

Machine learning and large language models for life cycle inventory compilation: Current situation and future developments

Spiros Gkousis*, Vasileia Vasilaki, Evina Katsou

Imperial College London, Department of Civil and Environmental Engineering, Exhibition Rd, South Kensington, London, SW7 2AZ, United Kingdom

HIGHLIGHTS

- Data-driven methods can help fill LCI data gaps.
- Different algorithms are applicable for different data sizes and structures.
- Language models can extract useful information from documents.
- Text-based matching can speed database selection.
- Combined AI approaches can complement traditional LCI practices.

ARTICLE INFO

Keywords:

Life cycle analysis
Life cycle inventory
Machine learning
Artificial intelligence
Large language models
Natural language processing

ABSTRACT

With increasing requirements for environmental accountability, Life Cycle Assessment (LCA) is becoming key for sustainability reporting. Nevertheless, significant challenges remain regarding data availability, especially for emerging and low-carbon energy technologies, for which Life Cycle Inventory (LCI) data are usually scarce and spread across studies and reports. Common LCI challenges concern the exploitation of available, smaller or larger, LCA datasets and the collection of LCA data from various sources when these are not found in LCA databases. This study explores machine learning (ML), natural language processing, and large language models (LLM) applications to tackle such challenges and estimate missing LCI data. A thorough review of suggested ML and artificial intelligence (AI) applications for LCI compilation is performed, complemented by case studies investigating ML and generative LLM methods to impute or gather missing LCI data for car-driving, power plants, and geothermal energy systems. ML methods can provide more reliable estimations than simple linear regression even for small datasets, while generative LLMs are found to effectively identify and extract LCI information from scientific papers. The potential of ML and AI methods to facilitate LCI compilation and enhance data reliability and availability for the LCA of emerging energy technologies is large, highlighting the crucial role such methods can play for decarbonization. Nevertheless, relevant applications remain in their infancy. More research is needed to construct robust frameworks for large-scale deployment to complement traditional LCI methods, and ensure correct usage of ML algorithms towards automated, accurate, and interpretable AI-assisted LCA.

1. Introduction

Life Cycle Assessment (LCA) is becoming a key pillar for achieving sustainable development goals [1], and a significant tool that supports decision-making towards decarbonized energy systems and circular resource use [2]. This is particularly critical for energy and renewable energy systems, where LCA is increasingly required to demonstrate the decarbonization potential of emerging technologies such as off-shore wind, deep geothermal, hydrogen, and next-generation batteries.

Requirements for environmental data reporting are adopted by several countries [3,4], while the environmental, social, and governance values of companies are increasing importance [5]. LCA is endorsed by the European Commission for estimating the environmental impacts of products and systems including renewable energy systems [6], buildings [7], and batteries [8]. Despite its relevance, LCA implementation is hindered due to a lack of LCA data, expert knowledge, and the associated costs [9]. The most critical stage of an LCA is the data collection during the Life Cycle Inventory (LCI) phase [10,11]. LCI data concern all

* Corresponding author.

Email address: s.gkousis@imperial.ac.uk (S. Gkousis).

<https://doi.org/10.1016/j.rser.2025.116577>

Received 16 July 2025; Received in revised form 27 November 2025; Accepted 29 November 2025

Available online 12 December 2025

1364-0321/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Nomenclature

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
GHG	Greenhouse Gases
GPR	Gaussian Process Regression
LCA	Life Cycle Assessment
LCI	Life Cycle Inventory
LLM	Large Language Models
LLR	Local Linear Regression

LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MRE	Mean Relative Error
NLP	Natural Language Processing
RAG	Retrieval Augmented Generation
RF	Random Forests
RR	Ridge Regression
XGB	XGBoost

inflows and outflows of a process, including water, energy, and raw materials consumed, emissions, waste, and products produced, and capital goods used (e.g., equipment). The LCI is separated into the foreground and background systems concerning, respectively, system specific and supply chain information. The accuracy of an LCA depends on how representative the data used for the investigated system are. Therefore, to increase the quality of LCA studies, reduce the costs, and widen LCA implementation, methods for reducing the time, complexity, and cost to compile the LCI are needed.

LCI information can be sought in several sources, including LCA databases. LCA databases may provide cross-sectoral and global coverage [12,13], or focus on specific countries [14,15] or sectors [16–18]. Navigating between numerous LCA databases and their unit processes can be a difficult task [19]. Also, LCA databases often focus on industrially mature, well-researched systems and processes and not on emerging technologies and products. Industrial reports [20–22] also contain LCI data, while for emerging technologies LCI information is usually reported in scientific LCA studies. In the latter case, and in the absence of process-specific data, LCA practitioners might use data reported in other LCA studies for the same or similar technologies with modifications. For example, to construct prospective LCA databases, the Premise software [23] uses LCI datasets for emerging technologies such as direct air capture and carbon capture and storage derived from other LCA studies. Similarly, Brandt used LCI values from natural gas and oil well drilling to perform LCA for natural hydrogen production [24]. Data scarcity is a common problem for the LCA of emerging technologies in the energy and decarbonization fields, where novel technologies such as green hydrogen, direct air capture, or battery recycling are developing rapidly but LCI data remain incomplete and scattered. In their review, Delval et al. [25] mentioned that of the twenty LCA studies on ocean-based CO₂ removal methods, most used literature or lab/pilot-scale data. Also, less than half of the thirty LCA studies on deep geothermal energy plants reviewed in [26] reported a primary LCI. In addition, Peeters et al. [27] showed that the majority of LCA studies on Li-ion batteries rely on datasets reported by a handful of initial studies, many of which involve lab-scale data. In such data-scarce cases, the accuracy of the LCA relies on finding useful sources of information and extracting the relevant data. Data scarcity is not only a methodological challenge but also a barrier to scaling renewable energy systems and low-carbon technologies, where reliable LCAs are essential for policy support and market adoption. Insufficient data can result in uncertainty about the environmental benefits of decarbonization technologies, delaying policy decisions and investment in cleaner alternatives.

1.1. LCA data estimation methods

Traditional LCA data estimation methods involve process simulations and calculations, extrapolation, and the usage of proxy data. In chemical manufacturing, Parvatkar and Eckelman suggested an eight-step hierarchy that emphasizes using actual plant data, performing process simulations, calculations, or stoichiometry, and as a last resort

using proxy secondary data [28]. This hierarchy was also considered for emerging technologies [29], general LCI data estimation [30] and aligns with what is described for the compilation of ecoinvent LCI datasets for chemical manufacturing [31]. For emerging technologies tested at a low scale, methods have been suggested to approximate LCI data at a higher scale of implementation. These include simplified process calculations [32], mapping lab-scale to industrial-scale processes [33], and scaling relationships based on statistically derived scaling factors [34]. Extrapolation methods using simple relationships of key influential parameters have also been suggested [35,36]. Nevertheless, using proxy data, with the appropriate modifications, has been suggested to be of similar quality to actual plant data, in cases where appropriate proxy datasets are available [28,37].

Process simulations are commonly used for estimating LCI data [38]. However, they require specialized software and/or domain-specific knowledge rendering them an unaffordable solution for small-medium enterprises, small research teams, early-stage technology developers, or organizations operating with limited capacity. Additionally, process models for emerging technologies are scarcely available and their development can require highly specialized personnel and assumptions. These are not always within the scope of an LCA study, which might opt to use data from a similar process to identify potential environmental hotspots and suggest mitigation measures, or to model an emerging technology as a small part of another process. For example, well drilling is a major environmental hotspot for deep geothermal energy production. However, models to calculate energy, steel, and cement consumption for different geological settings, depths and applications are not available and cannot be developed by an LCA practitioner, who would use data for wells developed for hydrocarbon production [24,39] or for geothermal wells developed elsewhere [40,41]. Similarly, literature values for the energy consumption during battery manufacturing would be used (e.g., [42,43]) since process models for battery plants are scarce or might require extensive data not available to the LCA practitioner (e.g., [44,45]). Moreover, process models might be costly/time-consuming and difficult to integrate with LCA software, hindering uncertainty and scenario analysis. In these cases, Machine Learning (ML) methods have been suggested to create surrogates of process models and facilitate LCA for multiple configurations [41,46,47].

Several researchers have highlighted ML, Natural Language Processing (NLP), and Large Language Model (LLM) applications as a promising avenues to tackle LCA data challenges and allow for widened, low-cost implementation of LCA. These methods are particularly valuable for renewable energy systems, where datasets are fragmented across studies, technical reports, and pilot-scale experiments. In these cases ML and LLM methods can support the LCI compilation with data-driven estimates of key LCI data and automate the search in relevant databases and the literature for LCI information about emerging energy systems.

ML methods for LCI data estimation have been discussed in previous review studies in limited detail. ML methods rely on creating data-driven models of systems (e.g., an emerging technology) that given a set of characteristics (features) can predict other characteristics (outputs) of

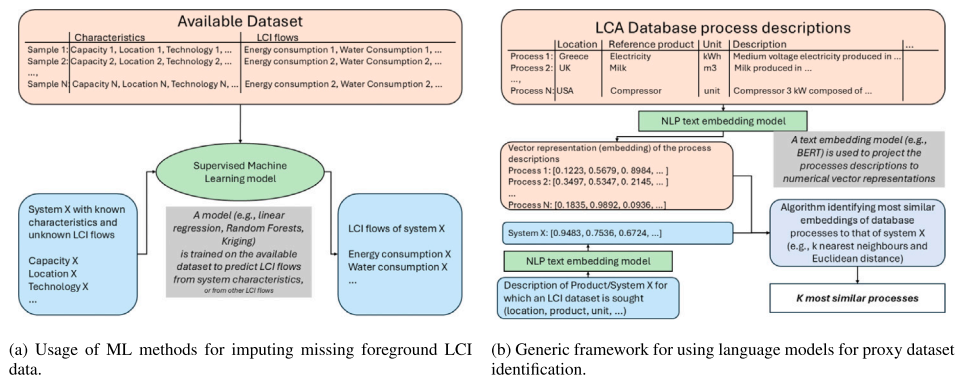


Fig. 1. Generic frameworks for ML (a) and language model (b) applications in LCA.

the system. To create these models a dataset relating the features to outputs is needed. In LCA context, this dataset can originate from lab experiments, process simulations, or from gathering data reported in the literature for similar processes. Fig. 1(a) depicts the process of using ML methods for LCI data generation. Kock et al. [30] focused on traditional LCI methods and briefly discussed ML methods for different LCA phases, without analyzing how they can be applied. Zargar et al. [48] reviewed data-driven, mechanistic, and prospective LCI methods including applications of ML and NLP to estimate missing data, find proxy datasets, and approximate process simulations. They identified a large potential for ML and AI methods in streamlining inventory data collection highlighting the need for further research. Martinez-Ramon et al. [49] focused on using ML methods to create surrogate models for process simulations that estimate LCA inputs. Also, Wright et al. [50] discussed emerging LCI methods noting that the successful application of ML methods depends on the amount of datasets available. Ghoroghi et al. [51] focused on applications of ML in LCA, with most reviewed studies concerning the impact assessment phase. Also, Neupane et al. [52] conducted a qualitative review of ML methods applied in LCA, identifying twenty studies using ML methods to approximate LCI inputs from experimental or literature data. Moreover, Romeiko et al. [53] focused on the methodological choices of ML models used in various LCA phases, mentioning that the sample size of the datasets used varied between 60–20,000 with a median of 540 samples.

Generative LLMs are increasingly used across diverse fields beyond LCA, to retrieve data from unstructured sources. For example, Zheng et al. [54] prompted ChatGPT to find the most relevant sections, summarize them, and extract chemistry-related information from 228 scientific papers. They reported an F1 score of over 90 % against manually extracted information. Furthermore, Dagdelen et al. [55] tested two generative LLMs (GPT and Llama) to extract information about material chemistry from abstracts of scientific papers. They evaluated the results by manual labeling, reporting, for GPT-3, relatively high recall (0.576–0.943), precision (0.4–1), and accuracy (0.5–0.943) values, depending on the information extracted (e.g., formula, acronym, description). Review studies demonstrating the application of NLP and generative LLMs for LCI compilation remain scarce. Preuss et al. [56] described potential LLM applications for LCA, including scraping databases and documents to gather LCI data, identifying emission factors, and facilitating the search of LCA databases. The authors presented a general conceptual framework for LLM-assisted LCA, without putting emphasis on current applications. Cornago et al. [57] conceptually designed a workflow to fine-tune LLMs with LCA data and use them to gather and evaluate data needed for LCA, without specifying how such a framework can be implemented. Also, Tu et al. [58] discussed the potential application of NLP and LLMs in LCA, highlighting their potential to extract inventory information from various document types. The authors provided a generic framework for the usage of LLMs for inventory modeling

based on Amazon tools. NLP and LLM methods for LCA applications, have been used to identify proxy datasets based on textual descriptions as in [59,60]. In this case, a language model is provided with text descriptions of (1) the investigated (query) system and (2) the LCIs available for similar systems and transforms them, via its tokenizer, into N-dimensional numerical vector representations (embeddings). Next, a similarity measure can be calculated between the vectors to identify the datasets that are most similar to each other. Fig. 1(b) depicts this process.

1.2. Scope of the study

Fig. 2 defines a typology of data challenges commonly encountered when searching for LCI data. These challenges involve seeking LCI data in (1) structured databases or (2) unstructured sources and are classified based on the size of the available data. Large structured datasets are typically available for well-established systems with robust reporting, such as direct emissions from conventional energy systems. For example, via its GHG reporting program, USEPA requires major GHG emitters in the US to report their emissions, which are published annually in various formats (e.g., by unit and fuel type) [61,62]. Similar data by plant and region can also be found from USEIA [63]. In contrast, structured datasets for emerging renewable energy technologies are scarce and data are often dispersed across unstructured sources. Data may be present in formats, such as PDF reports, websites, or online databases, requiring manual collection, making the extraction process time-consuming, costly, and error-prone, hindering the LCA of such emerging energy technology and thus potentially delaying their adoption.

Tackling data challenges and improving the quality and coverage of LCIs is key for reliable LCA of emerging clean energy technologies. Santucci et al. [64] identified significant data gaps in the current LCIs used for emerging, clean, hydrogen production technologies (e.g., water electrolysis) including missing values for LCI data and background datasets. A similar observation was made by Paul et al. [65] for battery recycling technologies, noting that the results of LCI studies can be influenced by data gaps and inconsistencies. For example, Muller et al. [66] showed that using outdated PV inventories led to 2.3–4.3 times higher GWP impact compared to an updated, industry-based inventory. Similarly, Pratiwi et al. [67] found that the calculated GWP impact for geothermal energy production in Geneva (electricity and heat) varied by 20–25 % (5–95 % percentiles) when data quality (reliability, completeness, temporal, geographic, and technological correlation) was considered. Such variations can impact which technologies appear environmentally favourable, influence investment priorities, and affect how quickly specific emerging solutions are integrated into energy-transition plans. When LCIs are outdated, incomplete, or not representative of current technologies and supply chains, LCAs may mis-rank competing technologies, resulting in misguided policy signals, slower scaling and deployment, or misaligned R&D and infrastructure investments. Integrating ML and LLM methods in LCI workflows can help alleviate

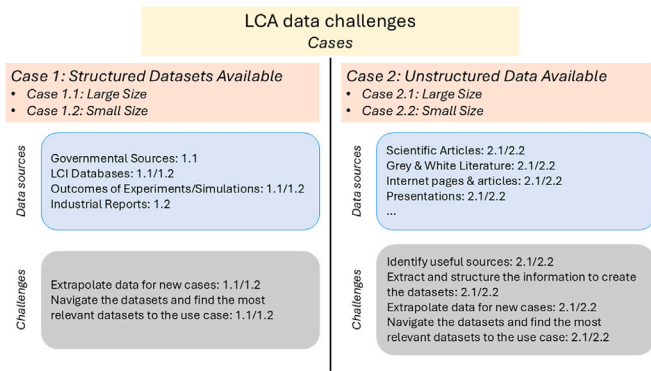


Fig. 2. Typology of LCI data challenges.

these issues, for example by using ML models to impute missing energy consumption data for battery manufacturing under different conditions, and LLMs to extract LCA data and technical parameters from different sources (i.e. industry reports, literature). By strengthening the completeness and representativeness of LCIs, these methods reduce the risk of misestimating the environmental impacts and benefits of emerging clean-energy options. This supports clearer comparisons between technologies and gives planners and policymakers more confidence in setting priorities for incentives, deployment pathways and investment.

This study explores how LCI data challenges (Fig. 2) can be alleviated by using ML and AI methods to enhance and complement traditional LCI compilation methods. It aims to provide a critical evaluation of the algorithms used and their technical specifications and bridge the gap between AI methods and LCA practitioners. It investigates separately ML and language model methods for LCI, identifies strengths and weaknesses of the different approaches, and suggests how they can complement each other. To analyze each case (ML and language models) a structured approach is followed by first exploring the relevant literature and recent developments, second suggesting frameworks for the utilization of ML and language model methods to facilitate LCI compilation, and third deriving observations based on what is reported in the literature and the results of the suggested frameworks. The outcomes of this study indicate that ML and LLM methods can complement traditional LCI methods and facilitate LCI compilation especially in cases governed by data-scarcity. This highlights the crucial role ML- and AI-assisted LCI methods can have for the environmental assessment of novel clean energy technologies, for which data-scarcity is pertinent, and thus for supporting sustainable development. Summarizing, this study contributes to the literature by:

1. Exploring how ML methods can be used to impute LCI data from previously reported information, describing technical specifications and bottlenecks,
2. Providing the first, to the authors' knowledge, review on language model methods for LCI compilation, discussing in detail the pertinent technical implementations and drawbacks,
3. Assessing concurrently ML and language model methods for LCI data revealing how they can supplement each other towards addressing LCA data challenges,
4. Proposing implementations and advancements for LCI-focused ML and AI methods to support the deployment and assessment of emerging technologies, such as renewable and low-carbon energy systems, where reliable LCI data are urgently needed.

2. Materials and methods

This section describes the methods followed to evaluate ML and language model applications for LCI compilation. First, ML methods for imputing missing LCI data using previously reported LCI datasets are

investigated via a literature review and two case studies testing five ML algorithms. Second, current language model applications are reviewed. This is followed by proposing a generic framework for using generative LLMs to extract LCI information, a promising and under-explored approach as indicated by the literature review, and testing it in a proof-of-concept example for an emerging energy technology (deep geothermal energy). The combined analysis highlights how ML and language models can complement each other towards unified robust ML-assisted LCI compilation frameworks.

2.1. Machine learning methods for LCI compilation

2.1.1. Literature Review Design

Three main applications of ML models for LCI data imputation are identified, (1) approximating process models, (2) training on data originating from experiments and technical databases for specific applications, and (3) training on data reported for similar, but independent, systems sourced from LCA databases and studies. Cases 1 and 2 have been explored in previous review studies [38,49,50,52,53] and are briefly discussed in this study, with emphasis on the, underexplored, case 3 which is becoming more relevant with the increasing adoption of environmental data reporting especially for emerging technologies.

Following principles of systematic reviewing methods to eliminate bias and increase coverage [68,69], the Google Scholar, Web of Science, and ScienceDirect databases were searched using keyword combinations relating to Machine Learning ('Machine Learning', 'Artificial Intelligence', 'ML', 'AI', 'Data-driven') and LCA ('LCA', 'Life Cycle Assessment', 'Life Cycle Analysis', 'Environmental Analysis'). The resulting articles were required to be in English, employ data-driven methods to estimate LCI flows, and elaborately describe the data-driven methods applied. A few selected articles were considered for cases 1 and 2, which have been explored in previous studies. To increase coverage, backward and forward snowball review methods were employed. Twenty-one articles on using ML methods for LCI data were included in the literature review, focusing on the ML models used, performance metrics calculated, field of application, dataset size and source, independent and dependent variables, and main conclusions regarding how ML models assist LCA. The insights from this review guided the selection of ML models and the design of the experimental pipeline in Section 2.1.2, ensuring that the models tested reflect the diversity of approaches identified in the literature while addressing the specific challenges of LCI datasets.

2.1.2. Case Study and ML Pipeline Design

Two case studies are investigated testing ML methods for estimating LCI data from previously reported datasets containing LCI information. These include a) a strongly linear dataset describing CO₂ emissions occurring during vehicle driving [70] and b) a non-linear dataset describing CO₂ emissions reported for US power plants [71]. Table 1 summarizes the ML models tested. The models were selected to represent widely used and well-explored ML algorithms with robust proof and documentation that represent different families of ML methods including linear (LR), regularized linear (RR), ensemble (RF, XGB), probabilistic (GRP), and neural network (ANN) models [72]. The diversity of the tested models allows to thoroughly assess the capability of ML models to work on LCI datasets that are governed by characteristics such as limited sample size, mixed data types, and potential non-linear relationships between variables. For example, RF and XGB can handle well categorical variables [73], while GPR is expected to work well with small datasets, and ANNs can excel in capturing complex relationships in non-linear datasets provided sufficient dataset size [74,75]. Furthermore, the models tested were identified as commonly used in the literature review performed in the previous section.

Fig. 3 depicts the ML pipeline applied in the two datasets. First, duplicate features, missing values, and outliers were removed in a data analysis step leading to 7385 and 5038 examples for the linear and

Table 1
Machine learning models tested for LCI data prediction.

ML model	Brief description	Python library used in this study
Linear Regression (LR)	Linear regression with multiple parameters as independent variables	scikit-learn [76]
Ridge Regression (RR)	Linear regression with L2 regularization	scikit-learn [76]
Random Forest (RF)	Ensemble method based on multiple Decision Trees	scikit-learn [76]
XGBoost (XGB)	Ensemble method based on boosted Decision Trees	XGBoost [77]
Gaussian Process Regression (GPR)	Non-parametric method modeling distributions over functions	scikit-learn [76]
Artificial Neural Networks (ANN)	Interconnected layers of artificial neurons embedded with an equation relating inputs to outputs	tensorflow [78]

non-linear datasets respectively. Categorical variables were encoded to integer values and skewed numerical variables were log-transformed. An 80–20 ratio was used to split the dataset into training and test sets. A consistent random sampling procedure was used for splitting the dataset into training and test sets and for the ten internal loops (Fig. 1), followed by the same steps for training each model in each loop. This procedure, along with the ten internal evaluations reduced the effect of data partitioning and enhanced the comparability of the results achieved with the different ML models. The ML models were trained for varying dataset sizes (20–2000 examples) and ten times for each size using consistent random sampling across models. The model performance was assessed with the R^2 and Mean Absolute Error (MAE) metrics and the average value and standard deviation for the ten loops were computed. As shown in Fig. 3 in each loop, normalization, feature selection, and hyper-parameter tuning using grid search cross validation were performed. Notably, for the ANN several architectures were manually tested, with the final architecture involving four layers of 64, 32, 16, and 1 nodes respectively, all using the relu activation function except for the output layer which used a linear function. The adam algorithm [79] was used for training with MAE as the loss function, allowing a maximum of 100 epochs. Sections S1 and S2 of the supplementary material provide a more detailed explanation of the ML pipeline implemented. Together, the critical review and case study results established the foundation for analysing and critically evaluating the performance of ML models for LCI data imputation, considering typical characteristics and challenges

of datasets in emerging energy technologies, including limited sample size, mixed data types, and complex non-linear relationships between parameters.

2.2. Language model methods for LCI compilation

Language model applications for LCI have focused on automating and reducing the cost of LCA to increase its application. To identify relevant studies, the Google Scholar, Web of Science, and ScienceDirect databases were searched using keyword combinations related to LCA ('LCA', 'Life Cycle Assessment', 'Life Cycle Analysis', 'Environmental Analysis') and NLP or LLMs ('Natural Language Processing', 'NLP', 'Generative LLM', 'Large Language Models'), following systematic reviewing methods [68,69]. Significant work on LCA automation is conducted outside of scientific journals. Thus, to capture the current state of development, the review is extended to include software tools, GitHub repositories, and reports and websites describing frameworks and language model applications for LCI compilation from both academic institutions and private companies. Only studies describing NLP and LLM applications and frameworks for the LCI phase were considered.

Fig. 1(b) depicts, the most frequent language model application identified, which concerns the automatic matching of foreground systems with background databases. Specific focus is placed on studies exploring such frameworks, since increasing interest in automation is observed in the LCA industry [80,81]. Additionally, generative LLM applications for LCI compilation are reviewed, including (1) matching text descriptions to background datasets and identifying emission factors, (2) directly generating LCI data, and (3) extracting data from documents. Finally, a framework is proposed and tested for utilizing generative LLMs to extract LCI information from unstructured documents. The generic framework is described in the next paragraph and is tested in a proof-of-concept example for extracting LCI information from four LCA studies on geothermal energy production. This analysis represents the first systematic review focusing on the use of language models for LCI compilation, bringing together both academic and industrial developments. The insights gained from this review guided the development of the generative LLM framework presented in the next section, which was designed to explore the largely under-investigated potential of language models for LCI information extraction from unstructured sources, given the limited number of studies addressing this capability to date.

2.2.1. Generative LLM framework for LCI information extraction

Fig. 4 presents the framework proposed for extracting LCI data from unstructured sources using generative LLMs. First, documents and sources that might contain LCI-related information for an investigated

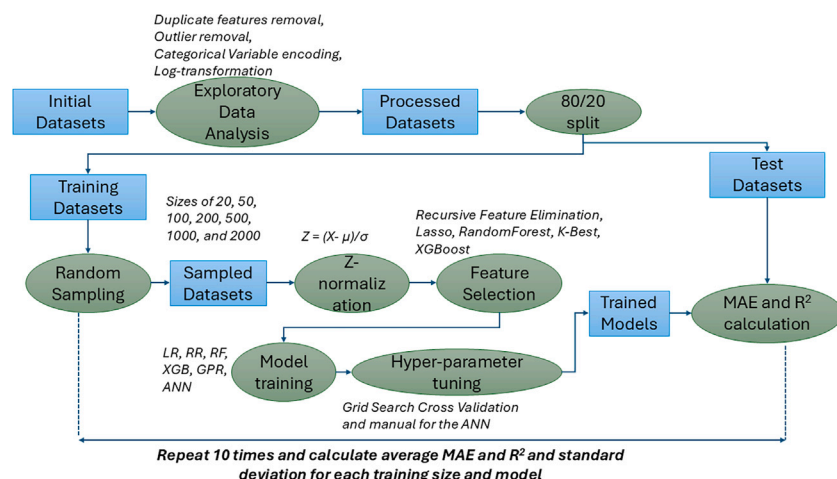


Fig. 3. Machine Learning pipeline applied for model evaluation.

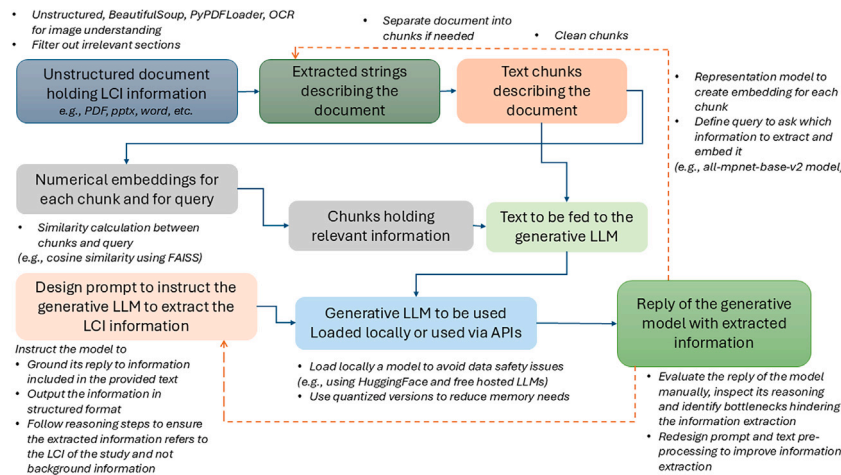


Fig. 4. Framework for utilizing generative LLMs to extract LCI information from unstructured documents.

use case are gathered. These might involve scientific papers, word documents (e.g., the supplementary materials of papers that usually contain valuable LCI information such as mass balances and inventory tables), presentations, grey/white literature reports, or websites. The collection of these sources can be done manually, or more automatically using a web-scraping application. Depending on the document type, different tools can be utilized to process the documents and extract the text (e.g., [82–84]). Document sections that are not expected to contain LCI information (e.g., introduction, conclusions, references) can be removed to allow more precise textual analysis and reduce the computational needs. Next, the extracted text can be separated into chunks to identify the most relevant ones to provide to the generative LLM for information extraction. To identify the most relevant text chunks, these can be embedded to numerical vectors using a language model along with a query describing the LCI information to be extracted. The most relevant text chunks can then be identified as those having the highest cosine similarity to the query. Alternatively, the entire text can be provided to the LLM, though this can lead to higher computational needs and less efficient information extraction especially for smaller language models. The framework follows a retrieval augmented generation (RAG) approach, grounding the responses (thus the extracted data) of the generative LLM exclusively on what is mentioned in the document. This ensures that the extracted values originate from the provided document and are traceable and verifiable. The text provided to the generative LLM is accompanied by a prompt. The prompt should contain domain knowledge to give context to the LLM and instruct it about what information to extract, to provide its reply solely based on the provided text and avoid hallucinations, to output the requested information in a structured format (e.g., json), and what steps to follow during information extraction. Other prompt engineering techniques such as few-shot prompting [85] and chain-of-thought [86] can be utilized. Extracting only LCI numerical data (e.g., electricity consumption), without the relevant process and methodological metadata, can limit the usage of extracted data. Therefore, the prompt should instruct the LLM to also extract both LCA and technical metadata such as the functional unit and boundaries used and the operational conditions of the system (e.g., capacity factor or temperature). Accompanying the LCI numerical values with such data ensures that the extracted information is well-defined and can be later evaluated and used without mistakes. This also allows the LCA practitioner to identify undefined data sources and exclude them from further consideration.

Next, the response of the generative LLM and the steps it took to achieve it, can be inspected to identify issues hindering information extraction and refine the prompt, query, and text pre-processing. This

human-in-the-loop step allows designing the application to target the extraction of numerical LCI data and required methodological metadata without being distracted by similar information mentioned in the documents. Generative LLMs can be used via provider APIs [87,88] or open models can be loaded locally [89]. The former results in low computing times, but can lead to data safety or high-cost issues. By contrast, locally deployed generative LLMs avoid these issues, but require access to compute which might be unavailable.

Framework testing. The framework suggested in Fig. 4 is tested on four LCA papers on geothermal energy extraction [39,90–92]. Nine types of information were extracted: Two numerical LCI flows (Steel and diesel consumed for well development), and seven related technical and methodological metadata (plant type, capacity, capacity factor, operational hours, and lifetime, well depth, and functional unit of the study), which were found in a previous study [26] to be crucial for the environmental performance of geothermal LCA plants and to be frequently reported in geothermal LCA studies. The extracted metadata consist different data types (e.g., numerical, textual) to demonstrate the capabilities of generative LLMs for meaningful LCI information extraction. Two generative LLMs (GPT4.1 and a locally deployed Qwen-14B) were tested and the all-mpnet-base-v2 text embedding model [93] was used to generate the text embeddings. The framework was tested once by feeding the LLMs the entire document, and once using only the text sections that had the highest cosine similarity to the following query: ‘How much is the steel (kg/m) and diesel (MJ/m or lt/m) consumption to drill one meter of well? How much is the capacity, lifetime, and annual operational hours of the plant? What functional unit is used?’. The FAISS library [94] was used to calculate the cosine similarity between the embeddings of the text sections and the query. To avoid proprietary issues, the GPT4.1 model was given only the most relevant sections of the document. The extracted information was compared with manually extracted information and classified as correct if they are semantically equivalent.

The selection of LCI data and associated metadata extracted in the exemplary case study did not exhaustively cover parameters determining the performance of geothermal systems, such as geological setting, permeability, thickness, geofluid flow rate, etc. This case study aims to investigate the potential of using generative LLMs to extract structured LCI-related data from unstructured sources across sectors, and facilitate LCA of emerging technologies for which LCI data are scarce. The same framework for LCI data extraction could be applied to other sectors or for detailed geothermal-related applications by adapting the prompt and query to extract more domain-specific parameters.

2.3. Integrated evaluation

The methodological structure adopted in this study linked systematic literature reviews, ML case studies, and proof-of-concept generative LLM framework testing to enable a coherent evaluation of state-of-the-art methods for LCI compilation in the energy sector. This ensured that the analyses and critical evaluation of the investigated techniques were both explanatory and comparative, allowing identification of dependencies between data characteristics, model behaviour, and methodological setup. It provides a solid foundation for the moving forward analysis, where the complementarity of ML and language model methods is examined and integrated frameworks are proposed to address LCI data challenges for emerging energy technologies.

3. Results and discussion

3.1. LCI data estimation using machine learning methods

Table 2 presents an overview of the twenty-one studies reviewed that applied ML methods to estimate LCI data. Ten studies [46,47,95–102] used ML to estimate process parameters, (e.g., process yields and fuel characteristics), that are used for LCA calculations. Another eleven studies [103–112] used ML methods to directly estimate LCI data (e.g., energy and material consumption) from other available LCI datasets. Notably, thirteen studies applied ML methods to facilitate the LCA of energy-related systems [46,47,95–97,99,100,102,105–107,109,113], with another three predicting energy consumption for chemical [110] or metal manufacturing [108,111].

3.1.1. Cases 1–2, machine learning models to approximate well-documented processes and simulations

Studies using ML to estimate process parameters used for LCA calculations mainly concern specific processes for which data are usually available in the literature and are governed by process laws. This facilitates ML modeling, since the training data can be representative of the process. Most of the relevant reviewed studies used over 150 samples and employed advanced ML algorithms with ANNs being used predominantly. High performance (e.g., $R^2 > 0.9$ [46,96,98]) are reported in these cases as the experiments and dimension space can be redesigned to facilitate ML modeling and create models that accurately predict values for narrower boundaries of design conditions. This applies also when ML models are used to approximate process simulations and use their outputs for LCA, with high R^2 (0.9–1 [114,115]) and low prediction error [116] being reported. Nevertheless, significant challenges remain. Complex simulation models or experiments need to be designed and ran multiple times to enable sufficient datasets for ML modeling. Representative features must be selected such that the relationships between features and outputs are properly captured and the model can generalize for unseen cases, avoiding overfitting and risking unreliable predictions when deployed. Domain knowledge and knowledge about the scope and limitations of the ML model are essential for ensuring its best use. For example, ANNs developed in [41] to predict the environmental impacts of a geothermal power plant from its design parameters cannot be used directly for other geological settings without considering that they were developed for a specific case. Also the transparency and robustness of the ML procedures followed by LCA practitioners needs to be ensured. Rankin et al. [117] reported that in several cases in which industrial ecology experts used ML methods, methodological and implementation errors were identified, while the lack of appropriate datasets also hindered conclusions derivation. Data availability when simulations or specific cases (cases 1–2) are targeted helps the usage of ML models. Nevertheless, the design and training of such models involves difficulties and requires domain, LCA, and ML knowledge such that the quality of data is ensured, appropriate features are selected, methodological errors during training are avoided (e.g., test data leaking), and the best use of the models is ensured.

3.1.2. Case 3, machine learning models to estimate LCI data from small datasets originating from unrelated sources

In the reviewed studies of case 3, the size and origin of the dataset varied, influencing the ML approaches used and conclusions derived. Two studies showed that non-parametric models such as local linear regression and ANNs achieved higher R^2 values than MLR for predicting LCI data such as CO₂ emissions of coal-fired power plants (0.49 vs 0.61 [109]) or energy consumption of chemical manufacturing (up to 0.74 vs less than 0.29 [110]) from characteristics of the power plant or chemical. MLR was also tested for predicting air emissions of dual-fuel engines in [105] achieving an R^2 of 0.4–0.6. The three studies mentioned used datasets of over 100 examples compiled from various sources including literature and on-site observations [105] or governmental reports [109], or LCA databases and in-house data [110]. Another three studies employed the kriging algorithm for relatively small datasets (e.g., less than 35 examples), gathered from the literature, industrial reports and LCA databases. Of these, two studies showed that kriging predicted LCI data (e.g., material and energy consumption) for wind [106] or hydropower [107] plant construction from plant characteristics with a lower MAE than linear regression, with only the former study reporting exact numerical results (e.g., MAE of 0.08 vs 0.32 and 0.09 vs 0.42 for the electricity and aluminum consumption). Also, the prediction accuracy improved with the co-kriging variant when correlated parameters were considered for model training. The third study [111] did not provide a comparison with linear regression, and focused on showing that introducing stochastic and Bayesian processes to kriging reduced the MAE (from 0.61 to 0.59 and 0.39) for predicting the energy consumption of metal powder production processes. Moreover, GPR, an algorithm with a strong resemblance to kriging [118], led to higher prediction accuracy for filling data gaps in datasets about fertilizer consumption than using the average value in [103]. Furthermore, two studies used ecoinvent unit datasets to train ML models that could predict data missing from them. One study [108] focused on manufacturing processes (315 examples) finding high R^2 values (close to 1), with RF and boosted trees performing the best of the tested models. The other study [112], used all available datasets (2546 examples) to train ML models to predict the missing LCI data from the remaining available data, with XGBoost leading to the highest R^2 (0.622 for 1 % of the LCI data missing, and 0.152 for 70 % of the data missing). An additional two studies adopted a distance-based approach (e.g., calculating the Minkowski distance) to identify the K most similar LCI datasets, over an available pool, to a case study, and use these to estimate missing LCI data for the case study (e.g., by taking the average value). One study used all ecoinvent unit datasets [104] finding that the prediction became unreliable for datasets with a lot of gaps (e.g., MPE 92 % for 1–20 % of LCI flows missing). In contrast, Meron et al. [113] used datasets gathered from the literature and showed that the LCI data estimated with this method led to more accurate GHG prediction than intuitively selecting an LCI dataset from a database.

Although interest in using ML methods for LCI data estimation is increasing, studies investigating such methods remain scarce. A trend indicating that ML models can perform better than conventional MLR methods is shown, with different models being selected for varying dataset size. For example, complex models such as ANNs and ensemble models are more frequently used for relatively larger datasets, while for smaller datasets models like kriging are more frequently encountered. However, the performance and suitability of ML models for LCI data imputation across varying dataset sizes and types remain underexplored, and more research is needed to provide robust recommendations for their usage.

3.1.3. Case study results and performance of ML models

This section assesses the performance of common ML algorithms for LCI data prediction based on the literature analysis described in the previous paragraph and on performance results for two test datasets directly assessing the suitability of algorithms across different dataset types

Table 2

Studies using machine learning algorithms to predict missing inventory data. Lit: Literature. GBR: Gradient Boosted Regression, MLR: Multiple Linear Regression, SVM: Support Vector Machines, ADB: AdaBoost, DT: Decision Trees, KR: Kernel Regression.

Study	ML models	Application	Dataset size	Dataset source	Independent variables	Dependent variables	Metrics	Conclusions
[95]	GBR, RF, SVM, XGB/ADB, ANN	Biomass gasification	312	Lit.	Process parameters	Gas properties	R^2 :0.24–0.97	GBR outperformed other models
[96]	Regression	Biodiesel production	33	Experiments	Process parameters	Biodiesel yield and HHV	R^2 :0.80–0.99	ML can be used to predict parameters used in LCA RF performed the best on predicting process parameters that were used for LCA calculations
[47]	MLR, RF, SVM, ANN	Hydrothermal treatment of biomass	800	Lit.	Process parameters	Yield, energy and carbon content of products	R^2 :0.16–0.90	
[97]	RF	Pyrolysis	538	Lit.	Feedstock and process parameters	Yield, energy, carbon, and N content of biochar	R^2 :0.45–0.87	
[103]	GPR	Food production	>2000 ¹	Gov. sources	Year, Location, Species	N fertilizer amount	MAPE:7.4–305 %	GPR can predict LCI data gaps and quantify prediction uncertainty
[104]	Link prediction	Ecoinvent processes	2546	Ecoinvent	LCI flows in the unit process	Missing LCI flows in the unit process	MPE 0–91 % ² .	Low MPE when less than 5 % data are missing
[98]	ANN, ANFIS	Sugarcane production	120	On site farm data	Energy, farming parameters	Production (in energy units)	R^2 :0.9–0.99, MAPE: 0.006–0.11 %	ANNs can predict parameters used for LCA calculations
[46]	ANN	Biomass pyrolysis	250	Lit.	Process parameters	Active carbon yield	R^2 :0.971	ANNs can predict parameters critical for LCA models.
[99]	RF, XGB, ANN	Biomass treatment	65	Lit.	Feedstock and process parameters	yield, chemical consumption	R^2 >0.7	ML models can predict key parameters used in LCA modeling
[105]	MLR	Dual fuel engines	114	Lit.	Operation and design parameters	GHG and air pollutant emissions	R^2 :0.39–0.59, MARE: 2.7–27 %	ML models can predict values to fill data gaps in process inventories
[106]	Kriging, co-kriging	Wind turbines	5	Ecoinvent	Wind turbine characteristics	Material and energy consumption	MAE of kriging lower than LR ³	Kriging and co-kriging outperform LR for estimating missing LCI data
[107]	Kriging, co-kriging	Hydropower	27	Lit.	Hydropower plant technical characteristics	Material and energy consumption	MAE of kriging lower than LR	Same as above
[100]	GBR, SVM, ANN, KR	Methanol production	1765	Lit.	Process parameters	Yield product parameters	R^2 :0.86	ML models can predict key process metrics to be used in LCA studies

(continued on next page)

Table 2 (continued)

Study	ML models	Application	Dataset size	Dataset source	Independent variables	Dependent variables	Metrics	Conclusions
[101]	ANN	Solid waste handling	245	Lit.	Process parameters and biomass specifications	Yield	R^2 :0.81–0.96	ANNs can predict variables critical for LCA studies
[102]	GPR, ANN, SVM, DT	Anaerobic digestion	167	Lit.	Process parameters	Yield and methane content	R^2 :0.82–0.84	ML models can be used to estimate parameters used in LCA studies
[108]	DT, RF, ADB, Gradient boosting	Manufacturing process	315	Ecoinvent	Process parameters, type, and technology	GHG emissions, energy, water, material consumption	R^2 close to 1	ML models can be used to estimate missing LCI data. Gradient boosting performed the best
[109]	MLR, LLR	Coal power plants	444	Lit.	Plant design and location parameters	CO ₂ emissions	R^2 :0.49 (MLR), 0.61 (LLR)	Regression models can be used to predict data gaps in LCI inventories
[110]	MLR, ANN	Chemical production	103	Ecoinvent, in house	Chemical composition parameters	Energy demand and impacts	MLR: R^2 <0.29, ANN: R^2 <0.74	ANNs outperformed MLR. They can be used to fill LCI gaps for chemical production.
[111]	Kriging, Bayesian Kriging	Metal-powder production	34	Lit., on site data	Process parameters	Energy consumption	MAE/MSE lower for bayesian kriging ⁴	Kriging models can predict missing LCI data.
[112]	DT, XGB, RF, SVM, ANN	Ecoinvent unit processes	2546	Ecoinvent	LCI flows in the unit process	Missing LCI flows in the unit process	R^2 :0.152–0.622, MPE:39–85 % (for 1–70 % missing data)	XGB and RF can estimate missing LCI data. Prediction error increases for more missing data.
[113]	Similarity-based prediction	Water supply/coal-fired systems	22/100	Lit.	Technical system characteristics	GHG emissions	N.A.	Similarity-based prediction outperformed arbitrarily selecting proxy datasets.

¹ The dataset was separated into much smaller datasets for different scenarios.² MPE was near 0 for 1–5 % of data missing, 40 % for 10 % of data missing, 91 % for 20 % of data missing.³ MAE of kriging prediction was lower than for LR for the consumption of: electricity (0.08 vs 0.32), aluminum (0.09 vs 0.42), copper (0.43 vs 0.72).⁴ MSE: 0.61, 0.59, and 0.39 and MAE: 0.62, 0.58, 0.48 for conventional, stochastic, and Bayesian stochastic kriging respectively.

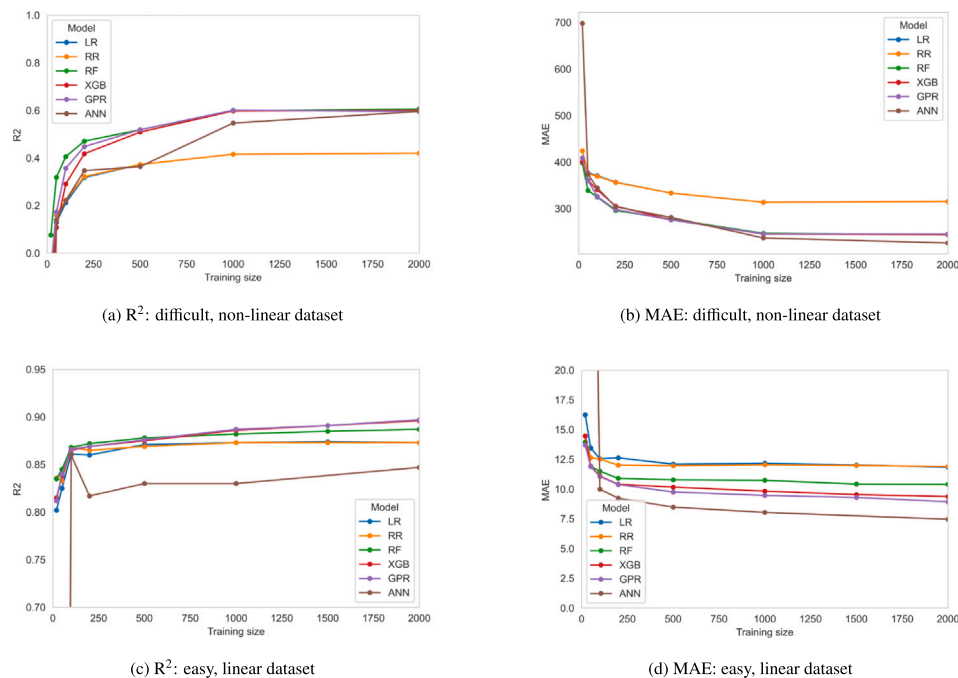


Fig. 5. Performance of ML algorithms on the test set against the dataset size for strongly non-linear (top) and strongly linear (bottom) datasets.

Table 3

Comparative performance of ML algorithms vs. linear regression. *Values worse than LR, with negative R^2 .

Dataset size	Dataset type	Metric	LR	RR	RF	XGB	GPR	ANN
Small (20-100)	Linear	R^2 MAE	0.80–0.8612.6–16.2	0.84–0.8712.5–14.5	0.83–0.8711.5–14.0	0.81–0.8611.1–14.5	0.8–0.9711.1–14	see*
	Non-linear	R^2 MAE	0.0–0.21371–425	0.0–0.22370–425	0.08–0.41325–398	0.0–0.29341–402	0.0–0.36326–409	see*
Large(200-2000)	Linear	R^2 MAE	0.86–0.8711.9–12.2	0.86–0.8711.9–12.0	0.87–0.8910.4–10.9	0.87–0.909.4–10.4	0.87–0.98.9–10.4	0.82–0.857.5–9.2
	Non-linear	R^2 MAE	0.32–0.42314–334	0.32–0.42314–334	0.47–0.61244–296	0.42–0.60245–306	0.45–0.6246–298	0.35–0.6226–304

(linear and non-linear) and sizes (20–2000 samples). Fig. 5 and Table 3 present the results of the case studies, comparing the performance of ML algorithms against linear regression.

For small datasets, (e.g., 20–100 samples), RR, RF, XGB, and GPR outperform LR, achieving 0.5–4.2 % higher R^2 for the linear dataset with the difference increasing for the non-linear dataset (up to 16.2 %). In contrast, ANNs perform poorly for small datasets possibly due to their large parameter space and overfitting risk. This observation is in line with what is reported in the literature, with non-parametric models such as GPR, kriging, and LLR being mainly employed to fill data gaps when scarce information is available, since they can lead to better data estimation than LR or using mean or proxy values [103,107,109]. For larger datasets (e.g., 200–2000 samples) the difference in performance between the ML algorithms and LR increases, with ANNs also outperforming LR. In fact for both datasets ANNs lead to the lowest MAE, while XGB and GPR achieve the highest R^2 . This is because the ANNs were trained using the MAE as loss function minimizing the error on most dataset points in expense of capturing the variance of points that lay further in relation to the mean. Notably, ANNs behave differently for small dataset size producing unstable predictions (negative R^2 values and oscillatory behavior). A similar pattern is shown in the reviewed literature, with ANNs and ensemble models (e.g., RF, XGB) performing well with relatively large datasets (e.g., over 100 examples) [46,108,110,112], and across various sectors such as chemical manufacturing or biomass pyrolysis. Notably, the results reported in Fig. 5 and Table 3 are the mean values over the ten loops that were run for each model and training dataset size (Fig. 3). The standard deviation of the R^2 and MAE over the ten training loops is relatively large for small datasets, especially for ANN, XGBoost, and GPR. This suggests that the model performance is

sensitive to the sampled training data, and thus caution should be paid when developing such models based on small datasets (e.g., to perform stratified sampling ensuring representativeness of the training dataset for the design space). Nevertheless, the standard deviation decreases for increasing training size for all models, indicating increased stability and reliability of the predictions. The detailed numerical results achieved for the two datasets can be found in section S3 of the SI.

The results indicate that ML models can impute missing LCI data better than more conventional linear regression methods, when scarce, unrelated data from similar systems are available. Depending on the type and size of the available dataset, the suitability and effectiveness of ML models can vary. In line with what reported in ML literature [74,75] complex ANN and ensemble models tend to be used for relatively large datasets, while algorithms such as kriging and GPR can be used even for small datasets outperforming MLR and other conventional methods. ANNs and ensemble models perform well with relatively large datasets (e.g., over 100 examples) and can work with various uncorrelated features. However, they have been tested in limited detail for estimating LCI data based on small datasets. Also, their complex implementation makes them essentially black-box models, providing difficult to interpret predictions, that can be problematic, when the approximated LCI flow causes a large portion of the environmental impacts. In contrast, kriging and GPR methods provide interpretable estimates, also quantifying the uncertainty of the prediction. These methods use covariance functions to capture spatial dependencies in the feature space putting more weight on the information held in neighboring data points when making their prediction [118]. They can thus model complex relationships, particularly when spatial correlations exist in the feature space, making them a good fit for datasets with such feature correlations. Introducing

Table 4
Advantages and disadvantages of used ML algorithms for missing LCI data imputation.

Method	Advantages	Disadvantages	Performance & Applicability
MLR	Simple, Interpretable, Good benchmark	Fails to capture non-linear relationships, Sensitive to outliers	Performs well on large linear datasets but not with complex systems.
Kriging & GPR	Captures spatial dependencies, Interpretable, Quantifies uncertainty	Sensitive to uncorrelated variables	Can perform better than MLR for small datasets, Better performance for correlated features
ANN and Ensemble models	Models complex relationships, Can handle non-linear patterns	Requires large datasets, Black-box model, Difficult to interpret	Best for large datasets with complex, non-linear relationships Performs well with medium-to-large datasets

randomness into kriging was shown to improve the estimation accuracy in [111], potentially because the randomness allowed the estimator to include variance not considered in conventional kriging. However, this result is specific to the test case investigated and more research is needed before making generalizations. In fact, in case of weakly correlated variables this approach could potentially lead to the introduced noise having a disproportionate influence on the prediction and its variance. Further research is needed to assess the performance of kriging and GPR on larger datasets of LCIs and when using uncorrelated features. Table 4 provides an indicative summary of the strengths and weaknesses of the different ML methods for missing LCI data estimation based on the literature reviewed and the analysis shown in Fig. 5.

3.2. Language model methods for LCI compilation

This section first reviews current developments in using language models for LCI compilation. The review first focuses on studies using language models for semantic matching between the foreground and background systems of an LCI to reduce costs and automate LCA (a commonly researched application: Table 5) and second on studies and applications that use generative LLMs to facilitate LCI compilation (e.g., for LCI data generation). Second, the results for the exemplary case study investigating generative LLM usage for LCI data extraction from LCA studies are presented. Finally, the observations from the literature review and case study results are discussed.

3.2.1. Analysis of existing language model applications

Table 5 summarizes six studies found in the literature exploring language models for automatic identification of background LCI datasets (e.g., Fig. 1b). Three studies led by the same authors used language models to match product descriptions to background LCI datasets. Balaji et al. [59] used the reference product of ecoinvent unit datasets and the all-mpnet-base-v2 language model to create numerical representations (embeddings) of the datasets. These were matched to embeddings created for product descriptions by calculating the cosine similarity between the embeddings to identify the closest ecoinvent dataset to the products Fig. 1(b). The authors introduced a classification layer imposing that the product and ecoinvent reference product share HS industry codes. The framework was tested on two datasets of 90 and 667 products, benchmarking its performance against manual selection. For the larger dataset, introducing the intermediate layer led to precision values ranging between 3.2–54.1 %, 68.7–99.7 % and 57.2–75 % for, respectively, identifying a matching dataset, identifying cases where no relevant ecoinvent dataset was present, and overall precision. The respective precision values when the intermediate layer was omitted were lower (9.7 %, 85.8 % and 66.6 %). Nevertheless, the precision values achieved by humans matching the datasets manually were higher (75.5 %, 76.1 %, and 76.4 %), especially for the matching precision, indicating that further development is needed for the algorithm to reach human-level LCA. Notably, most errors originated from semantic mistakes of the language model and errors on the HS code classification. Similar results were shown when the same framework was used for input-output LCA using the USEIO database [119]. In a next study, Balaji

et al. [120] used generative LLMs to refine the product descriptions and as a re-ranker that suggests the most relevant dataset of the ten with the highest cosine similarity (Fig. 6). They tested the framework using the Claude 3 Sonnet generative LLM and gte-large text embedding model on six datasets (four IO-LCA and two process LCA), considering the matching correct if a human would agree with it. The precision of the algorithm ranged between 71–97 % for the six datasets. A similar framework was suggested by Bekavac and Volk [60] who developed an ontology for ecoinvent datasets and employed a generative LLM to explain the algorithm's selection. They used the framework to perform LCA of clothing and textiles without however mentioning quantitative results. Another two studies used similar frameworks to automate the matching of building information modeling (BIM) outputs to Okobaudat LCI datasets [121] and Andalusia database entries [122], reporting respectively, matching accuracy of 79–88 % and prediction error of 25 %.

Moreover, generative LLMs have been directly applied to facilitate LCI compilation, including finding emission factors from background datasets, predicting LCI data, and extracting LCI data from documents. For the former, Gu et al. [123] used Claude 3.5 and GPT-4o to match BIM outputs to Chinese Life Cycle Database datasets, and to calculate the building's embodied carbon. Claude 3.5 led to accuracy, recall, precision, and F1 scores of 0.84, 0.71, 0.70, and 0.69 for finding the database entries, with the error on the embodied carbon calculation compared to manual LCA ranging between 0.6–27 %. Similarly, Präger et al. [124] used GPT-4o-mini to match material descriptions derived from tender documents for construction activities to Okobaudat database entries finding a 0.01–76.8 deviation for the GWP of the building compared to conventional LCA methods. A similar framework was conceptually described in [125], and in [126] where open-source code for such applications was provided, without presenting case study results. Two studies were found using generative LLMs to directly predict LCI data. Luo et al. [127] prompted five generative LLMs to act as experts steel, fabric, and batteries production and generate the LCI for their manufacturing. They mentioned that of the five LLMs, GPT models provided LCIs closest to those generated by human-experts without however providing numerical results. Moreover, Kumar et al. [128] fine-tuned Llama-2 using five-thousand LCA studies and five-hundred pairs of questions-answers related to LCA of methanol production. The fine-tuned LLM was asked to directly predict various LCI flows, such as water and energy consumption for standard methanol production, leading to an error range of 3–44 %. Finally, generative LLMs have been suggested for LCI information extraction from documents aiming to reduce the manual effort needed to gather LCI data. Two studies reported using generative LLMs for extracting LCI information from documents regarding biologically active building materials [129] and sustainability data from reports, papers, and monitoring systems [130], without mentioning specific results on the information extraction or the architecture of the frameworks used. Moreover, the Tiangong initiative [131] offers open-source AI-powered solutions to facilitate searching for sustainability-related information in documents. At the time of writing, a detailed documentation is not available, and the following description of the Tiangong framework is based on the open-source code available on GitHub [132]. The approach

Table 5
Studies using language models to identify proxy background datasets.

Study	ML models	Application	Query text	Database entries text	Dataset	Similarity metric	Performance Metrics	Conclusions
[120]	gte-large, Claude 3.5-Sonnet ^a	Background dataset identification	Product description	Ecoinvent reference product description	Six datasets ^a	Cosine	Precision (first recommendation: 71–97 %, in top 10 recommendations: 73–100 %)	Semantic matching reduces the time to find proxy datasets. Generative LLMs can help refine descriptions
[60]	ada-002, Mixtral-8x7B-v0.1	Background dataset identification	Product description	Ecoinvent LCI datasets ^b	LCA for textiles	Euclidean	Not reported	The proposed method reduced the time needed for LCA.
[59]	all-mpnet-base-2	Background dataset identification	Product description and attributes	Ecoinvent reference product description	Two datasets (90 and 667 products)	Cosine	For the small and large datasets respectively ^c : Matching precision: 13.6–45.5 %, 6.9–54.1 %, No-Match precision: 55.9–98.5 %, 68.7–99.7 %, Overall precision: 53.3–81.1 %, 57.2–75.0 %	The proposed matching algorithm reduces the manual effort for LCA.
[119]	Same approach as in [59] but for IO-LCA and testing 40,000 products. The proposed approach reaches results comparable to human performance.							
[121]	spacy, BERT, GermaNET	Match BIM outputs to Okobaudat database entries	Names and properties of BIM outputs	Okobaudat entries description	Five building case studies	Cosine	BERT-NLP leading to 79–88 % similarity score and 78.1 % correct matching accuracy	The proposed method can reduce the time to find LCA datasets for BIM outputs
[122] ^c	Bag-of-Words	Building sector	Building elements materials sescription	Andalucia entries description	Building rehabilitation	Cosine	Estimated error of prediction matching: 25 %	The algorithm reduced the time needed to search the Andalucia database ^d by 80–90 %

^a Several text embedding models and generative LLMs were tested with the performance of the algorithm ranging. Claude 3.5 Sonnet (generative LLM) and gte-large embedding model led to the highest performance. The datasets tested were: procurement products (3980), invoices (2159), invoices (1121), heavy equipment (1803), food ingredients (1956), packaging (195 unique entries).

^b The authors reported using an own-defined ontology for ecoinvent datasets.

^c Matching: finding the most relevant Ecoinvent dataset. No-Match: Identifying correctly that a relevant Ecoinvent dataset is not present.

^d The Andalucia database is not an LCA database but contains information that can be subsequently used for LCI compilation.

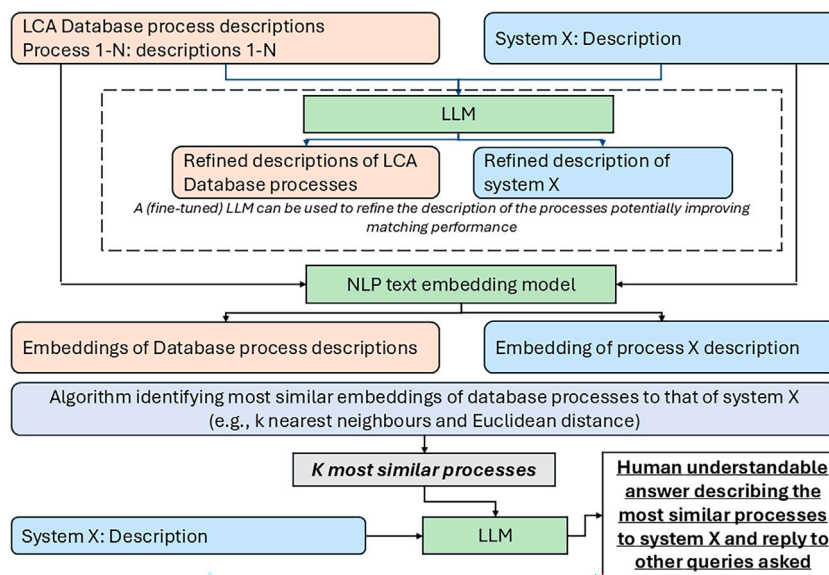


Fig. 6. Generic RAG framework for LLM usage in LCI.

followed is similar to that described in Section 2.2.1, with documents being processed to extract the text, language models used to embed the text, and a vector database used to store these embeddings. Given a user query, semantic matching is used to identify the most relevant documents and sections, which are given to a generative LLM that provides natural language interpretation and recommendations. However, specific results have not, to the authors' knowledge, been reported.

Increasing interest in using language models for facilitating LCI compilation in numerous manners is observed. Automating the identification of background LCI datasets based on semantic matching and embedding models has been tested in different case studies, with the findings indicating that the time and manual effort needed to search LCA databases can be greatly reduced. Nevertheless, significant errors and variations were reported suggesting that more research is needed before these frameworks reach human-level performance for LCA. Generative LLMs have been also suggested to find background datasets or even directly predicting LCI data. However, generative LLM usage for such numerical calculations can raise transparency and replicability issues, while the errors reported in the studies reviewed ranged widely without robust justification for these variations. Regardless, generative LLMs can facilitate LCI compilation by helping to extract useful data from unstructured sources as already done in other fields [54,55]. Such applications can largely help the LCA of emerging decarbonization technologies for which LCI datasets are usually spread across scientific publications and industrial reports. Despite some efforts made (e.g., [131]) actual results and frameworks for generative LLM usage to extract LCI information from unstructured sources have not been, to the authors' knowledge, reported. In the next section, the framework depicted in Fig. 4 is tested.

3.2.2. LCI information extraction from PDFs using generative LLMs

In this section the framework proposed for LCI information extraction using generative LLMs in Section 2.2.1 is tested for four LCA papers on geothermal energy extraction and to extract nine types of information, including numerical LCI data (steel and diesel consumed) and case study related data (e.g., plant type and functional unit). Table 6 presents the information extracted for the steel and diesel consumed. The rest of the results can be found in the SI (Section 5), along with the description of the PDF pre-processing applied and the prompt given to the generative LLMs (Section 4). Both generative LLMs effectively extracted LCI information (steel and diesel consumed) and other methodological and case study related information needed to define the extracted numerical LCI

data. This result highlights the potential of generative LLMs to extract LCI data and related metadata from unstructured sources. The information extracted from the PDFs from the generative LLMs was compared with manually extracted information. Small deviations are observed between the replies of the generative models with some data being missed in some cases. This happens mainly for the well depth and plant capacity. The deviations occur due to the articles investigating multiple case studies with varying depths and capacity and because the prompt used put more emphasis on the steel and diesel consumption, which are the main LCI flows sought. In fact, the replies of the generative LLMs for the latter are more consistent. The only inconsistency observed, regards the steel consumption value for the study of Frick et al. [90], where the local Qwen-14B model, when provided with the entire document, only reported the value for low-alloyed steel, missing the value for high-alloyed steel, due to a reported range of 80–100 kg/m for the steel consumption in the paper.

The information extraction results (Table 6 and Section 5 in SI) demonstrate that generative LLMs can effectively be used to extract LCI information from scientific papers. Notably, more local generative LLMs were tested (for example, Qwen-0.6B, Qwen-8B), with larger models taking more time to run but also providing better responses. Some inefficiencies were identified (e.g., amount of high-alloyed steel not extracted in one case), which can be tackled by redesigning the prompt, using a larger model, providing examples to the generative LLM, and by giving the model more targeted and smaller tasks. In addition, in this study, tables were converted to strings and figures were disregarded. A more delicate handling of these information types could improve the LCI information extraction. Moreover, the LLM was not provided with the supplementary information of the articles, which usually in LCA articles, includes mass and energy balances and detailed inventory tables. In this case, the supplementary information would have to be pre-processed and the prompt would need to be refined accordingly. Such actions could improve the accuracy of the LCI information extraction and need to be further investigated.

However, the scope of this analysis is limited to showcase that generative LLMs can effectively be used to gather LCI data demonstrating the capabilities of generative LLMs for LCI information extraction towards facilitating the LCA of novel technologies, rather than presenting production-ready applications or providing in-depth geothermal assessment. In the proof-of-concept example explored the generative LLM was instructed to extract along with the numerical LCI data (steel and

Table 6
LCI data extraction with generative LLMs.

Study	Data	Gpt4.1 most relevant sections	Qwen14B full document	Qwen14B most relevant sections
[90]	Steel	'69.1 kg/m (low alloyed), 34.0 kg/m (high alloyed)'	'69.1 kg/m'	'69.1 kg/m and 34.0 kg/m'
[90]	Diesel	'7492 MJ/m per well'	'7492 MJ/m'	'7492 MJ/m'
[91]	Steel	'111.3 kg/m'	'111.3 kg/m'	'111.3 kg/m'
[91]	Diesel	'4000 MJ/m'	'4000 MJ/m'	'4000 MJ/m'
[92]	Steel	'Injection well: 84.30 kg/m, Production well: 76.46 kg/m, Production leg 2: 46.83 kg/m'	'84.30 kg/m (injection), 76.46 kg/m (production)'	'84.30 kg/m, 76.46 kg/m, 46.83 kg/m'
[92]	Diesel	'7.23 GJ/m'	'7.23 GJ/m'	'7.23 GJ/m'
[39]	Steel	'100.200 kg/m'	'100.200 kg/m'	'100.200 kg/m'
[39]	Diesel	'53.100 kg/m'	'53.100 kg/m'	'53.100 kg/m'

diesel consumption), metadata defining the case study, including the functional unit, the type of the geothermal plant, and the well depth, demonstrating that generative LLMs can extract the information needed to define the LCI data and ensure their correct usage. The same framework can be applied for other technological domains by modifying the queries and prompts to target domain-specific parameters. For geothermal systems, these parameters could include geological characteristics such as reservoir lithology, porosity, thickness and permeability and plant characteristics such as geofluid output temperature and well-head pressure, all defined within a domain-dedicated ontology. However, the development of such ontologies or the creation of geothermal-related LCI databases goes beyond the scope of this study which used the investigated example as proof-of-concept to demonstrate that generative LLMs can be used to extract LCI numerical data the relevant contextual information needed.

3.2.3. Analysis of Literature Review and Case Study Results

Language models have large potential to enhance LCI compilation. Semantic embedding-based matching frameworks can help reduce the time and cost required to search LCA databases by using text descriptions to identify background datasets. High matching precision values for such frameworks have been reported (e.g., higher than 70 % [120]) indicating that datasets that an LCA practitioner would select can be identified by semantic matching frameworks. Such frameworks use embedding models that produce consistent results for the same input, ensuring reproducible matching results. Nevertheless, the precision of these frameworks can only be evaluated against the background datasets an LCA practitioner would choose from the database. Therefore, even if 100 % accuracy would be achieved this would contain the subjectivity of the ground-truth dataset used, even if that would be compiled via questioning multiple experts (e.g., in [59]). For example, if a background dataset for steel is sought, low-alloyed, unalloyed, or high-alloyed steel are options an LCA practitioner might choose, while another might suggest seeking more site-specific data than those reported in an LCA database. The decision would also be affected by the contribution of the steel consumption to the overall impacts of the system. Therefore, such frameworks cannot substitute the LCA practitioners. However the high matching precision values reported in the literature indicate that they can be used complementarily for LCA applications to allow for efficient database search that considers semantic information reported in the description of the datasets, especially in cases that the practitioner lacks extensive domain-knowledge. This can allow the practitioner to quickly search multiple databases, with the semantic algorithm suggesting the *K* most semantically similar datasets and explaining the suggestions via explainable AI methods to reduce bias in the selection process [133,134] and actively support background dataset identification.

Using generative LLMs to identify the most relevant background datasets (as in [123,124]) creates replicability and transparency issues, since the replies of a generative LLM cannot be traced and will vary due to model stochasticity, temperature settings, and prompt formulation. The latter limitations govern also practices in which generative LLMs

are used to directly predict numerical LCI data (e.g., [127,128]). More specifically, LCI data generated by generative LLMs cannot be a resort for formal assessments since the reply is not traceable or replicable. The generative LLM will reply based on what previously reported to generate a new possible value, without respecting physical or chemical laws or following process-modeling principles. Errors in data generated this way were reported to vary widely (3–44 % in [128], or 1–98 % for embodied carbon calculations in which generative LLMs were used to identify emissions factors in [124]) without explaining these variations, possibly because the replies of the generative LLM could not be traced. Therefore, even if a generative LLM would predict an LCI value very close to the actual one (or even equal), this should be considered with caution for informal usage.

Generative LLMs have great potential to facilitate LCI compilation, when used to extract LCI information from unstructured sources to reduce the time needed by the LCA practitioner to find LCA data. This case is especially relevant for novel energy technologies which are not represented in LCA databases, and for which primary data are difficult to find. Employing frameworks for extracting LCI data using generative LLMs (Fig. 4) can effectively reduce the time needed for LCI data gathering, enabling LCA of emerging technologies. The information extraction framework in Fig. 4 eliminates hallucinations and grounds the LLM replies on the provided document via a RAG manner. Furthermore, it can, along with numerical LCI values, extract additional metadata that define the case study such as functional units, lifetime set, and technical characteristics of the system. This allows later evaluation of the extracted data and their exclusion in case the associated metadata are not aligned with the requirements set by the LCA practitioner (e.g., if process designs are not reported or differ what is sought). The ability of generative LLMs to capture such contextual information poses them as a promising solution for LCA information extraction enabling provenance tracking and metadata annotation to enhance reliability of the extracted information. This aligns with recent developments across fields [55,131,135,136] and in the industrial setting [137,138]. Nevertheless, human experts should be involved in the data extraction process to first, design the information extraction framework (prompt, query) to ensure that the numerical LCI data and the necessary metadata are extracted, and second, to ensure that data to be used for further analysis are well-defined and reliable (e.g., by inspecting the selected documents). For example, generative LLMs can be used to scrap multiple documents and annotate them with the extracted information. The LCA practitioner can decide on what documents to inspect further based on the annotations and verify the extracted information greatly reducing the time needed for LCA.

The reviewed methods show great potential for reducing the cost of sustainability reporting. However, the development of these methods is still in its early stages. Further work is needed to improve the accuracy of the algorithms, ensure their reproducibility, and wider their adoption by providing user-friendly tools for non-experts. More research is required to establish best practices for applying generative LLMs in LCA. It is crucial that a human expert evaluates the LLM's recommendations

Table 7
Overview of methods to identify relevant background LCI datasets to a query system.

Method	Advantages	Disadvantages	Performance & Applicability
NLP (Natural Language Processing)	Reduces reliance on domain knowledge, Can handle unstructured text data	Accuracy varies, Requires good-quality text data for matching	Effective for datasets with complex or minimal structure, Still under development for widespread LCA use
Generative LLM	Enhances semantic matching, Reduces manual effort in data gathering and curation	Computationally intensive, Lack of transparency in decision-making, Unsuitable for numerical LCI estimations	Good for refining descriptions and improving matches, Still in early stages of integration into LCA workflows, Best used in RAG frameworks for extracting LCA related information from unstructured data

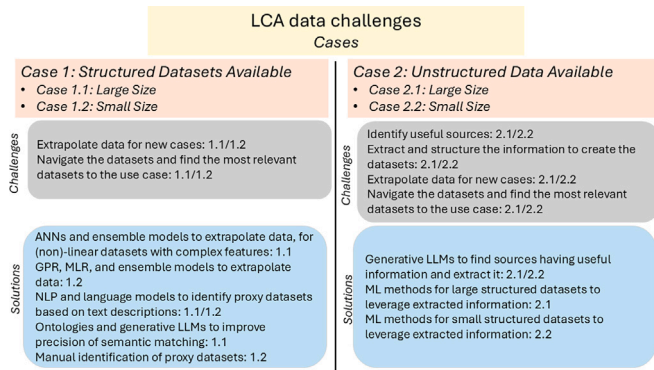


Fig. 7. Approaches to mitigate LCI data challenges.

and reasoning, before using them in LCA studies, as LLMs cannot be held responsible for the study's outcomes [56]. Table 7 describes an indicative summary of strengths and weaknesses of using language models for LCI compilation derived by the literature review and analysis performed in the previous paragraphs.

3.3. Moving forward

3.3.1. Typology of data challenges in LCI and methods to address them

Fig. 2 defined four types of LCI data challenges. This study indicates that ML and LLM methods can alleviate these challenges with different strategies being more appropriate for each challenge. Fig. 7 presents solutions derived from the analysis of this study. For large datasets, ML algorithms such as ANNs, XGBoost, and RF can excel in extrapolating missing data for new cases. When these data are only available in unstructured format, generative LLMs can be employed to extract this information in structured format making it usable by ML algorithms extracting also relevant metadata to ensure correct usage of the extracted data. When small datasets are available, different ML algorithms such as kriging and GPR, can be more efficient to extrapolate missing LCI data (Table 4), while for linear datasets, linear regression can be also a good option for estimating missing LCI data. In case the information is sought in unstructured sources, generative LLMs can be used to extract it similarly to what described for larger datasets. In addition, when the most similar LCI dataset for a product or system is sought, NLP methods (Fig. 1(b)) can excel in identifying the most relevant dataset without the heavy domain knowledge needs. By contrast, when only a few datasets are available manual identification can be more efficient.

Integration of ML and language model methods. The approaches described in Fig. 7 rely on using generative LLMs to extract LCI information from unstructured sources, NLP and language models to identify relevant background datasets for an investigated system, and ML algorithms to extrapolate missing LCI data from the available information. These three approaches can complement each other leading to more robust frameworks for AI-assisted LCI compilation. Generative LLMs are,

currently, not suitable for performing numerical calculations, a task in which ML methods excel. Similarly, using generative LLMs to identify relevant proxy datasets can lead to replicability issues. Nevertheless, generative LLMs can be used to gather the datasets needed for the ML algorithms by extracting LCI numerical data and associated methodological metadata and guiding their design. Major challenges regarding ML algorithms and semantic proxy selection methods can be, respectively, the selection of meaningful features and the design of text descriptions. Feature selection is significant for the design of an ML algorithm since the features need to allow the model to capture actual physical relationships between the inputs and outputs. In this regard, generative LLMs can complement ML by guiding feature selection for the ML algorithms as already reported for other scientific fields [139,140]. This approach is also similar to when generative LLMs were used to redesign text descriptions to improve semantic matching of background and foreground processes in [120]. Fig. 8 describes an integrated framework in which generative LLMs, ML algorithms, and semantic matching frameworks complement each other towards a robust LCI compilation pipeline. In the proposed integrated framework, generative LLMs complement ML algorithms by providing data sources (numerical LCI data and associated metadata) and guiding feature selection, while ML algorithms supplement the inefficiency of generative models to perform reliable, transparent, and reproducible calculations. More specifically, the LCI information derived by the LLM from the documents or when the LLM is used to identify proxy background datasets can have gaps or mistakes in numerical values, or their values might need updating to account for temporal or regional differences. These numerical estimations, cannot be made by the generative LLM itself. In contrast, ML models trained in pre-existing verified data for similar products or processes can be employed to provide numerical estimations, impute missing data, or update outdated data. This way ML algorithms can complement LLM-based LCI compilation to make it reproducible and transparent.

Incomplete inventories can significantly influence the LCA results of emerging and fast-evolving energy and decarbonization technologies (Paul et al. [65], Santucci et al. [64], Muller et al. [66]). As a result, comparative assessments may reflect data gaps rather than actual environmental differences. This can occur when one technology has complete data while another relies on approximations (i.e. when the technology is new or with underdeveloped inventories). For example, when comparing geothermal with biomass-fired heating systems, data on drilling or geological information may be lacking compared to well-documented biomass heating. Similarly, when investigating emerging fuels (e.g., hydrogen) and carbon-removal options against mature fossil-based systems, data for the latter can be more complete, compared to data available for the former systems.

Integrated ML and LLM approaches can address these issues and facilitate the adoption of emerging low-carbon energy technologies which are usually underrepresented in LCA databases and for which data may be sparse in the literature. Integrated LLM-ML approaches can search scientific publications and reports to extract missing LCI parameters and use ML models to impute values when needed, allowing more balanced and transparent comparisons. This can strengthen the environmental

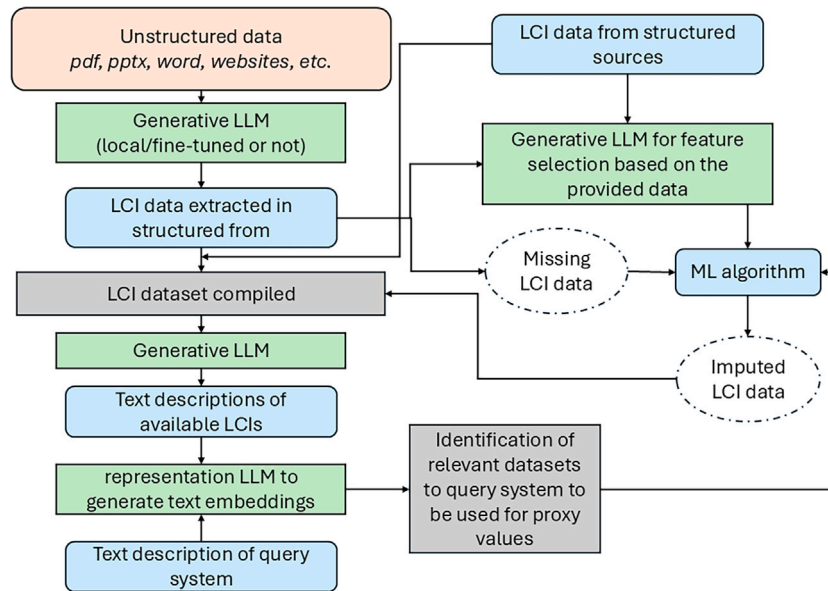


Fig. 8. Framework integrating ML algorithms and LLMs for streamlined LCI compilation.

assessment and support the adoption of emerging low-carbon technologies, while improving communication between technology design, sustainability analysis and policy-making.

3.3.2. Limitations and future research

ML- and AI- driven applications for LCI compilation are still under active development and are not to replace traditional LCI methods, such as using plant data or performing process simulations, but to complement them. For example, generative LLMs can gather data that would have to be gathered manually by the LCA practitioner when investigating technologies not represented in traditional LCA databases, such as emerging energy and decarbonization technologies (e.g., Direct Air Capture and Carbon Capture and Utilization units). This can significantly reduce the time needed for the LCA. The gathered information would still need human validation, though the time needed to validate will be lower than gathering all information from scratch. Also, ML models are shown to produce more reliable estimations for missing LCI data than using the average or arbitrarily selecting a proxy value. However, the dataset used should be sufficiently diverse and dense. Also, the features used should be selected using domain knowledge, to allow the ML model to capture actual physical correlations between the independent and dependent variables, generalize to unseen cases and provides reliable predictions. Furthermore, this study discussed that generative LLMs cannot, at the time of writing, be used to directly predict missing LCI data as such predictions would be untraceable and unpredictable leading to unreliable results.

ML- and AI-driven LCI compilation methods should be employed when sufficiently diverse and representative datasets, appropriate features, and knowledge to perform ML modeling is available. Data-driven methods are used to create models that represent realistic conditions. Therefore, when physics-based models or real data are available these should be prioritized. However, in cases that novel technologies are investigated for which process models are not available and cannot be developed (e.g., due to lack of knowledge or funds) or in cases where the missing LCI flows concern only specific parts of an investigated system and a detailed process model is too expensive to be developed, data-driven methods can provide a resort if relevant datasets are available. Nevertheless, the costs of gathering data and developing the models should be considered and compared to the cost of creating a process model or acquiring real data from a similar plant.

More research is needed to construct robust ML- and AI-driven LCI frameworks and enable their adoption by LCA practitioners. For example, LCA software providers could provide such solutions, also explaining the outcomes of the results and related uncertainty. Additionally, domain-specific databases reporting LCI flows and the variables driving them could be developed to enable finding datasets to be used for ML model training. Future research could also investigate agentic frameworks for generative LLM usage for LCI development. For example, generative LLMs can process natural language descriptions provided by LCA practitioners to frame the machine learning problem based on the available data and models, and then pass this information to dedicated algorithms for model training. Similar frameworks are already employed for optimization and contain safety layers to ensure proper problem statement [141]. Moreover, future research efforts could put emphasis on integrating ML and language AI methods to complement the weaknesses of each other as for example described in the framework of Fig. 8. Demonstrating the actual usage of such frameworks and providing open-source solutions for these could greatly help further understanding the applicability and hindrances of ML and AI methods for LCA and widen their adoption.

The proposed LCI data extraction framework uses generative LLMs to enable the structured extraction of both numeric LCI data and accompanying metadata describing the methodology, assumptions, and context of each source. This distinction is important, as accurate extraction does not necessarily mean that the original data are reliable. By capturing methodological metadata, the framework supports a first layer of quality control that allows users to trace how each value was derived and decide whether the data are suitable for further use. For example, if key metadata required to evaluate the quality of extracted LCI data, are missing, unclear or incompatible for the intended use, these data can be excluded. Future work should focus on the development of systematic approaches to assess the quality and reliability of the original sources from which these data are derived as this affects the trustworthiness of the extracted values. With the increasing use of generative LLMs across fields [55,131,135–138], future research should combine expert-in-the-loop validation of the extracted data [142], metadata extraction and verification, statistical analysis, or classification techniques to evaluate both the extraction process and credibility of the underlying data.

Knowledge graph structures have been suggested for structuring LCI datasets and databases [143–145], without, however, having been integrated with ML or AI frameworks. Such structures can facilitate LCI

automation since they allow for more meaningful representation of data than traditional relational databases [58]. Future research should investigate developing domain-specific ontologies for LCA data representation and use these for knowledge graph structuring. Such ontologies could integrate LCI numerical data, LCA methodological data (e.g., functional unit), and domain- or process-specific data that define the dataset. A generative LLM could then populate these metadata fields by parsing literature with frameworks similar to that proposed in this study. Therefore, domain and LCA experts should collaborate to develop metadata ontologies for different domains allowing the efficient numerical and contextual representation of LCI datasets. The geothermal energy example investigated in this study indicates that generative LLMs can extract both numerical and contextual LCI data, though remained in a proof-of-concept stage. Future work should expand its scope by integrating such domain-specific LCI data ontologies, including for example, for geothermal energy, geological parameters such as reservoir permeability and porosity suggested by domain experts, and expanded across sectors.

4. Conclusions

This study explores how ML and AI algorithms can help solve data challenges in LCI compilation, by investigating a variety of methods proposed in the literature and presenting specific case studies. This combined literature and case-study investigation for ML and language model methods highlights their strengths and weaknesses and enables identifying synergies towards unified AI-assisted LCI frameworks. The literature review and case study on ML methods showed that these can impute LCI data from previously reported information with lower errors than conventional alternatives such as linear regression. Different ML models perform better for smaller or larger datasets, while caution should be taken to ensure reliability and representativeness of the datasets and features used, and to consider the uncertainties and trade-offs introduced compared to traditional LCI data estimation methods. Nevertheless, when sufficient datasets are available and the ML algorithms are designed appropriately, they can provide reliable estimates for missing LCI data. Domain-specific guidelines for ML model development to estimate LCI data could expedite the adoption of such methods by LCA practitioners.

The literature review on language model methods for LCI compilation covered semantic matching and generative LLM methods and provides the first, to the authors' knowledge, review on this emerging topic. It showed that language models excel in searching LCA databases and automating the identification of background datasets based on textual descriptions. This can help non-domain experts identify proxy datasets and perform LCA with reduced time and cost requirements, while maintaining transparency of the analysis. Generative LLMs are found to effectively search documents for LCI data and extract the latter, greatly reducing the time needed for the data gathering phase of LCA. Also, it is shown that generative LLMs can extract both numerical LCI data and methodological and case-study specific metadata that help to correctly define and use the former. However, such applications are still in early stage of development and production-ready solutions are yet to be developed. Domain knowledge needs to be incorporated in the matching procedure by for example developing and utilizing domain-specific ontologies for representing the LCI datasets of different sectors. These ontologies can then be filled by using generative LLMs for LCI information extraction. The definition of such ontologies can also help in the standardization and harmonization procedures for LCI data gathered from diverse sources to ensure completeness and interoperability. Best practices for using generative LLMs for LCI data gathering need also to be defined to eliminate model hallucinations, and ensure the safety, transparency, and explainability of the data extracted, documents processed, and processing procedures employed.

The combined analysis of ML and LLM language model methods allowed demonstrating clear integration pathways. These two methodological families can complement each other, with LLMs enhancing

data discovery from unstructured sources and database search, and ML methods performing reliable and interpretable numerical estimations for missing data. This synergy lays the foundation for unified, AI-assisted LCI frameworks. The implementation of such standardized and robust AI-assisted LCI compilation workflows can facilitate the early stage assessment of novel renewable and low-carbon technologies, which are hindered by data scarcity. Improving the completeness and consistency of LCI datasets, for example through integrated ML-LLM frameworks, can strengthen evidence-based assessment, reduce uncertainty and support more reliable decisions about technology design and policy. By enabling continuous updating of LCI data for such technologies with data coming from experiments, pilot-cases, actual field data, or data reported in the literature, ML and LLM tools can provide the environmental assessment with updated data that were missing and help guiding the design of sustainable energy technologies.

More research is needed to ensure best usage of ML- and AI method for LCI compilation and develop robust ML-/AI-driven LCA frameworks. ML and LLM solutions need to be integrated to supplement each other's inefficiencies and lead to more comprehensive and transparent solutions that can be deployed for production. Also open-source solutions are needed to demonstrate the usage of ML and AI in LCA and make the frameworks transparent, trustworthy, and replicable. Other technologies such as agentic frameworks and knowledge graphs should also be explored to further enhance AI-assisted LCA solutions. Notably, this study focused on the application of ML and AI methods for facilitating the inventory compilation in LCA studies. However, similar data challenges are encountered in other sustainability assessment methods such as Life Cycle Costing, where data on energy, materials, water consumption, and waste generation should be gathered and matched to their respective costs. The findings of this study are transferable. ML methods can estimate missing quantitative data using available datasets, while language model techniques can identify and extract the required information from various literature sources. This can concern cost data and material quantities, which are essential for costing analysis. More research is needed to extend the ML and LLM applications investigated in this study for LCA, to other sustainability assessment methods, such as Life Cycle Costing and Life Cycle Sustainability Analysis, to broaden their application for novel low-carbon energy systems, but also across products and systems, and allow their implementation for facilitating the sustainable transition.

CRedit authorship contribution statement

Spiros Gkousis: Conceptualization, Formal Analysis, Software, Writing – Original Draft, Writing – Review & Editing, Visualization.
Vasileia Vasilaki: Conceptualization, Writing – Review & Editing.
Evina Katsou: Conceptualization, Writing – Review & Editing, Funding Acquisition, Supervision.

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

The authors have received funding from the STREAMS project which is funded by the European Union under the grant agreement number 101137771.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:[10.1016/j.rser.2025.116577](https://doi.org/10.1016/j.rser.2025.116577).

Data availability

Data will be made available on request.

References

- [1] Backes J, Traverso M. Life cycle sustainability assessment as a metrics towards sdgs agenda 2030. *Curr Opin Green Sustain Chem* 2022;38:100683. <https://doi.org/10.1016/j.cogsc.2022.100683>
- [2] Hellweg S, Benetto E, Huijbregts MAJ, Verones F, Wood R. Life-cycle assessment to guide solutions for the triple planetary crisis. *Nat Rev Earth Environ* 2023;4:471–86. <https://doi.org/10.1038/s43017-023-00449-2>
- [3] EPD. Environmental product declarations; 2025. <https://www.environdec.com/all-about-epds/the-epd>
- [4] EC. Environmental footprint methods. Calculating the environmental impact of products and services; 2024. https://green-forum.ec.europa.eu/green-business/environmental-footprint-methods_en
- [5] WEF. How harnessing the power of smes and mid-sized companies can shape the future of sustainability; 2022. <https://www.weforum.org/stories/2022/04/smes-future-of-sustainability/>
- [6] EU. Eur-lex, document 32023l241; 2023. <https://eur-lex.europa.eu/eli/dir/2023/2413/oj/eng>
- [7] EU. Eur-lex, document 32024l1275; 2024. <https://eur-lex.europa.eu/eli/dir/2024/1275/oj/eng>
- [8] EU. Eur-lex, document 32023r1542; 2023. <https://eur-lex.europa.eu/eli/reg/2023/1542/oj/eng>
- [9] Gomez-Garza R, Güereca LP, Padilla-Rivera A, Ibarra AA. Barriers and enablers of life cycle assessment in small and medium enterprises: a systematic review. *Environ Dev Sustain* 2024. <https://doi.org/10.1007/s10668-024-05622-1>
- [10] Palsson A, Riise E. Performing the inventory in the lca study; 2011. <https://tosca-life.info/getting-started-guides/life-cycle-assessment/how-to-perform-an-lca/phases-in-an-lca-study/inventory/>
- [11] Saavedra-Rubio K, Thonemann N, Crenna E, Lemoine B, Caliendo P, Laurent A. Stepwise guidance for data collection in the life cycle inventory (lci) phase: Building technology-related lci blocks. *J Clean Prod* 2022;366:132903. <https://doi.org/10.1016/j.jclepro.2022.132903>
- [12] Steubing B, Wernet G, Reinhard J, al. E. The ecoinvent database version 3 (part ii): analyzing lca results and comparison to version 2. *Int J Life Cycle Assess* 2016;21. <https://doi.org/10.1007/s11367-016-1109-6>
- [13] Sphera. A quick guide to sphera's 2024.2 lca databases update; 2024. <https://sphera.com/resources/blog/a-quick-guide-to-spheras-2024-2-lca-databases-update/>
- [14] NREL. U.S. life cycle inventory database; 2012. <https://www.lcacommons.gov/nrel/search>
- [15] AusLCI. The Australian national life cycle inventory database initiative; 2011. <https://www.auslci.com.au/>
- [16] Carbon Minds. Measure and reduce your supply chain emissions with our life cycle database; 2024. <https://www.carbon-minds.com/products/data/carbon-footprint-and-lca-data/>
- [17] AGRIBALYSE. Homepage & news welcome to the agribalyse® documentation. 2024. <https://doc.agribalyse.fr/documentation-en>
- [18] Blonk. Agri-footprint, the world's leading source of environmental footprint data for the agri-food sector. 2024. <https://blonksustainability.nl/agri-footprint>
- [19] Martínez-Rocamora A, Solís-Guzmán J, Marrero M. LCA databases focused on construction materials: a review. *Renew Sustain Energy Rev* 2016;58:565–73. <https://doi.org/10.1016/j.rser.2015.12.243>
- [20] WSA. Life cycle inventory data and eco-profiles. 2025. <https://worldsteel.org/wider-sustainability/life-cycle-thinking/life-cycle-inventory-data-and-eco-profiles/>
- [21] Cobalt Institute. Cobalt institute LCA data. 2025. <https://lca-data.cobaltinstitute.org/>
- [22] EAA. Our environmental profile reports. 2024. <https://european-aluminium.eu/blog/environmental-profile-reports/>
- [23] Sacchi R, Terlouw T, Siala K, Dirnhaichner A, Bauer C, Cox B, Mutel C, Daioglou V, Luderer G. Prospective environmental impact assessment (premise): a streamlined approach to producing databases for prospective life cycle assessment using integrated assessment models. *Renew Sustain Energy Rev* 2022;160:112311. <https://doi.org/10.1016/j.rser.2022.112311>
- [24] Brandt AR. Greenhouse gas intensity of natural hydrogen produced from subsurface geologic accumulations. *Joule* 2023;7(8):1818–31. <https://doi.org/10.1016/j.joule.2023.07.001>
- [25] Delval MH, Thonemann N, Henriksson PJG, Tanzer SE, Behrens P. Life cycle assessment of ocean-based carbon dioxide removal approaches: a systematic literature review. *Renewable Sustain Energy Rev* 2025;224:116091. <https://doi.org/10.1016/j.rser.2025.116091>
- [26] Gkousis S, Welkenhuysen K, Compennolle T. Deep geothermal energy extraction, a review on environmental hotspots with focus on geo-technical site conditions. *Renewable Sustain Energy Rev* 2022;162:112430. <https://doi.org/10.1016/j.rser.2022.112430>
- [27] Peters JF, Baumann M, Zimmermann B, Braun J, Weil M. The environmental impact of li-ion batteries and the role of key parameters – a review. *Renewable Sustain Energy Rev* 2017;67:491–506. <https://doi.org/10.1016/j.rser.2016.08.039>
- [28] Parvatkar AG, Eckelman MJ. Comparative evaluation of chemical life cycle inventory generation methods and implications for life cycle assessment results. *ACS Sustain Chem Eng* 2019;7(1):350–67. <https://doi.org/10.1021/acssuschemeng.8b03656>
- [29] van der Giesen C, Cucurachi S, Guinée J, Kramer GJ, Tukker A. A critical view on the current application of lca for new technologies and recommendations for improved practice. *J Clean Prod* 2020;259:120904. <https://doi.org/10.1016/j.jclepro.2020.120904>
- [30] Köck B, Friedl A, Serna Loaiza S, Wukovits W, Mihalyi-Schneider B. Automation of life cycle assessment—a critical review of developments in the field of life cycle inventory analysis. *Sustainability* 2023;15(6). <https://doi.org/10.3390/su15065531>
- [31] Hirschier R, Hellweg S, Capello C, al. E. Establishing life cycle inventories of chemicals based on differing data availability (9 pp). *Int J Life Cycle Assess* 2005;10. <https://doi.org/10.1065/lca2004.10.181.7>
- [32] Piccinno F, Hirschier R, Seeger S, Som C. From laboratory to industrial scale: a scale-up framework for chemical processes in life cycle assessment studies. *J Clean Prod* 2016;135:1085–97. <https://doi.org/10.1016/j.jclepro.2016.06.164>
- [33] Simon B, Bachtin K, Kiliç A, Amor B, Weil M. Proposal of a framework for scale-up life cycle inventory: a case of nanofibers for lithium iron phosphate cathode applications. *Integr Environ Assess Manag* 2016;12(3):465–77. <https://doi.org/10.1002/ieam.1788>
- [34] Caduff M, Huijbregts MAJ, Koehler A, Althaus H-J, Hellweg S. Scaling relationships in life cycle assessment. *J Ind Ecol* 2014;18(3):393–406. <https://doi.org/10.1111/jiec.12122>
- [35] Canals LMI, Azapagic A, Doka G, Jefferies D, King H, Mutel C, Nemecek T, Roches A, Sim S, Stichnothe H, Thoma G, Williams A. Approaches for addressing life cycle assessment data gaps for bio-based products. *J Ind Ecol* 2011;15(5):707–25. <https://doi.org/10.1111/j.1530-9290.2011.00369.x>
- [36] Pradeleix L, Roux P, Bouarfa S, Bellon-Maurel V. Multilevel environmental assessment of regional farming activities with life cycle assessment: tackling data scarcity and farm diversity with life cycle inventories based on agrarian system diagnosis. *Agric Syst* 2022;196:103328. <https://doi.org/10.1016/j.agsy.2021.103328>
- [37] Erakka M, Baumann M, Helbig C, Weil M. Systematic review of scale-up methods for prospective life cycle assessment of emerging technologies. *J Clean Prod* 2024;451:142161. <https://doi.org/10.1016/j.jclepro.2024.142161>
- [38] Ferdius J, Bensebaa F, Hewage K, Bhowmik P, Pelletier N. Use of process simulation to obtain life cycle inventory data for LCA: a systematic review. *Clean Environ Syst* 2024;14:100215. <https://doi.org/10.1016/j.cesys.2024.100215>
- [39] Wang Y, Du Y, Wang J, Zhao J, Deng S, Yin H. Comparative life cycle assessment of geothermal power generation systems in china. *Resour Conserv Recycl* 2020;155:104670. <https://doi.org/10.1016/j.resconrec.2019.104670>
- [40] Zhang R, Wang G, Shen X, Wang J, Tan X, Feng S, Hong J. Is geothermal heating environmentally superior than coal fired heating in china? *Renewable Sustain Energy Rev* 2020;131:110014. <https://doi.org/10.1016/j.rser.2020.110014>
- [41] Gkousis S, Braimakis K, Nimmegheers P, Karellas S, Compennolle T. Multi-objective optimization of medium-enthalpy geothermal organic rankine cycle plants. *Renew Sustain Energy Rev* 2025;210:115150. <https://doi.org/10.1016/j.rser.2024.115150>
- [42] Liu W, Liu H, Liu W, Cui Z. Life cycle assessment of power batteries used in electric bicycles in china. *Renewable Sustain Energy Rev* 2021;139:110596. <https://doi.org/10.1016/j.rser.2020.110596>
- [43] Kallitsis E, Korre A, Kelsall G, Kupfersberger M, Nie Z. Environmental life cycle assessment of the production in china of lithium-ion batteries with nickel-cobalt-manganese cathodes utilising novel electrode chemistries. *J Clean Prod* 2020;254:120067. <https://doi.org/10.1016/j.jclepro.2020.120067>
- [44] Jinasaena A, Burheim OS, Strømman AH. A flexible model for benchmarking the energy usage of automotive lithium-ion battery cell manufacturing. *Batteries* 2021;7(1). <https://doi.org/10.3390/batteries7010014>
- [45] Lechner M, Kollenda A, Bendzuck K, al. E. Cost modeling for the gwh-scale production of modern lithium-ion battery cells. *Commun Eng* 2024;3. <https://doi.org/10.1038/s44172-024-00306-0>
- [46] Liao M, Kelley S, Yao Y. Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. *ACS Sustain Chem Eng* 2020;8(2):1252–61. <https://doi.org/10.1021/acssuschemeng.9b06522>
- [47] Cheng F, Porter MD, Colosi LM. Is hydrothermal treatment coupled with carbon capture and storage an energy-producing negative emissions technology? *Energy Convers Manag* 2020;203:112252. <https://doi.org/10.1016/j.enconman.2019.112252>
- [48] Zargar S, Yao Y, Tu Q. A review of inventory modeling methods for missing data in life cycle assessment. *J Ind Ecol* 2022;26(5):1676–89. <https://doi.org/10.1111/jiec.13305>

- [49] Martínez-Ramón N, Calvo-Rodríguez F, Iribarren D, Dufour J. Frameworks for the application of machine learning in life cycle assessment for process modeling. *Clean Environ Syst* 2024;14:100221. <https://doi.org/10.1016/j.cesys.2024.100221>
- [50] Wright M, Tan ECD, Tu Q, Martins A, Parvatkar AG, Yao Y, Sunol A, Smith RL. Life cycle inventory availability: status and prospects for leveraging new technologies. *ACS Sustain Chem Eng* 2024;12(34):12708–18. <https://doi.org/10.1021/acssuschemeng.4c02519>
- [51] Ghoroghi A, Rezgui Y, Petri I, al. E. Advances in application of machine learning to life cycle assessment: a literature review. *Int J Life Cycle Assess* 2022;27. <https://doi.org/10.1007/s11367-022-02030-3>. <https://link.springer.com/article/10.1007/s11367-022-02030-3#citeas>
- [52] Neupane B, Belkadi F, Formentini M, Rozière E, Hilloulou B, Abdolmaleki SF, Mensah M. Machine learning algorithms for supporting life cycle assessment studies: an analytical review. *Sustain Prod Consum* 2025;56:37–53. <https://doi.org/10.1016/j.spc.2025.03.015>. <https://www.sciencedirect.com/science/article/pii/S235255092500065X>
- [53] Romeiko XX, Zhang X, Pang Y, Gao F, Xu M, Lin S, Babbitt C. A review of machine learning applications in life cycle assessment studies. *Sci Total Environ* 2024;912:168969. <https://doi.org/10.1016/j.scitotenv.2023.168969>. <https://www.sciencedirect.com/science/article/pii/S0048969723075988>
- [54] Zheng Z, Zhang O, Borgs C, Chayes JT, Yaghi OM. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *J Am Chem Soc* 2023;145(32):18048–62. PMID: 37548379. <https://doi.org/10.1021/jacs.3c05819>
- [55] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, Persson KA. Structured information extraction from scientific text with large language models. *Nature Communications* 2024;15:1418. <https://doi.org/10.1038/s41467-024-45563-x>
- [56] Preuss N, Alshehri AS, You F. Large language models for life cycle assessments: opportunities, challenges, and risks. *J Clean Prod* 2024;466:142824. <https://doi.org/10.1016/j.jclepro.2024.142824>
- [57] Cornago S, Ramakrishna S, Low JSC. How can transformers and large language models like chatgpt help LCA practitioners? *Resour Conserv Recycl* 2023;196:107062. <https://doi.org/10.1016/j.resconrec.2023.107062>
- [58] Tu Q, Guo J, Li N, Qi J, Xu M. Mitigating grand challenges in life cycle inventory modeling through the applications of large language models. *Environ Sci Technol* 2024;58(44):19595–603. <https://doi.org/10.1021/acs.est.4c07634>
- [59] Balaji B, Vunnavu VSG, Domingo N, Gupta S, Gupta H, Guest G, Srinivasan A. Flamingo: environmental impact factor matching for life cycle assessment with zero-shot machine learning. *ACM J Comput Sustain Soc* 2023;1(2):1–23. <https://doi.org/10.1145/3616385>
- [60] Bekavac L, Volk L. Towards an integrated semantic approach to GHG accounting. *Jan* 2024. <https://doi.org/10.13140/RG.2.2.28050.22728>
- [61] USEPA. Greenhouse gas reporting program (GHGRP). 2025. <https://www.epa.gov/ghgreporting>
- [62] USEPA. Greenhouse gas reporting program (ghgrp), data sets. 2025. <https://www.epa.gov/ghgreporting/data-sets>
- [63] USEIA. Emissions by plant and by region. 2024. <https://www.eia.gov/electricity/data/emissions/>
- [64] Santucci V, Eynard U, Candelaresi D, Mathieux F. Enhancing life cycle thinking in emerging sectors: the example of hydrogen technologies and the opportunities of the life cycle data network. *Int J Life Cycle Assess* 2025;30:1607–22. <https://doi.org/10.1007/s11367-025-02482-3>
- [65] Paul D, Pechancová V, Saha N, Pavelková D, Saha N, Motiei M, Jamatia T, Chaudhuri M, Ivanichenko A, Venher M, Hrbáčková L, Saha P. Life cycle assessment of lithium-based batteries: Review of sustainability dimensions. *Renew Sustain Energy Rev* 2024;206:114860. <https://doi.org/10.1016/j.rser.2024.114860>. <https://www.sciencedirect.com/science/article/pii/S1364032124005860>
- [66] Müller A, Friedrich L, Reichel C, Herceg S, Mittag M, Neuhaus DH. A comparative life cycle assessment of silicon pv modules: Impact of module design, manufacturing location and inventory. *Sol Energy Mater Sol Cells* 2021;230:111277. <https://doi.org/10.1016/j.solmat.2021.111277>. <https://www.sciencedirect.com/science/article/pii/S0927024821003202>
- [67] Pratiwi A, Ravier G, Genter A. Life-cycle climate-change impact assessment of enhanced geothermal system plants in the upper rhine valley. *Geothermics* 2018;75:26–39. <https://doi.org/10.1016/j.geothermics.2018.03.012>. <https://www.sciencedirect.com/science/article/pii/S0375650517302912>
- [68] Muench S, Guenther E. A systematic review of bioenergy life cycle assessments. *Appl Energy* 2013;112:257–73. <https://doi.org/10.1016/j.apenergy.2013.06.001>. <https://www.sciencedirect.com/science/article/pii/S0306261913005084>
- [69] Dolan SL, Heath GA. Life cycle greenhouse gas emissions of utility-scale wind power. *J Ind Ecol* 2012;16(s1):S136–54. <https://doi.org/10.1111/j.1530-9290.2012.00464.x>
- [70] Batuhan S. Vehicle CO2 emissions dataset. 2024. <https://www.kaggle.com/datasets/brsahan/vehicle-co2-emissions-dataset/data>
- [71] EIA. Emissions by plant and by region. 2023. <https://www.eia.gov/electricity/data/emissions/>
- [72] Ma Y, Xu P, Li M, Ji X, Zhao W, Lu W. The mastery of details in the workflow of materials machine learning. *NPJ Comput Mater* 2024;10(10). <https://doi.org/10.1038/s41524-024-01331-5>. <https://www.nature.com/articles/s41524-024-01331-5>
- [73] Au TC. Random forests, decision trees, and categorical predictors: The absent levels problem. *arXiv:1706.03492*. 2018.
- [74] Goldin MA, Virgili S, Chalk M. Scalable gaussian process inference of neural responses to natural images. *Proc Natl Acad Sci* 2023;120(34):e2301150120. <https://doi.org/10.1073/pnas.2301150120>
- [75] Kim J, Lee Y, Lee M-H, Hong S-Y. A comparative study of machine learning and spatial interpolation methods for predicting house prices. *Sustainability* 2022;14(15). <https://doi.org/10.3390/su14159056>. <https://www.mdpi.com/2071-1050/14/15/9056>
- [76] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12(Oct):2825–30.
- [77] xgboost developers. DMLC XGBoost. 2022. <https://xgboost.readthedocs.io/en/stable/changes/index.html>
- [78] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. TensorFlow: large-scale machine learning on heterogeneous systems, software available from tensorflow.org 2015. <https://www.tensorflow.org/>
- [79] Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv:1412.6980*. 2017.
- [80] Makersite. Using AI for cradle-to-grave product lifecycle analysis (LCA); 2025. <https://makersite.io/insights/using-ai-for-cradle-to-grave-product-lifecycle-analysis-lca>
- [81] OneClickLCA. AI-powered sustainability software for data-driven insights. 2025. <https://oneclicklca.com/why-us/capabilities/powered-by-ai>
- [82] Unstructured. Unstructured: Your genai has a data problem. 2025. <https://unstructured.io/>
- [83] Langchain. Pypdfloader. 2025. https://python.langchain.com/docs/integrations/document_loaders/pypdfloader/
- [84] Richardson L. Beautiful soup documentation; 2025. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- [85] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. *arXiv:2005.14165*. 2020.
- [86] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*. 2023.
- [87] Open AI. Vector embeddings. 2025. <https://platform.openai.com/docs/guides/embeddings>
- [88] M.AI. Frontier AI. In your hands. 2023. <https://mistral.ai/>
- [89] HuggingFace. Hugging face: the AI community building the future. 2025. <https://huggingface.co/>
- [90] Frick S, Kaltschmitt M, Schröder G. Life cycle assessment of geothermal binary power plants using enhanced low-temperature reservoirs. *Energy* 2010;35(5):2281–94. <https://doi.org/10.1016/j.energy.2010.02.016>. <https://www.sciencedirect.com/science/article/pii/S0360544210000708>
- [91] Lacirignola M, Blanc I. Environmental analysis of practical design options for enhanced geothermal systems (egs) through life-cycle assessment. *Renew Energy* 2013;50:901–14. <https://doi.org/10.1016/j.renene.2012.08.005>. <https://www.sciencedirect.com/science/article/pii/S096014811200465X>
- [92] Paulillo A, Cotton L, Law R, Striolo A, Lettieri P. Geothermal energy in the uk: The life-cycle environmental impacts of electricity production from the united downs deep geothermal power project. *J Clean Prod* 2020;249:119410. <https://doi.org/10.1016/j.jclepro.2019.119410>. <https://www.sciencedirect.com/science/article/pii/S0959652619342805>
- [93] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush AM. Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*. 2020.
- [94] META. Faiss. 2025. <https://ai.meta.com/tools/faiss/>
- [95] Ascher S, Wang X, Watson I, Sloan W, You S. Interpretable machine learning to model biomass and waste gasification. *Bioresour Technol* 2022;364:128062. <https://doi.org/10.1016/j.biortech.2022.128062>. <https://www.sciencedirect.com/science/article/pii/S0960852422013955>
- [96] Corral-Bobadilla M, Lostado-Lorza R, Sabando-Fraile C, Íñiguez Macedo S. An artificial intelligence approach to model and optimize biodiesel production from waste cooking oil using life cycle assessment and market dynamics analysis. *Energy* 2024;307:132712. <https://doi.org/10.1016/j.energy.2024.132712>. <https://www.sciencedirect.com/science/article/pii/S0360544224024861>
- [97] Cheng F, Luo H, Colosi LM. Slow pyrolysis as a platform for negative emissions technology: an integration of machine learning models, life cycle assessment, and economic analysis. *Energy Convers Manag* 2020;223:113258. <https://doi.org/10.1016/j.enconman.2020.113258>. <https://www.sciencedirect.com/science/article/pii/S0196890420308013>
- [98] Kaab A, Sharifi M, Mobli H, Nabavi-Pelesaee A, wing Chau K. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. *Sci Total Environ* 2019;664:1005–19. <https://doi.org/10.1016/j.scitotenv.2019.02.004>. <https://www.sciencedirect.com/science/article/pii/S0048969719304838>
- [99] Long F, Liu H. An integration of machine learning models and life cycle assessment for lignocellulosic bioethanol platforms. *Energy Convers Manag* 2023;292:117379. <https://doi.org/10.1016/j.enconman.2023.117379>. <https://www.sciencedirect.com/science/article/pii/S0196890423007252>
- [100] Ghafarian Nia SA, Shahbeik H, Shafizadeh A, Rafiee S, Hosseinzadeh-Bandbafha H, Kiehbaddrounizhad M, Sheikh Ahmad Tajuddin SAF, Tabatabaei M,

- Aghbashlo M. Machine learning-driven optimization for sustainable CO₂-to-methanol conversion through catalytic hydrogenation. *Energy Convers Manag* 2025;325:119373. <https://doi.org/10.1016/j.enconman.2024.119373>. <https://www.sciencedirect.com/science/article/pii/S0196890424013141>.
- [101] Omidkar A, Alagumalai A, Li Z, Song H. Machine learning assisted techno-economic and life cycle assessment of organic solid waste upgrading under natural gas. *Appl Energy* 2024;355:122321. <https://doi.org/10.1016/j.apenergy.2023.122321>. <https://www.sciencedirect.com/science/article/pii/S0306261923016859>.
- [102] Hajabdollahi Ouderji Z, Gupta R, McKeown A, Yu Z, Smith C, Sloan W, You S. Integration of anaerobic digestion with heat pump: Machine learning-based technical and environmental assessment. *Bioresour Technol* 2023;369:128485. <https://doi.org/10.1016/j.biortech.2022.128485>. <https://www.sciencedirect.com/science/article/pii/S0960852422018181>.
- [103] Dai T, Jordaan SM, Wemhoff AP. Gaussian process regression as a replicable, streamlined approach to inventory and uncertainty analysis in life cycle assessment. *Environ Sci Technol* 2022;56(6):3821–29. <https://doi.org/10.1021/acs.est.1c04252>.
- [104] Hou P, Cai J, Qu S, Xu M. Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environ Sci Technol* 2018;52(9):5259–67. <https://doi.org/10.1021/acs.est.7b05366>.
- [105] Meng F, LaFleur C, Wijesinghe A, Colvin J. Data-driven approach to fill in data gaps for life cycle inventory of dual fuel technology. *Fuel* 2019;246:187–95. <https://doi.org/10.1016/j.fuel.2019.02.124>.
- [106] Moreau V, Bage G, Marcotte D, Samson R. Estimating material and energy flows in life cycle inventory with statistical models. *J Ind Ecol* 2012;16(3):399–406. <https://doi.org/10.1111/j.1530-9290.2012.00459.x>.
- [107] Moreau V, Bage G, Marcotte D, Samson R. Statistical estimation of missing data in life cycle inventory: an application to hydroelectric power plants. *J Clean Prod* 2012;37:335–41. <https://doi.org/10.1016/j.jclepro.2012.07.036>.
- [108] Saad M, Zhang Y, Jia J, Tian J. Decision tree-based approach to extrapolate life cycle inventory data of manufacturing processes. *J Environ Manag* 2024;360:121152. <https://doi.org/10.1016/j.jenvman.2024.121152>.
- [109] Heilmann ZJN, Venkatesh A, Hauck M, Schipper AM, Karuppihar R, Laurenzi LJ, Huijbregts MAJ. How to address data gaps in life cycle inventories: a case study on estimating CO₂ emissions from coal-fired electricity plants on a global scale. *Environ Sci Technol* 2014;48(9):5282–89. PMID: 24749645. <https://doi.org/10.1021/es500757p>.
- [110] Wernet G, Hellweg S, Fischer U, Papadokonstantakis S, Hungerbühler K. Molecular-structure-based models of chemical inventories using neural networks. *Environ Sci Technol* 2008;42(17):6717–22. PMID: 18800554. <https://doi.org/10.1021/es7022362>.
- [111] Xiao H, Gao B, Yu S, al. E. Life cycle assessment of metal powder production: a bayesian stochastic kriging model-based autonomous estimation. *Auton Intell Syst* 2024;4. <https://doi.org/10.1007/s43684-024-00079-5>.
- [112] Zhao B, Shuai C, Hou P, Qu S, Xu M. Estimation of unit process data for life cycle assessment using a decision tree-based approach. *Environ Sci Technol* 2021;55(12):8439–46. <https://doi.org/10.1021/acs.est.0c07484>.
- [113] Meron N, Blass V, Thoma G. Selection of the most appropriate life cycle inventory dataset: new selection proxy methodology and case study application. *Int J Life Cycle Assess* 2020;25:771–83. <https://doi.org/10.1007/s11367-019-01721-8>.
- [114] Olafasakin O, Chang Y, Passalacqua A, Subramaniam S, Brown R, Wright M. Machine learning reduced order model for cost and emission assessment of a pyrolysis system. *Energy Fuels* 2021;35(12):9950–60. <https://doi.org/10.1021/acs.energyfuels.1c00490>.
- [115] Nguyen TH, Nong D, Paustian K. Surrogate-based multi-objective optimization of management options for agricultural landscapes using artificial neural networks. *Ecol Model* 2019;400:1–13. <https://doi.org/10.1016/j.ecolmodel.2019.02.018>. <https://www.sciencedirect.com/science/article/pii/S0304380019300870>.
- [116] Huntington T, Baral NR, Yang M, Sundstrom E, Scown CD. Machine learning for surrogate process models of bioproduction pathways. *Bioresour Technol* 2023;370:128528. <https://doi.org/10.1016/j.biortech.2022.128528>. <https://www.sciencedirect.com/science/article/pii/S0960852422018612>.
- [117] Rankin K, Hawes J, van Lierde S, Donati F, Tu Q. Ai for ie - standardization and best practices. 2025. <https://isie2025.exordo.com/programme/presentation/7>.
- [118] Marinescu M. Explaining and connecting kriging with gaussian process regression. *arXiv:2408.02331* [Preprint]. 2024.
- [119] Balaji B, Venkata SGV, Geoffrey G, Kramer J. Caml: Carbon footprinting of household products with zero-shot semantic text similarity. In: *Proceedings of the ACM Web Conference 2023, WWW '23*. New York, NY, USA: Association for Computing Machinery; 2023. p. 4004–14. <https://doi.org/10.1145/3543507.3583882>.
- [120] Balaji B, Ebrahimi F, Domingo NGG, Vunava VSG, Faridee A-Z, Ramalingam S, Gupta S, Wang A, Gupta H, Belcastro D, Axten K, Hakian J, Kramer J, Srinivasan A, Tu Q. Emission factor recommendation for life cycle assessments with generative AI. *Environ Sci Technol* 2025. PMID: 40117648. <https://doi.org/10.1021/acs.est.4c12667>.
- [121] Forth K, Abualdenien J, Borrmann A. Calculation of embodied GHG emissions in early building design stages using bim and NLP-based semantic model healing. *Energy Build* 2023;284:112837. <https://doi.org/10.1016/j.enbuild.2023.112837>.
- [122] Gachkar S, Gachkar D, Ghofrani E, García Martínez A, Angulo Bahón C. Text-based algorithms for automating life cycle inventory analysis in building sector life cycle assessment studies. *J Clean Prod* 2025;486:144448. <https://doi.org/10.1016/j.jclepro.2024.144448>. <https://www.sciencedirect.com/science/article/pii/S0959652624038976>.
- [123] Gu X, Chen C, Fang Y, Mahabir R, Fan L. CECA: an intelligent large-language-model-enabled method for accounting embodied carbon in buildings. *Build Environ* 2025;272:112694. <https://doi.org/10.1016/j.buildenv.2025.112694>. <https://www.sciencedirect.com/science/article/pii/S0360132325001763>.
- [124] Präger L, Woytowicz J, Reitberger R, Lang W. LCA-based calculation of GHG protocol scope 3: a bottom-up approach to determine GHG emissions of the construction activity of municipalities. *Build Environ* 2025;285:113502. <https://doi.org/10.1016/j.buildenv.2025.113502>. <https://www.sciencedirect.com/science/article/pii/S0360132325009758>.
- [125] Chen L, Silvennoinen H, De Wolf C, Hall D, Van Mele T, Block P. Towards automated building life cycle assessments: a novel approach using large language models and the compas framework. *arXiv:2410.08288*. 2025.
- [126] Kallitisis E, Offer G, Edge J. ARIA: artificial intelligence for sustainability assessment. 2025. <https://doi.org/10.31223/X5JF17>. <https://eartharxiv.org/repository/view/9250/>.
- [127] Luo B, Liu J, Deng Z, Yuan C, Yang Q, Xiao L, Xie Y, Zhou F, Zhou W, Liu Z. AutoPCF: a novel automatic product carbon footprint estimation framework based on large language models. *Proceedings of the AAAI Symposium Series* 2024;2(1):102–06. <https://doi.org/10.1609/aaais.v2i1.27656>. <https://ojs.aaai.org/index.php/AAAI-SS/article/view/27656>.
- [128] Kumar A, Sen A, Bakshi B. Sustain-GPT: a large language model for creating a structural database from unstructured text resources to develop life cycle inventories. 2024. <https://aiche.confex.com/aiche/2024/meetingapp.cgi/Paper/691830>.
- [129] Turhan D. Life cycle assessment for the unconventional construction materials in collaboration with a large language model. In: *Digital Design Reconsidered - Proceedings of the 41st Conference on Education and Research in Computer Aided Architectural Design in Europe*; 2023. p. 39–48. <https://doi.org/10.52842/conf.eacaed.2023.2.039>.
- [130] Li TT, Liang R, Shang Y, Ding CX, Hua Y, Wang Z, Alhazmi M. SustainLLM: Ai-driven lifecycle sustainability assessment and energy transition optimization. *Sustain Energy Technol Assess* 2025;82:104475. <https://doi.org/10.1016/j.seta.2025.104475>. <https://www.sciencedirect.com/science/article/pii/S221313882503066>.
- [131] Tiangong. Welcome to tiangong initiative. 2025. <https://www.tiangong.earth/>.
- [132] Tiangong. Tiangong-AI-chatgpt-plugin. 2025. <https://github.com/linancn/TianGong-AI-ChatGPT-Plugin>.
- [133] Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decis Anal J* 2023;7:100230. <https://doi.org/10.1016/j.dajour.2023.100230>. <https://www.sciencedirect.com/science/article/pii/S27726622300070X>.
- [134] IBM. What is explainable AI? 2025. <https://www.ibm.com/think/topics/explainable-ai>.
- [135] Khan AA, Hasan MT, Kemell KK, Rasku J, Abrahamsson P. Developing retrieval augmented generation (RAG) based LLM systems from PDFs: an experience report. *arXiv:2410.15944*. 2024.
- [136] Garcia GL, Manesco JRR, Paiola PH, Miranda L, de Salvo MP, Papa JP. A review on scientific knowledge extraction using large language models in biomedical sciences. *arXiv:2412.03531*. 2024.
- [137] Neha F, Bhatti D, Shukla DK. Retrieval-augmented generation (RAG) in healthcare: a comprehensive review. *AI* 2025;6(9). <https://doi.org/10.3390/ai6090226>. <https://www.mdpi.com/2673-2688/6/9/226>.
- [138] Hong E, Ong YN. Leveraging rag-powered LLMs for analytical tasks. 2024. <https://engineering.grab.com/transfoming-the-analytics-landscape-with-RAG-powered-LLM>.
- [139] Jeong DP, Lipton ZC, Ravikumar P. LLM-select: Feature selection with large language models. *arXiv:2407.02694*. 2024.
- [140] Li D, Tan Z, Liu H. Exploring large language models for feature selection: a data-centric perspective. *arXiv:2408.12025*. 2024.
- [141] Li S, Kulkarni J, Menache I, Wu C, Li B. Towards foundation models for mixed integer linear programming. *arXiv:2410.08288*. 2025.
- [142] Eldridge E, Trowbridge E, Berns B, Heidt J, Stenkowski N, Adams C, Mack CD. From text to insight: Methods for validation of NLP-extracted real-world oncology data using a human-in-the-loop electronic review tool. *J Clin Oncol* 2025;43(16 suppl):23291. PMID: https://doi.org/10.1200/JCO.2025.43.16_suppl.e23291.
- [143] Saad M, Zhang Y, Tian J, Jia J. A graph database for life cycle inventory using neo4j. *J Clean Prod* 2023;393:136344. <https://doi.org/10.1016/j.jclepro.2023.136344>.
- [144] Peng T, Gao L, Agbozo RSK, Xu Y, Svyarenko K, Wu Q, Li C, Tang R. Knowledge graph-based mapping and recommendation to automate life cycle assessment. *Adv Eng Inform* 2024;62:102752. <https://doi.org/10.1016/j.aei.2024.102752>.
- [145] Wang Y, Tao J, Liu W, Peng T, Tang R, Wu Q. A knowledge-enriched framework for life cycle assessment in manufacturing. *Procedia CIRP* 2022;105:55–60, the 29th CIRP Conference on Life Cycle Engineering, April 4–6, 2022, Leuven, Belgium. <https://doi.org/10.1016/j.procir.2022.02.010>.