Sustainable Production and Consumption 67 (2026) 428–444

Contents lists available at ScienceDirect

Sustainable Production and Consumption

journal homepage: www.elsevier.com/locate/spc

Smarter LCA: Using large language models to predict manufacturing processes and their uncertainties☆

Alexandra Belyaeva [a,b,*], Christoph Helbig [b], Torsten Hummen [a]

[a] Robert Bosch GmbH, Bosch Research, Robert-Bosch-Campus 1, 71272, Renningen, Germany
[b] Ecological Resource Technology, University of Bayreuth, Universitätsstr. 30, Bayreuth, Germany

ARTICLE INFO

Editor: Dr Rodrigo Salvador

Keywords:
Prospective Life Cycle Assessment
Large language model
Mechanical engineering
Sustainability
Uncertainty

ABSTRACT

Prospective Life Cycle Assessment offers significant potential to guide sustainable product development by influencing design in the early stages, when the flexibility for impactful changes is greatest. However, its practical application is often constrained by the limited availability and quality of foreground life cycle inventory data. For mechanically engineered products, key manufacturing-related decisions that determine material waste and energy demand are frequently unavailable, and while expert estimations can fill these gaps, this approach is time-consuming and limited by specialist availability. To overcome this challenge, we present an automated, uncertainty-aware method for generating screening-level missing manufacturing foreground data. Our approach leverages large language models, starting with a fine-tuned Bidirectional Encoder Representations from Transformers model that predicts part group classifications directly from short part names. These predicted groups are then mapped to specific manufacturing process datasets in ecoinvent 3.9.1 using a deterministic categorical mapping rule, with the large language model used to extract manufacturing parameters from process descriptions. A key feature of our method is the automated generation of pedigree matrix scores to quantify prediction confidence and mapping quality. This uncertainty is then propagated using Monte Carlo simulation, providing transparent decision support. An illustrative industrial case study shows that the pipeline can predict plausible proxies for manufacturing processes from limited textual data, enabling a more robust and timelier prospective Life Cycle Assessment when crucial manufacturing information remains unknown.

1. Introduction

Prospective Life Cycle Assessment (pLCA) is a forward-looking approach to environmental assessment that models emerging or early-stage technologies under future, full-scale deployment conditions (Arvidsson et al., 2018; Cucurachi et al., 2018). Its goal is to inform sustainable design choices before technologies become locked into mass production. It plays a crucial role in the early stages of product development, when designers have the greatest flexibility to influence environmental outcomes (Bhander et al., 2003). Incorporating Life Cycle Assessment (LCA) at this stage enables optimization of design choices to reduce environmental burdens and can also lead to cost savings by

Abbreviations: AC, Aluminum cast parts; AI, Artificial Intelligence; API, Application Programming Interface; BART, Bidirectional and Auto-Regressive Transformers; BERT, Bidirectional Encoder Representations from Transformers; BOM, Bill of Materials; CAD, Computer-Aided Design; CO_2-eq, Carbon dioxide equivalent; CPC, Central Product Classification; DF, Drop forged parts; DIN, Deutsches Institut für Normung (German Institute for Standardisation); ECE, Expected Calibration Error; F1, Harmonic mean of precision and recall; GHG, Greenhouse Gas; GLO, Global (ecoinvent geography); GPT, Generative Pretrained Transformer; GSD, Geometric Standard Deviation; GWP, Global Warming Potential; GWP100, Global Warming Potential over a 100-year time horizon; IPCC, Intergovernmental Panel on Climate Change; ISIC, International Standard Industrial Classification; LCA, Life Cycle Assessment; LCI, Life Cycle Inventory; LCIA, Life Cycle Impact Assessment; LLM, Large Language Model; MC, Monte Carlo; ML, Machine Learning; MRL, Manufacturing Readiness Level; PCF, Product Carbon Footprint; PL, Parts made from thermoplastics; pLCA, Prospective Life Cycle Assessment; pLCI, Prospective Life Cycle Inventory; RER, Europe (ecoinvent geography); SD, Stamped, drawn, bent parts; STEP, Standard for the Exchange of Product model data; T5, Text-to-Text Transfer Transformer; TF-IDF, Term Frequency-Inverse Document Frequency; TR, Turned parts without finishing; TRf, Turned parts with finishing; TRL, Technology Readiness Level; UUID, Universally Unique Identifier.

☆ This article is part of a Special issue entitled: 'AI for Sustainability' published in Sustainable Production and Consumption.

* Corresponding author at: Robert Bosch GmbH, Bosch Research, Robert-Bosch-Campus 1, 71272, Renningen, Germany.
E-mail addresses: alexandra.belyaeva@de.bosch.com (A. Belyaeva), christoph.helbig@uni-bayreuth.de (C. Helbig), torsten.hummen@de.bosch.com (T. Hummen).

https://doi.org/10.1016/j.spc.2026.07.008
Received 1 April 2026; Received in revised form 5 July 2026; Accepted 6 July 2026
Available online 9 July 2026

avoiding the need for late-stage redesigns (Ramani et al., 2010; Schöggl et al., 2017).

The data-intensive nature of the LCA method makes it especially well-suited for assessing existing products with abundant data (Wright et al., 2024). However, its application to emerging products is challenging due to a lack of comprehensive data, particularly regarding foreground life cycle inventory (LCI) data gaps (Bergerson et al., 2020; Thonemann et al., 2020). In the manufacturing phase of mechanically engineered products, these gaps often involve missing information, such as details on manufacturing processes, associated material waste and energy use, and auxiliary inputs, including consumables.

Missing manufacturing-process information is consequential because the selected process or process sequence affects both process energy demand and the gross material input required to obtain a given net part mass. Process-specific scrap, yield losses, and reprocessing can cause material inputs to exceed the mass retained in the final part; consequently, foreground LCI models based only on net part mass may underestimate upstream material requirements and associated environmental burdens (Allwood et al., 2011).

In practice, obtaining this information is difficult because modern manufacturing occurs within globalized, highly distributed value chains, where key production stages are performed by specialized suppliers. As a result, many manufacturing processes fall outside the direct control of the final producer and are reported as Scope 3 emissions under the GHG Protocol (World Resources Institute and World Business Council for Sustainable Development, 2011). These supplier-related emissions often constitute a major component of a company's carbon footprint (Hertwich and Wood, 2018), yet detailed information on process routes and associated material flows is typically unavailable. This lack of transparency directly contributes to the foreground LCI data gaps that hinder precise modeling in pLCA.

Existing approaches to foreground LCI data gap filling include expert elicitation, rule-based systems, geometry-based machine learning, and LLM-assisted workflows, as reviewed in Section 2. However, two gaps remain relevant for early-stage assessments of mechanically engineered products. First, procurement-stage textual data, such as short part names, remain underexplored as a primary source for inferring manufacturing-process information, although such data are often available earlier and at larger scale in industrial procurement systems than supplier-specific process data or detailed CAD models. Second, automated foreground-data generation is commonly evaluated as a prediction or mapping task, whereas mechanisms for converting model-derived confidence into pedigree-matrix-compatible data-quality indicators remain underdeveloped.

This study addresses these gaps through an operational, uncertainty-aware pipeline to generate manufacturing-process proxies from procurement-stage part names. The novelty lies in the integration of three methodological elements. First, a hierarchical fine-tuned Bidirectional Encoder Representations from Transformers (BERT) classifier predicts manufacturing-relevant part groups from short textual part descriptions. Second, the predicted part groups are mapped to ecoinvent 3.9.1 manufacturing process proxies using a constrained, auditable mapping procedure based on structured manufacturing parameters. Third, prediction probability and proxy-mapping quality score are translated into pedigree matrix reliability and technological-correlation scores, allowing the uncertainty introduced by automated foreground-data generation to be propagated through Monte Carlo simulation.

The research questions addressed in this study are as follows:

- How can LLMs be used to infer manufacturing-relevant part groups from procurement-stage part names, in order to fill foreground LCI data gaps in early-stage assessments of mechanically engineered products?
- How can the confidence of such LLM-based inference be systematically translated into pedigree matrix indicators for uncertainty propagation in pLCA?

The method is developed for mechanically engineered products, whose manufacturing routes are represented in the training data and in the ecoinvent database. In principle, the same workflow could be applied to other part types, provided that a sufficiently large and consistently labeled set of part names and a compatible proxy database are available. The resulting pipeline is demonstrated on an industrial case-study product to illustrate the end-to-end workflow from bill of materials (BOM) input to manufacturing-process proxy assignment, pedigree-based uncertainty propagation, and product carbon footprint (PCF) output. The case study demonstrates the operational feasibility of the approach.

Accordingly, the contribution of this paper is an operational methodological integration: a BERT-based part-group classifier, a constrained ecoinvent proxy-mapping procedure, and pedigree-based uncertainty propagation are combined into a single pipeline for screening-level pLCA. The integration is developed and demonstrated within a single company's internal procurement-classification ecosystem. It is therefore best understood as a company-specific proof-of-concept automation pipeline for a defined industrial setting, rather than as a method already validated for general prospective LCA practice. The proposed pipeline supports screening-level pLCA, operationalized here as prospective cradle-to-gate PCF, when supplier-specific manufacturing data are unavailable; it does not replace expert review, supplier-specific data collection, or full prospective scenario modeling for background systems. The mapping of model-derived confidence metrics onto pedigree-matrix indicators is a pragmatic operational reinterpretation that enables the uncertainty associated with automated foreground assumptions to be represented within an established LCA uncertainty workflow. It does not establish a theoretical or conceptual equivalence between machine-learning confidence and pedigree-based data quality.

The remainder of this paper is organized as follows. Section 2 (Literature Review) reviews existing methods for pLCA and LLM applications. Section 3 (Methods) presents the proposed method and its implementation as an integrated pipeline. Section 4 (Results) reports results on the BERT model training, LLM-assisted mapping to ecoinvent datasets, and a mechanically engineered case-study product. Section 5 (Discussion) discusses the findings and their limitations. Section 6 (Conclusions) presents the main findings and outlines directions for future research.

2. Literature review

Future-oriented LCA studies are often referred to as prospective, ex-ante, or anticipatory LCA. Arvidsson (2023) distinguishes these concepts based on three key dimensions: real time, maturity, and causality. LCA that models environmental impacts into the future is generally referred to as pLCA. Ex-ante LCA is a specific type of pLCA that assesses technologies at early stages of development but models them as fully mature in the future. Process upscaling refers to translating laboratory- or pilot-scale material and energy flows to industrial-scale conditions (Erakca et al., 2024). Anticipatory LCA shares similarities with ex-ante LCA but additionally incorporates stakeholder engagement. In this study, the term pLCA is used to reflect the absence of technology upscaling in the scope and to follow the terminology recommendations made by Arvidsson et al. (2023).

2.1. Methods for prospective life cycle assessment

Assessing the environmental impact of products during early development can substantially reduce future environmental burdens (Moni et al., 2020). Early-stage design retains high flexibility, allowing for meaningful design modifications with relatively low cost (Bergerson et al., 2020). The lower the product readiness level, the greater the potential to choose more sustainable technological pathways. pLCA supports the comparison of design alternatives and enables the evaluation of the environmental performance of emerging technologies

relative to mature ones (Bergerson et al., 2020; van der Giesen et al., 2020). It is also instrumental in identifying environmental hotspots (Moni et al., 2020). Because design changes are less expensive early in development, pLCA helps guide decisions before technical lock-in occurs. For the industry, this allows identification of environmentally motivated cost reductions while avoiding or minimizing redesign. In this way, pLCA supports a first-time-right approach by reducing costly iteration loops.

The future-oriented perspective makes pLCA especially relevant in sectors with long development cycles. Applying pLCA during early design helps engineers select options that remain environmentally optimal under future conditions (Grenz et al., 2023). Recent reviews show rapid growth in pLCA applications: Nurdiawati et al. (2025) report that approximately 75% of reviewed pLCA case-study articles were published between 2018 and 2024, with applications spanning energy systems, batteries and energy storage, mobility, industrial manufacturing, waste management, biobased products, water technologies, and food systems. These examples show that pLCA is increasingly applied where early technology or design choices may have long-term environmental consequences.

Conducting a pLCA on a technology still under development introduces unique challenges beyond those of retrospective LCA. A systematic review by Thonemann et al. (2020) identifies three main categories of challenges: comparability, data, and uncertainty. Comparability refers to the meaningful comparison between emerging and incumbent technology. Because the technology is still in development, data availability and quality are often poor. Additionally, if only lab-scale data exist, scaling up those data to an industrial level is necessary (Erakca et al., 2024). Uncertainty is an inherent aspect of LCA in general, but its nature and implications differ significantly in pLCA. In retrospective LCA, uncertainties primarily stem from measurement error, variability in real-world systems, model structure, and methodological choices (Rosenbaum et al., 2018). These are often categorized by their source as parameter, model, scenario, or epistemic uncertainties (Rosenbaum et al., 2018). Uncertainty related to inaccurate, non-representative, or no inventory data in the LCI phase is defined as parameter uncertainty (Rosenbaum et al., 2018). In parallel, uncertainty can be categorized by its nature as epistemic (also called systematic) and aleatory (also called stochastic or random). Epistemic uncertainty arises from incomplete knowledge or data limitations and can, in principle, be reduced with more information. In contrast, aleatory uncertainty stems from inherent variability or randomness in a system and is generally irreducible (Tan et al., 2025).

In pLCA, uncertainty extends beyond quantifiable statistical measures (Olsen et al., 2018). As van der Giesen et al. (2020) emphasize, pLCA often deals with uncertainty about unknown futures and even "unknown unknowns" – factors or consequences not yet identified or understood at the time of assessment.

To ensure robust results, quantifying and propagating uncertainty is a crucial part of LCA, particularly the LCI phase. Methods for addressing uncertainty can be broadly categorized into quantitative, qualitative, and mixed approaches (Leroy and Froelich, 2010; Lloyd and Ries, 2007). Monte Carlo simulation is a quantitative method that is based on the propagation of uncertainty by repeatedly sampling from probability distributions assigned to input parameters, thereby generating a distribution of possible outcomes (Tan et al., 2025). Qualitative methods use tools such as the pedigree matrix to assess data quality. The pedigree matrix approach, used for the LCI phase (Leroy and Froelich, 2010) and first introduced in the LCA context by Weidema and Wesnæs (1996), is a matrix with data quality indicators that allow LCA practitioners to assess LCI data quality by assigning quality scores to five criteria: reliability, temporal, geographical, technological correlation, and completeness. These scores are subsequently translated into uncertainty factors (Ciroth et al., 2016). Mixed approaches combine qualitative assessments with quantitative techniques, such as propagation of pedigree matrix-derived uncertainties with Monte Carlo simulations (Tan et al., 2025).

A review by Nurdiawati et al. (2025) indicates that uncertainty analysis is still frequently absent in pLCA studies. Where uncertainty analysis is included, Monte Carlo simulation is reported as the most commonly used method and is often supplemented by pedigree-matrix-based qualitative data-quality assessment. This makes the pedigree matrix particularly relevant for automation: it is already compatible with established LCA uncertainty workflows, but its standard use relies on expert judgment (Ciroth et al., 2016).

In pLCA, addressing gaps in prospective life cycle inventory (pLCI) data is crucial for accurate environmental impact assessments. Data availability depends on the product's maturity level, which can be categorized as Technology Readiness Level (TRL) (Tzinis, 2015) or Manufacturing Readiness Level (MRL) (U.S. Department of Defense, 2025) (Table 1).

In the early stages of product development, meaning at low TRL and MRL, significant data gaps include unknown manufacturing processes, material waste percentages, production locations, and transportation details. This missing information poses challenges to precise environmental evaluations.

In the absence of data, a straightforward approach is to consult domain experts to estimate the missing LCI or upscale a system to the production level (Erakca et al., 2024; Tsoy et al., 2020). While effective, this approach can be resource-intensive and may not always be feasible due to constraints in expert availability.

Table 1
Production stages and respective Technology Readiness Levels and Manufacturing Readiness Levels (MRL) from (Erakca et al., 2024).

Production stages			Technology readiness level	Manufacturing readiness level
Laboratory	Conceptual	1	Basic principles observed and reported	Basic manufacturing implications identified
Pilot	development	2	Technology concept or application formulated	Manufacturing concepts identified
		3	Experimental and analytical critical function and characteristic proof of concept	Manufacturing proof of concept developed
		4	Component or breadboard validation in a laboratory environment	Capability to produce the technology in a laboratory environment
	Technology development	5	Component or breadboard validation in a relevant environment	Capability to produce prototype components in a production-relevant environment
Pilot		6	System or subsystem model or prototype demonstrated in a relevant environment	Capability to produce a prototype system or subsystem in a production-relevant environment
	Engineering development	7	System prototype demonstration in an operational environment	Capability to produce systems, subsystems, or components in a production representative environment
	Small-scale production	8	Actual system completed and "flight qualified" through test and demonstration	Pilot line capability demonstrated; ready to begin low-rate initial production
Industrial		9	Actual system "flight proven" through successful mission operations	Low-rate production demonstrated; capability in place to begin full-rate production
	Mass production	10	–	Full rate production demonstrated, and lean production practices in place

To overcome this issue, various approaches are employed to automate the LCI stage. Within the DfC-Industry project, a rule-based approach utilizing a decision tree was applied to estimate the primary manufacturing processes of mechanically engineered parts (Kusch et al., 2024). The decision tree input parameters were extracted from STEP files that describe the product model data for these parts. While the approach performed adequately for standard cases, it struggled with rare or non-standard parts because a simple decision tree was unable to represent the full complexity of manufacturing-process selection (Kusch et al., 2024).

Alternatively, ML models offer a more flexible and robust approach, capable of identifying patterns and predicting missing data with greater accuracy (Ben-David and Frank, 2009). However, the effectiveness of ML models depends on the availability of large, high-quality datasets for training, as well as careful model design and validation (Gennatas et al., 2020). Among the possible ML approaches to predict the manufacturing process of a part are models trained on data extracted from 3D files to classify parts by manufacturing processes (Zhao et al., 2020) or to recognize design features (Zhang et al., 2018), which can subsequently be used to infer suitable manufacturing processes. Wang and Rosen (2023) combine heat kernel signatures with convolutional neural networks to predict one of four selected manufacturing processes based solely on the shapes of mechanically engineered parts represented as polygonal mesh data. To address limitations in real-world training data, they augment their dataset with synthetically generated parts based on representative shape features for each process, while still relying on a substantial set of real components. These approaches demonstrate the potential of geometry-based ML for manufacturing process prediction; however, they rely on detailed 3D representations and large, specialized datasets. As a result, they are not directly applicable in situations where such geometric models are unavailable and sufficiently large training datasets are lacking.

2.2. Application of large language models in LCA

When only textual data are available, LLMs and related natural language processing methods can be used to estimate missing LCI values. Preuss et al. (2024) discuss LLM-related opportunities across the LCA workflow, including LCI data extraction, semantic matching, database querying, and accelerated extraction of inventory data from heterogeneous sources. Gkousis et al. (2026) review the broader field of machine learning, natural language processing, and LLM methods for LCI compilation, showing that these methods can support missing-data estimation, document-based information extraction, and matching foreground descriptions to background database entries.

These applications indicate that LLMs are particularly relevant in contexts where structured inventory data are incomplete or unavailable, but textual descriptions, such as material specifications, process descriptions, or part names, remain accessible. This is typically the case in early-stage product development, where detailed manufacturing information has not yet been finalized.

Recent academic work provides examples of how language models can support LCA automation. LLMs and related natural language processing methods have been used to assess product carbon footprints and construct inventories. AutoPCF, for example, proposes an LLM- and deep-learning-based framework for cradle-to-gate PCF modeling, covering emission inventory determination, activity-data collection, emission-factor matching, carbon-emission estimation, and verification (Luo et al., 2024a). Other work focuses on semantic matching between product or process descriptions and environmental datasets. Flamingo uses natural-language machine learning to identify environmental impact factors from text descriptions and constrains the matching space through sector-classification information (Balaji et al., 2023). FAULDIER further demonstrates LLM-assisted LCI construction by using LLMs to map non-standardized raw LCI entries to database-compatible process and elementary-flow names, while also highlighting remaining limitations related to mapping accuracy, reproducibility, and confidence metrics (Lazar, 2026). These approaches demonstrate the potential of language models for automating parts of LCI and PCF workflows, but they primarily address inventory construction or database matching rather than the propagation of model-derived prediction confidence through established LCA uncertainty frameworks.

EcoSemantic illustrates LLM-assisted direction in commercial LCA software by connecting AI assistants, including Claude, to professional LCI databases and enabling natural-language queries of ecoinvent data (Gaete, 2026). In parallel, commercial AI-supported LCA platforms such as Makersite and Niatsu indicate increasing industrial adoption of automated product-footprint workflows, including BOM enrichment, database mapping, digital-twin construction, emission-factor matching, portfolio-scale footprint calculation, and quality or confidence ratings (Makersite, 2026; Walker and Tresch, 2026). These examples indicate industrial demand for scalable LCA workflows, but they do not address the specific methodological gap targeted here: linking confidence in automated foreground-data generation to pedigree-matrix-based uncertainty propagation.

Because these applications rely on different types of language models, the distinction between encoder-based classification models and generative LLMs is relevant for positioning the architecture used in this study.

LLMs belong to the broader family of deep learning models, i.e., machine learning (ML) methods based on multi-layer neural networks that learn representations with multiple levels of abstraction, typically implemented using transformer architectures. Modern LLMs are built on the transformer architecture, which consists of an encoder, a decoder, or a combination of both (Mienye et al., 2025). The encoder converts text into high-dimensional embedding vectors, while the decoder generates or reconstructs textual output (Vaswani et al., 2017). A central concept in transformer models is the token, which represents the smallest unit of text, like a word, a word fragment, or a symbol, processed by the model. LLMs operate by transforming sequences of tokens into contextual embeddings that capture their semantic relationships (Mienye et al., 2025; Haslett and Cai, 2025).

Depending on architectural design, transformer-based language models can be categorized as autoencoding (encoder-only), autoregressive (decoder-only), or sequence-to-sequence (encoder-decoder) models (Borders and Volkova, 2021; Shao et al., 2024). In addition, models differ in how they process contextual information: unidirectional models attend only to past tokens, whereas bidirectional models incorporate both left and right context when encoding text (Sun, 2024).

Encoder-only models are designed for text understanding rather than text generation (Gillioz et al., 2020). They produce dense contextual embeddings that are well-suited for tasks such as classification, semantic similarity, and named-entity recognition (Chen et al., 2022; Zhao et al., 2024). Because these models are optimized for representation learning and operate with a fixed output label space when combined with a classification head, they are particularly effective for multi-class classification tasks involving short and sparse textual inputs (Neumann et al., 2025), which aligns with the task addressed in this study.

BERT (Bidirectional Encoder Representations from Transformers) is a representative example of an encoder-only architecture. It is a bidirectional language model, meaning it considers both left and right context when encoding tokens. The model has been pre-trained on vast amounts of text data from Wikipedia and the Toronto Book Corpus (Devlin et al., 2018). It has demonstrated strong performance in multi-class classification tasks (Ilic et al., 2022). Several architectures have been derived from BERT, including lightweight variants such as DistilBERT (Sanh et al., 2019), larger and more computationally intensive models such as RoBERTa (Liu et al., 2019), and domain-specific adaptations such as PatentBERT that are fine-tuned on specialized corpora (Lee and Hsiang, 2019).

A major advantage of encoder-only architectures is their computational efficiency and strong performance on structured downstream

tasks such as classification and information extraction. Recent benchmarks show that fine-tuned encoder models such as RoBERTa-large can match or outperform large decoder-only LLMs in classification settings, particularly when decoder models are used without task-specific fine-tuning (Zhao et al., 2024).

Decoder-only models are designed for text generation and modeling sequential dependencies. They predict the next token in a sequence based only on previous tokens (i.e., unidirectional training). A well-known example of this type is the GPT (Generative Pretrained Transformer) family (Brown et al., 2020; OpenAI, 2023).

Recent autoregressive LLMs, such as Google's Gemini (Gemini Team, 2023) or OpenAI's GPT, and Anthropic's Claude (*The Claude 3 Model Family: Opus, Sonnet, Haiku*, 2024) are capable of generating coherent text and performing complex semantic reasoning over longer contexts. However, like other generative models, they are prone to hallucinated outputs when applied without constraints, which limits their direct applicability in analytical workflows unless appropriate safeguards, such as constrained prompting, structured outputs, and validation mechanisms, are employed (Luo et al., 2024b).

Sequence-to-sequence models are used for tasks that involve transforming one sequence into another, such as machine translation, summarization, and question answering. These models first encode an input sequence and then decode it into a target sequence. Examples include T5 and BART (Lewis et al., 2019; Raffel et al., 2019).

This distinction is central to the method developed in this paper. A direct generative-LLM approach could, in principle, infer a manufacturing process from a part name and select an LCI proxy in a single step. However, such an approach makes it difficult to separate classification uncertainty, proxy-matching quality, and database coverage limitations. The proposed pipeline, therefore, separates these tasks: a supervised BERT classifier predicts manufacturing-relevant part groups from short part names, while a constrained generative LLM is used only for bounded extraction of structured manufacturing parameters from textual descriptions. The final ecoinvent proxy mapping and technological-correlation score are assigned by explicit rules rather than by unconstrained LLM judgment.

The literature reviewed in Sections 2.1 and 2.2 identifies two gaps relevant to early-stage pLCA and screening-level PCF studies of mechanically engineered products. First, procurement-stage textual data remain underexplored as a source of inference for manufacturing-process information compared with expert elicitation, rule-based systems, and geometry-based ML. Second, automated foreground-data generation is commonly evaluated by prediction accuracy, mapping accuracy, or computation speed, whereas the associated model confidence is less often linked to pedigree-matrix-based uncertainty propagation. The method developed in this paper addresses these gaps by integrating part-name-based manufacturing-process inference, constrained ecoinvent proxy mapping, and pedigree-based uncertainty propagation into a single operational pipeline.

3. Methods

This study focuses on predicting missing foreground inventory data for mechanically engineered product parts during the early development stage, corresponding to TRLs 4–6 (Table 1). At this stage, a preliminary product design exists, typically represented by technical drawings, 3D models, and an initial BOM. While the final assembly location within the company may be known, the main manufacturing processes, expected scrap or material loss rates, and upstream supply chain details are typically not yet determined. These unknown parameters lead to gaps in the foreground pLCI data, hindering robust pLCA modeling.

The proposed method aims to produce screening-level proxies for manufacturing processes, each with an explicit data-quality score and a propagated uncertainty, for use in early-stage cradle-to-gate PCF studies where supplier-specific process information is not yet available. Its purpose is to make the missing manufacturing assumptions explicit and to focus expert effort on components that most affect the result. It is not intended to provide exact process accounting or to replace supplier-data collection and expert review. The demonstration is limited to a single task: predicting the dominant main manufacturing process for mechanically engineered parts from short procurement part names within an internal classification system.

The missing data addressed in this study concern the manufacturing processes required to produce individual product parts. According to DIN 8580, manufacturing processes are classified into the main groups: primary shaping, forming, separating, joining, coating, and changing material properties (Deutsches Institut für Normung, 2022). Primary shaping comprises processes that create a solid body from shapeless material (e.g., casting, molding), whereas forming processes reshape an existing body without changing its mass (e.g., forging, stamping, drawing). Separating processes remove material to generate the final geometry (e.g., machining, turning, milling). In this work, selected processes from these main groups, specifically those that are material- and/or energy-intensive, are jointly referred to as main manufacturing processes. Joining, coating, and changing material properties processes, as defined in DIN 8580, are excluded from the scope of this study.

To address this challenge, a pretrained BERT (bert-base-uncased) model is fine-tuned in a supervised multi-class classification setting using a domain-specific textual dataset consisting of short part names and their manually assigned manufacturing groups (Devlin et al., 2018). The manufacturing group names contain hints about the main manufacturing process applied to those parts. After classification, the predicted manufacturing groups are mapped to the corresponding ecoinvent 3.9.1 (Wernet et al., 2016) datasets using a deterministic categorical mapping rule, with Claude Sonnet 4.6 (*The Claude 3 Model Family: Opus, Sonnet, Haiku*, 2024) for parameter extraction assistance from part group and ecoinvent process description, to assess manufacturing process proxy selection confidence.

The combined model predicts the most likely main manufacturing process directly from a part's textual description. Because each part group is associated with a main manufacturing process, the predicted group label enables automated completion of missing process-related foreground inventory information during early product development, while preserving associated uncertainty metrics for later propagation.

3.1. Training dataset

The training dataset consists of part names of mechanically engineered parts, classified according to an internal Robert Bosch GmbH procurement system. This classification system organizes components hierarchically: high-level part fields represent major manufacturing process families, which are further subdivided into more specific part groups that typically correspond to dominant manufacturing processes. Table 2 presents an anonymized sample of the training dataset.

Table 3 summarizes the selected part fields included in the training dataset, the number of associated part groups per field, and the number of part names available for each field. While the internal classification system defines 913 unique part groups across these fields, only 172 groups are represented in the training dataset due to the availability of labeled examples and are therefore used for model training.

Table 2
Anonymized excerpt of part fields, part groups, and part names in the training dataset.

Part field	Part group	Example part name
Parts made from thermoplastics	Injection-molded parts made from thermoplastics	Elec. cable w/ connection
Castings	Aluminum casting, misc.	Corrugated box, neutral - 200x200x210
Stamped, drawn, bent parts, made of metal	Stamped, drawn, bent parts, sheet metal housing: steel	Inlet connector

Table 3
Dataset with part fields, part groups, and numbers of parts representing each part field.

Part field	Number of part groups	Number of part names
Aluminum cast parts (AC)	48	8198
Stamped, drawn, bent parts (SD)	46	50,624
Turned parts without finishing (TR)	15	6272
Turned parts with finishing (TRf)	26	8342
Drop forged parts (DF)	7	2547
Parts made from thermoplastics (PL)	30	15,564
Total selected parts fields	**172**	**91,547**

3.2. Selection of large language model

The prediction task in this study involves assigning each part name to exactly one part group, making it a single-label, multi-class classification problem (Er et al., 2016). Because the inputs consist of short and often unstructured technical descriptions, the model must extract subtle semantic cues from abbreviations, naming conventions, and dimensional references.

For this purpose, the pretrained BERT model (bert-base-uncased) is selected. As discussed in Section 2.2, encoder-only architectures are well-suited for multiclass classification tasks, making BERT an appropriate choice for predicting part fields and part groups from technical part names. Apart from that, the resulting classifier provides prediction probabilities that form the basis for the uncertainty quantification in this study (Section 3.5).

Predicting part groups further enables the automated assignment of corresponding process datasets from ecoinvent version 3.9.1. Because this mapping step requires evaluating the technological plausibility of candidate processes, the assessment is performed using a deterministic rule that compares the part group and the candidate ecoinvent datasets across four aspects: process group, process subtype, material family, and dataset scope. To refine parameter extraction from textual descriptions of part groups and ecoinvent datasets, a large language model, Claude Sonnet 4.6, is used.

3.3. BERT model fine-tuning

To exploit the hierarchical structure of the procurement system, the classification is decomposed into two stages (see Fig. 1). This simplifies the overall task and reduces the effective number of classes each model must discriminate.

In the first stage, BERT is fine-tuned to classify part names into one of the six part fields (Table 3). This stage captures broad distinctions among high-level component families (e.g., cast, turned, and stamped parts). The model learns characteristic naming patterns at this aggregated level, providing an interpretable intermediate prediction. Because part fields differ substantially in vocabulary and naming conventions, this stage improves robustness and reduces confusion when predicting fine-grained categories.

In the second stage, six field-specific BERT models are trained, one for each part field. Each model predicts the part group within its respective field. This approach exploits the fact that each part name belongs to exactly one part field, reducing the effective number of classes per model from 172 to between 7 and 48.

During fine-tuning, each field-specific BERT model learns to associate textual patterns in part names with part groups. Because each part group in the internal classification is linked to a dominant manufacturing process, the predicted part group can serve as a proxy for the underlying process for LCI purposes.

Details of the BERT architecture and hierarchical classifier setup are reported in Supplementary Information (SI) Sections S1.2.1–S1.2.2, the hyperparameter configuration in SI Section S1.2.3 and Table S2, and the training procedure and calibration settings in SI Sections S1.3.1–S1.3.4 and Tables S3–S4.

3.4. Mapping with ecoinvent datasets

The mapping between internal part groups and ecoinvent 3.9.1 process datasets follows two stages (Fig. 2): (1) categorical parameter extraction that is applied to both the part groups and the candidate ecoinvent reference products, and a deterministic proxy-mapping rule. For each part group, the rule identifies the eligible candidate datasets, assigns a categorical pedigree class to each, and selects the proxy with the lowest pedigree class (lower is better; class 2 is the strongest match) by lexicographic comparison on the underlying relation states. Parameter extraction uses a rule-based pass supplemented by a constrained language model pass (Claude Sonnet 4.6) for the cases the rules leave open. The full vocabulary, decision table, and worked examples are documented in SI Sections S2.3–S2.6 (Tables S7–S11).

To preserve traceability, the mapping procedure distinguishes

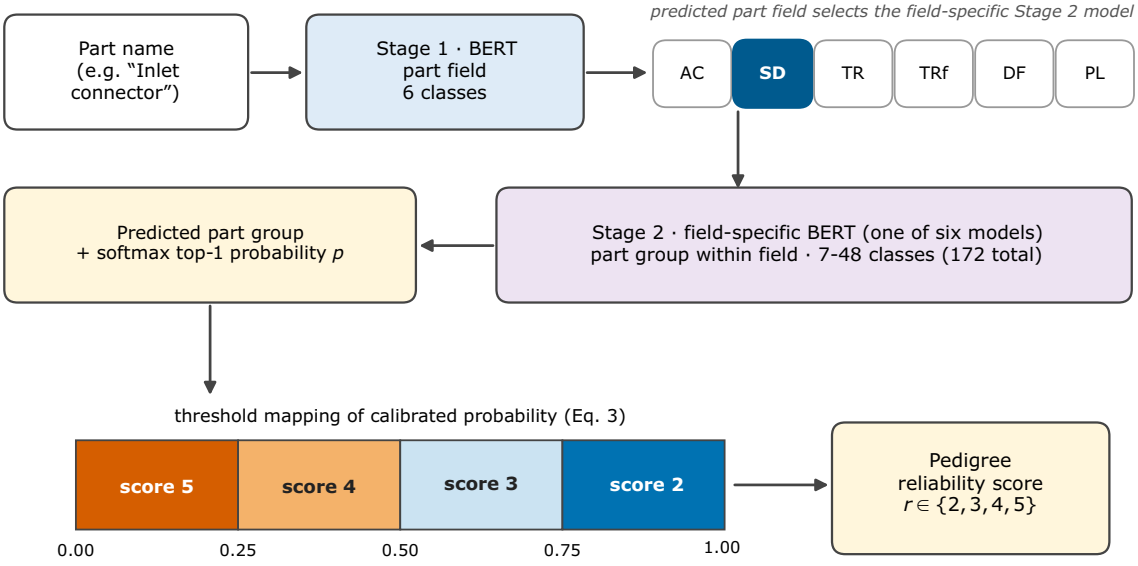

Fig. 1. Two-stage BERT classification and reliability-score assignment.

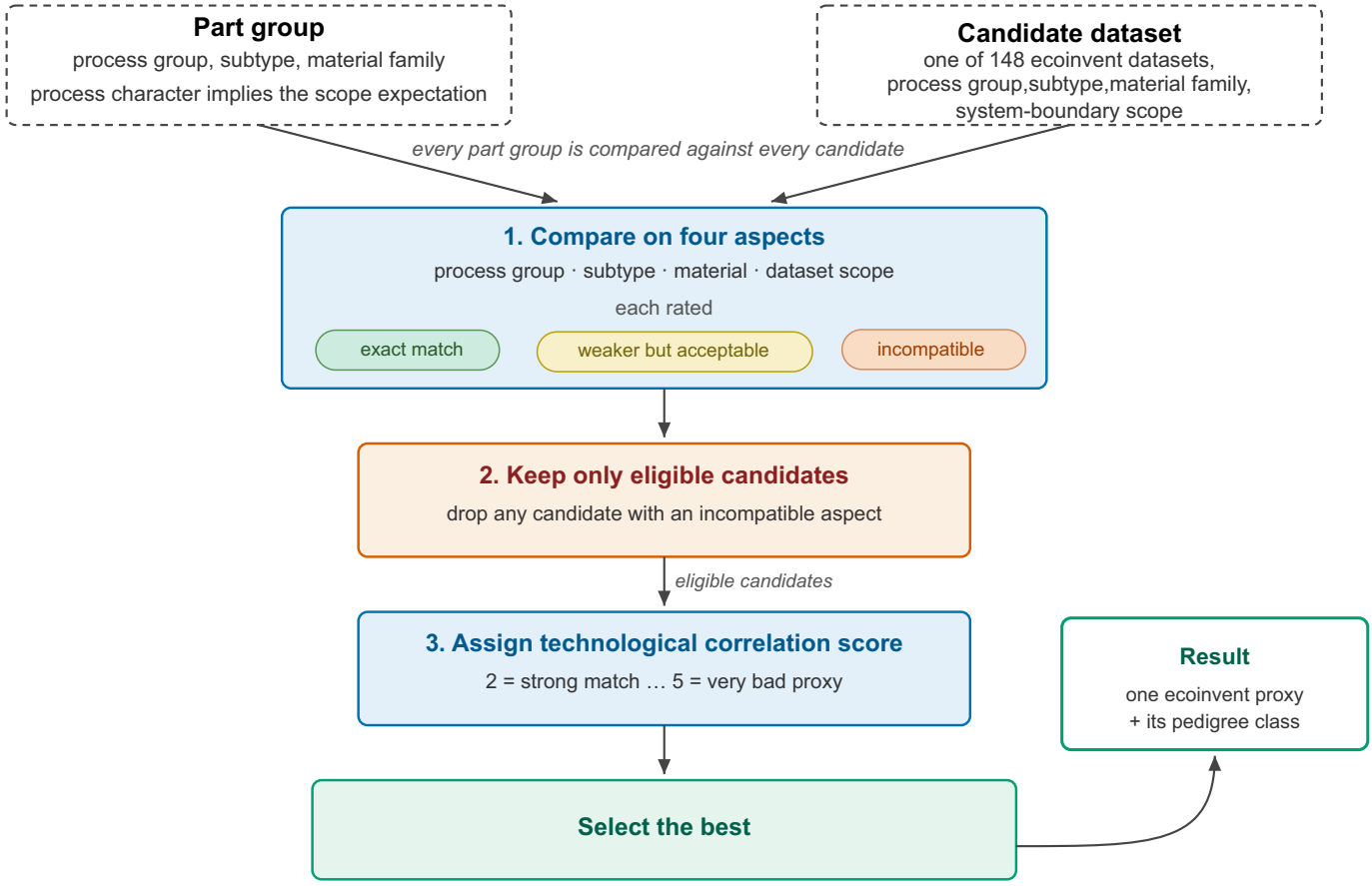

Fig. 2. Mapping a part group to an ecoinvent proxy dataset.

between source-derived information, rule-derived classifications, and model-inferred assumptions. Source-derived information includes part-group labels, ecoinvent activity names, process descriptions, classification codes, and quantitative exchanges. Rule-derived classifications are obtained by applying fixed compatibility rules to these fields. Claude Sonnet 4.6 is used only when a structured parameter cannot be reliably extracted from textual descriptions by rules. Such outputs are treated as documented assumptions, not as independent source evidence. For each inferred field, the extracted value, extraction origin, input text, model identifier, and diagnostic information are stored. The final ecoinvent proxy and technological-correlation score are then assigned by the deterministic mapping rule.

3.4.1. Parameter extraction

Both the part groups and the candidate datasets are first described with the same small set of parameters: the manufacturing process group (e.g., turning, casting, or stamping), the process subtype within that group (e.g., die-casting or fine-blanking), the material family (e.g., aluminum or thermoplastic), and three coarse indicators of precision, energy intensity, and auxiliary intensity. For part groups, these parameters are read from the procurement description in two passes: a rule-based pass for unambiguous cases and an LLM-assisted pass (Claude Sonnet 4.6, temperature 0, output restricted to the allowed vocabulary) for cases the rules leave open. For candidate datasets, the parameters are read from the ecoinvent activity name, its ISIC/CPC classification, and the electricity and heat quantities declared in the inventory. Crucially, the language model is used only to read parameters, never to choose the proxy; the choice itself is made by the rule described below. Every parameter value is stored with its origin (rule or model) and a short evidence span, so that any later decision can be traced back to the text it came from. The full vocabulary and the exact prompts are given in SI Section S2.3 (Table S7) and Section S2.7.

3.4.2. Proxy selection

For every possible pairing of a part group with an ecoinvent candidate dataset, the rule assigns ordered relationship labels for process group, process subtype, material family, and dataset scope. These labels range from exact or close matches to weak but still defensible proxy relationships and incompatible relationships. The labels are assigned according to the ordered state sets in SI Section S2.4 and Table S8, with process-group, material, scope, specialty-dataset, and coverage-gap rules specified in SI Sections S2.4.1–S2.4.5. These relation states determine candidate eligibility and the technological-correlation class; their ordinal ranks are used only for deterministic lexicographic tie-breaking among candidates in the same class.

A candidate is eligible only if all four relationships are at an acceptable label; any incompatible relationship excludes it. The complete eligibility rules, including a small number of documented exceptions for specific process families, are tabulated in SI Section S2.4 (Tables S8–S10). Where no eligible candidate exists for a part group, no substitute is forced; the group is flagged for review instead.

Each eligible pair is assigned a technological-correlation pedigree class on the proxy scale used in this study. Class 2 denotes the strongest available proxy, classes 3 and 4 denote progressively weaker but still defensible proxy relationships, and class 5 denotes a very weak proxy or a review flag rather than a verified process representation. Score-5 mappings are retained for transparency and should trigger manual review.

When several candidates share the same best pedigree class, the selected proxy is determined by a fixed lexicographic order: closer process-group match, closer subtype match, closer material match, more compatible scope, and finally the reference-product name, which serves

only as a stable identifier to guarantee a unique and repeatable result. Formally, the selected proxy e for part group n is

$$e(n) = \text{arg min over } e \in E(n) \text{ of } \tau(n, e) \tag{1}$$

$$\tau(n, e) = (\pi, \rho_g, \rho_s, \rho_m, \rho_\sigma, id) \tag{2}$$

where $E(n)$ is the set of eligible candidate datasets for part group n, and the tuple $\tau(n, e)$ is compared left to right (lexicographically). Here π is the pedigree class; ρ_g, ρ_s, ρ_m, and ρ_σ are the ordinal positions of the process-group, subtype, material, and scope relations, respectively (smaller means closer); and *id* is the reference-product name, used as a stable identifier.

Because no numerical weights or thresholds are used in this selection, the result is fully determined by the input parameters and is exactly repeatable: running the mapping again on the same parameter tables produces the same selections. The ordering encodes one deliberate modeling choice – for casting, a closer match on process subtype is preferred over a closer match on material, because the energy intensity of casting ecoinvent datasets varies far more between subtypes (for instance, lost-wax versus die-casting) than between metals. This choice, a step-by-step example, and the supporting diagnostic analyses are documented in SI Section S2.5, Section S2.6, and Section S2.8.

Every selection is stored together with the four relationship labels, the pedigree class, the eligibility outcome (eligible or excluded), the origin of each input parameter, and the rank of the selected candidate among the alternatives. This record enables a retrospective audit of every individual mapping decision. Consistent with this, the language model is restricted to parameter extraction, while the selection itself is made by the deterministic rule.

All 172 part-group mappings are manually validated against their eligible-candidate sets on two axes: whether the selected proxy is the most appropriate available option, and whether the assigned pedigree class is justified. Agreement is reported as the proportion of accepted proxy selections and the proportion of concordant pedigree classes, together with the linear-weighted Cohen's κ on the ordinal 2–5 scale.

3.5. Integration of predictive model performance into pedigree matrix scoring

This study integrates ML model performance and proxy mapping quality into the pedigree matrix by assigning scores to two indicators: reliability (reflecting BERT prediction probability) and technological correlation (reflecting ecoinvent proxy quality).

The reliability indicator is derived from the classifier's confidence in its own prediction. For each part, the Stage 2 model returns a top-1 probability for the predicted part group, which is interpreted as an estimate of the likelihood that the prediction is correct. Before use, these probabilities are temperature-scaled: each field-specific model is assigned a single calibration factor, fitted on a separate validation set rather than on the training or test data, to improve the agreement between predicted probabilities and observed correctness. Because a model prediction is not a physical measurement, reliability score 1 ("verified data based on measurements") does not apply, and the probability is mapped onto the remaining scores 2 to 5. In this study, reliability is therefore used in a restricted operational sense. It denotes the calibrated statistical reliability of the predicted part-group label under the training and validation data distribution. The assigned reliability score should therefore be interpreted as the evidential strength of an automated foreground assumption, not as source reliability in the same sense as measured data or expert-confirmed process information. This distinction is why score 1 is excluded, even for high-probability predictions.

The probability scale is divided into four equal intervals, with the most confident interval receiving the strongest score. Let p denote the temperature-scaled top-1 probability assigned by the part-field-specific Stage 2 classifier. The pedigree reliability score r is assigned as:

$$r(p) = \begin{cases} 2 & p \geq 0.75 \\ 3 & 0.50 \leq p < 0.75 \\ 4 & 0.25 \leq p < 0.50 \\ 5 & p < 0.25 \end{cases} \tag{3}$$

The mapping is monotone, so a higher probability always yields a lower (better) reliability score. A discrete mapping is used rather than a continuous one to remain consistent with the structure of the pedigree matrix, whose reliability factors are themselves defined in discrete steps.

A mapping of this kind is only meaningful if the probabilities themselves are trustworthy – a property known as calibration. A classifier is calibrated when predictions issued with a given probability are correct at approximately that rate; for example, predictions made with a probability near 0.7 should be correct in roughly 70% of cases. This is not automatic: neural-network classifiers are frequently overconfident, so calibration must be verified rather than assumed. It is summarized with the expected calibration error (ECE) – the average gap between predicted probability and observed accuracy across probability bins – for which a value near zero indicates that the probabilities can be read directly as reliabilities,

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{4}$$

where B_m is the set of predictions in the probability bin m, n is the total number of predictions, $\text{acc}(B_m)$ is the observed accuracy in the bin, and $\text{conf}(B_m)$ is the mean predicted probability in the bin (Guo et al., 2017). In this study, $M = 15$ equal-width bins are used. The ECE is used to test whether the probabilities can support ordered reliability mapping; it does not make the predictions equivalent to measured or supplier-verified foreground data. The mapping is then checked against observed classifier error rates, and the influence of the cut points is tested by repeating the analysis with alternative thresholds. These results are reported in Section 4.2 and SI Sections S4.2–S4.4.

The technological correlation score is assigned directly by the proxy-mapping rule described in Section 3.4.

3.6. Integrated pLCA pipeline

The methodological components described above in Section 3.3 (BERT classification), Section 3.4 (categorical ecoinvent mapping), and Section 3.5 (pedigree matrix scoring) are integrated into an automated pipeline that calculates prospective cradle-to-gate PCF from early-stage BOM data. Fig. 3 illustrates the complete pipeline architecture and data flow, Table 4 operationalizes the pipeline by listing the input, operation, output, uncertainty score, and reproducibility artifact for each step.

3.6.1. Pipeline processing steps

The pipeline processes multi-level BOMs through four sequential steps. In BOM ingestion (Step 1), the system parses hierarchical assembly structures and extracts part identifiers, names, quantities, and mass specifications. In BOM processing (Step 2), leaf components (parts without child assemblies) are identified, cumulative weights are calculated accounting for assembly quantities at each hierarchy level, and specific components can be manually included or excluded, producing a flattened list of components with their total masses. In part group classification (Step 3), the hierarchical BERT system is applied to each component: Stage 1 predicts the part field, and Stage 2 applies the corresponding field-specific model to predict the part group. For each prediction, the prediction probability is used to determine the pedigree reliability score. In ecoinvent process mapping (Step 4), predicted part groups are matched to their pre-computed ecoinvent 3.9.1 proxy processes, retrieving the ecoinvent activity UUID and the technological correlation score derived from the mapping confidence assessment.

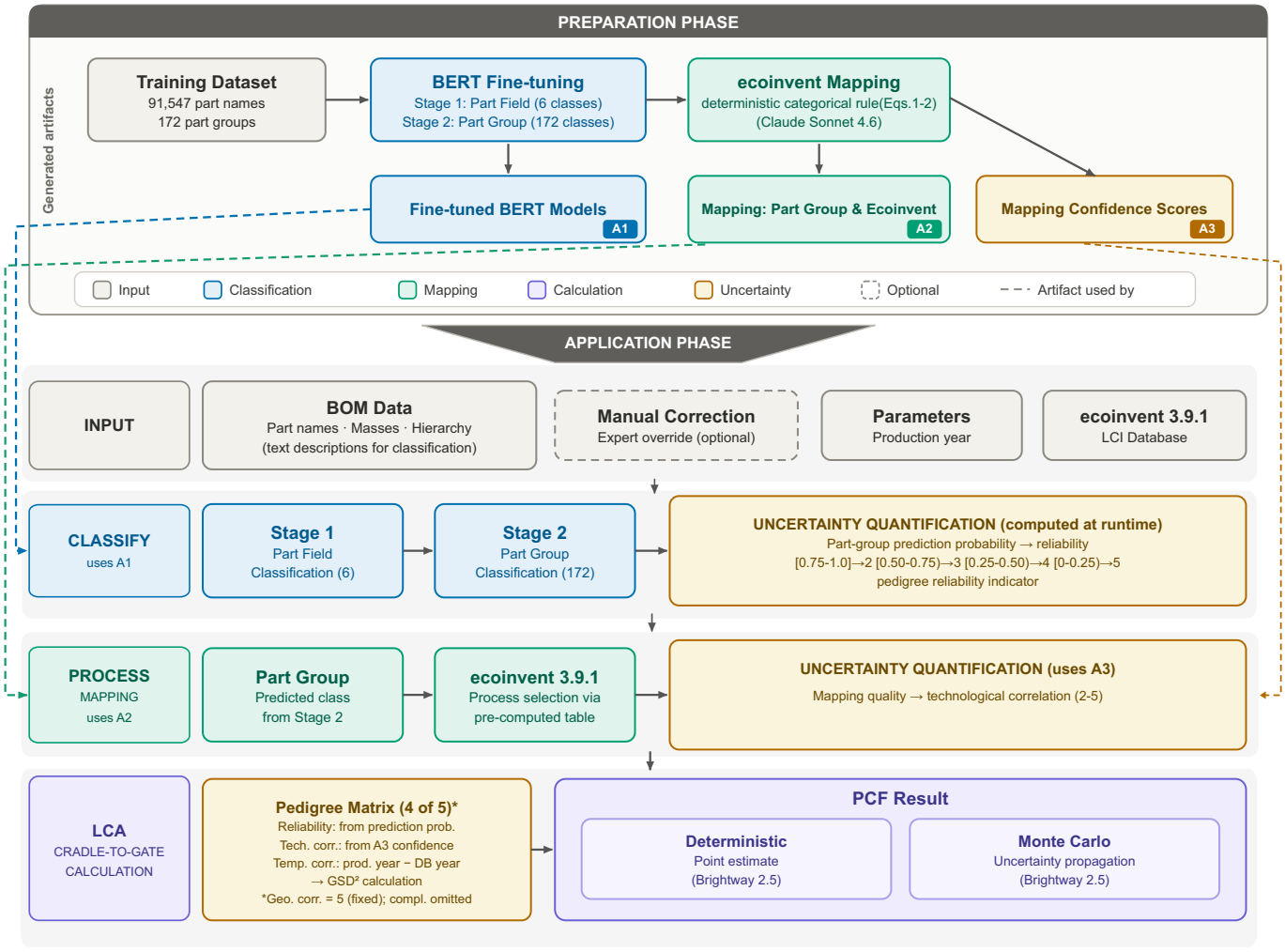

Fig. 3. Schematic representation of the pLCA pipeline for Product Carbon Footprint calculation.

3.6.2. *Uncertainty quantification through pedigree matrix*

The pedigree matrix aggregates uncertainty information from multiple pipeline stages into five indicators that parameterize the uncertainty distribution for each technosphere flow:

- Reliability: Derived from BERT classification probability score (Section 3.5). A higher probability yields lower uncertainty (score 2), while a lower probability increases uncertainty (up to score 5).
- Technological correlation: assigned by the deterministic categorical ecoinvent mapping rule (Section 3.4). Strong matches (identical process group with compatible subtype, material, and scope) receive a score of 2, while very weak proxies – or part groups for which no eligible candidate exists – receive a score of 5 and are flagged for review.

Temporal correlation is calculated from the difference between the ecoinvent reference year and the user-specified planned production year, following standard pedigree matrix temporal scoring. The geographical correlation score is assigned a fixed value of 5, and the completeness score is omitted. These indicators are combined to calculate the geometric standard deviation (GSD) for each technosphere flow, using the pedigree-matrix uncertainty factors established for ecoinvent. The LCA calculation is performed using Brightway 2.5 (Mutel, 2017), an open-source Python framework for LCA.

The pipeline produces two levels of results. First, a deterministic PCF is calculated using nominal exchange values, yielding a single-point estimate in kg CO_2-eq per functional unit. Second, a Monte Carlo simulation propagates uncertainty through the LCA model using the pedigree matrix approach (Ciroth et al., 2016; Weidema, 2013). For each foreground technosphere exchange, the assigned pedigree scores are translated into uncertainty factors and used to obtain a geometric standard deviation (Ciroth et al., 2016). Background exchanges from ecoinvent 3.9.1 retain their original uncertainty distributions and are sampled independently during simulation. Exchange amounts are then sampled from lognormal distributions parameterized by these geometric standard deviations, and the simulation is performed using Brightway 2.5. The pipeline reports both the deterministic PCF and the Monte Carlo simulation distribution results. Technical implementation details are provided in SI Section S3.

3.7. *Direct LLM-baseline*

To test whether the structured architecture is justified relative to a simpler LLM alternative, a direct Claude Sonnet 4.6 mapping baseline is evaluated on a stratified 150-row sample from the Stage 2 test set. The baseline gives the LLM only the raw or normalized procurement part name and asks it to select one ecoinvent 3.9.1 activity from the same candidate pool used by the proposed mapping procedure. It therefore bypasses the BERT classifier, the internal part-group taxonomy, and the deterministic part-group-to-ecoinvent mapping rule.

The baseline is compared with two deterministic references. The proposed-pipeline mapping is the ecoinvent activity selected after BERT

Table 4
Operationalization and reproducibility artifacts of the proposed pLCA pipeline.

Step	Input	Operation/decision rule	Output passed forward	Uncertainty or quality score generated	Reproducibility
1. BOM ingestion	Multi-level BOM with part identifiers, part names, quantities, masses, and part hierarchy. Planned production year is inputted manually.	Parse accepted column names, standardize quantity and mass fields, and preserve the assembly hierarchy.	Structured BOM with validated identifiers, masses, quantities, hierarchy, and production year.	No pedigree score; missing required fields are flagged before calculation.	BOM schema and parsing implementation in SI Section S3.1.1 and Table S12.
2. BOM processing	Structured BOM from Step 1.	Identify leaf components, propagate quantities through the hierarchy, calculate cumulative component masses per functional unit, and apply manual include/exclude rules where required.	Flattened component list with net mass, quantity per functional unit, material information, and part name.	No pedigree score in the current implementation; this step defines the foreground exchanges to which later scores are attached.	Leaf-component identification, hierarchy handling, and weight propagation in SI Section S3.1.1; output fields in SI Section S3.1.7.
3.1 Part-field classification	Part name from the flattened BOM.	Stage-1 BERT classifier predicts one of six broad part fields.	Predicted part field used to select the corresponding field-specific Stage-2 model.	No pedigree score.	Data split, architecture, hyperparameters, and Stage 1 results in SI Sections S1.1.1–S1.4.1 and Tables S2–S5; direct-LLM baseline in SI Sections S5.1–S5.6.
3.2 Part-group classification	Part name and predicted part field.	The field-specific Stage 2 BERT model predicts the manufacturing-relevant part group. The calibrated top-1 probability is retained.	Predicted part group and calibrated probability.	Pedigree reliability score: $p >= 0.75$ -> 2; $0.50 <= p < 0.75$ -> 3; $0.25 <= p < 0.50$ -> 4; $p < 0.25$ -> 5.	Stage 2 training and performance in SI Sections S1.3.3–S1.4.2 and Table S6; temperature calibration in SI Section S1.3.4 and Table S4; probability-to-reliability validation in SI Sections S4.2–S4.4, Tables S14–S16, and Fig. S3; baseline comparison in SI Section S5.6 and Table S18.
4. Ecoinvent proxy mapping	Predicted part group and ecoinvent 3.9.1 candidate process datasets.	Compare structured manufacturing parameters, including process group, subtype, material family, and dataset scope. Claude is used only for bounded parameter extraction where rule-based extraction is insufficient; final proxy selection is deterministic.	Selected ecoinvent activity UUID with proxy-quality flag; score-5 mappings are retained as weak-proxy or review flags.	Pedigree technological-correlation score from deterministic proxy-match class, restricted to scores 2–5.	Candidate-pool construction and input tables in SI Sections S2.1–S2.2; mapping vocabulary in SI Section S2.3 and Table S7; compatibility and selection rules in SI Sections S2.4–S2.5 and Tables S8–S11; validation and sensitivity outputs in SI Section S2.8.
5. Foreground LCI assembly	Flattened BOM, material data, selected ecoinvent proxy, activity UUID, production year, and assigned pedigree scores.	Construct foreground exchanges for material inputs and manufacturing-process services in Brightway 2.5.	Deterministic foreground LCI model for cradle-to-gate PCF calculation.	Temporal score from reference-year difference; geographical score fixed at 5; completeness omitted in current implementation.	Ecoinvent activity lookup in SI Section S3.1.3; separating-process handling in SI Section S3.1.4; pedigree implementation and output fields in SI Sections S3.1.5 and S3.1.7, including Table S13.
6. Uncertainty parameterization	Foreground exchanges with reliability, technological correlation, temporal, and geographical scores.	Translate pedigree scores into uncertainty factors and combine them into a geometric standard deviation for each foreground technosphere exchange.	Parameterized uncertainty model for Monte Carlo simulation.	GSD per exchange.	Pedigree indicator implementation in SI Section S3.1.5 and Table S13; calibration, reliability-band validation, and threshold sensitivity in SI Sections S4.2–S4.4, Tables S14–S16, and Fig. S3.
7. PCF calculation and triage	Deterministic foreground model, background ecoinvent model, and exchange-level uncertainty distributions.	Calculate deterministic PCF and propagate uncertainty with Monte Carlo simulation. Components with low reliability, poor technological correlation, or high impact contribution are flagged for expert review.	PCF result, uncertainty interval, contribution analysis, and review flags.	Optional review-priority flag based on reliability score, technological-correlation score, and PCF contribution.	Pipeline output structure in SI Section S3.1.7; direct-LLM baseline sample design, run configuration, and results in SI Sections S5.3–S5.6 and Tables S17–S18.

predicts the part group. The true-group mapping is the deterministic mapping rule applied to the true internal part group. Because this reference is not a unique ecoinvent ground truth, the metric is reported as exact activity-name concordance rather than accuracy against a uniquely correct process.

4. Results

4.1. Training dataset characteristics

The training dataset comprises 91,547 procurement part names distributed across 6 part fields and 172 part groups. The dataset exhibits severe class imbalance, with the most frequent part group containing 15,283 samples and the least frequent containing only 13 (imbalance ratio 1176:1). The median number of samples per class is 103, compared to a mean of 532, indicating a long-tailed distribution in which a small number of classes dominate the dataset. Such an imbalance significantly affects model training and evaluation, particularly for less frequent part groups. Additionally, substantial lexical overlap between part names across different fields presents further modeling challenges. A TF-IDF cosine similarity analysis reveals notable overlap between aluminum casting (AC) and thermoplastics (PL) part names (0.38), as well as between turning with finishing (TRf) and drop forged (DF) parts (0.30), indicating that semantic distinctions between fields can be subtle. Class distribution and similarity figures are provided in SI Section S1 (Figs. S1 and S2).

4.2. BERT classification performance

The Stage 1 model classifies part names into six part fields. Training converged after 8 epochs with early stopping triggered by validation loss. The model achieves an overall accuracy of 0.844 and a weighted F1 of 0.849 (Table 5). All part fields achieve F1 scores above 0.72, with aluminum casting and stamped metal parts reaching the highest F1 of 0.888. Lower performance for thermoplastics (F1 = 0.752) and drop forged parts (F1 = 0.727) reflects greater lexical overlap with other categories and smaller training set sizes. Per-class metrics are reported in SI Section S1.4.1 and Table S5.

In Stage 2, six field-specific models classify part names into part groups within each field, for a total of 172 classes. Performance varies substantially across part fields (Table 5), ranging from a weighted F1 of 0.821 for turning without finishing (TR), which benefits from fewer classes and more distinct naming conventions, to 0.397 for thermoplastics (PL), where high lexical similarity among part names and severe class imbalance limit classification performance. The discrepancy between weighted and macro F1 scores in most fields confirms that models perform better on frequent classes than on rare ones.

The reliability mapping is evaluated on the held-out test set (13,733 part names not seen during training). Overall, across all part fields, the temperature-scaled Stage 2 probabilities are well calibrated: the expected calibration error is 0.072, meaning the stated probability and the observed accuracy differ by about 7 percentage points on average. In the corresponding reliability diagram, the bars lie close to the diagonal of perfect calibration – stated confidence closely tracks actual accuracy, rather than sitting below it as it would for an overconfident model (SI Section S4, Fig. S3a).

Grouped by reliability score, the predictions show declining accuracy and rising error rate from one band to the next, as the mapping assumes. Within the predicted field, the part-group error rate is 17.2% at score 2, 35.9% at score 3, 64.1% at score 4, and 79.6% at score 5 (corresponding accuracies 0.83, 0.64, 0.36, and 0.20; SI, Section S4, Fig. S3b). Two features make this more than a simple decline. The bands fall in the correct order with no reversals, and each band's measured accuracy lies within the probability interval that defines it – for example, score-2 predictions (probability $\geq$0.75) are correct 83% of the time. Each band contains at least 815 predictions, and the 95% confidence intervals of neighboring bands do not overlap: the difference in accuracy between adjacent bands is larger than the statistical margin of error of each, so the separation is a real effect rather than sampling noise.

The most confident band is still not error-free, at about 83% correct, so score 2 should not be read as verified process information; score 1 remains excluded, as noted above, because model predictions are not measured against foreground data. This indicator reflects only the classifier's confidence in the part group within its predicted field. An incorrect field assignment is a distinct error source that this score does not capture; its frequency is given separately by the Stage 1 field accuracy of 0.844 reported above.

The separation does not depend on the particular choice of thresholds. The thresholds do determine which predictions fall into which band, and moving them shifts the share of predictions in a band by up to 14 percentage points; nevertheless, repeating the analysis with five alternative threshold settings, and with data-driven thresholds placed at quantiles of the probability distribution, preserves the ordering of the four bands in every case. Despite the redistribution of predictions, the population-average reliability factor changes by at most 3.0%, and the median GSD is unchanged to two decimal places. The bands, therefore, correspond to distinct and correctly ordered levels of reliability, and the uncertainty they imply is stable against the exact choice of thresholds. The full per-field calibration, per-band statistics, and sensitivity results are given in SI Section S4 (Tables S14–S16).

Table 5
Hierarchical BERT classification performance. Stage 1 classifies part names into six part fields; Stage 2 assigns part groups within each field.

	Accuracy	Weighted F1	Macro F1	Classes	Test samples
Stage 1 (Part Field)	0.844	0.849	0.815	6	13,733
Stage 2: TR	–	0.821	0.673	15	942
Stage 2: TRf	–	0.708	0.794	26	1247
Stage 2: SD	–	0.685	0.601	46	7597
Stage 2: DF	–	0.567	0.553	7	384
Stage 2: AC	–	0.468	0.517	48	1227
Stage 2: PL	–	0.397	0.469	30	2336
Stage 2 (Overall)	–	0.624	0.601	172	13,733

4.3. Ecoinvent mapping results

The mapping procedure is applied to the 172 in-scope part groups. Each part group is compared with the 148 ecoinvent 3.9.1 candidate datasets, and one proxy is retained using the deterministic mapping rule. For 170 of the 172 part groups, the selections use seven distinct ecoinvent reference products, reflecting the fact that many internal part groups share the same closest available manufacturing-process proxy. For the remaining two part groups, no admissible proxy exists in the candidate pool; these are flagged as coverage gaps (class 5), with the nearest inadmissible dataset listed for diagnostic purposes only.

Most mappings receive moderate or strong technological correlation scores. In total, 63 of 172 part groups received a score of 2, 82 received a score of 3, 25 received a score of 4, and 2 received a score of 5. Thus, 145 of 172 mappings (84.3%) receive a score of 2 or 3. Score-5 mappings are retained as explicit weak-proxy or review flags rather than interpreted as ordinary automatic assignments.

Match quality varies by part field. Thermoplastic part groups are mapped most directly to injection molding. Turned parts mapped to available turning datasets, with score differences mainly reflecting machining mode and material variant. Aluminum casting groups mapped to casting proxies with penalties where ecoinvent material or subtype coverage was approximate. Forged groups were represented by the documented forming fallback, while sheet-metal groups mapped mainly to deep-drawing press datasets. These differences reflect the structure of

the available ecoinvent candidate pool rather than the existence of a unique correct proxy for each internal part group.

The selected mappings are manually reviewed for process-family suitability, exact ecoinvent variant selection, and technological correlation score. The selected process family was accepted for 170 of 172 in-scope part groups (98.8%). At the exact ecoinvent-variant level, the selected proxy was accepted for 98 of 172 groups (57.0%); in most remaining cases, the reviewer retained the same broad process family but preferred a more representative variant. The assigned technological-correlation score was accepted for 144 of 172 groups (83.7%), with a linear-weighted Cohen's κ of 0.79 on the ordinal 2–5 scale. The main variant-level refinements concerned lower-press-force deep-drawing datasets for smaller stamped or drawn parts and primarily dressing or metal-specific turning datasets for selected turned-part groups. These refinements indicate where the deterministic tie-break rule could be improved.

4.4. Direct-LLM baseline comparison

A direct Claude Sonnet 4.6 baseline is evaluated to test whether the structured mapping pipeline could be replaced by a single constrained LLM call operating directly on short procurement part names. The LLM selected from the same ecoinvent candidate pool but did not receive the internal part-group label, material class, geometry, supplier information, or manufacturing parameters.

Exact activity-name concordance between the direct LLM and the true-group mapping was 5/150 cases, or 3.3%. In contrast, the proposed structured pipeline agreed with the true-group mapping in 72.5% of mapped cases. These values are reported as concordance between proxy-selection strategies, not as accuracy against a unique correct ecoinvent process, because several proxies may sometimes be defensible for one part group.

The result indicates that, under the tested part-name-only input conditions, a single direct LLM call does not reproduce the structured proxy-selection strategy. The LLM's self-assigned technological-correlation scores are also not used for uncertainty propagation because they represent model self-assessment rather than calibrated or rule-derived pedigree information.

4.5. Case study

The proposed framework is demonstrated on six selected components of an electric machine: rotor shaft, bearing carrier A, bearing carrier B, stator, rotor, and housing (see Fig. 4). The functional unit is defined as one set of these six components, with the system boundary covering their cradle-to-gate manufacturing. Aligned with early-stage pLCA assumptions, no specific manufacturers or geographic sourcing information is used, and the projected manufacturing year is 2027. Environmental impact is assessed using the IPCC 2021 method for climate change, expressed as global warming potential over a 100-year horizon (GWP100), in kg CO_2-eq. As the assessment is limited to climate change rather than to multiple environmental impact categories, the results are reported as a PCF rather than a full LCA. Material emission factors represent the cradle-to-gate impact of semi-finished material inputs (e.g., steel rods) and are taken from an internal material database. The manufacturing process impacts are calculated using the ecoinvent proxy datasets selected by the pipeline and are added to the material impacts to obtain the total cradle-to-gate PCF.

The hierarchical BERT model predicts part groups solely from textual part descriptions, returning the top five predictions with corresponding probabilities. The highest-probability class is automatically selected, and its associated ecoinvent process is programmatically integrated into the pLCA model using Brightway 2.5. Pedigree matrix reliability and technological correlation scores are assigned as described in Section 3.5.

Table 6 compares the expert-assigned manufacturing processes with the automated predictions and reports the corresponding pedigree scores. The automated pipeline correctly identifies casting processes for both bearing carriers but assigns a material variant (brass) rather than the aluminum die casting specified by the expert. It predicts deep-drawing processes for the rotor and stator – a sheet-metal forming family close to the expert-assigned stamping. For the housing, the model predicts casting rather than the expert-assigned extrusion process.

Fig. 5 presents the Monte Carlo distribution of the automated PCF result (1000 iterations). The simulation median is 375.3 kg CO_2-eq, with a 90% confidence interval of [372.7, 379.4] kg CO_2-eq. Because the material emission factors are identical in the manual and automated calculations, the entire difference between the two totals falls in the

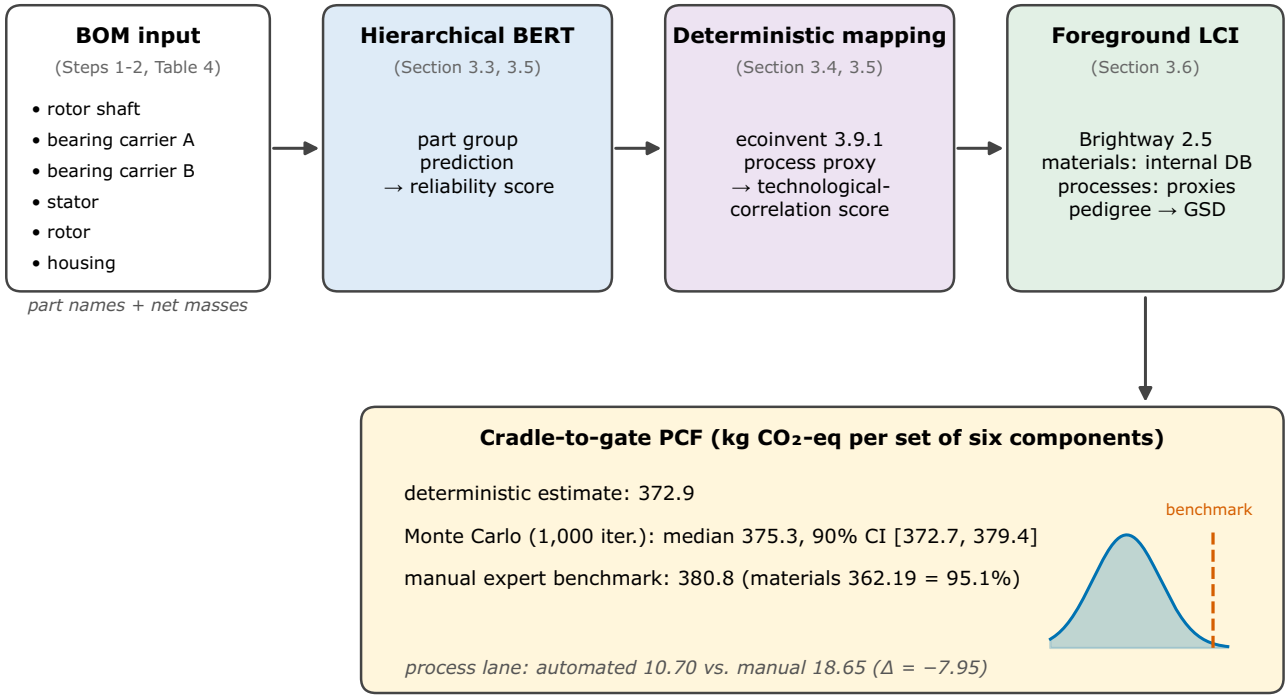

Fig. 4. Schematic representation of the pipeline application of a case study.

Table 6
Comparison of manual and automated process assignments.

Part name	Manual process	Automated prediction	Pedigree (technological correlation)	Pedigree (reliability)	Δ (auto – manual [CO_2-eq])
Rotor shaft	Turning	Impact extrusion of steel, hot, 1 strokes	3	3	+3.11
Bearing carrier A	Aluminum pressure die casting	Casting, brass	3	4	−3.92
Bearing carrier B	Aluminum pressure die casting	Casting, brass	3	3	−3.40
Housing	Extrusion	Casting, brass	3	5	−1.22
Rotor	Stamping (high-speed press)	Deep drawing, steel, 10000 kN press, automode	2	2	−1.91
Stator	Stamping (high-speed press)	Deep drawing, steel, 10000 kN press, automode	2	2	−0.62
Total	–	–	–	–	−7.95

manufacturing-process lane, which is the quantity the pipeline predicts. The automated process estimate of 10.70 kg CO_2-eq captures about 57% of the expert process contribution of 18.65 kg CO_2-eq (Table 6), a difference of 7.95 kg CO_2-eq. This gap arises because the expert reference model breaks each component into several sequential manufacturing steps, whereas the pipeline assigns a single dominant ecoinvent process to each component. Consequently, the automated total of 372.9 kg CO_2-eq lies about 2% below the manual benchmark of 380.8 kg CO_2-eq, which falls just outside the upper bound of the confidence interval.

A decomposition of the manual benchmark shows that material inputs dominate the cradle-to-gate PCF for this product: material inputs account for 362.19 kg CO_2-eq (95.1%), while manufacturing processes contribute only 18.65 kg CO_2-eq (4.9%). Because the material lane is reproduced exactly, the screening-level total is relatively insensitive to the process predictions. At the component level, however, the process deviations are bidirectional (Table 6): the rotor shaft is over-estimated by 3.11 kg CO_2-eq, because a heavy forming proxy (impact extrusion) is applied where the expert process is light turning, while the remaining components are under-estimated, most strongly the die-cast bearing carriers. The small process share seen here reflects this material-intensive product rather than the method itself: the automated process lane is expected to be more decision-relevant for manufacturing-intensive parts, where processing accounts for a larger share of the cradle-to-gate footprint, and its screening output serves to triage which components warrant detailed process data.

5. Discussion

This study contributes an operational proof-of-concept workflow for uncertainty-aware foreground data generation in early-stage pLCA and, specifically, cradle-to-gate PCF studies. Rather than proposing a general pLCA framework, the method addresses a specific bottleneck: missing manufacturing-process information for mechanically engineered parts when only BOM data and short part names are available. The results are discussed from four perspectives: the validity of translating model-derived confidence into pedigree matrix scores, the interpretation of

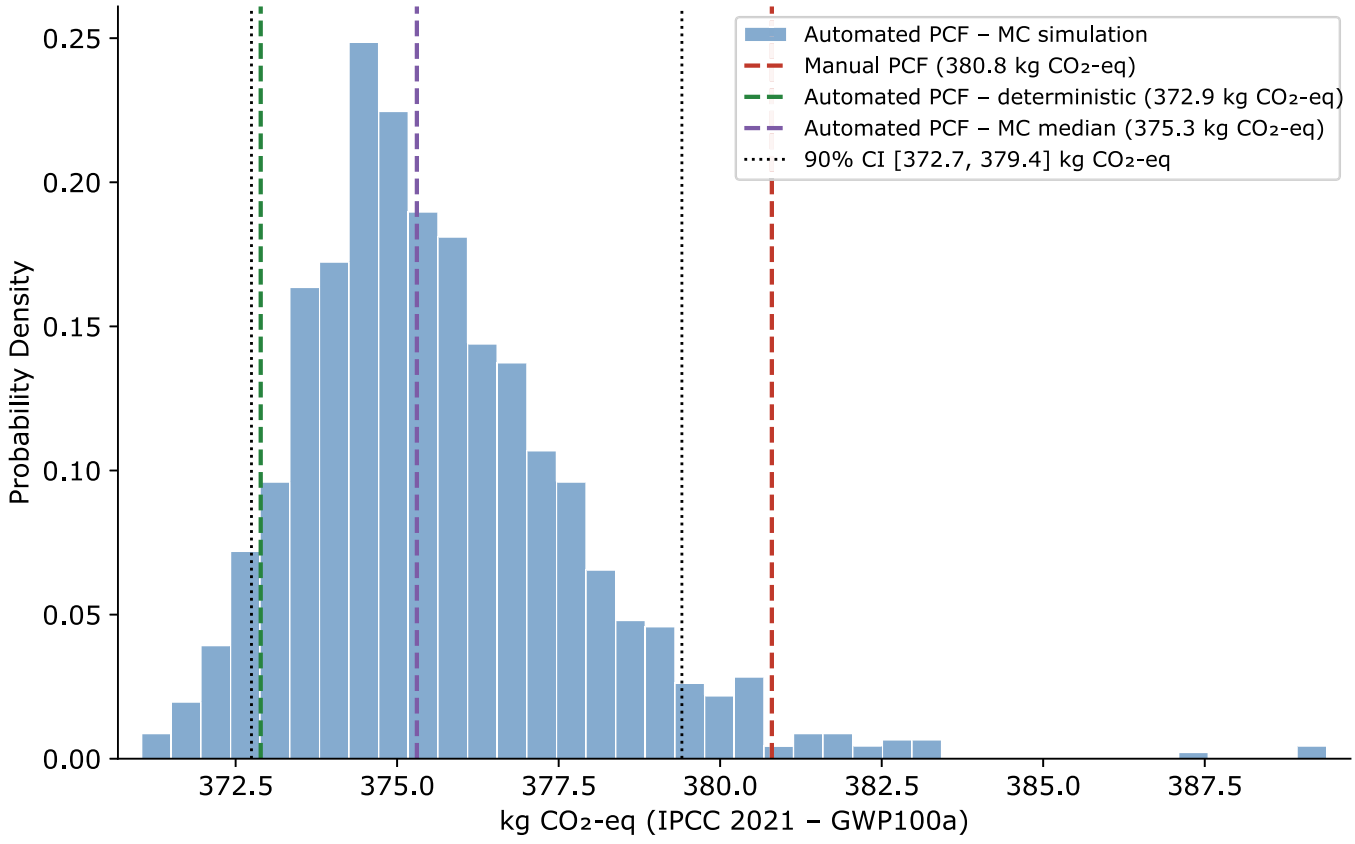

Fig. 5. Monte Carlo distribution of automated PCF with manual benchmark comparison.

the industrial case study as an operational demonstration, the decision-support value of uncertainty-aware process proxies, and the limitations of applying the pipeline beyond the demonstrated case.

5.1. Validity of automated pedigree scoring

The central methodological contribution is the translation of model-derived information into indicators for the pedigree matrix. This is useful for uncertainty propagation, but it also changes the interpretation of the pedigree scores. In conventional pLCA practice, pedigree scores are often assigned through practitioner judgment, expert elicitation, or documented data sources. Such judgments are valued not only because experts may have relevant technical knowledge but also because their assumptions can be discussed, challenged, and revised within the study's goal and scope.

The reliability score used in this study has a narrower meaning. It should be interpreted as prediction reliability, not as source reliability in the original pedigree-matrix sense. A high BERT probability indicates that the model has learned a strong association between a part name and a part group in the training data distribution. It does not mean that the actual manufacturing process has been observed, verified, or confirmed by a supplier. This distinction is visible in the case study: confident automated assignments can still differ from the expert benchmark when the short part name does not contain enough information to identify the actual production route.

The empirical justification for using BERT probability as a reliability indicator is calibration. After temperature scaling, the reliability bands are ordered consistently with observed error rates on the held-out test set: higher-probability predictions are more often correct, and lower-probability predictions are more often wrong. This supports the use of probability bands as an ordered indicator of uncertainty. It is important not to overstate what this calibration justifies: good calibration is established only on an internal held-out test set, and it does not by itself demonstrate that the resulting confidence values correspond to real-world process fidelity, supplier-specific production choices, or missing manufacturing details. Calibration is therefore used here only as evidence that the probabilities can be ordered into reliability bands, not as evidence that a confident prediction reflects the actual production route of a given part. However, calibration only supports statistical reliability under the assessed data distribution. It does not remove deeper epistemic uncertainty about future supplier choices, omitted secondary operations, or changes in production route.

The technological correlation score has a different epistemic status. It is not derived from an LLM confidence score. It is assigned by a deterministic ecoinvent proxy-mapping rule that compares structured parameters such as process group, process subtype, material family, and dataset scope. This makes the score reproducible and auditable, but it still serves as a proxy judgment of technological representativeness. Manual validation of the mapping table is therefore an essential part of the method, not merely an optional quality check.

The baseline comparison reinforces this distinction: a direct LLM call can return plausible proxy choices, but its self-assigned pedigree scores do not reproduce the deterministic technological-correlation scores used by the proposed pipeline. Therefore, LLM self-reported confidence or pedigree is treated as a diagnostic output only, whereas uncertainty propagation relies on calibrated BERT probabilities and rule-derived proxy classes.

The proposed workflow should consequently be understood as a screening and triage method rather than a replacement for expert elicitation. Its value lies in converting missing manufacturing information into explicit, reproducible, and uncertainty-bearing assumptions. Expert effort can then be directed to the components where it is most needed: high-impact items, low-reliability predictions, weak technological-correlation scores, or process choices that are important for design decisions.

The linear mapping of probability thresholds (0.75, 0.50, 0.25) to pedigree scores 2–5 has been empirically validated on the test set (Section 3.5, SI Sections S4.3–S4.4, Tables S15–S16, and Fig. S3b): observed end-to-end classification accuracy decreases monotonically across the four bins, confirming that the thresholds order reliability consistent with the model's actual error rates, and the conclusion is robust to alternative cut-point choices.

5.2. Case study interpretation

The case study should be interpreted as an operational demonstration rather than a statistical validation, given its size. The deviation between the automated and manual PCF estimates at the total level is modest (about 2%), reflecting the dominance of material inputs: manufacturing processes account for only 4.9% of the cradle-to-gate PCF (Section 4.5), so even substantial process-level mismatches move the total only slightly. The decision-support value of the pipeline is therefore primarily at the component and data-quality level rather than at the total-PCF level. In early-stage pLCA, the relevant question is often not whether every manufacturing-process proxy is exact, but whether missing supplier information prevents a timely screening-level estimate and which assumptions require further review. The pipeline enriches each BOM entry with a manufacturing-process proxy, a reliability score, a technological-correlation score, and propagated uncertainty, which together support a triage workflow: proxy assignments with acceptable reliability and technological correlation can be used for screening, whereas components with low scores can be flagged for expert review or supplier follow-up. In the case study, for example, the housing is assigned a casting proxy although the expert benchmark uses extrusion; its elevated reliability score (5) signals limited confidence in the part-group prediction and would flag the component for manual review before design decisions are made.

At the component level, the manufacturing-process contribution is systematically under-estimated: the automated estimate (10.70 kg CO_2-eq) recovers about 57% of the expert process total (18.65 kg CO_2-eq). The principal reason is structural – the pipeline assigns a single dominant process to each component, whereas the expert reference model breaks each component down into a sequence of process steps (e.g., a primary forming operation followed by machining or finishing). The omitted secondary steps account for most of the −7.95 kg CO_2-eq process gap. Two routes could narrow it. Where a part group implies specific downstream operations, selected secondary steps could be added deterministically as part of the proxy assignment. Where future process chains are still too vague to specify – the typical situation in early-stage pLCA – the residual incompleteness can instead be expressed as uncertainty: by lowering the technological-correlation score of the assigned proxy, or by activating the completeness indicator of the pedigree matrix.

This finding is based on a single product with six components and cannot be generalized. The component-level process errors are bidirectional and partly cancel: the rotor shaft is over-estimated by 3.11 kg CO_2-eq, because a heavy forming proxy is applied where the actual process is light turning, while the remaining components – most strongly the die-cast bearing carriers – are under-estimated, so the total appears closer than the component-level process predictions actually are. For products where manufacturing contributes more to the total impact, such as energy-intensive forming or precision machining, both the proxy mismatches and the omitted secondary steps would have a larger effect, and the deviation could be considerably greater. Furthermore, the expert benchmark itself carries unquantified uncertainty. The case study demonstrates that the approach produces plausible, interpretable results for one product, but it is not statistically validated.

5.3. Limitations

The first limitation concerns the classifier. The BERT models learn statistical associations from the training data and may not identify the

true process route for individual parts, especially when a part name is short, ambiguous, or belongs to an underrepresented class. Stage 2 part-group classification remains the main bottleneck, with a weighted F1 score of 0.624 across 172 part groups. The method, therefore, requires a sufficiently large and consistently labeled procurement dataset.

The second limitation is the information content of the input data. Procurement part names often omit material, geometry, dimensions, tolerance class, and intended production volume. These attributes can be decisive in manufacturing process selection. For example, the same generic term, such as "housing," can refer to plastic injection molding, aluminum casting, sheet-metal forming, or extrusion depending on material and geometry. This limitation cannot be solved by model choice alone; it requires additional structured inputs from the BOM, CAD metadata, material databases, or procurement specifications.

The third limitation concerns process-chain completeness. The current pipeline assigns one dominant manufacturing process to each component. The case study shows that this can underestimate manufacturing impacts when the expert model includes secondary operations such as machining, finishing, or cleaning. Future versions should therefore distinguish between dominant-process assignment and process-chain construction. Where secondary steps are implied by the part group, they could be added by deterministic rules. Where secondary steps are unknown, the missing process-chain information should be represented through lower technological-correlation scores, an explicit completeness score, or a separate sensitivity scenario.

The fourth limitation concerns the representation of material losses. Several manufacturing processes, especially separating processes such as turning or machining, require information on gross material input and cut-off mass. Early product development BOMs usually provide net part mass, not the gross input required before material removal. The present workflow, therefore, does not yet automate process-specific yield, scrap, or cut-off estimation. This is important because manufacturing-route selection affects not only process energy but also upstream material demand. Future work should link predicted processes to gross-to-net material factors, ideally using part geometry, process rules, or company-specific scrap data.

The fifth limitation concerns the ecoinvent proxy space. Some manufacturing categories are represented only by generic or imperfect process datasets in ecoinvent 3.9.1, and manual validation identified cases where different material defaults for turning or a lower-force deep-drawing dataset would be more appropriate than the deterministic tie-break result. Such cases should be treated as review priorities rather than hidden automation errors. Custom process datasets or parameterized process models would improve representativeness when ecoinvent provides no suitable dataset.

A distinct issue arises even when ecoinvent provides finer-grained variants: the correct process family is identified, but the appropriate variant is not selected, primarily because part size is not among the extracted parameters. The deep-drawing press-force class is the clearest example: the deterministic tie-break defaults to a single generic variant (the 10,000 kN press), whereas the required tonnage scales physically with blank area, sheet thickness, and material strength, so smaller parts are better represented by lower-force datasets. A more rigorous treatment would derive a size or mass descriptor – from the bill-of-materials mass, characteristic part dimensions, or bounding-geometry analysis of CAD/STEP models, as already noted above for material-loss estimation – and map it to the available ecoinvent press-force classes through the established deep-drawing and blanking force relations, so that the variant is chosen on a physical basis rather than by identifier tie-break. The same descriptor would refine other size-sensitive proxies, such as turning datasets distinguished by workpiece diameter. Extracting part size was outside the scope of the present pipeline and is left to future work.

The Claude-assisted parameter extraction step depends on a proprietary API. This dependence is limited by design because the language model is used only for bounded parameter extraction, where rule-based extraction is insufficient, and not for final proxy selection. Nevertheless, exact reproducibility requires archiving the model identifier, prompts, allowed vocabularies, extracted parameter values, evidence spans, and cached outputs. Future use with another model or API version should repeat the mapping validation.

Finally, the pipeline is benchmarked only against a part-name-only direct-LLM baseline, which does not establish that it compares favorably with stronger alternatives. Systematic comparison against more competitive baselines – a structured rule-based approach, a non-transformer supervised classifier, a variant without the pedigree-based uncertainty translation, and more elaborate agentic or multi-stage LLM designs – is left to future work.

6. Conclusions

This study demonstrates that short procurement part names can be used to generate screening-level manufacturing-process proxies for early-stage cradle-to-gate PCF studies when supplier-specific process data are not yet available. The proposed workflow combines hierarchical BERT-based part-group classification, deterministic mapping of predicted part groups to ecoinvent 3.9.1 process proxies, and pedigree-based uncertainty propagation in a Brightway implementation. The generative LLM is used only for bounded parameter extraction where rule-based extraction is insufficient; final proxy selection and technological-correlation scoring are assigned by deterministic rules.

The results show both the potential and the current limits of this approach. The Stage 1 classifier reached 0.844 accuracy and 0.849 weighted F1, while the more difficult Stage 2 part-group classification reached 0.624 weighted F1 across 172 classes. Temperature-scaled Stage 2 probabilities are sufficiently calibrated to support an ordinal reliability mapping, with a pooled expected calibration error of 0.072. The ecoinvent mapping step produced mostly usable proxies, with 145 of 172 part groups (84.3%) receiving technological-correlation scores of 2 or 3. Manual validation accepted the selected process family for 170 of 172 groups (98.8%), the exact ecoinvent variant for 98 of 172 groups (57.0%), and the assigned technological-correlation score for 144 of 172 groups (83.7%).

The industrial case study illustrates the value of the method as a screening and triage tool rather than as a substitute for supplier data or expert modeling. The automated deterministic PCF estimate was 372.9 kg CO_2-eq compared with a manual benchmark of 380.8 kg CO_2-eq, while the Monte Carlo median was 375.3 kg CO_2-eq with a 90% uncertainty interval of [372.7, 379.4] kg CO_2-eq. The small total deviation is partly explained by the dominance of material inputs, which account for 95.1% of the benchmark PCF. At the process level, however, the automated model captured only 10.70 kg CO_2-eq of the expert process contribution of 18.65 kg CO_2-eq, primarily because it assigns a single dominant process to each component and omits secondary operations.

The method is therefore most useful for making missing manufacturing assumptions explicit, assigning reproducible data-quality scores, propagating uncertainty, and identifying components that require expert review or supplier follow-up. It is less suitable as a stand-alone tool for final PCF reporting, detailed process-chain modeling, or products whose environmental profile is strongly driven by manufacturing energy or process-specific scrap.

Future work should focus on four priorities: improving Stage 2 classification through larger and more balanced labeled datasets; integrating additional structured inputs such as material class, geometry, CAD metadata, or production volume; extending the pipeline from dominant-process assignment to process-chain and gross-material estimation; and validating the method across more products, companies, and manufacturing domains.

CRediT authorship contribution statement

Alexandra Belyaeva: Writing – original draft, Visualization,

Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christoph Helbig:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Torsten Hummen:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Claude Opus 4.8 (Anthropic) in order to improve the readability and language of the manuscript, and to support the technical preparation of author-designed figures. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Funding statement

Funded by the Open Access Publishing Fund of the University of Bayreuth.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was conducted at Robert Bosch GmbH, Corporate Research. The authors gratefully acknowledge Robert Bosch GmbH for providing industrial datasets used in this study. We thank our colleagues at Robert Bosch GmbH: Aleksei Beloshapko and Hugo Mesquita for exchange datasets and providing feedback.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.spc.2026.07.008.

References

Allwood, J.M., Ashby, M.F., Gutowski, T.G., Worrell, E., Jan. 2011. Material efficiency: a white paper. Resour. Conserv. Recycl. 55 (3), 362–381. https://doi.org/10.1016/j.resconrec.2010.11.002.

Arvidsson, R., 2023. Prospective, Anticipatory and Ex-Ante – What's the Difference? Sorting Out Concepts for Time-Related LCA [Online]. Available: https://research.chalmers.se/en/publication/535660.

Arvidsson, R., et al., Dec. 2018. Environmental assessment of emerging technologies: recommendations for prospective LCA. J. Ind. Ecol. 22 (6), 1286–1294. https://doi.org/10.1111/jiec.12690.

Arvidsson, R., Svanström, M., Sandén, B.A., Thonemann, N., Steubing, B., Cucurachi, S., Dec. 2023. Terminology for future-oriented life cycle assessment: review and recommendations. Int. J. Life Cycle Assess. https://doi.org/10.1007/s11367-023-02265-8.

Balaji, B., et al., Dec. 2023. Flamingo: environmental impact factor matching for life cycle assessment with zero-shot machine learning. ACM J. Comput. Sustain. Soc. 1 (2), 1–23. https://doi.org/10.1145/3616385.

Ben-David, A., Frank, E., Apr. 2009. Accuracy of machine learning models versus 'hand crafted' expert systems – a credit scoring case study. Expert Syst. Appl. 36 (3), 5264–5271. https://doi.org/10.1016/j.eswa.2008.06.071.

Bergerson, J.A., et al., Feb. 2020. Life cycle assessment of emerging technologies: evaluation techniques at different stages of market and technical maturity. J. Ind. Ecol. 24 (1), 11–25. https://doi.org/10.1111/jiec.12954.

Bhander, G.S., Hauschild, M., McAloone, T., Dec. 2003. Implementing life cycle assessment in product development. Environ. Prog. 22 (4), 255–267. https://doi.org/10.1002/ep.670220414.

Borders, T., Volkova, S., Apr. 2021. An Introduction to Word Embeddings and Language Models. INL/EXT-21-61935-Rev000, 1773690. https://doi.org/10.2172/1773690.

Brown, T.B., et al., 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165.

Chen, Y., Mikkelsen, J., Binder, A., Alt, C., Hennig, L., 2022. A Comparative Study of Pre-trained Encoders for Low-Resource Named Entity Recognition. https://doi.org/10.48550/ARXIV.2204.04980.

Ciroth, A., Muller, S., Weidema, B., Lesage, P., Sep. 2016. Empirically based uncertainty factors for the pedigree matrix in ecoinvent. Int. J. Life Cycle Assess. 21 (9), 1338–1348. https://doi.org/10.1007/s11367-013-0670-5.

Cucurachi, S., van der Giesen, C., Guinée, J., 2018. Ex-ante LCA of emerging technologies. Proc. CIRP 69, 463–468. https://doi.org/10.1016/j.procir.2017.11.005.

Deutsches Institut für Normung, 2022. DIN 8580 – Manufacturing Processes – Terms and Definitions.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805.

Er, Meng Joo, Venkatesan, R., Wang, Ning, Oct. 2016. An online universal classifier for binary, multi-class and multi-label classification. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, pp. 003701–003706. https://doi.org/10.1109/SMC.2016.7844809.

Erakca, M., Baumann, M., Helbig, C., Weil, M., Apr. 2024. Systematic review of scale-up methods for prospective life cycle assessment of emerging technologies. J. Clean. Prod. 451, 142161. https://doi.org/10.1016/j.jclepro.2024.142161.

Gaete, C., 26 February 2026. "EcoSemantic, the First AI-Native Platform for Sustainability Analysis," presented at the LCA DF 92 - AI in LCA: Innovations, Applications, and Challenges, Empa Akademie, Dübendorf. Accessed: May 01, 2026. [Online]. Available: https://lca-forum.ch/forum?tx_news_pi1%5BactbackPid%5D=1&tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Bnews%5D=91&cHash=69acafdf32fc351091a0a2310d85f045.

Gemini Team, et al., 2023. Gemini: A Family of Highly Capable Multimodal Models. https://doi.org/10.48550/ARXIV.2312.11805.

Gennatas, E.D., et al., Mar. 2020. Expert-augmented machine learning. Proc. Natl. Acad. Sci. U. S. A. 117 (9), 4571–4577. https://doi.org/10.1073/pnas.1906831117.

Gillioz, A., Casas, J., Mugellini, E., Khaled, O.A., Sep. 2020. Overview of the Transformer-based Models for NLP Tasks. presented at the 2020 Federated Conference on Computer Science and Information Systems, pp. 179–183. https://doi.org/10.15439/2020F20.

Gkousis, S., Vasilaki, V., Katsou, E., Mar. 2026. Machine learning and large language models for life cycle inventory compilation: current situation and future developments. Renew. Sustain. Energy Rev. 228, 116577. https://doi.org/10.1016/j.rser.2025.116577.

Grenz, J., et al., Jun. 2023. Integrating prospective LCA in the development of automotive components. Sustainability 15 (13), 10041. https://doi.org/10.3390/su151310041.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On Calibration of Modern Neural Networks. https://doi.org/10.48550/ARXIV.1706.04599.

Haslett, D.A., Cai, Z.G., Apr. 2025. How much semantic information is available in large language model tokens? Trans. Assoc. Comput. Linguistics 13, 408–423. https://doi.org/10.1162/tacl_a_00747.

Hertwich, E.G., Wood, R., 2018. The growing importance of scope 3 greenhouse gas emissions from industry. Environ. Res. Lett. 13 (10), 104013. https://doi.org/10.1088/1748-9326/aae19a.

Ilic, E., Garcia Martinez, M., Souto Pastor, M., 2022. A review of text classification models from Bayesian to transformer. In: Swiss Text Analytics Conference.

Kusch, A., et al., Oct. 2024. DfC-Industry Design for Circularity – Operationalisierung in der industriellen Produktentwicklung. INEC, Hochschule Pforzheim, Pforzheim.

Lazar, L., Jun. 2026. Do you speak LCA? FAULDIER: a framework for large language model assisted life cycle inventories in life cycle assessment. SoftwareX 34, 102602. https://doi.org/10.1016/j.softx.2026.102602.

Lee, J.-S., Hsiang, J., 2019. PatentBERT: Patent Classification with Fine-Tuning a Pre-trained BERT Model. https://doi.org/10.48550/ARXIV.1906.02124.

Leroy, Y., Froelich, D., 2010. Qualitative and quantitative approaches dealing with uncertainty in life cycle assessment (LCA) of complex systems: towards a selective integration of uncertainty according to LCA objectives. IJDE 3 (2), 151. https://doi.org/10.1504/IJDE.2010.034862.

Lewis, M., et al., 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. https://doi.org/10.48550/ARXIV.1910.13461.

Liu, Y., et al., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692.

Lloyd, S.M., Ries, R., Jan. 2007. Characterizing, propagating, and analyzing uncertainty in life-cycle assessment: a survey of quantitative approaches. J. Ind. Ecol. 11 (1), 161–179. https://doi.org/10.1162/jiec.2007.1136.

Luo, B., et al., 2024a. AutoPCF: a novel automatic product carbon footprint estimation framework based on large language models. AAAI-SS 2 (1), 102–106. Jan. https://doi.org/10.1609/aaaiss.v2i1.27656.

Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., Dudek, G., 2024b. Hallucination Detection and Hallucination Mitigation: An Investigation. https://doi.org/10.48550/ARXIV.2401.08358.

Makersite, 2026. Makersite GmbH [Online]. Available: https://makersite.io/.

Mienye, I.D., Jere, N., Obaido, G., Ogunruku, O.O., Esenogho, E., Modisane, C., Sep. 2025. Large language models: an overview of foundational architectures, recent trends, and a new taxonomy. Discov Appl Sci 7 (9), 1027. https://doi.org/10.1007/s42452-025-07668-w.

Moni, S.M., Mahmud, R., High, K., Carbajales-Dale, M., Feb. 2020. Life cycle assessment of emerging technologies: a review. J. Ind. Ecol. 24 (1), 52–63. https://doi.org/10.1111/jiec.12965.

Mutel, C., Apr. 2017. Brightway: an open source framework for life cycle assessment. JOSS 2 (12), 236. https://doi.org/10.21105/joss.00236.
Neumann, J., Lange, R., Susanti, Y., Färber, M., 2025. Optimizing Small Transformer-Based Language Models for Multi-Label Sentiment Analysis in Short Texts. https://doi.org/10.48550/ARXIV.2509.04982.
Nurdiawati, A., Mir, B.A., Al-Ghamdi, S.G., Jun. 2025. Recent advancements in prospective life cycle assessment: current practices, trends, and implications for future research. Resour. Environ. Sustain. 20, 100203. https://doi.org/10.1016/j.resenv.2025.100203.
Olsen, S.I., Borup, M., Andersen, P.D., 2018. Future-oriented LCA. In: Hauschild, M.Z., Rosenbaum, R.K., Olsen, S.I. (Eds.), Life Cycle Assessment. Springer International Publishing, Cham, pp. 499–518. https://doi.org/10.1007/978-3-319-56475-3_21.
OpenAI, et al., 2023. GPT-4 Technical Report. https://doi.org/10.48550/ARXIV.2303.08774.
Preuss, N., Alshehri, A.S., You, F., Aug. 2024. Large language models for life cycle assessments: opportunities, challenges, and risks. J. Clean. Prod. 466, 142824. https://doi.org/10.1016/j.jclepro.2024.142824.
Raffel, C., et al., 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://doi.org/10.48550/ARXIV.1910.10683.
Ramani, K., et al., Sep. 2010. Integrated sustainable life cycle design: a review. J. Mech. Des. 132 (9), 091004. https://doi.org/10.1115/1.4002308.
Rosenbaum, R.K., Georgiadis, S., Fantke, P., 2018. Uncertainty management and sensitivity analysis. In: Hauschild, M.Z., Rosenbaum, R.K., Olsen, S.I. (Eds.), Life Cycle Assessment. Springer International Publishing, Cham, pp. 271–321. https://doi.org/10.1007/978-3-319-56475-3_11.
Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. https://doi.org/10.48550/ARXIV.1910.01108.
Schöggl, J.-P., Baumgartner, R.J., Hofer, D., Jan. 2017. Improving sustainability performance in early phases of product design: a checklist for sustainable product development tested in the automotive industry. J. Clean. Prod. 140, 1602–1617. https://doi.org/10.1016/j.jclepro.2016.09.195.
Shao, M., Basit, A., Karri, R., Shafique, M., 2024. Survey of different large language model architectures: trends, benchmarks, and challenges. IEEE Access 12, 188664–188706. https://doi.org/10.1109/ACCESS.2024.3482107.
Sun, Y., Feb. 2024. The evolution of transformer models from unidirectional to bidirectional in natural language processing. ACE 42 (1), 281–289. https://doi.org/10.54254/2755-2721/42/20230794.
Tan, E.C.D., Tu, Q., Martins, A.A., Yao, Y., Sunol, A., Smith, R.L., May 2025. Uncertainty in inventories for life cycle assessment: state-of-the-art, challenges, and new technologies. Environ. Prog. Sustain. Energy, e14644. https://doi.org/10.1002/ep.14644.
The Claude 3 Model Family: Opus, Sonnet, Haiku, 2024. Anthropic, Model Card [Online]. Available: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
Thonemann, N., Schulte, A., Maga, D., Feb. 2020. How to conduct prospective life cycle assessment for emerging technologies? A systematic review and methodological guidance. Sustainability 12 (3), 1192. https://doi.org/10.3390/su12031192.
Tsoy, N., Steubing, B., van der Giesen, C., Guinée, J., Sep. 2020. Upscaling methods used in ex ante life cycle assessment of emerging technologies: a review. Int. J. Life Cycle Assess. 25 (9), 1680–1692. https://doi.org/10.1007/s11367-020-01796-8.
Tzinis, I., 2015. Technology readiness level [Online]. Available: http://www.nasa.gov/directorates/heo/scan/engineering/technology/technology_readiness_level.
U.S. Department of Defense, 2025. Manufacturing Readiness Level (MRL) Deskbook [Online]. Available: https://www.dodmrl.com/MRL_Deskbook_2025.pdf.
van der Giesen, C., Cucurachi, S., Guinée, J., Kramer, G.J., Tukker, A., Jun. 2020. A critical view on the current application of LCA for new technologies and recommendations for improved practice. J. Clean. Prod. 259, 120904. https://doi.org/10.1016/j.jclepro.2020.120904.
Vaswani, A., et al., 2017. Attention is All you Need, 30 [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
Walker, C., Tresch, J., 2026. "AI-Powered Tool for Footprinting," presented at the LCA DF 92 - AI in LCA: Innovations, Applications, and Challenges, Empa Akademie, Dübendorf, Feb. 26, 2026. Accessed: May 01, 2026. [Online]. Available: https://lca-forum.ch/forum?tx_news_pi1%5BactbackPid%5D=1&tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Bnews%5D=91&cHash=69acafdf32fc351091a0a2310d85f045.
Wang, Z., Rosen, D., Dec. 2023. Manufacturing process classification based on heat kernel signature and convolutional neural networks. J. Intell. Manuf. 34 (8), 3389–3411. https://doi.org/10.1007/s10845-022-02009-9.
Weidema, B.P., 2013. Overview and methodology: Data quality guideline for the ecoinvent database version 3 [Online]. Available: https://vbn.aau.dk/en/publications/overview-and-methodology-data-quality-guideline-for-the-ecoinvent.
Weidema, B.P., Wesnæs, M.S., Jan. 1996. Data quality management for life cycle inventories—an example of using data quality indicators. J. Clean. Prod. 4 (3–4), 167–174. https://doi.org/10.1016/S0959-6526(96)00043-1.
Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., Weidema, B., Sep. 2016. The ecoinvent database version 3 (part I): overview and methodology. Int. J. Life Cycle Assess. 21 (9), 1218–1230. https://doi.org/10.1007/s11367-016-1087-8.
World Resources Institute, World Business Council for Sustainable Development, 2011. Corporate Value Chain (Scope 3) Accounting and Reporting Standard. Supplement to the GHG Protocol Corporate Accounting and Reporting Standard [Online]. Available: https://ghgprotocol.org/corporate-value-chain-scope-3-standard.
Wright, M. Mba, et al., Aug. 2024. Life cycle inventory availability: status and prospects for leveraging new technologies. ACS Sustain. Chem. Eng. 12 (34), 12708–12718. https://doi.org/10.1021/acssuschemeng.4c02519.
Zhang, Z., Jaiswal, P., Rai, R., Aug. 2018. FeatureNet: machining feature recognition based on 3D convolution neural network. Comput. Aided Des. 101, 12–22. https://doi.org/10.1016/j.cad.2018.03.006.
Zhao, C., Dinar, M., Melkote, S.N., Apr. 2020. Automated classification of manufacturing process capability utilizing part shape, material, and quality attributes. J. Comput. Inf. Sci. Eng. 20 (2), 021011. https://doi.org/10.1115/1.4045410.
Zhao, H., Chen, Q.P., Zhang, Y.B., Yang, G., 2024. Advancing Single and Multi-task Text Classification through Large Language Model Fine-tuning. https://doi.org/10.48550/ARXIV.2412.08587.