappings. We compare the AI-selected activity against expert ground truth and score whether the mapper has chosen a defensible mapping. We also deduct points if the mapper chooses a particularly bad mapping. We run the full mapper tests weekly and rerun them before merging any major decision points to the mapper prompt or underlying LLM, to ensure quality improves or does not degrade.

Performance: As of January 2026, the Auto-Mapper was evaluated on $N = 1198$ test cases spanning external benchmarks and production data (curated representative samples and real customer materials). For non-vague inputs, the system achieved 91.2% defensible mapping rate overall, with performance varying by category: 98.9% on the Amazon Parakeet external benchmark and 85.9% on production data. For vague inputs (ambiguous or incomplete material descriptions), performance dropped significantly to 59.7%, demonstrating that input data quality is a primary driver of mapping accuracy. When coupled with the quality score that accurately flags nearly all indefensible mappings as "low" quality, this suffices to reduce human review burden to consist only of the materials that are most relevant to the GHG analysis. While the eventual goal is to approach 100%, we recognize that i) human experts generally perform worse than our auto-mapper, ii) there is ambiguity about what constitutes a defensible mapping as it depends on the use case and context, and iii) we provide the match quality scores (described more elsewhere in this article) to properly flag any material matches that may require additional human review or revisions.

Dataset: We have curated an expert-labeled dataset that includes input of material name, description, and supplier industry. The expected output is one or more ecoinvent activities that are defensible proxy mappings for the given input. There are also banned substrings for which the mapper should absolutely never map to; for example, mapping "silicon" to "silicone" would be extremely problematic. These test cases and labels were developed manually with Watershed's experienced climate scientists. We developed this dataset by first creating different test categories on which we wanted to understand the model's performance—what we term "unit evaluations" and

describe further in Criteria 1.2—and then by expanding with a more representative set of anonymized materials that Watershed customers have uploaded. We further expanded it by including Amazon's published dataset, to which we corrected and expanded the labels.

Table S1: Auto-Mapper Performance Summary

Test Category	**N**	**Non-vague Performance**	**Vague Input Performance**
Amazon Parakeet (ext. benchmark)	281	98.9% ($n = 275$)	83.3% ($n = 6$)
Production Data	704	85.9% ($n = 574$)	55.4% ($n = 130$)

3.2 Criterion 1.2: Consistency Across Material Types and Input Formats

Criterion: Mapping performance should remain stable across different material industry categories and levels of complexity. If it is not stable, this should be disclosed and caveated appropriately if the model is used for materials that are outside its domain expertise.

Paper details: A distribution of benchmark performance by industry classification may be used to demonstrate this consistency. Additional considerations that arise from actual company material data may be used to generate additional benchmarks and evaluations. For example, company material data is often provided with abbreviations or industry-specific terms and acronyms, in different languages, or with misspellings.

What we do: We maintain labeled validation sets and expert-reviewed mappings. We compare the AI-selected activity against expert ground truth and score whether the mapper has chosen a defensible mapping. We also deduct points if the mapper chooses a particularly bad mapping. We run the full mapper tests weekly and rerun them before merging any major decision points to the mapper prompt or underlying LLM, to ensure quality improves or does not degrade.

Performance: Current performance across different input formats shows the model properly accounts for abbreviations, synonyms, grade information, close sounding but different items, percentages, special characters, and accounts for industry information ($> 90\%$). This makes sense since foundational LLMs are trained are large amounts of data and are familiar with all of this type of text. On the industry-specific evaluations, performance is largely consistent across categories. The more ambiguous categories—for example, where there is not one clear ecoinvent match such as with processed metals; or where there is a hierarchy decision needed for whether to map to most important components or most representative components of, say a catalyst—show some variability. We supplement any such ambiguous mappings with our separate quality review scoring (described further below) which highlights materials that may need human review or revisions.

Dataset: We have curated an expert-labeled dataset that includes input of material name, description, and supplier industry. The expected output is one or more ecoinvent activities that are defensible proxy mappings for the given input. There are also banned substrings for which the mapper should absolutely never map to; for example, mapping "silicon" to "silicone" would be extremely problematic. These test cases and labels were developed manually with Watershed's experienced climate scientists. We developed this dataset by first creating different test categories on which we wanted to understand the model's performance—what we term "unit evaluations"—which evaluate different material types and input formats as required by this criteria. We further expanded it by including Amazon's published dataset to which we corrected and expanded the labels.

As of January 2026, this dataset includes the following categories and number of test cases (Table S2).

Table S2: Auto-Mapper performance by test category (Jan 2026)

Test category	**What it tests**	**N**	**Performance**
amazon_parakeet	Representative performance on an externally published dataset (largely food)	281	99%
production	Production data (curated samples and real customer materials)	704	80%
packaging	Packaging materials	78	91%
catalysts	Catalysts (small quantities, high emissions)	40	100%
composite	Lightly processed materials (material + process)	14	93%
identity	Input materials labeled with ecoinvent activities	13	100%
abbreviation	Abbreviations in material inputs	9	100%
grade	Material grade affects activity selection	8	88%
synonyms	Synonyms for the library material	8	100%
org_context	Organization context affects mapping	8	100%
flash_context	Additional context affects mapping	8	88%
multi_ingredient_food	Multi-ingredient food products	6	83%
earlier_mistakes	Previously flagged performance issues	5	100%
close_sounding_but_different	Materials that sound similar but differ	3	100%
special_characters	Special characters in input	3	100%
percentage	Percentages affecting activity selection	2	100%
recycled	Recycled materials	2	100%
Total		**1198**	**87%**

3.3 Criterion 1.3: Appropriate Granularity and Proxy Approach

Criterion: Where exact matches aren't available, the systems should apply appropriate proxy selection logic that aligns with LCA best practices. The system should map to the most specific appropriate activity dataset rather than defaulting to overly broad categories.

Paper details: Carbon accounting and materials mapping have no universally accepted playbook; practitioners improvise with ad-hoc heuristics, making results hard to reproduce or audit. By adapting a clear, step-by-step method for turning messy material descriptions into consistent mappings and codifying these rules, the resulting approach can be consistent, explainable, and audit-ready. This process framework and example test cases of how the system follows the framework should be provided to assist verifiers, auditors, or reviewers in understanding the system.

What we do: We have provided our model with the approach to take for proxy mapping, including criteria to follow. The labeled expected outputs described in criteria 1.1 and 1.2 are a result of experts following these same criteria, which is how we evaluate performance. In the step after the mapped material is selected, the match quality indicator model is provided with material match and emissions match criteria upon which to evaluate the quality of the match. The December 2025 mapper criteria include the following:

- Material Type and Properties Accuracy (Critical)
- Emissions Similarity (Critical)
- Form Factor and Physical State

- Degree of fabrication
- Specificity Alignment
- Plausibility and Constraints
- Common-Sense Material Category Matching
- Material Composition Factors

3.4 Criterion 1.4: Repeatability

Criterion: Small variations in product descriptions or specifications should not result in dramatically different mappings. In addition, the same material should generally be mapped consistently.

Paper details: Given the non-deterministic nature of AI/ML methods, this metric will rarely be 100% consistent, but it should be high enough to be useful for decision makers and at least as consistent as human experts. This can be demonstrated with repeatability evaluations that consider test cases across various material types, where the same material or similar materials are mapped multiple times. Metrics include majority-pick rate, normalized entropy, and the number of unique predictions per material across multiple runs.

What we do: We maintain labeled validation sets and expert-reviewed mappings. Watershed runs the mapper 10 times for each input material. We then compare the results of the 10 mappings to assess the consistency of the selected material(s). We run the repeatability tests before merging any major decision points to the mapper prompt or underlying LLM, to ensure repeatability improves or does not degrade.

Performance: The performance aligns well with scientific expectations, given the non-deterministic behavior of LLMs. Repeatability correlates with match quality; where match quality is low, repeatability is low, and where match quality is high, repeatability is high. Where there is a single unambiguous match, repeatability approaches 100%.

Dataset: Watershed's data and climate scientists curated a repeatability dataset with subcategories that measure performance across different types of mappings. The input includes material name, description, and supplier industry. The categories that we are interested in performance include expert-selected ambiguous sectors, identity (e.g., if told what material to map to, does the mapper choose that material), and anonymized customer materials split between low and high match quality scores.

Table S3 demonstrates the correlation between input ambiguity and output entropy: identity inputs (exact matches to ecoinvent activity names) yield perfectly repeatable outputs (entropy = 0.00), while low quality score inputs (ambiguous descriptions) yield high entropy (0.36), indicating the AI system correctly signals uncertainty when input data is ambiguous.

Table S3: Repeatability performance by category (Jan 2026)

Test category	N	Repeatability score	Normalized entropy	Unique predictions
Identity	106	1.00	0.00	1.00
High quality score	101	0.91	0.24	1.43
Low quality score	51	0.83	0.36	2.00

3.5 Criterion 1.5: Match Quality Indication (material-level)

Criterion: The system should display to the user the uncertainty level and highlight poor matches between the input material and the mapped proxy dataset; these would be places where, if material, the user should attempt to augment their data input or move to Case 2 carbon footprinting.

Paper details: Many types of uncertainty affect GHG analyses. For proxy mapping to a single dataset, a qualitative uncertainty framework such as DQI can suffice to inform users of relative quality. A key additional uncertainty arises from match quality between the dataset and input material, which is often overlooked in LCA but can strongly impact emissions representation.

What we do: A key feature of Watershed's proxy mapping is the material match score and emissions match scores that are surfaced in product. We run a second model that judges the quality of each selected mapping for each given input material and input context, against a specified rubric. Other sections of this article provide further guidance on when low scores are acceptable and how to take action to improve the quality of mappings.

Rubric: The match quality evaluation is based on rubrics developed by Watershed's data and climate scientists. The two aspects—emissions match and material match—are meant to assist with different use cases for mapping: GHG inventory accuracy/completeness for reporting, and procurement, respectively. The rubric includes the following considerations as of December 2025:

Material Match Analysis:

- Material Type and Properties Accuracy (critical)
- Form Factor and Physical State
- Processing State and Functionality
- Specificity Alignment
- Plausibility and Constraints

GHG Emissions Match Analysis:

- Material Composition Factors
- Processing and Manufacturing
- Degree of fabrication

3.6 Criterion 1.6: Mapping Methodology Documentation (system-level)

Criterion: Explanation of how the AI system interprets product characteristics and selects appropriate activity datasets.

Paper details: Provide a process framework and example test cases so verifiers and auditors can understand and review the system's mapping logic and its consistent application.

What we do: The mapping system is documented in the main paper. The input data including the details and descriptions fields are passed to the model. The model then selects an appropriate proxy material. The quality reviewer assesses the quality of that matching. All of these results are passed back to the user.

3.7 Criterion 1.7: Data Source Identification

Criterion: Documentation of the activity-based datasets used (e.g., ecoinvent version) and any supplementary data sources.

Paper details: Clearly identify dataset names and versions used in mappings so reviewers can reproduce results and understand scope and updates over time.

What we do: Every selected mapping is clearly described with its source. This is included at the graph-level and within the citations for corporate footprint syncing.

3.8 Criterion 1.8: Version Control

Criterion: Tracking of mapping algorithm versions and activity dataset versions to ensure reproducibility.

Paper details: Maintain version metadata for both the AI system and the data libraries so that outputs can be tied back to specific configurations and re-run as needed after updates.

What we do: Watershed uses a standard software engineering git-based workflow, wherein each change to the code is committed to a branch, reviewed, and then merged; and prior changes can be reopened, reverted, or simply viewed/audited via git. In addition, each version of our material library is saved with a timestamp and can be revisited if needed. As described in Criterion 1.7 the provenance of any given material mapping is provided at the graph-level and citation-level.

3.9 Criterion 1.9: Decision Logic Transparency

Criterion: Visibility into key decision points and thresholds used in the mapping process.

Paper details: Provide insight into how candidate activities are ruled in or out and what thresholds or heuristics govern selection, so reviewers can assess logic quality.

What we do: The mapping selection process, including the criteria and quality judgment rubrics, are described above. Each selected material includes text descriptions of the mapping quality that can be reviewed.

3.10 Criterion 1.10: Audit Trail

Criterion: Ability to trace how specific product characteristics led to particular activity dataset selections. Given the non-deterministic character of LLM-based systems, this audit trail may be at the bulk level or on example datasets rather than specific to each individual product or material.

Paper details: Provide page-level or sample-level traces that show inputs, intermediate candidates, reasoning highlights, and final selection to support reviewer understanding even when item-level determinism is not feasible.

What we do: In addition to the match quality logic that is user-facing, there are additional data available that could be provided on a case-by-case basis for verification, if needed. Watershed uses an LLM tracing system that includes records for all tool calls and decisions. The amount of data is overwhelming to read and not always as useful as the summary, which is why it is not all passed through to the users. However, it is used frequently as part of the data and climate science development and improvement of our modeling pipeline.

3.11 Criterion 1.11: Continuous Improvement

Criterion: Description of mechanisms for incorporating feedback to improve mapping accuracy over time.

Paper details: Explain how evaluation results, expert reviews, and user feedback are fed back into the system to reduce errors and address regressions across releases.

What we do: Watershed reruns the mapper benchmarking evaluations weekly and uses them to evaluate the potential performance gains or regressions associated with model or context changes. A few recent examples of this are as follows:

- Alpha customers that purchase packaging materials noticed that there were material + process combinations that they thought may be a better fit than the materials selected by Watershed's auto-mapper. We added new composite materials to the library, updated evaluation labels to account for them, and upon verifying performance promoted these updates to the generally available version of the auto-mapper.
- When evaluating against the Amazon Parakeet dataset, Watershed scientists noticed that the mapper often mapped meat products (e.g., packaged beef) to non-meat activities (e.g., tofu) because of logic within the criteria that made the level of processing appear more relevant than the common-sense background material similarity. In response, we revised the prompt criteria and reevaluated performance. Performance improved substantially, and we then merged the revised criteria.

3.12 Criterion 1.12: Position in GHGP Hierarchy

Criterion: Clear positioning of this approach within the GHG Protocol Scope 3.1 data quality hierarchy, demonstrating improvement over broad sector averages while acknowledging that Cases 2 or 3 are needed to move further up the hierarchy.

Paper details: State explicitly where proxy mapping sits in the hierarchy, how it compares to spend-based and supplier primary data, and when to advance to automated modeling (Case 2) or standards-compliant pathways (Case 3).

What we do: Document that Case 1 improves over broad sector averages and when users should move to Case 2 or supplier PCFs. Auto-mapping from Case 1 falls into the "average-data method" of the GHG Protocol hierarchy.

4 Case 2: Advanced Modeling System Criteria

4.1 Criterion 2.1: Reasonable Precision and Benchmarking Across Components

Criterion: Component-level and end-to-end benchmarking and evaluations, where feasible. There is rarely ground truth GHG emissions data available. Additionally, experts do not always agree on how to generate a system model, LCI, or LCA for a given product. However, benchmarking is still useful across many aspects of assumptions and calculations to demonstrate rigor and demonstrate system improvements.

Paper details: Carbon footprints and the underlying activity estimates and energy and processing components can be benchmarked against datasets. Full carbon footprints can be compared to results from published EPDs, licensed data such as ecoinvent, and literature values such as The Carbon Catalogue. Targets may be framed with median error and P90 error where verified PCFs exist, paired with checks to avoid "right for wrong reasons."

What we do: Watershed has developed a full suite of component-level and end-to-end benchmarking and evaluations to improve modeling performance. These include component-level checks on process energy, yields, transport; end-to-end comparisons vs EPDs and literature; guardrails on unit conversions, allocation decisions, and mass balance. For each benchmark, we establish a model performance baseline and in some cases, a target against a metric such as a reasonable output of a PCF, where we set a goal that 90% of the AI generated output fits in this range.

The end-to-end benchmarking is rerun regularly and before every major model or system change, while some of the component datasets are run in a targeted manner when there are changes that either are targeting improvements in their areas or might inadvertently affect their performance. Other transient evaluation datasets are developed as needed for specific feature testing, such as formula evaluations, mass to unit conversion validity, and industry code classifications.

Performance: As of January 2026, most graphs that are produced by Watershed's advanced AI model meet benchmarks for total emissions, top 3 emissions hotspots, energy input rates, and mapping. Performance has continued to improve as our experts improve the context, structure, and internal validation steps of the models.

In end-to-end benchmarking against 269 EPDs, the Advanced Modeling System achieved a median relative error of 33%. The 80% confidence interval for relative error spans [6%, 98%], with 86% of test cases falling within 80% of the benchmark value. Analysis of prediction bias shows the system is approximately unbiased at the median (median signed error: −2%), with symmetric over- and under-estimation (48% and 52% of predictions, respectively). This lack of systematic bias is desirable for decision-support applications where both over- and under-counting carry risks. The most extreme over-prediction occurs in cases with complex allocation issues (e.g., cashmere) and cases with limited publicly available data to inform the modeling assumptions (e.g., specialty chemical manufacturing). In addition, the current system has a limitation that it is unable to implement recycled content for certain alloys within ecoinvent; therefore, residual errors remain regardless of the quality of the rest of the model in cases where this content is relevant. For context, EPDs for identical products can themselves vary by over 50% under different but equally compliant modeling assumptions, owing to combined differences in reference service life, allocation methods, electricity mix, and background databases (Gelowitz and McArthur, 2017; Konradsen et al., 2024). In the context of screening-level LCA—where process variance can span orders of magnitude—this precision is sufficient for strategic hotspot identification.

Additional performance statistics are provided as follows.

The Predicted vs Expected plot (Figure S2) shows the error in kgCO2e across our test cases. It demonstrates that there is not systematic error that varies by magnitude of emissions and that the distribution by product categories is broad.

The signed relative error by category (Figure S3) demonstrates that there is some variation in error by product category but that the median error is of a consistent magnitude.

The score heatmap by category (Figure S4) shows the Advanced Model graph reviewer's pass rates across the validation categories and product categories. This demonstrates that there is no apparent

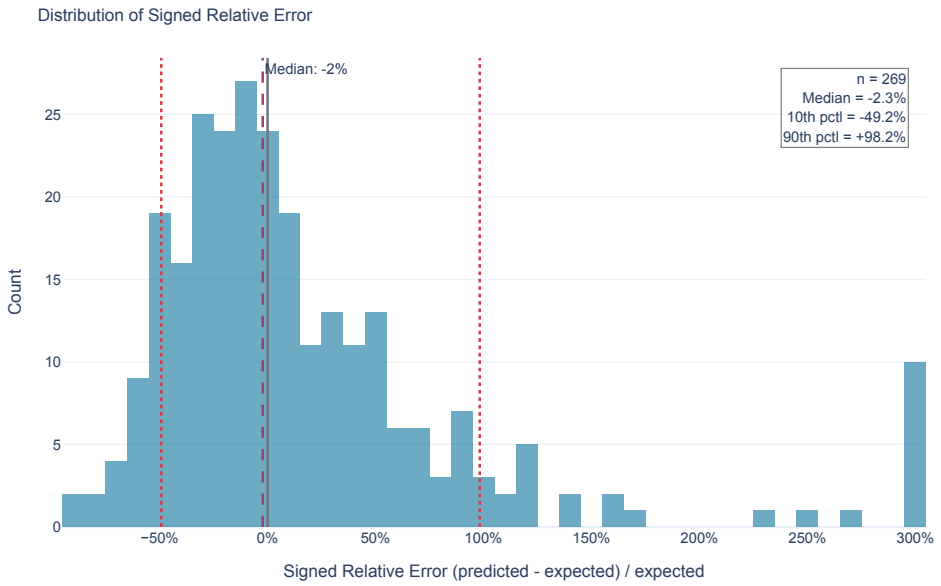

Figure S1: Advanced Model end-to-end results: Signed relative error distribution

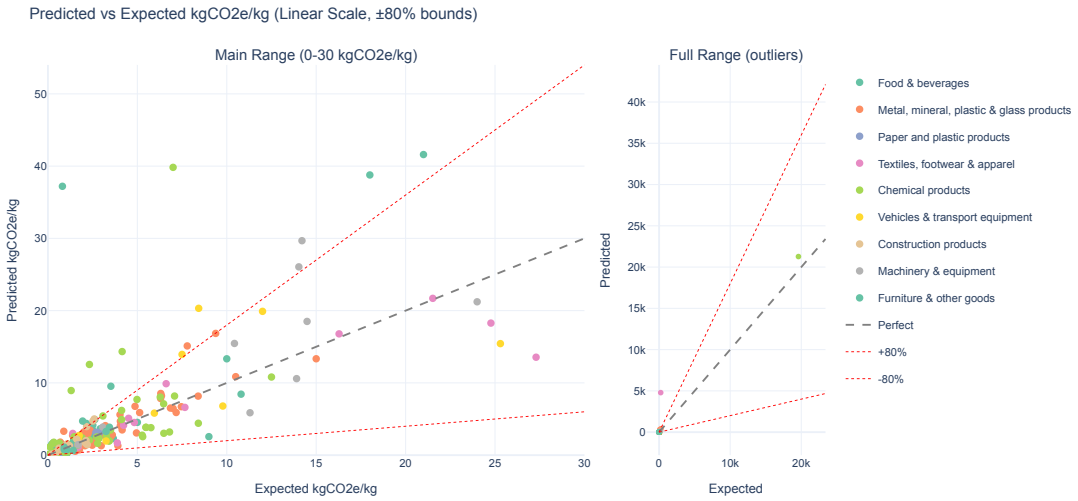

Figure S2: Predicted vs Expected kgCO2e/kg

bias between product categories; that overall a high proportion of checks pass—thus reducing the review burden of these models—and that there is continued room for improvement in model system development, particularly for input rates where much data is proprietary.

Benchmarks are useful to provide guidance on where to focus model improvement efforts and how to assess reasonableness. However, because methodological choices and choice of background database influence the results substantially, they are insufficient as sole metrics. Because there is not "ground truth" GHG emissions data for full supply chains, we focus on usefulness and reasonability (as well as transparency). The goal here is to develop one reasonable model, and then track improvements.

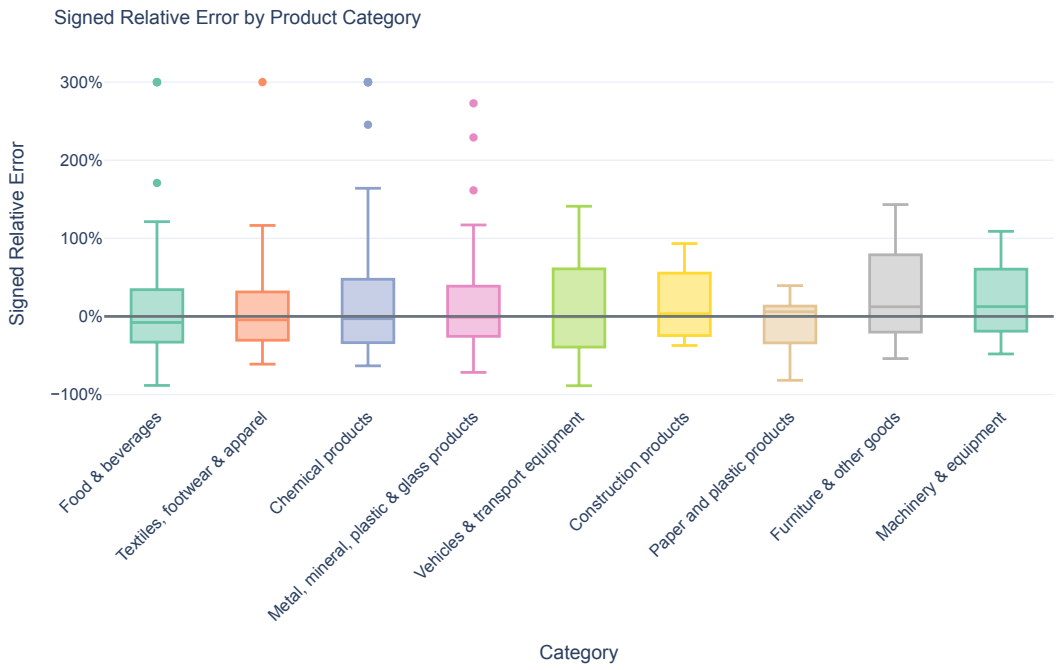

Figure S3: Signed relative error by category

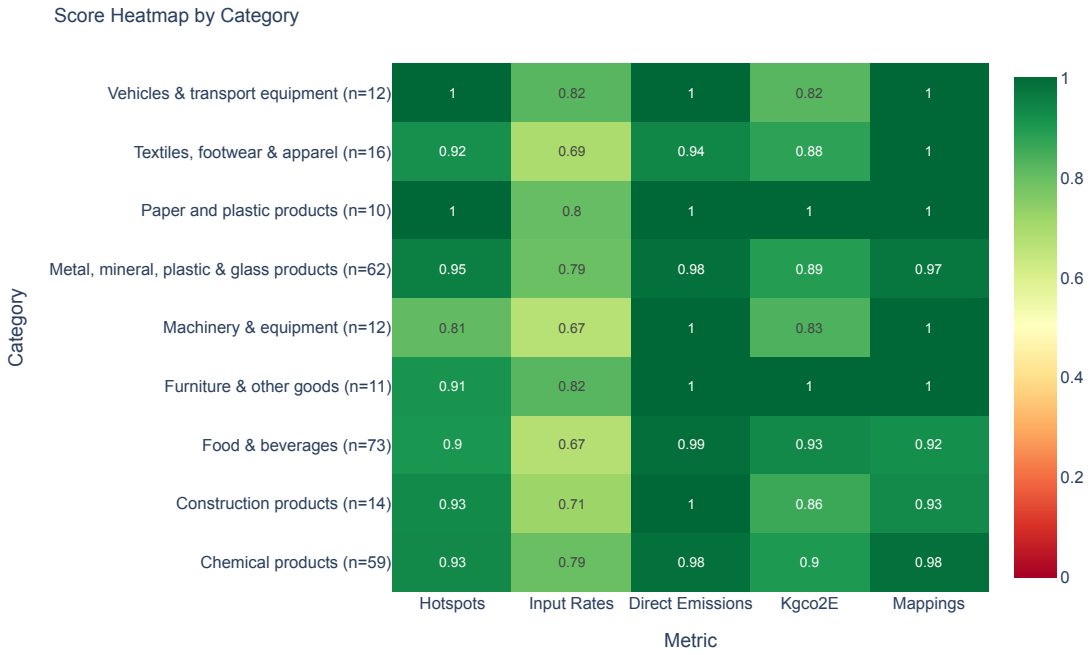

Figure S4: Score heatmap by category

In addition to the graph-level benchmarks, every graph also includes component-level benchmarks and citations that are clearly visible to users.

Component Validity: Many EPDs report only total carbon footprints, not the intermediate calculations that produce them—making it impossible to mathematically validate model components (e.g., yield rates, energy inputs) against the EPD benchmarks. To address this, we developed an "LLM Judge" graph reviewer approach: for each model output, an independent AI research agent searches the web and literature for credible reference values (e.g., industry reports, technical specifications, academic papers), following a source hierarchy that prioritizes peer-reviewed, government, and manufacturer data. A human reviewer then spot-checks a sample of these AI-retrieved benchmarks to confirm reliability.

This approach validates five model components: total kgCO2e, emissions hotspots, key input rates, proxy mappings, and direct emissions. For each component, the system reports whether the model output falls within acceptable bounds of the reference value (typically a factor of 2), along with the supporting citations. This design serves two audiences: system developers can use component-level pass rates to identify systematic weaknesses, while end users preparing the PCFs can prioritize manual review on flagged components. The key advantage of this approach is generalizability—it does not require a curated ground-truth dataset for every product category. Pass rates on the 269 test cases are as follows:

Table S4: Component Validity of Advanced Model ($N = 269$)

Validation category	**Description of review approach**	**Pass rate**
Total kgCO2e	Check whether the total emissions are within a factor of 2 of a trustworthy reference value.	90%
Hotspots	Identify the top 3 emissions sources in the Advanced Model. Determine what the typical top 3 upstream emissions sources should be for this material based on its production process. Compare the number of sources that align and whether each hotspot's cumulative emissions align with a trustworthy reference source.	93%
Key input rates	Evaluate whether input rates for hotspot materials are realistic.	75%
Proxy mappings	Evaluate whether the top 3 mappings to ecoinvent activities produce less than a factor of 2 error in emissions for the respective mapping.	96%
Direct emissions	If direct emissions of any foreground node are greater than 5% of overall emissions, check whether the emissions are within a factor of 2 of a trustworthy reference.	99%

The system pass rates mean that for most models, limited human review is required to assess the defensibility of the model. The input rates pass less frequently than the other categories. We hypothesize that this is because the input rates (which include material quantities as well as energy) can vary widely depending on the exact production process and facility, and therefore accurate references to compare with can be difficult to find or primary data may be most warranted for these cases.

Datasets: Watershed data and climate scientists and software engineers have curated datasets to check each major decision point or component of the advanced PCF model as well as end-to-end results. The end-to-end dataset uses environmental product declarations (EPDs) from EPD International across all industries and sectors covered there, in addition to some manually-curated

PCFs or values from scientific literature. In all end-to-end test cases, inputs include the material name and description, while expected outputs include the emissions (kgCO2e) for cradle-to-gate lifecycle stages. In addition to the end-to-end benchmarks and graph reviewer categories described above, other exemplar evaluations include:

Table S5: Additional evaluation categories

Evaluation	**What it represents**
New materials created when needed	Evaluates whether materials are created when needed, and existing activities are mapped to when there is a good library match
Reproducibility	Evaluates the reproducibility of end-to-end PCFs where the same input is run multiple times

4.2 Criterion 2.2: Learning Capability

Criterion: Demonstrated improvement in accuracy and insight generation as more data is processed, via quality indicator improvements.

Paper details: This can be shown via a narrative or examples for the results of actions taken on the match quality and uncertainty indicators; as more data is added, the quality should improve and the uncertainty should be reduced. The approach depends on how uncertainty is incorporated and propagated through the system.

What we do: Watershed never trains our models on customer data, and one customer's data will never be shown or influence the decisions for the AI for another customer. The way that our system improves over time is that Watershed's data and climate scientists observe any errors or take feedback from customers or the AI graph reviewer and use these inputs to identify system improvement opportunities. Then, we seek examples in the literature to support improvements. As the system improves, the performance on the evaluations described in section 2.1 improves.

This learning capability and ability to improve system performance over time has resulted in substantial improvements to overall quality. Watershed's initial demonstrations of AI-powered production graphs in August 2024 looked compelling but had a high error rate. Through constant iteration and system refactoring, as of January 2026 we can now target the vast majority of graphs to meet our quality bar.

4.3 Criterion 2.3: Uncertainty Indication (material-level)

Criterion: Indication of confidence level, uncertainty ranges, or sensitivity for key datapoints.

Paper details: Modeled materials and products should make clear to the user which assumptions or nodes have the biggest uncertainties driving the overall uncertainty and sensitivity. This could be a simple indicator, or it could be numeric. PCFs contain many sources of uncertainty. A goal is to reduce these with a focus on where data can enable directionally correct comparative decisions. Lacking clear ground-truth data or sufficient sample sizes, it may be infeasible to achieve true statistically supported conclusions. Recent literature provides useful insights such as that the most relevant source of uncertainty for certain types of chemical production is in facility-level process specification, more than allocation method choices. A key goal is to make a useful and actionable set of results rather than a mathematically rigorous theoretical model.

What we do: Watershed scientists have experimented with several ways of displaying, calculating, and propagating uncertainty. The way that the model is set up includes "markets" which may be fulfilled by more than one potential material, geography, or manufacturing approach; these approaches are weighted by the probability of their occurrence given the supply chain of the described input material. When a user has primary data, they can then edit these markets to collapse the uncertainty. The AI assistant within the graph editor can also provide some insights as to the most impactful nodes or branches that comprise a material's emissions. All AI-generated nodes include benchmarking with citations and indications of the relative match and representativeness of the benchmarks; these can be used to understand highly uncertain parameters.

We continue to solicit feedback as to how to best highlight uncertainty in the product in ways that can be insightful without being overwhelming or too technical for users.

4.4 Criterion 2.4: Key Data Sources (system-level)

Criterion: Documentation of major data sources and databases used.

What we do: All sources used in every graph, along with all formulas and assumptions, are documented and visible in the relevant graph nodes as of January 2026. All leaf nodes map to ecoinvent activities, but because there are other assumptions and inputs (e.g., processing energy use, component breakdowns, direct emissions sources) it is important that these sources are also documented.

4.5 Criterion 2.5: AI Methodology

Criterion: High-level explanation of the AI approaches used to generate insights beyond simple mapping.

What we do: Watershed's advanced mapper is a multi-agent architecture with information-gathering and formula tools and independent validation agents. It can generate synthetic BOMs, estimate masses and process flows, map to EF library, and aggregate to cradle-to-gate emissions and energy totals.

4.6 Criterion 2.6: Summary Methodology (calculation approach)

Criterion: High-level explanation of calculation approach.

What we do: Sections 2.1 through 2.5 provide these narratives.

4.7 Criterion 2.7: Limitations Statement

Criterion: Transparent communication of system limitations and appropriate levels.

What we do: The system is described clearly in this SI and the parent paper. Any AI-generated assumptions are clearly cited with source: AI nodes. In addition, Watershed's sustainability advisory teams work with companies that are using the advanced AI model to clearly communicate use cases.

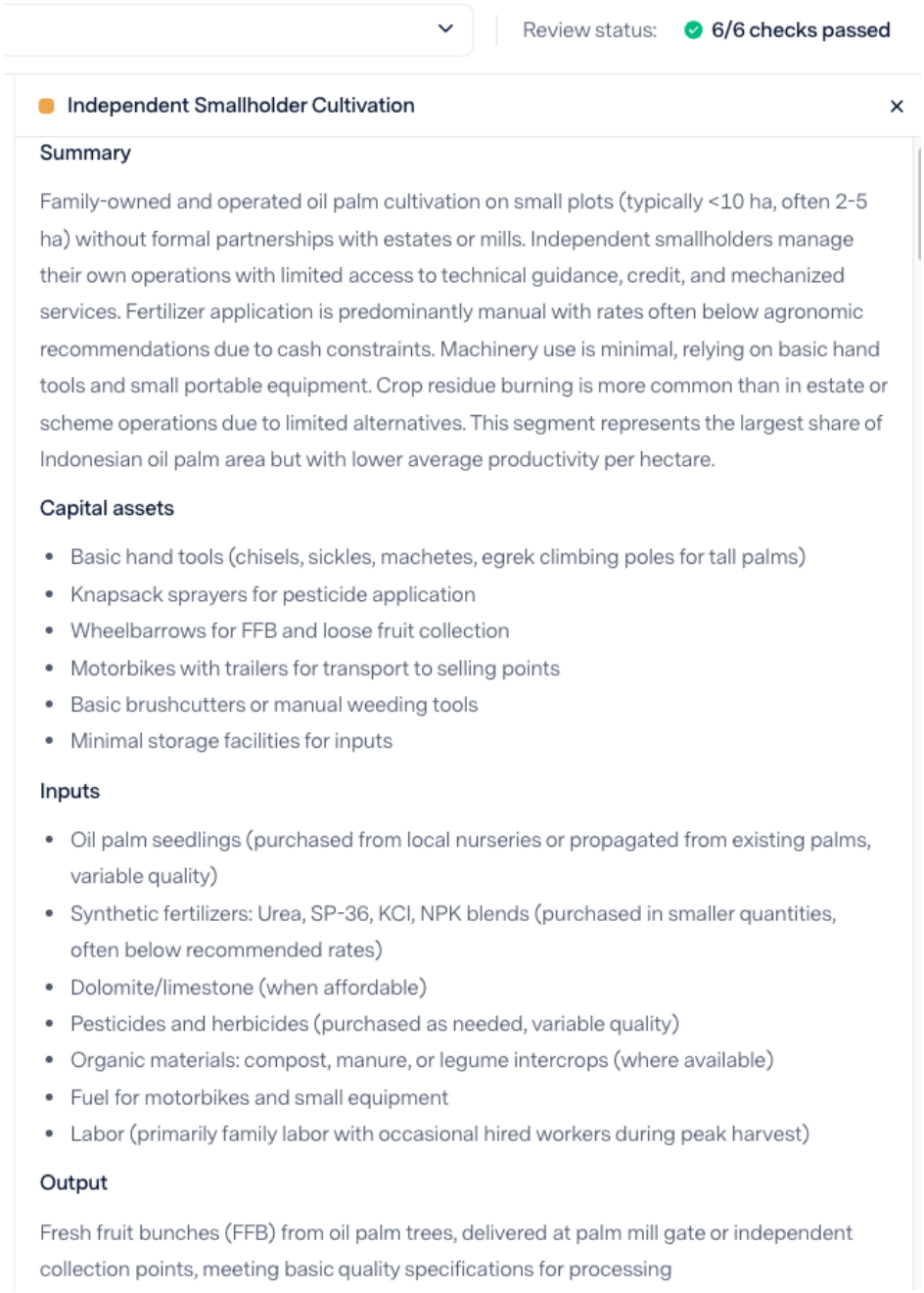

Figure S5: Example output of the advanced modeling system: process overview for independent smallholder oil palm cultivation. The system generates a detailed process description including a summary of the production process, capital assets, input materials, and output specifications. All content is AI-generated with web-sourced information.

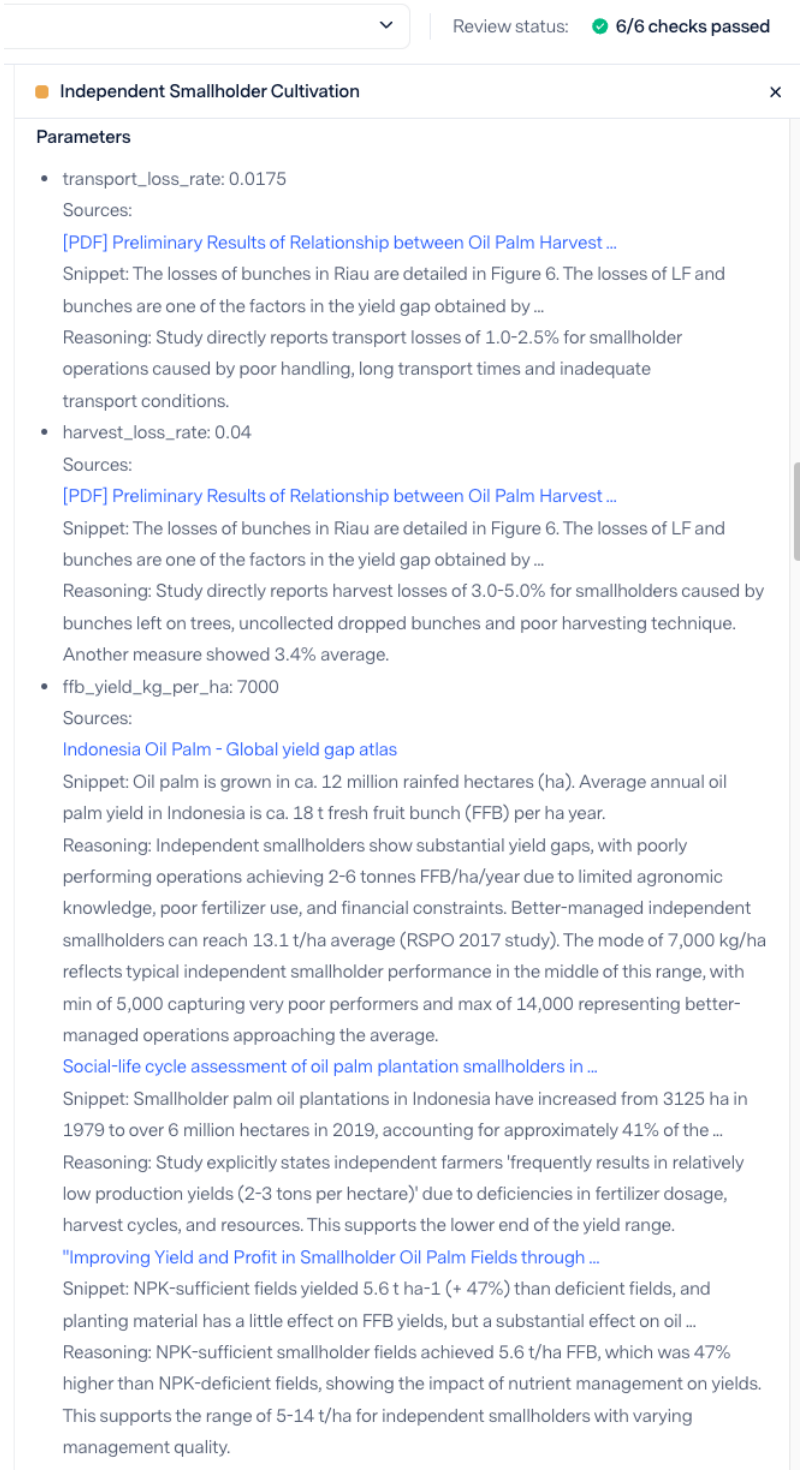

Figure S6: Example output of the advanced modeling system: AI-researched parameters with citations for independent smallholder oil palm cultivation. Each numeric parameter (e.g., transport loss rate, harvest loss rate, FFB yield) includes web-sourced references with relevant snippets and reasoning chains explaining how the value was derived. The review status indicator shows that all validation checks passed.

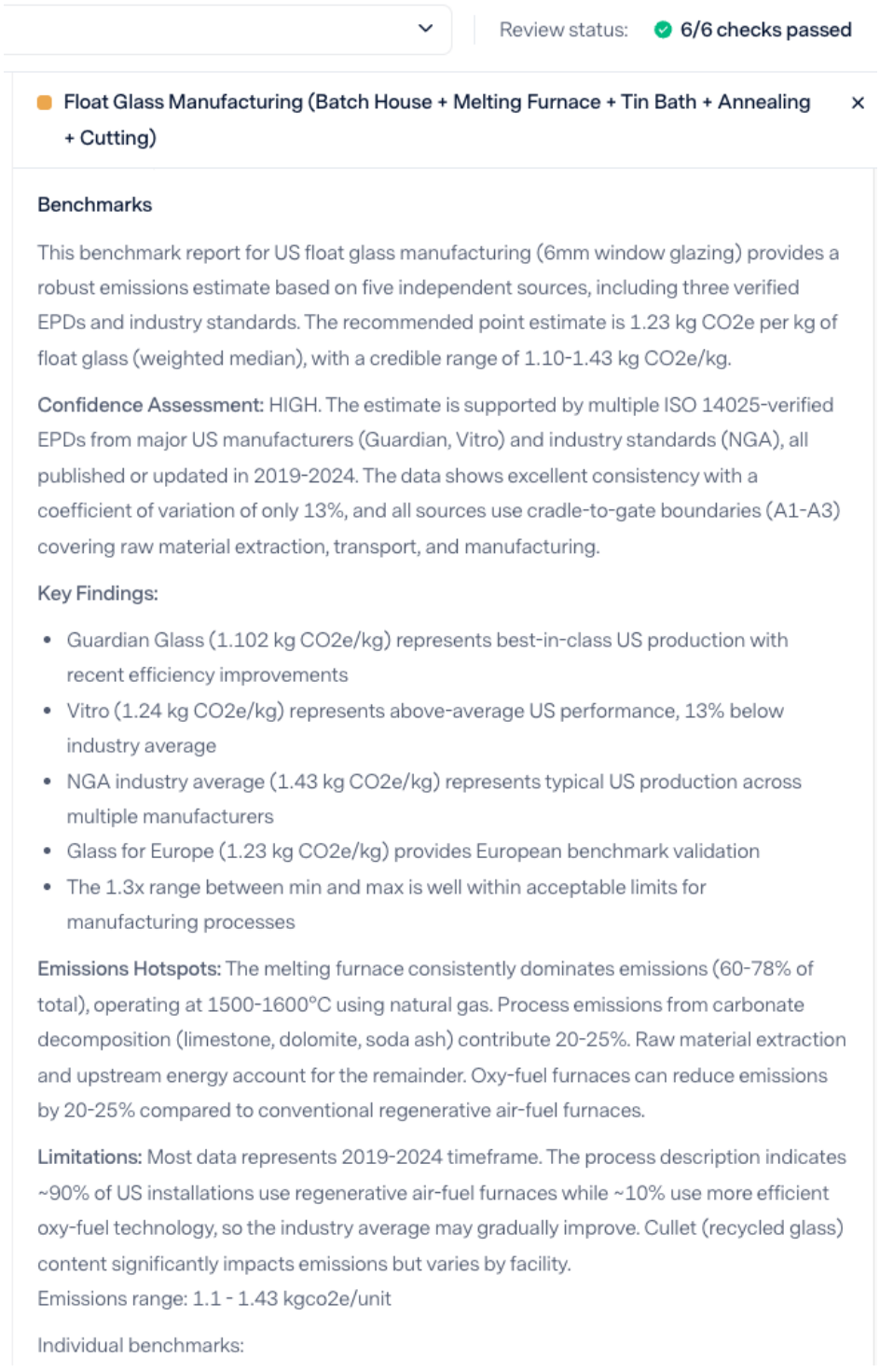

Figure S7: Example output of the advanced modeling system: benchmark report for float glass manufacturing. The system compiles independent emissions benchmarks from verified EPDs and industry standards, provides a confidence assessment, identifies emissions hotspots, and documents limitations. The recommended point estimate and credible range are derived from multiple sources.

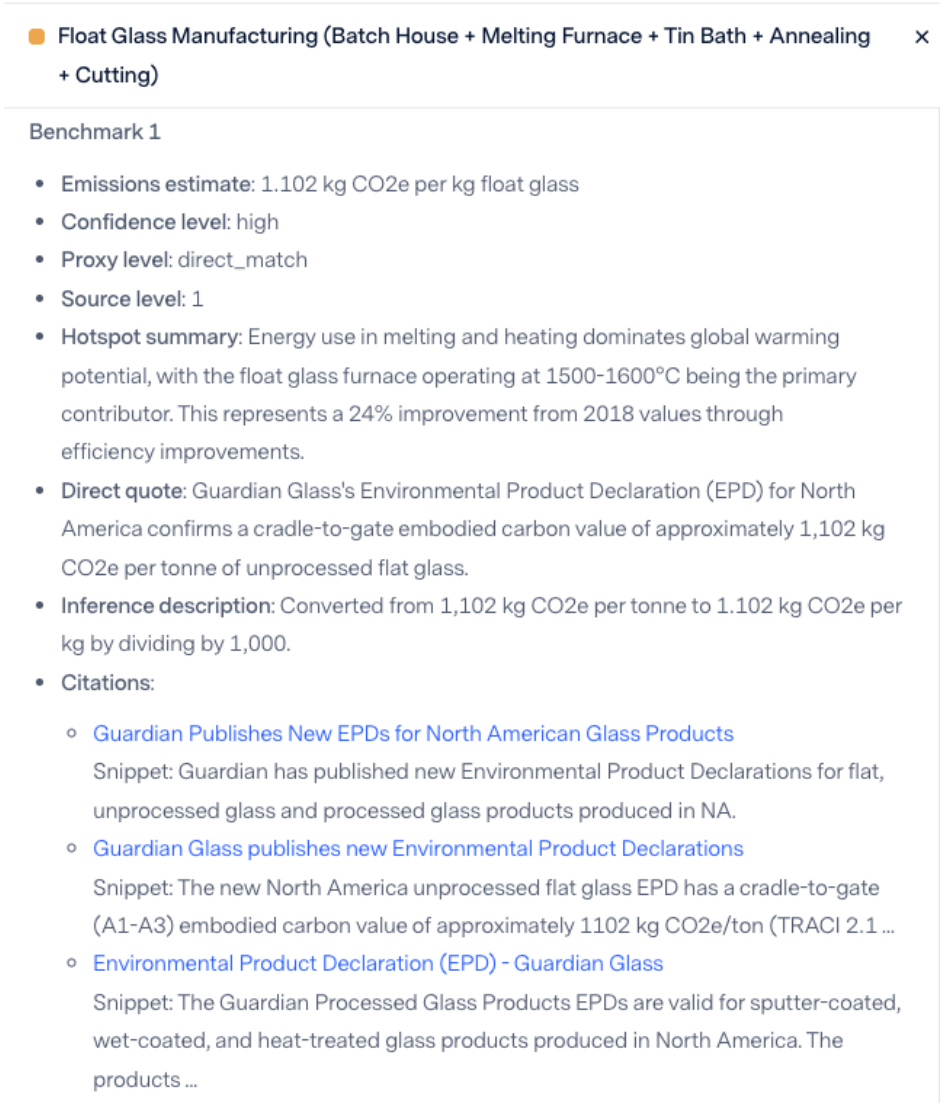

Figure S8: Example output of the advanced modeling system: detailed individual benchmark entry. Each benchmark includes the emissions estimate, confidence and proxy levels, a hotspot summary, direct quotes from source documents, inference descriptions explaining unit conversions or adjustments, and full citations with links to original sources.

4.8 Criterion 2.8: Major Assumptions (material-level)

Criterion: Recording of significant assumptions that drive results.

What we do: In addition to the significant assumptions that drive results, all other assumptions are also recorded per node, including yields, locations, and process choices.

4.9 Criterion 2.9: Hotspot Identification (material-level)

Criterion: Clear indication of major emissions sources and improvement opportunities.

What we do: Watershed's user interface is meant to quickly surface dominant emissions contributors

and actionable levers. At a material level, the graph is constructed with emissions-weighted bars that allow the user to quickly identify and trace hotspots where they could revise assumptions or evaluate scenarios. The AI assistant can also assist the user in finding these hotspots in a more interactive way.

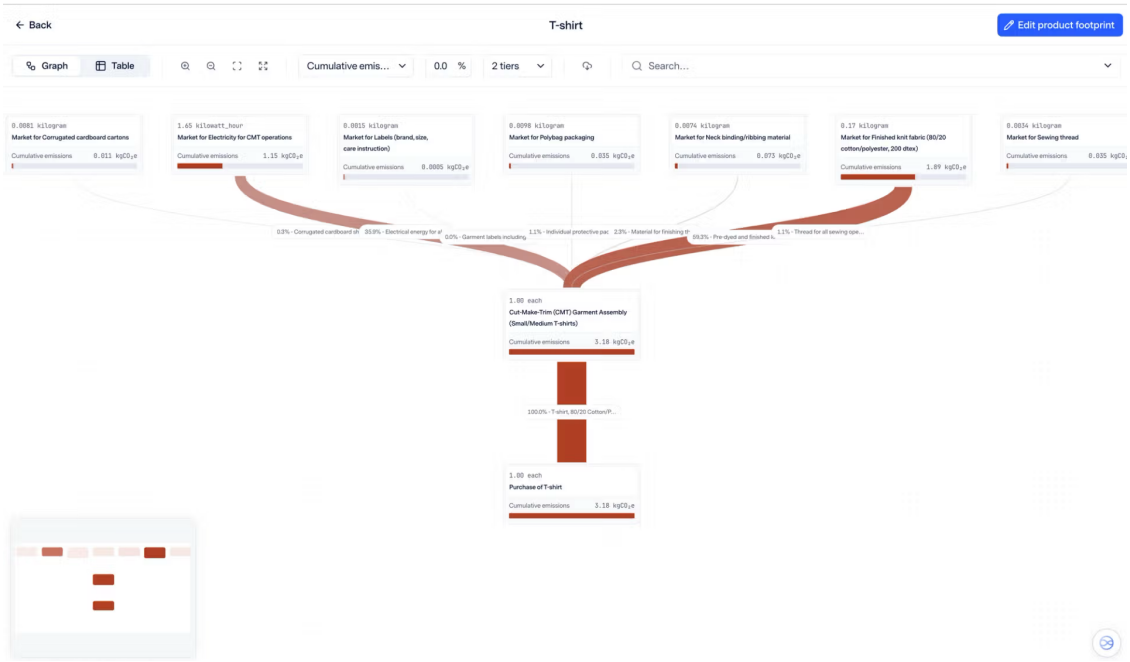

Figure S9: Example user interface view of a production graph

At the material portfolio level, the business intelligence (BI) level can be used to quickly aggregate and compare hotspots among any dimension.

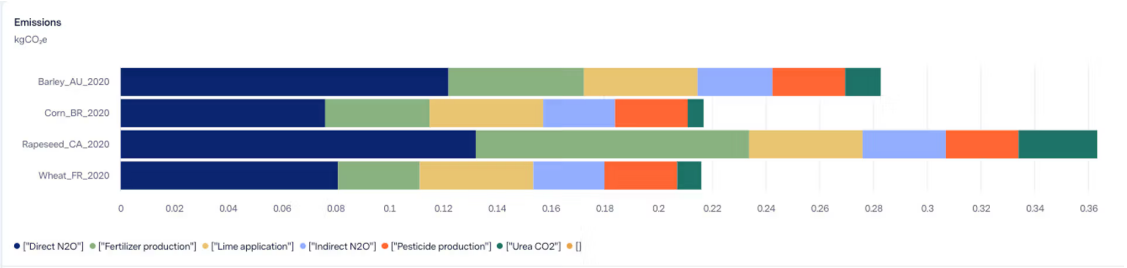

Figure S10: Example comparison BI view for row crop models