Supplementary Material for: Evaluation Criteria for AI-Assisted Product Carbon Footprinting Systems: The Cases of Mapping and Supply Chain Modeling

Online Resource 1: System Architecture and Criteria Demonstration

Shaena Ulissi[1*] Andrew Dumit[1] P. James Joyce[1] Jacob Feintzeig[1]
Krishna Rao[1] Steven Watson[1] Sangwon Suh[1*]

April 2026

[1]Watershed Technology Inc., San Francisco, CA, USA
[*]Corresponding authors: shaena@watershedclimate.com; sangwon@tsinghua.edu.cn

1 Introduction

This Online Resource accompanies the paper *Evaluation Criteria for AI-Assisted Product Carbon Footprinting Systems: The Cases of Mapping and Supply Chain Modeling*. It provides (a) architectural details for the AI-assisted carbon footprinting (AI-CF) systems described in the case study (Section 4 of the main paper), (b) the methodology and datasets used for each criterion, and (c) a criterion-by-criterion demonstration of how the systems can be evaluated against the criteria proposed in Section 3 of the main paper. The systems are proprietary commercial products of Watershed Technology Inc. The scientific contribution of the main paper is the evaluation criteria framework, which is system-agnostic. This resource is provided for transparency regarding the specific case study systems and can serve as a reference for how the criteria may be used to assess AI-CF system and output quality. As shown below, the criteria can be operationalized and can identify specific areas where system performance or output quality may be improved.

2 System Architectures

This section provides architectural details for the two AI-CF systems evaluated in the case study. Both systems are deployed commercial products; the descriptions here are intended to provide sufficient detail to show how the systems can be evaluated against the criteria, without revealing proprietary implementation details such as prompts or orchestration logic. The Auto-Mapper architecture is described in full in a separate JIE manuscript (currently under review) (Dumit et al., 2026a); key results are summarized here. The Advanced Modeling System architecture is described in greater detail in this resource than in the main paper.

2.1 System A: The Auto-Mapper (Case 1)

Objective: Automate the mapping of unstructured procurement text to specific emission factors in an LCI database (e.g., ecoinvent v3.11).

Architecture: The system utilizes a tool-using agentic framework built on Retrieval-Augmented Generation (RAG). It employs a dual-agent workflow to enforce a separation of concerns. The mapping agent ingests raw text, augments it with domain context (e.g., resolving "PP" to "Polypropylene" based on the reporting company's sector), and utilizes RAG to query the vector-indexed material library. The Quality Reviewer (Judge) Agent operates independently and acts as a critic. It scores the Mapping Agent's selection against a strict rubric, generating Material Match and Emissions Match scores. The Reviewer does not see the Mapper's internal chain of thought, which prevents self-correction bias and provides an independent assessment of the output's fitness for use. The Quality Reviewer evaluates matches based on critical dimensions including Material Type and Properties Accuracy, Emissions Similarity, Form Factor, Degree of Fabrication, and Specificity Alignment. The system processes input data comprised of unstructured procurement text, material names, descriptions, and supplier data.

The Pedigree Matrix approach as outlined in Weidema and Wesnæs (1996) and ISO 14044 has been widely used by LCA practitioners for data quality and data fitness scoring. However, the Pedigree Matrix was designed for expert assessment of individual datasets against generic quality dimensions (temporal, geographical, technological representativeness), not for evaluating the specific match quality between an unstructured input description and a proxy dataset at the scale of thousands of materials. Its indicators do not distinguish between a mapping that is correct on material type but wrong on emissions profile (e.g., virgin vs. recycled aluminum) and one that captures emissions well but misidentifies the material category. Therefore, the Quality Reviewer implements a two-dimensional scoring rubric—Material Match and Emissions Match—that extends the Pedigree concepts into dimensions specific to AI-assisted proxy selection, enabling automated quality assessment at scale while capturing distinct failure modes that single-dimension scoring would miss (Dumit et al., 2026a). The system is built to be flexible: the quality dimensions and scoring rubric can be modified and tested against the benchmark datasets to evaluate alternative criteria configurations, enabling systematic experimentation with different quality assessment approaches.

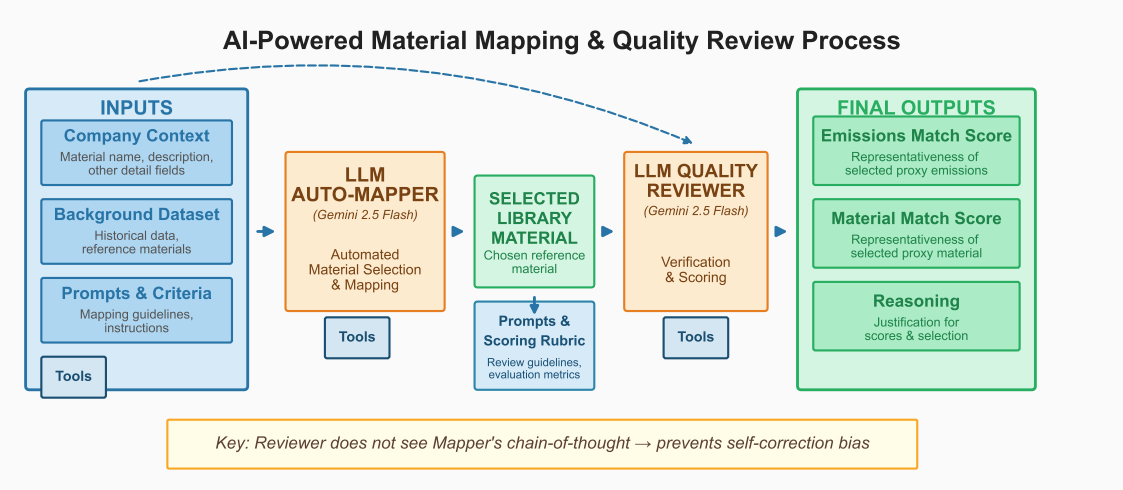


Figure S1: Schematic of the Auto-Mapper system

Key Design Decisions:

- **Separation of generation from evaluation:** The Mapping Agent and Quality Reviewer are architecturally independent. The Reviewer never sees the Mapper's chain of thought, preventing self-correction bias where a model rationalizes its own errors. This design pattern—separating generation from evaluation to improve reliability in domains where errors are plausible but consequential—is validated both for LCI mapping (Dumit et al., 2026a) and as a general approach to reliable AI deployment (Dumit et al., 2026b). Empirically, the independent judge achieves superior calibration compared to self-assessment: while ranking performance is similar (AUROC 0.848 vs. 0.842), the independent judge achieves a 0.3% error rate at the HIGH quality tier compared to 2.1% for the self-judge—a seven-fold improvement that enables reliable auto-acceptance of high-confidence mappings (Dumit et al., 2026a).
- **No abstention:** The system always produces a mapping rather than declining to answer. When no good match exists, the Quality Reviewer flags the mapping with a low match quality score, directing it to human review.
- **Two-dimensional scoring:** The Quality Reviewer produces separate Material Match and Emissions Match scores. This captures distinct failure modes—a mapping can be correct on material type but wrong on emissions profile (e.g., virgin vs. recycled aluminum), or vice versa. Single-dimension scoring would miss these distinctions.

Models: As of January 2026, the Mapping Agent uses Gemini 2.5 Flash and the Quality Reviewer uses Claude Haiku 4.5 (Dumit et al., 2026a). Model and database versioning (including ecoinvent v3.11) are maintained for reproducibility and regression detection.

Cross-References: A detailed evaluation of the Auto-Mapper architecture, including ablation studies demonstrating the value of agentic iteration (+22 percentage points over single-shot RAG) and selective review methodology, is presented in (Dumit et al., 2026a). At a 20% review budget, the judge captures 67% of all mapping errors, compared to 37–40% for heuristic baselines. Confidence score calibration is further analyzed in (Dumit et al., 2026a).

2.2 System B: The Advanced Modeling System (Case 2)

Objective: Generate a full Screening PCF by constructing a directed acyclic graph (DAG) representing the production supply chain for complex products.

Architecture: This system employs a multi-agent Researcher architecture that systematically decomposes a product. In the decomposition phase, the system breaks down high-level products (e.g., Bicycle) into sub-components (Frame, Chain, Tires) and further into raw materials across multiple supply chain tiers. It initializes processes, parameters, and formulas used to represent the production process. Specialized agents estimate mass balances, material composition, and processing inputs (energy, heat) for each node in the graph, using diverse information retrieval tools and validation and benchmarking checks. Validation agents use benchmarking as well as deterministic (mathematical) tests to check validity and solve equations. Similar to the Auto-Mapper's Quality Reviewer, an independent validation agent critiques the generated graph against external benchmarks, checking for logical consistencies such as yield rates and the presence of expected hotspots (e.g., energy intensity in aluminum smelting). The output is a fully linked production graph rather than a static emission factor. The system models a cradle-to-gate (or, if applicable, cradle-to-site) boundary.

Materials, processes, and energy inputs are mapped to ecoinvent v3.11 (cut-off approach) “total CO2e, excluding biogenic” AR6 GWP-100 emission factors.

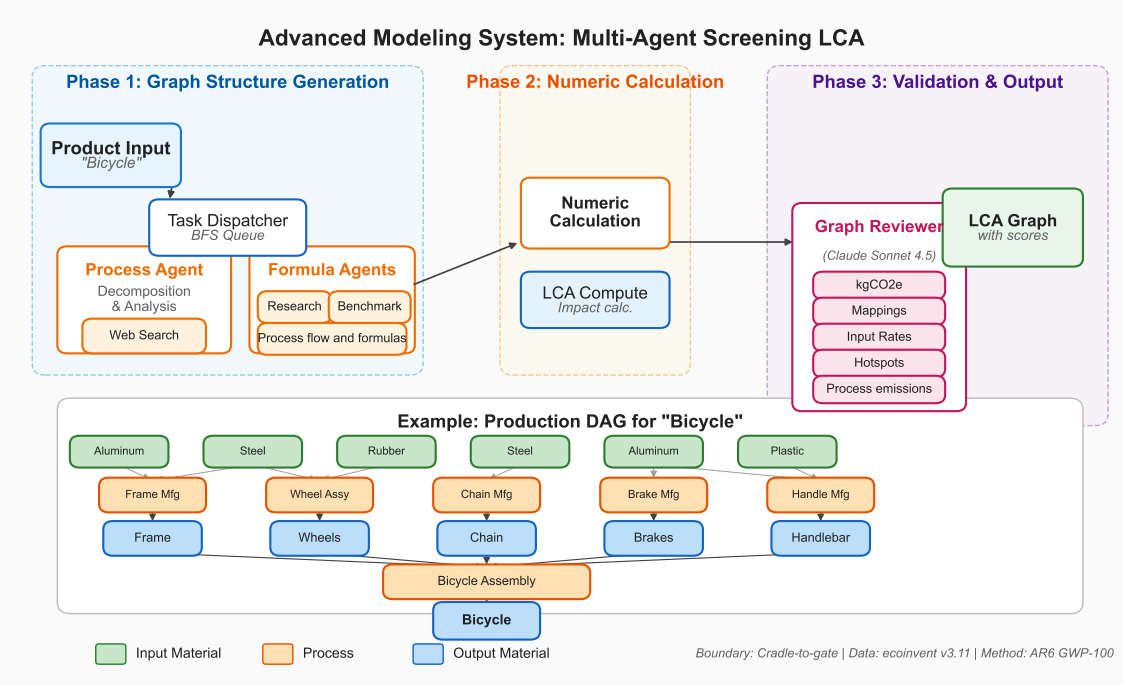


Figure S2: Schematic representation of the Advanced Model system.

Agent Pipeline Stages:

The Advanced Modeling System operates through five sequential stages, each handled by specialized agents with distinct tools and validation checks:

Table S1: Advanced Modeling System agent pipeline stages

#	Stage	Description	Key Outputs
1	Product Decomposition	Breaks down high-level products into subcomponents and raw materials across multiple supply chain tiers. The agent constructs a hierarchical bill of materials (BOM) by identifying constituent parts, their relationships, and the manufacturing processes that connect them.	Hierarchical BOM structure
2	Parameter Research	Specialized agents estimate mass, energy, yields, and other process parameters using web and literature sources. Each parameter is accompanied by citations, source snippets, and reasoning chains explaining how the value was derived.	Cited parameters with reasoning chains
3	Formula Construction & Validation	Deterministic checks verify mathematical consistency: mass balance across nodes, unit conversion validity, energy balance, and formula correctness. These checks are algorithmic (not LLM-based), ensuring that the graph's quantitative relationships are internally consistent regardless of LLM variability.	Validated process formulas
4	Emission Factor Mapping	Leaf nodes in the production graph are mapped to ecoinvent v3.11 activities using the Auto-Mapper approach (Section 2.1). This ensures that the same quality controls, match quality scoring, and proxy selection logic apply to EF assignments within the Advanced Model.	EF assignments with match quality scores
5	Graph Review & Benchmarking	An independent validation agent reviews the completed graph against external benchmarks sourced from EPDs, industry reports, and academic literature. It checks total emissions, hotspot alignment, key input rates, proxy mappings, and direct emissions, reporting pass/fail for each component with citations.	Pass/fail per component with citations

Key Design Decisions:

- **Deterministic validation layer:** Stage 3 uses algorithmic checks (mass balance, unit conversions) wherever feasible rather than LLM judgments, providing a mathematical safety net that catches errors regardless of LLM variability.
- **Independent graph reviewer:** Similar to the Auto-Mapper's separation of generation and evaluation, the Graph Reviewer (Stage 5) operates independently from the agents that constructed the graph, preventing self-rationalization of errors.

- **Structured output with full provenance:** Every node in the output graph includes its parameter values, source citations, reasoning chains, and validation status. This structured format enables both automated quality checks and targeted human review.

Models: As of January 2026, the Advanced Modeling System primarily uses Claude Opus 4.5 for agent reasoning and graph construction. Model and database versioning are maintained for reproducibility.

Model Selection and Flexibility: Both the Auto-Mapper and the Advanced Modeling System are architected so that the specific LLM version and provider can be changed as new models are released. When evaluating a candidate model change, performance is measured against the benchmark datasets and criteria described in Section 3 of the main paper. The selection process prioritizes the lowest-cost (and likely least energy-intensive) model that achieves equivalent or better performance to the incumbent. Because these systems are deployed in an enterprise context with strict data privacy and data residency requirements, the set of eligible model providers is constrained; not all commercially available LLMs meet these requirements. The criteria and evaluation datasets described in this resource are designed to be model-agnostic and can be applied to any candidate system, facilitating systematic comparison when models or providers change.

3 Validation Datasets

3.1 The Ground Truth Problem

A central challenge in validating AI for PCF is the absence of clear ground truth. Unlike image classification, where an image definitively either "contains cat" or "does not contain cat", LCI mapping often involves subjective expert judgment regarding the "best available" proxy. Studies such as (Courtat et al., 2025), (Konradsen et al., 2024), and (Kuczenski et al., 2021) have demonstrated that even experienced practitioners can arrive at different but expert-accepted choices for the same product, and that the representativeness of secondary LCI datasets can significantly influence product ratings. Instead, what statistical approaches aim to measure is precision. But precision alone does not capture the overall fitness question. A highly precise number derived from an outdated or technologically irrelevant process is of low quality. Conversely, a number with wider uncertainty bounds that accurately reflects the specific technology and geography of a supplier may be of higher quality because it is more representative.

3.2 New Error Modes from AI

The integration of generative AI into sustainability workflows introduces risks that do not exist in deterministic software systems. Unlike deterministic calculation engines, LLMs generate outputs based on statistical patterns learned from training data, introducing inherent variability and the potential for several distinct error modes. These include *fabricated mappings* (selecting a nonexistent database entry), *misapplied facts* (correctly extracting a parameter from a source but applying it to the wrong context), *invented citations* (attributing a claim to a source that does not contain it), and *incorrect reasoning* (applying a valid fact through flawed logic, e.g., using a per-unit value without adjusting for yield). The term "hallucination" is often applied broadly to all of these, but in practice some error modes are structurally preventable while others require independent verification.

3.3 Evaluation Datasets

Few publicly available datasets exist for evaluating AI-CF systems, and no standardized benchmark suite has been established for this domain. To evaluate conformance with the evaluation criteria—particularly Criterion 1.1 (Benchmark Performance) and Criterion 2.1 (Reasonable Precision and Component-Level Benchmarking)—we assembled evaluation datasets from a combination of internally curated test cases, anonymized production data, and external sources. These datasets span a range of evaluation granularity, from narrow unit-level tests analogous to software unit tests to full end-to-end comparisons against published product carbon footprints:

- **Unit Evaluations (Deterministic).** A suite of specific test cases designed to evaluate distinct skills, such as identifying chemical synonyms (e.g., Acetone vs. Propanone), handling abbreviations, and respecting negative constraints (e.g., ensuring "Silicon" is never mapped to "Silicone"). These tests have unambiguous correct answers and serve as regression checks for core capabilities.
- **Production Data Sets (Probabilistic).** Anonymized production data ($N = 704$ combining curated representative samples and real customer materials) from diverse customer portfolios—spanning consumer goods, manufacturing, and food—were used to test performance against the semantic noise and ambiguity typical of real-world procurement data. Test cases were tagged as *vague* when input descriptions lacked sufficient specificity for unambiguous material identification—for example, internal SKU codes (e.g., "SKU-44821"), unclear abbreviations (e.g., "FLV NAT CHOC PB"), or generic descriptors (e.g., "plastic part"). This tagging enables analysis of how input data quality affects mapping accuracy and supports the entropy analysis described in the main paper.
- **External Benchmarks.** The study utilized the Amazon Parakeet dataset (Balaji et al., 2025), a public set of product descriptions and emission factors ($N = 281$). To ensure rigorous evaluation, internal experts re-reviewed and corrected labels in this dataset to align with internal standards, creating a high-fidelity benchmark (Dumit et al., 2026a).
- **End-to-End Benchmarks (Modeling).** For Case 2, generated graphs were compared against 269 Environmental Product Declarations (EPDs) from EPD International and other sources, serving as a proxy for ground truth in full-product modeling. The EPDs were selected via stratified sampling across 9 product categories (Chemical products, Construction products, Food & beverages, Furniture & other goods, Machinery & equipment, Metal/mineral/plastic/glass products, Paper and plastic products, Textiles/footwear/apparel, and Vehicles & transport equipment) to ensure broad coverage. Input data provided to the AI system included product name and description extracted from each EPD. The emissions intensities were normalized to per-kilogram or per-unit declared units. A breakdown of the EPDs by product category is provided in Online Resource 2.

We acknowledge that the vague/non-vague tagging and expert labels in these datasets are currently produced by internal annotators, which introduces some risk of overfitting to the materials and patterns present in our background dataset. External benchmarks such as the Amazon Parakeet dataset provide a partial check against this risk. The development of standardized, publicly available benchmark datasets for AI-CF evaluation remains an important area for future work. We hope to contribute to and benefit from such community efforts; such datasets could enable cross-system

comparisons and help establish community-wide performance baselines for different aspects of AI-assisted carbon footprinting.

Evaluation datasets are expanded as modeling capabilities are extended into new areas or as new test categories are needed to detect regressions.

4 Criteria Conformance Demonstration

This section demonstrates how the case study systems may be evaluated against the criteria in the paper. The Auto-Mapper is an example of the mapping case (Case 1) and targets the criteria in Tables 1 and 2 of the paper. The Advanced Modeling System 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 describe how the systems may be evaluated against the criteria.

4.1 Case 1: Auto-Mapper Criteria

Criterion 1.1: Benchmark Performance **Criterion:** AI-generated mappings should demonstrate their ability to select expert-accepted proxy LCI datasets when tested against a validation set of products with expert-labeled acceptable 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. If a labeled dataset were to be made available such as (Balaji et al., 2025; Dumit et al., 2024), 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 below 100% may be acceptable where multiple expert-accepted proxy mappings exist. A system with a lower performance may be acceptable only if automated mappings are used for items that are not material to a given company's purchased-materials footprint or if low-quality matches are routed to human review. Acceptable performance thresholds and materiality are context-dependent and should be defined by the reporting framework in use (e.g., under the GHG Protocol, materiality is typically assessed based on a material's contribution to total reported emissions). 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 system undergoes a change that could affect output quality, including switching LLM versions or providers, significant prompt or context modifications, and architectural changes to the agent pipeline. Because LLM-based systems can drift with provider-side model updates, periodic revalidation is recommended even absent intentional changes. The frequency of revalidation may be constrained by the cost of running evaluation suites.

Approach in this system: 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 an expert-accepted mapping. Mappings to banned substrings (clear errors) receive additional score penalties. 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) using *expert-accepted accuracy*: the fraction of test cases where the AI-selected mapping matches any expert-annotated acceptable option (Dumit et al., 2026a). This metric accounts for the subjective nature of proxy selection, where multiple choices may be equally valid; 24.8% of items have two expert-accepted mappings, 7.3% have three, and 3.1% have four or more (Dumit et al., 2026a). For non-vague inputs, the system achieved 91.2% expert-accepted 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 non-expert-accepted mappings as "low" quality, this supports a selective review workflow focused on materials most relevant to the GHG analysis. While the eventual goal is to approach 100%, we note that i) expert disagreement on proxy selection is well documented (Courtat et al., 2025; Kuczenski et al., 2021), ii) there is inherent ambiguity about what constitutes an expert-accepted mapping as it depends on the use case and context, and iii) the match quality scores (described elsewhere in this resource) flag outputs that may require additional human review.

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 expert-accepted 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 an error. These test cases and labels were developed manually with 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 customers have uploaded. We further expanded it by including Amazon's published dataset, to which we corrected and expanded the labels.

Table S2: Auto-Mapper Performance Summary

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

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 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.

Approach in this system: 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 an expert-accepted mapping. Mappings to banned substrings (clear errors) receive additional

score penalties. 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 ($> 90\%$). On industry-specific evaluations, performance is largely consistent across categories. Categories with greater ambiguity—for example, processed metals where there is not one clear ecoinvent match, or catalysts where a hierarchy decision is needed between most important and most representative components—show more variability. Ambiguous mappings are flagged by the quality review scoring (described below) for human review.

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 expert-accepted 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 an error. These test cases and labels were developed manually with 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 S3).

Table S3: Auto-Mapper performance by test category (Jan 2026)

Test category	**What it tests**	**N**	**Perf.**
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%**

Criterion 1.3: Appropriate Granularity and Proxy Approach **Criterion:** Where exact matches are not available, the system should apply appropriate proxy selection logic that aligns with LCA best practices.

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 adopting 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.

Approach in this system: 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

Criterion 1.4: Repeatability **Criterion:** Small variations in product descriptions or specifications for inputs with unambiguous matches should not result in dramatically different mappings. In addition, the same material should generally be mapped consistently and, if it is not, should be flagged as uncertain.

Paper details: Given the non-deterministic nature of AI/machine learning (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 for inputs with unambiguous matches. 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. Recent work has established that entropy over sampled LLM outputs is a meaningful signal for output reliability (Farquhar et al., 2024) and that output variability can be decomposed into input ambiguity, knowledge gaps, and decoding randomness (Taparia et al., 2026).

Approach in this system: We maintain labeled validation sets and expert-reviewed mappings. The system 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: The 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 S4 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 S4: Repeatability performance by category (Jan 2026)

Test category	N	Repeat. score	Norm. entropy	Unique pred.
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

Criterion 1.5: Match Quality Indication (material-level) **Criterion:** The system should display to the user an uncertainty or match-quality indication and highlight poor matches between the input material and the mapped proxy dataset. If material, these are cases where the user should improve the input data, conduct human review, or move to Case 2 carbon footprinting.

Paper details: Many types of uncertainty affect GHG analyses, including estimation uncertainty, parameter uncertainty, statistical uncertainty, model uncertainty, systematic uncertainty, and scientific uncertainty. There are many potential paths to display and calculate uncertainty in LCA, e.g., (Mendoza Beltran et al., 2018; Qiao et al., 2025; Cullen et al., 2024). To date, many existing LCA datasets have used a Pedigree matrix or data quality indicators (DQIs). Parametric Monte-Carlo propagation and model-centred approaches are more data-heavy approaches that may result in more statistically significant results. In addition, global sensitivity analysis to quantify which inputs dominate footprint variance may be useful (Adams and Allacker, 2025). The GHG Protocol Technical Working Groups suggest that uncertainty and data quality will be seen with greater importance in the upcoming standards update (GHG Protocol Technical Working Group, 2025). For proxy mapping, an additional uncertainty arises from the match between the input material description and the selected proxy dataset. Match quality refers here to the degree of correspondence between the input description and the selected proxy LCI dataset, considering material properties, form, processing state, specificity, and expected emissions profile. For proxy mapping to a single dataset, a qualitative uncertainty or data-quality framework may suffice to inform the user of the relative quality of each mapping. We found that match quality is often overlooked in LCA but can have a strong impact on emissions representation (Dumit et al., 2026a).

Approach in this system: A key feature of the proxy mapping system is the material match score and emissions match scores that are presented to the user. A second model 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 the 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

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.

Approach in this system: The mapping system is documented in the main paper and in Section 2.1 of this resource. 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.

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.

Approach in this system: Every selected mapping is clearly described with its source. This is included at the graph-level and within the citation metadata associated with each mapping.

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.

Approach in this system: The system 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.

Criterion 1.9: Decision Logic Documentation **Criterion:** Documentation of 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.

Approach in this system: 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.

Criterion 1.10: Output-Level Review Trail **Criterion:** Ability to inspect how specific product characteristics contributed to particular activity dataset selections. Given the non-deterministic character of LLM-based systems, this review trail may include structured summaries, sampled traces, and example datasets rather than deterministic reconstruction of every internal model step.

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.

Approach in this system: 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. The system uses an LLM tracing system that includes records for all tool calls and decisions. Full traces are available for verification but are primarily used for internal development and debugging.

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.

Approach in this system: The team reruns the mapper benchmarking evaluations weekly and uses them to evaluate the potential performance gains or regressions associated with model or context changes. Revalidation is triggered by any change that could affect output quality, including LLM version or provider switches, prompt or context modifications, and architectural changes. Because LLM providers may update models without notice, periodic revalidation occurs even absent intentional system changes. In addition, the Auto-Mapper evaluation suite runs automatically on a weekly schedule to catch any system drift early, independent of whether any changes were made. A few recent examples of this are as follows:

- Early users purchasing packaging materials noticed that there were material + process combinations that they thought may be a better fit than the materials selected by the 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, the team 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.

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).

Approach in this system: Auto-mapping from Case 1 falls into the "average-data method" of the GHG Protocol hierarchy, improving over broad sector averages. When higher-quality data is needed, users should advance to Case 2 (automated modeling) or supplier-provided PCFs.

4.2 Case 2: Advanced Modeling System Criteria

Criterion 2.1: Reasonable Precision and Component-Level Benchmarking **Criterion:** Component-level and end-to-end benchmark checks should be performed where feasible. There is rarely ground-truth GHG emissions data available, and experts do not always agree on system boundaries or other methodological choices for a given product. Benchmarking remains useful for assessing assumptions and calculations, identifying systematic weaknesses, and demonstrating 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 (EPD International, 2024; Smart EPD, 2026), licensed data such as ecoinvent (ecoinvent Association, 2024; Wernet et al., 2016), and literature values such as The Carbon Catalogue (Meinrenken et al., 2022). Within a modeled product or material, sub-processes and components can be benchmarked against literature values or calculated values. For example, theoretical and practical energy minimums for industrial processes can be quantified using methodology proposed in Bolson et al. (2025). A potential target for full product footprints is median error and P90 error below specified thresholds when benchmarked against verified PCFs (where available), plus clear assessments of subcomponents. This is an area where publicly available datasets and leaderboards or competitions might be useful for assessing and calibrating different AI-CF systems. Care would be needed to ensure there is no cross-contamination of training data and evaluation data.

Approach in this system: We have 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 performance baseline and, in some cases, a target—for example, that 90% of outputs fall within a factor of 2 of documented reference values.

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, component-level benchmark check results are reported in Table S5 below.

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). 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. EPDs are used as the primary benchmark because they represent the most widely available, third-party verified product carbon footprint data. The benchmark dataset is pre-filtered to ensure comparability: only EPDs reporting cradle-to-gate (A1–A3) emissions on the same declared unit as the AI model output are included. As a consequence, some residual methodological differences between the AI model and any given EPD are expected and acceptable; achieving zero median error is not the goal, since even EPDs for the same product under different PCRs can diverge substantially. While EPDs report only total emissions and cannot validate intermediate calculations, they serve as an essential reasonableness check for end-to-end system output.

For context, substantial variability is inherent in LCA practice even among human experts. EPDs for identical products can 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). A comparison of six expert teams modeling cotton cultivation from the same input data found climate change results varying by 44% even excluding land use change (Textile Exchange, 2026). Applying different national LCA frameworks to an identical office building produced total greenhouse gas results ranging from 10 to 71 kg CO_2-eq/m^2/year (Frischknecht et al., 2019). In the context of screening-level LCA—where process variance can span orders of magnitude (Meinrenken et al., 2022; Henriksson et al., 2015)—the system's 33% median error is within the range of human practitioner disagreement documented in the studies above.

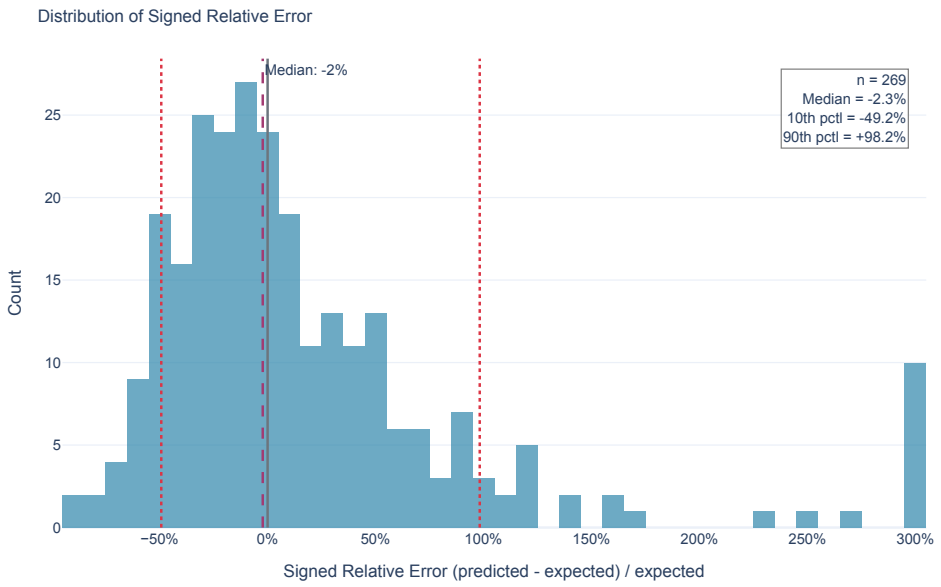


Figure S3: Advanced Model end-to-end results: Signed relative error distribution

Additional performance statistics are provided as follows.

The Predicted vs Expected plot (Figure S4) 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.

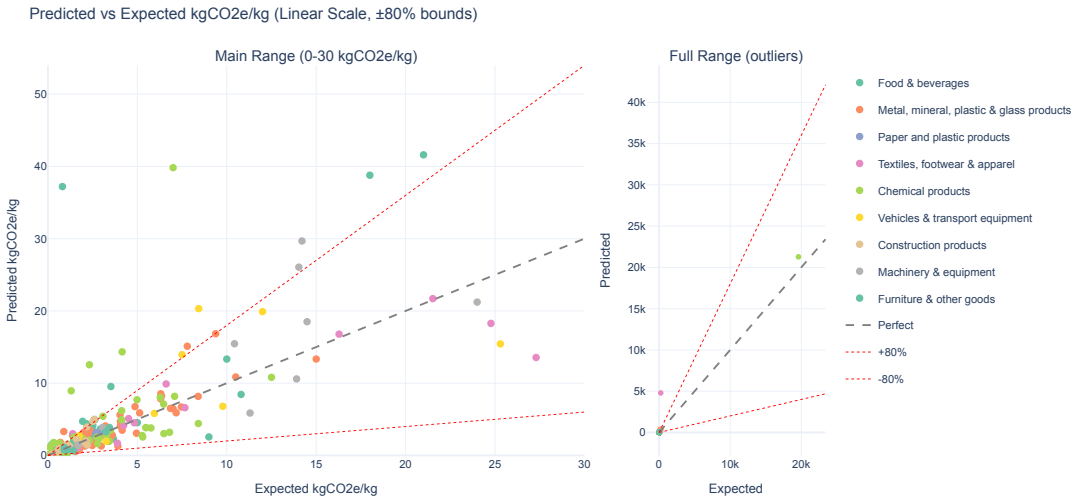


Figure S4: Predicted vs Expected kgCO2e/kg

The signed relative error by category (Figure S5) demonstrates that there is some variation in error by product category but that the median error is of a consistent magnitude.

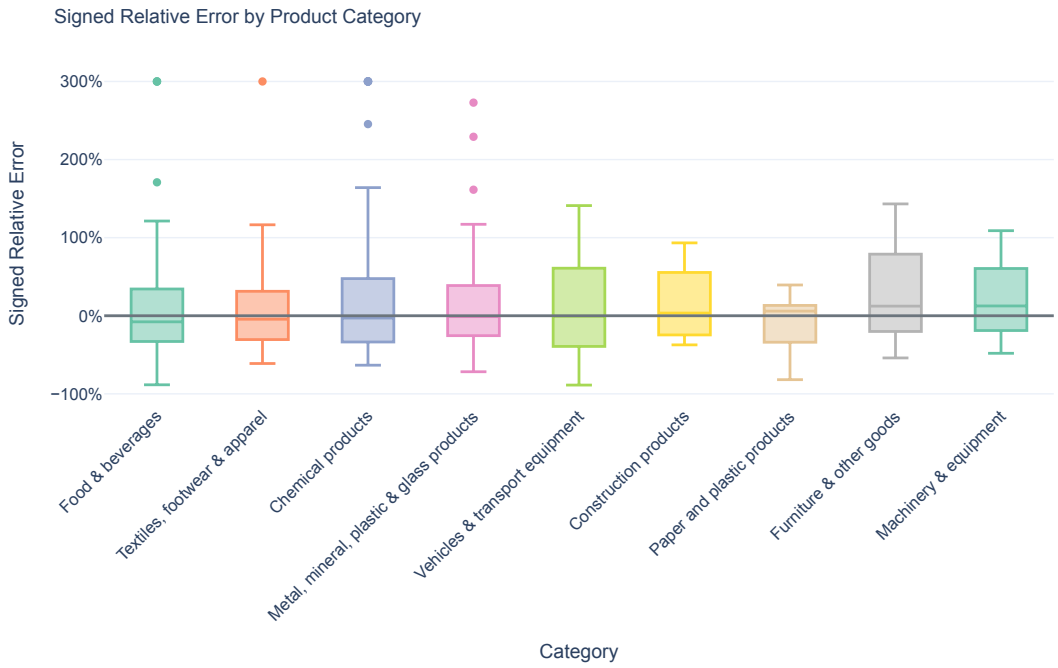


Figure S5: Signed relative error by category

The score heatmap by category (Figure S6) shows the Advanced Model graph reviewer's pass rates across the validation categories and product categories. Pass rates are broadly consistent across product categories. Key input rates show lower pass rates than other categories, which may reflect the difficulty of finding reliable reference values for proprietary process parameters.

Figure S6: Score heatmap by category

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, benchmarks are used to assess reasonableness and transparency rather than absolute accuracy. The goal is to establish a reasonable baseline model and track improvements over time.

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

Component-Level Benchmark Checks: 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 documented 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 (the material and energy consumption parameters that drive emissions for hotspot processes, such as electricity consumption per kilogram of product, raw material yield rates, and processing energy requirements), 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. A factor of 2 was selected as the acceptability threshold based on the documented variability of EPDs for identical products (>50%, (Gelowitz and McArthur, 2017; Konradsen et al., 2024)), precedent in screening-level LCA where order-of-

magnitude accuracy is often the initial goal, and practical experience that at this threshold, the component validation meaningfully distinguishes systematic errors from acceptable process variation. 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. Applying these checks to the system on the 269 test cases results in the following differences by component:

Table S5: Component-Level Benchmark Checks of Advanced Model ($N = 269$)

Validation category	**Description of review approach**	**Share within ref. bounds**
Total kgCO2e	Check whether the total emissions are within a factor of 2 of a documented 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 documented 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 documented reference.	99%

The input rates pass less frequently than the other categories. This likely reflects the fact that input rates (material quantities and energy) can vary widely depending on the specific production process and facility, making accurate reference values difficult to find. Primary data may be most warranted for these parameters.

Datasets: The 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 (EPD International, 2024) and other published EPD databases such as Smart EPD (Smart EPD, 2026), 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 S6: 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

Criterion 2.2: Learning Capability **Criterion:** Demonstrated improvement in accuracy, quality indicators, and reviewability as the system is improved or additional information becomes available.

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. There are many potential approaches to demonstrate a useful learning capability. It is related to uncertainty and will depend on how uncertainty is incorporated into and propagated through the system. Further work on the usefulness and best practices for decisive corporate action would be valuable to refine the approach.

Approach in this system: The system never trains 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 the 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.

The initial AI-powered production graphs in August 2024 appeared plausible but had a high error rate upon systematic evaluation—illustrating why structured evaluation criteria, rather than surface-level inspection, are necessary for AI-CF systems. Through iterative refinement of context, validation checks, and system architecture, error rates have decreased to the levels reported in Criterion 2.1.

Criterion 2.3: Uncertainty Indication (material-level) **Criterion:** Indication of confidence level, uncertainty ranges, sensitivity, or other diagnostic uncertainty indicators for key datapoints and assumptions.

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 a plethora of 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, which Henriksson et al. (2015) suggests is required for PCFs to be used to make decisions. 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 (Cullen et al., 2024). A key goal is to make a useful and actionable set of results rather than a mathematically rigorous theoretical model.

Approach in this system: The team has experimented with several ways of displaying, calculating, and propagating uncertainty. The way that the model is set up includes "markets" which may

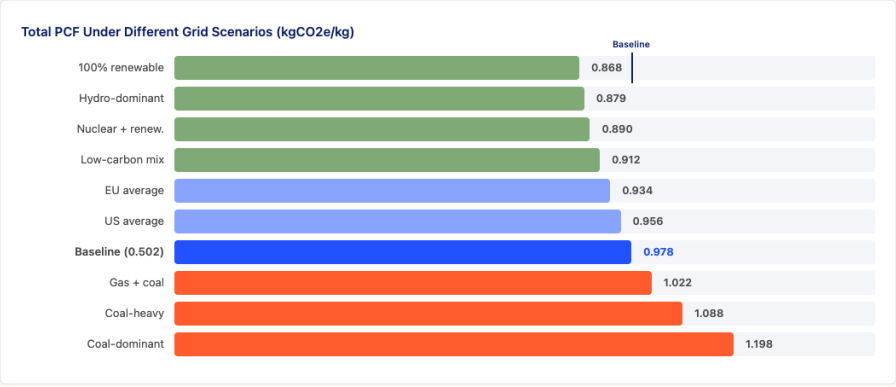
be fulfilled by more than one potential material, geography, or manufacturing approach; these approaches are weighted by the probability of their occurrence given the supply chain of the described input material. When a user has primary data, they can then edit these markets to collapse the uncertainty. 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.

In addition, we have developed sensitivity analysis capabilities that allow users to understand how key assumptions affect results. For example, electricity intensity is a major driver of emissions for many manufactured products, and regional electricity grid mixes can vary by an order of magnitude. The system can display how the total product carbon footprint changes under different electricity intensity assumptions, helping users identify where primary data collection would most reduce uncertainty.

3. Sensitivity Analysis: Grid Emissions Intensity

This analysis holds electricity consumption constant at 0.22 kWh/kg and varies the effective grid emissions factor across a range of representative scenarios. The non-electricity portion of the PCF (0.868 kgCO2e/kg) is held constant.

Scenario	Grid EF (kgCO2e/kWh)	Elec. Emissions (kgCO2e)	Total PCF (kgCO2e/kg)	Δ from Baseline	Δ%
100% renewable	0.000	0.000	0.868	−0.111	−11.3%
Hydro-dominant (Norway, Iceland)	0.050	0.011	0.879	−0.100	−10.2%
Nuclear + renewables (France, Sweden)	0.100	0.022	0.890	−0.088	−9.0%
Low-carbon mix (Canada)	0.200	0.044	0.912	−0.066	−6.8%
EU average	0.300	0.066	0.934	−0.045	−4.5%
US average	0.400	0.088	0.956	−0.022	−2.3%
Baseline (current model)	**0.502**	**0.111**	**0.978**	—	—
Gas + coal mix	0.700	0.154	1.022	+0.044	+4.5%
Coal-heavy (India, China avg.)	1.000	0.220	1.088	+0.110	+11.2%
Coal-dominant (Mongolia, South Africa)	1.500	0.330	1.198	+0.220	+22.5%



Key finding: The total range of grid intensity sensitivity is **0.330 kgCO2e/kg** (from 0.868 to 1.198), representing a 33.7% spread around the baseline. Moving from the current grid mix to the US average grid (0.400 kgCO2e/kWh) would yield a modest 2.3% reduction, suggesting the current model already assumes a moderately clean grid. Relocating manufacturing to a hydro-dominant grid would save 10.2%.

Figure S7: Example sensitivity analysis showing the effect of electricity grid emissions intensity assumptions on total product carbon footprint. The analysis holds electricity consumption constant and varies the grid emission factor across representative scenarios, demonstrating the range of outcomes and helping users identify where primary data collection would most reduce uncertainty.

Further work is needed on how to present uncertainty information effectively to non-specialist users.

Criterion 2.4: Key Data Sources (system-level) **Criterion:** Documentation of major data sources and databases used.

Approach in this system: 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.

Figure S8: 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.

Criterion 2.5: AI Methodology **Criterion:** High-level explanation of the AI approaches used to generate insights beyond simple mapping.

Approach in this system: The system architecture and agent pipeline stages are described in Section 2.2 of this resource.

Criterion 2.6: Summary Methodology (calculation approach) **Criterion:** High-level explanation of calculation approach.

Approach in this system: Section 2 (System Architectures) and the Case 2 criteria assessments in Section 4.2 of this resource provide these narratives.

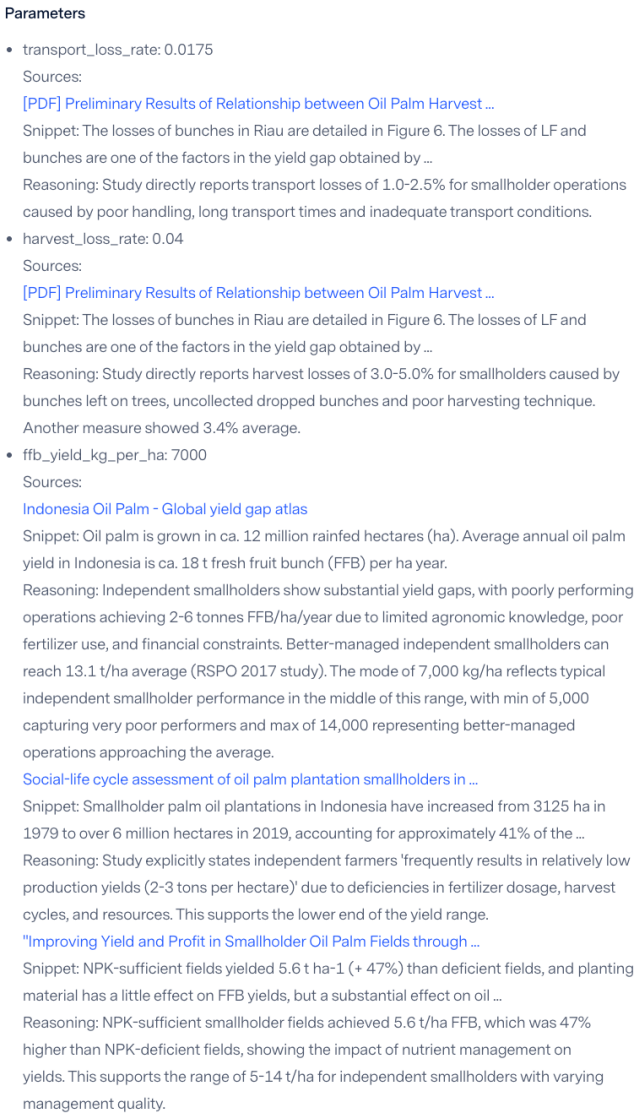


Figure S9: Example output of the advanced modeling system: AI-researched parameters with citations for independent smallholder oil palm cultivation. Each numeric parameter (e.g., transport loss rate, harvest loss rate, FFB yield) includes web-sourced references with relevant snippets and reasoning chains explaining how the value was derived. The review status indicator shows that all validation checks passed.

Criterion 2.7: Limitations Statement **Criterion:** Transparent communication of system limitations and appropriate use cases.

Approach in this system: The system's intended use cases and boundary conditions are described in this resource and the main paper. AI-generated assumptions are documented with source citations at each node. The sustainability advisory team communicates system limitations and appropriate use cases to users.

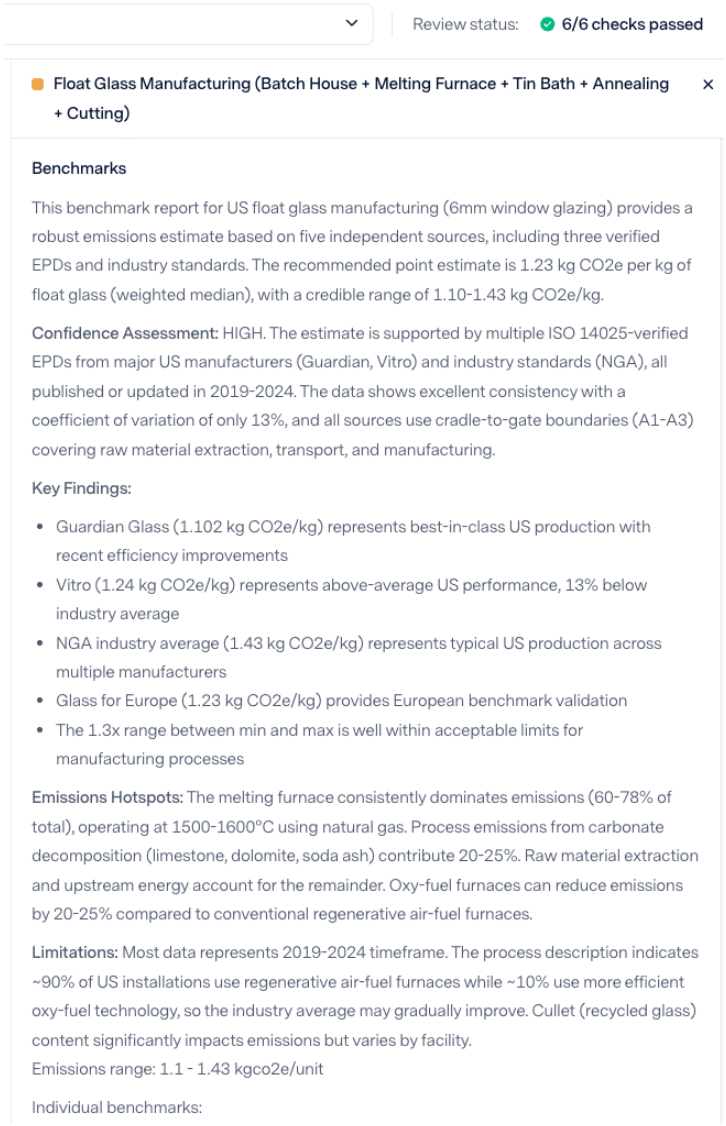


Figure S10: Example output of the advanced modeling system: benchmark report for float glass manufacturing. The system compiles independent emissions benchmarks from verified EPDs and industry standards, provides a confidence assessment, identifies emissions hotspots, and documents limitations. The reported point estimate and credible range are derived from multiple sources.

Criterion 2.8: Major Assumptions (material-level) **Criterion:** Recording of significant assumptions that drive results.

Approach in this system: In addition to the significant assumptions that drive results, all other assumptions are also recorded per node, including yields, locations, and process choices.

Criterion 2.9: Hotspot Identification (material-level) **Criterion:** Clear indication of major emissions sources and improvement opportunities.

Approach in this system: The production graph interface displays emissions-weighted nodes that allow users to identify dominant emissions contributors and trace hotspots to specific assumptions

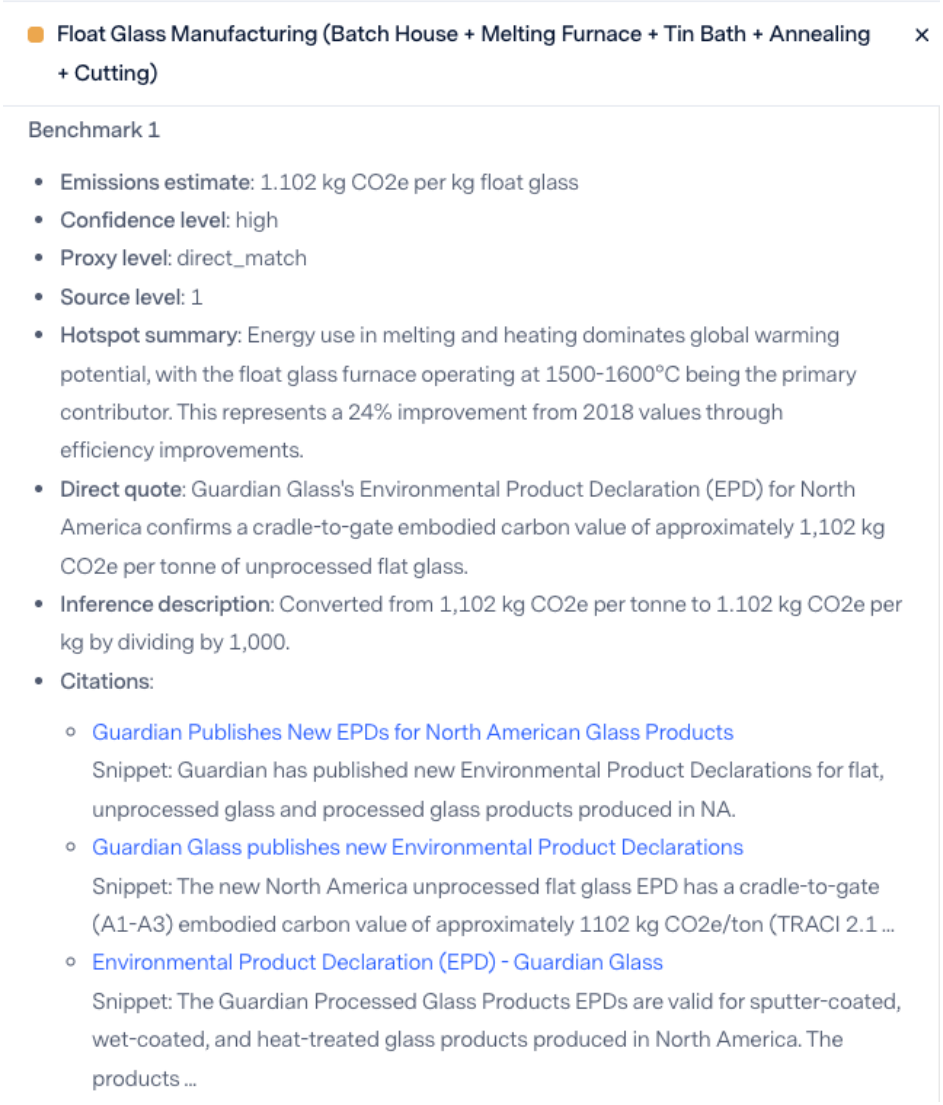


Figure S11: Example output of the advanced modeling system: detailed individual benchmark entry. Each benchmark includes the emissions estimate, confidence and proxy levels, a hotspot summary, direct quotes from source documents, inference descriptions explaining unit conversions or adjustments, and full citations with links to original sources.

or process parameters (Figure S12).

At the material portfolio level, hotspots can be aggregated and compared across product categories (Figure S13).

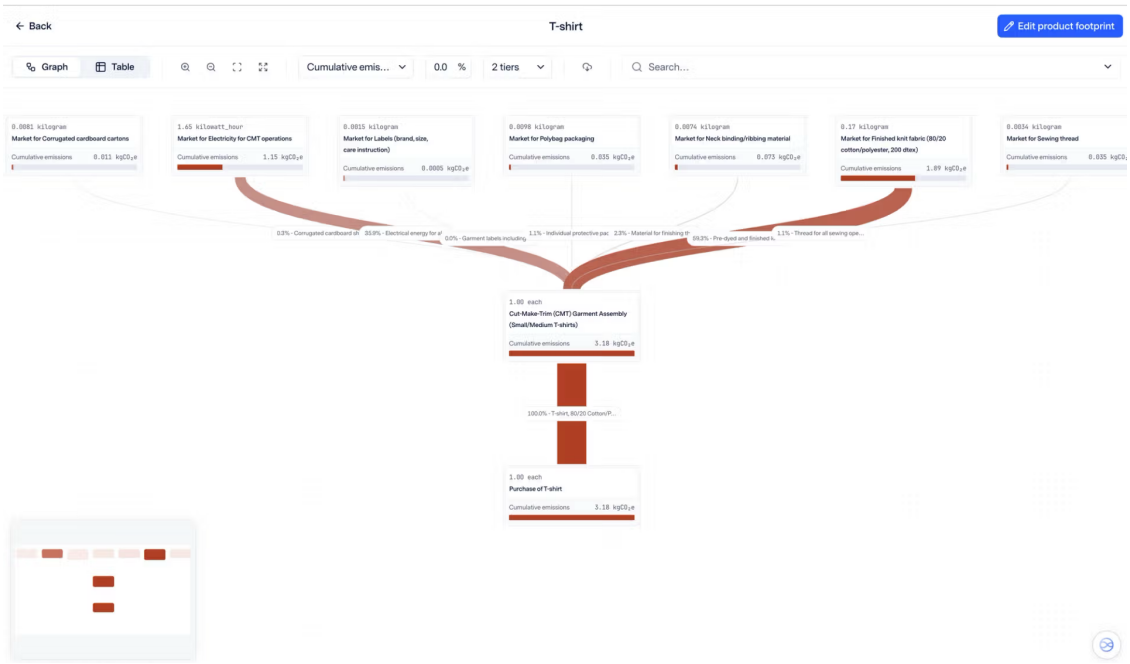


Figure S12: Example user interface view of a production graph

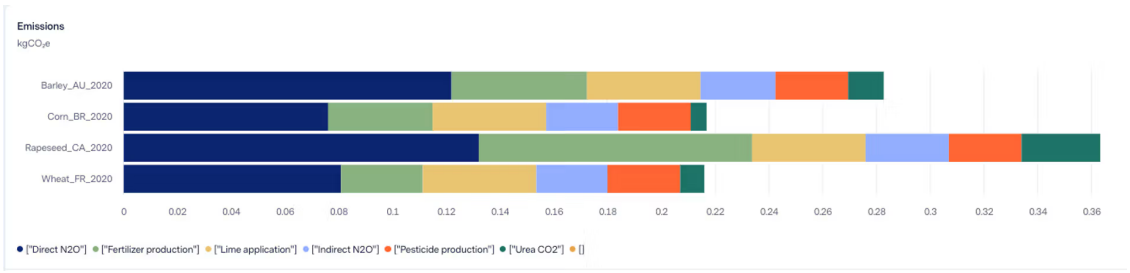


Figure S13: Example comparison BI view for row crop models

5 Third-Party Review

5.1 Review Process

To further validate the Advanced Modeling System's outputs, an independent third-party environmental firm has been engaged to review a sample of AI-generated production graphs. The reviewers are experienced LCA practitioners independent of Watershed. The review evaluates whether the system's outputs—including process decomposition, parameter estimates, source citations, and emissions calculations—are reasonable and methodologically sound from an LCA perspective.

The review is currently underway; quantitative inter-rater statistics comparing the third-party reviewer's assessments to the Graph Reviewer's automated assessments will be reported in a follow-up study. Preliminary results are encouraging: from a sample of 20 AI-generated production graphs that included 4 planted errors (introduced without informing the reviewer), the third-party expert correctly identified all 4 planted errors. Additionally, for 5 production graphs where Watershed's

customers independently requested third-party review, the expert's assessments concurred with the Graph Reviewer's automated ratings in all cases. The review rubric used by the third-party reviewer is documented in Section 5.2.

5.2 Review Rubric

The third-party reviewer evaluates each production graph against a structured rubric, designed to assess whether a production graph is useful and methodologically sound rather than exact relative to a specific standard. The rubric uses a "factor of two" threshold throughout: the AI-estimated emissions should be at least half and less than double a documented reference value. Methodology choices are standardized: cradle-to-gate boundary, GWP100, cut-off allocation, biogenic CO_2 treated as zero, and ecoinvent v3.11 default land-use-change emission factors.

Check 1: Overall emissions reasonableness. Are the cradle-to-gate emissions within a factor of two of a documented reference value? If yes, the reviewer documents the reference value and source and the review concludes. If no, the reviewer documents the discrepancy and continues to deeper checks. An answer of "it depends" is not permitted; if information is missing or uncertain, the reviewer errs on the side of continuing the review. Figure S14 shows the table view used to assess overall emissions.

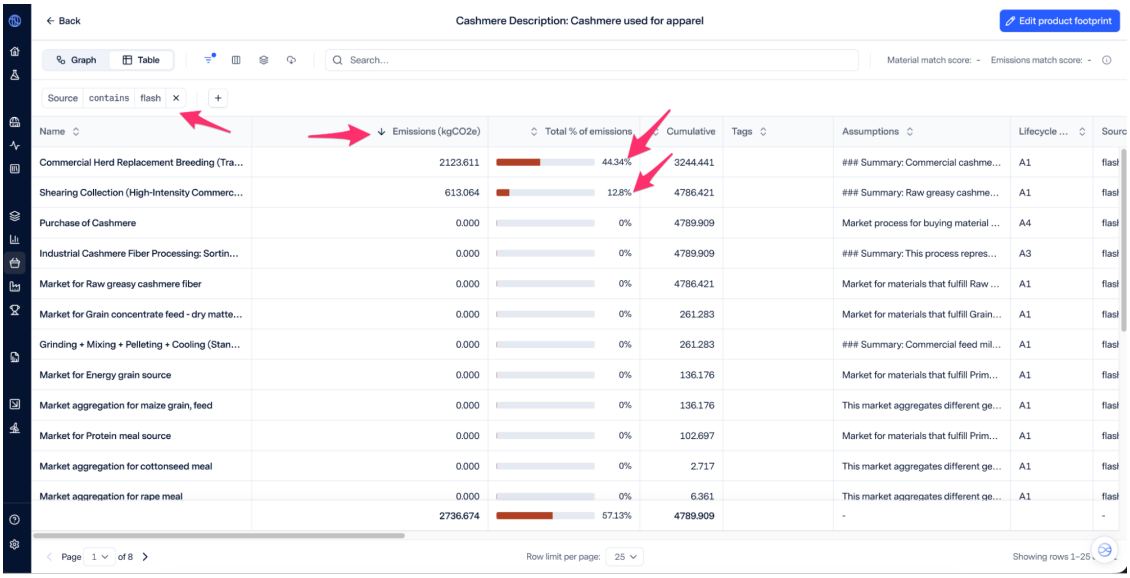


Figure S14: Table view of a production graph showing emissions, cumulative contributions, and node types. The reviewer uses this view to assess overall emissions reasonableness against documented reference values.

Check 2: Hotspot analysis. Hotspots are defined as the nodes with the largest cumulative emissions that are at least three tiers deep in the production graph and are distinct from each other (i.e., one is not a subset of another).

- *Check 2.1: Hotspot emissions.* Are the top three hotspots' cumulative emissions each within a factor of two of a documented reference? Figure S15 illustrates hotspot identification in a production graph for calcium citrate powder, with the three largest emissions contributors annotated.

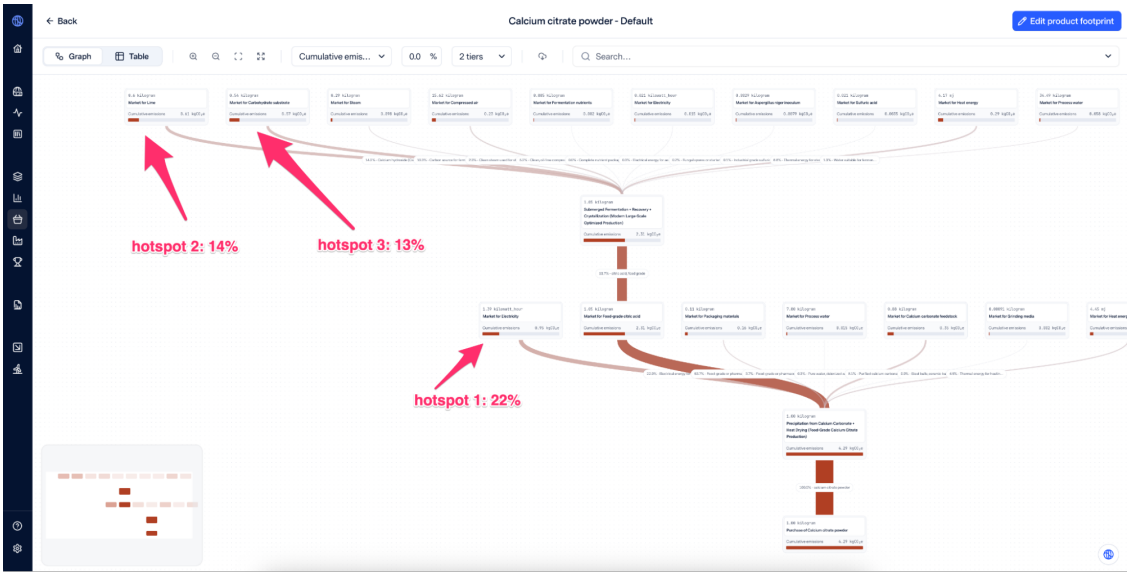


Figure S15: Graph view of a production graph with the top three emissions hotspots annotated, showing their cumulative emissions contributions (22%, 14%, and 13% of total).

- *Check 2.2: Hotspot quantities.* For each hotspot, is the key input quantity (e.g., kg of input per kg of output) within a factor of two of a documented source? For example, if the graph shows 1.05 kg of food-grade citric acid as input to a precipitation process to output 1 kg of calcium citrate, the reviewer checks whether this input rate is plausible. Figure S16 shows a node-level detail view used to inspect process parameters and input quantities.

Figure S16: Node-level detail view showing process inputs, parameters, and cumulative emissions for a specific production step (Precipitation from Calcium Carbonate + Heat Drying).

Check 3: Direct emissions plausibility. If the direct emissions of any single node exceed 5% of overall emissions, are those direct emissions within a factor of two of a documented reference? This

check is particularly useful for agricultural products where good reference data is often available showing the breakdown of emissions sources (e.g., enteric fermentation, fertilizer application, land use change).

Figure S17: Example of a reviewer comment documented on a production graph node. The reviewer notes that overall emissions do not appear reasonable and cites a specific EPD as the documented reference value.

What constitutes a documented reference value A documented reference value is a verifiable, citable source that can validate whether an AI-generated product footprint is reasonable. The reference should be for the same or closely related material or process, from a reputable source, and used with appropriate judgment about similarity and adjustments. Sources, in rough order of preference:

1. **Environmental Product Declarations (EPDs):** Product-specific, third-party verified LCA data. Allowed if product specifications match. Not allowed: using an EPD for polycarbonate to validate polyethylene terephthalate without valid justification that the manufacturing processes and raw materials are similar.

2. **Peer-reviewed LCA literature:** Published studies in journals such as *Journal of Cleaner Production* or *The International Journal of Life Cycle Assessment*. The range or midpoint may be used as a reference. Not allowed: cherry-picking a single value that pertains to a sub-process rather than the overall production process.

3. **Industry benchmarks from trade associations:** Sector-specific average data from reputable industry groups (e.g., World Steel Association benchmarks). Appropriate for common commodity materials. Not allowed: applying generic benchmarks to specialty alloys or highly customized products.

4. **Secondary databases (e.g., GREET, GaBi, ecoinvent):** Well-documented background data. Allowed when the technology mix is appropriate. Not allowed: using reference values more than three decades old if the technology mix has changed significantly.

5. **Industry reports from consultancies:** Reports from firms providing broad emissions ranges. Appropriate as sanity checks when better data is unavailable. Not allowed: treating broad ranges as precise validation.

6. **Company sustainability reports:** Self-reported data from manufacturers. Allowed if the company is reputable and data appears reasonable. Not allowed: using marketing materials that cite only "reduced emissions" without absolute values.

Proxy references when an exact match is unavailable When an exact reference is not available, the reviewer may use a proxy reference with clearly documented limitations. Four scenarios are defined:

1. *Similar material, same process:* A reference for a different grade of the same material (e.g., technical-grade citric acid used as proxy for food-grade) is allowed if the production processes are identical or very similar, with the limitation noted.

2. *Related material, same product category:* A reference for a related material in the same category (e.g., calcium carbonate for calcium citrate) is allowed if the reviewer can reason about the additional processing steps and adjust accordingly. The adjusted range is used as a plausibility check, not a direct comparison.

3. *Similar process, different material:* A process-level reference (e.g., fermentation energy use from citric acid production applied to PLA fermentation) is allowed with caution for validating process-level parameters, but not the full product footprint.

4. *Upstream components:* Summing component-level references (e.g., FR-4 laminate + electronic components + assembly for a printed circuit board) is allowed if data is available for all major parts. Not allowed if a major component is missing from the sum.

Allowable transformations and adjustments The following transformations are permitted: unit conversions, geographic scaling for energy-intensive processes based on grid carbon intensity differences, process yield adjustments (scaling proportionally), purity or grade differences with documented reasoning, and material density conversions with clear density data.

The following transformations are not permitted: changing the process technology type (e.g., adjusting blast furnace steel to estimate electric arc furnace steel), cross-sector analogies (e.g., pharmaceutical emissions for semiconductor validation), adjustment factors exceeding 2× in either direction (indicating the reference is too dissimilar), or combining incompatible scopes (e.g., cradle-to-grave references to validate cradle-to-gate estimates).

The rubric dimensions align with the Graph Reviewer's automated validation categories (Table S5), enabling direct comparison of human and automated assessments.

References

N Adams and K Allacker. Parameter sensitivity and data uncertainty assessment of the cradle-to-gate environmental impact of state-of-the-art passive daytime radiative cooling materials. *Environmental Sciences Europe*, 37(1):53, 2025. doi: 10.1186/s12302-025-01093-x. URL https://doi.org/10.1186/s12302-025-01093-x.

Bharathan Balaji et al. Emission factor recommendation for life cycle assessments with generative AI. *Environmental Science & Technology*, 59(18):9113–9122, 2025. doi: 10.1021/acs.est.4c12667. URL https://doi.org/10.1021/acs.est.4c12667.

Natanael Bolson, Luke Cullen, and Jonathan Cullen. A robust framework for estimating theoretical minimum energy requirements for industrial processes. *Energy*, 322:135411, 2025. doi: 10.1016/j.energy.2025.135411. URL https://doi.org/10.1016/j.energy.2025.135411.

Martin Courtat, Patrick James Joyce, Sarah Sim, et al. Ensuring consistent data quality for environmental rating ecolabels with representative secondary datasets. *The International Journal of Life Cycle Assessment*, 30:2780–2793, 2025. doi: 10.1007/s11367-025-02545-5. URL https://doi.org/10.1007/s11367-025-02545-5.

Luke Cullen et al. Reducing uncertainties in greenhouse gas emissions from chemical production. *Nature Chemical Engineering*, 1(4):311–322, 2024. doi: 10.1038/s44286-024-00047-z. URL https://doi.org/10.1038/s44286-024-00047-z.

Andrew Dumit, Krishna Rao, Shaena Ulissi, Steven Watson, Jacob Feintzeig, P. James Joyce, and Shuhan Bao. Quality-aware automation for LCI database mapping. *Journal of Industrial Ecology*, 2026a. doi: 10.21203/rs.3.rs-9285034/v1. URL https://doi.org/10.21203/rs.3.rs-9285034/v1. Preprint.

Andrew Dumit, Krishna Rao, Shaena Ulissi, Steven Watson, P. James Joyce, Shuhan Bao, Jacob Feintzeig, and Sangwon Suh. Owl: Separating generation from evaluation to detect plausible failures in lifecycle inventory mapping. In *ICLR 2026 Workshop on Agentic AI in the Wild: From Hallucinations to Reliable Autonomy*, 2026b. URL https://openreview.net/forum?id=z9ZB5sqoLM.

Andrew Dumit et al. ATLAS: A spend classification benchmark for estimating scope 3 carbon emissions. In *Climate Change AI, NeurIPS 2024*, 2024. URL https://www.climatechange.ai/papers/neurips2024/70.

ecoinvent Association. ecoinvent database version 3.11, 2024. URL https://ecoinvent.org.

EPD International. EPD library, 2024. URL https://www.environdec.com/library.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Rolf Frischknecht, Harpa Birgisdottir, Chang U. Chae, Thomas Lützkendorf, Alexander Passer, Erik Alsema, et al. Comparison of the environmental assessment of an identical office building with national methods. In *IOP Conference Series: Earth and Environmental Science*, volume 323, page 012037, 2019. doi: 10.1088/1755-1315/323/1/012037. URL https://doi.org/10.1088/1755-1315/323/1/012037.

M. D. C. Gelowitz and J. J. McArthur. Comparison of type III environmental product declarations for construction products: Material sourcing and harmonization evaluation. *Journal of Cleaner Production*, 157:125–133, 2017. doi: 10.1016/j.jclepro.2017.04.133. URL https://doi.org/10.1016/j.jclepro.2017.04.133.

GHG Protocol Technical Working Group. Scope 3 presentation, 2025. URL https://ghgprotocol.org/sites/default/files/2025-06/S3-Meeting3-Presentation-20250529_1.pdf.

Patrik J. G. Henriksson et al. Product carbon footprints and their uncertainties in comparative decision contexts. *PLoS ONE*, 10(3):e0121221, 2015. doi: 10.1371/journal.pone.0121221. URL https://doi.org/10.1371/journal.pone.0121221.

Frederik Konradsen et al. Same product, different score: how methodological differences affect EPD results. *The International Journal of Life Cycle Assessment*, 29(2):291–307, 2024. doi: 10.1007/s11367-023-02246-x. URL https://doi.org/10.1007/s11367-023-02246-x.

Brandon Kuczenski et al. Prototypes for automating product system model assembly. *The International Journal of Life Cycle Assessment*, 26(3):483–496, 2021. doi: 10.1007/s11367-021-01870-9. URL https://doi.org/10.1007/s11367-021-01870-9.

Christoph J. Meinrenken et al. The carbon catalogue, carbon footprints of 866 commercial products from 8 industry sectors and 5 continents. *Scientific Data*, 9(1):87, 2022. doi: 10.1038/s41597-022-01178-9. URL https://doi.org/10.1038/s41597-022-01178-9.

Angelica Mendoza Beltran, Valentina Prado, David Font Vivanco, et al. Quantified uncertainties in comparative life cycle assessment: What can be concluded? *Environmental Science & Technology*, 52(4):2152–2161, 2018. doi: 10.1021/acs.est.7b06365. URL https://doi.org/10.1021/acs.est.7b06365.

Yaning Qiao, Xia Wen, Shirui Liu, Songtao Lv, and Liang He. Stochastic analysis for comparing life cycle carbon emissions of hot and cold mix asphalt pavement systems. *Resources, Conservation and Recycling*, 212:107881, 2025. doi: 10.1016/j.resconrec.2024.107881. URL https://doi.org/10.1016/j.resconrec.2024.107881.

Smart EPD. Smart EPD database, 2026. URL https://www.smartepd.com.

Aditya Taparia, Ransalu Senanayake, Kowshik Thopalli, and Vivek Narayanaswamy. The anatomy of uncertainty in LLMs. In *ICLR 2026 Workshop on I Can't Believe It's Not Better (ICBINB)*, 2026. URL https://openreview.net/forum?id=0GYclsjLUb.

Textile Exchange. Life cycle assessment for cotton: Model comparison, 2026. URL https://textileexchange.org/knowledge-center/documents/life-cycle-assessment-for-cotton-model-comparison/.

Bo P. Weidema and Marianne S. Wesnæs. Data quality management for life cycle inventories—an example of using data quality indicators. *Journal of Cleaner Production*, 4(3–4):167–174, 1996. doi: 10.1016/S0959-6526(96)00043-1. URL https://doi.org/10.1016/S0959-6526(96)00043-1.

Gregor Wernet, Christian Bauer, Bernhard Steubing, Jürgen Reinhard, Emilia Moreno-Ruiz, and Bo Weidema. The ecoinvent database version 3 (part I): overview and methodology. *The International Journal of Life Cycle Assessment*, 21(9):1218–1230, 2016. doi: 10.1007/s11367-016-1087-8. URL https://doi.org/10.1007/s11367-016-1087-8.