Environmental Science & Technology

pubs.acs.org/est

Perspective

Mitigating Grand Challenges in Life Cycle Inventory Modeling through the Applications of Large Language Models

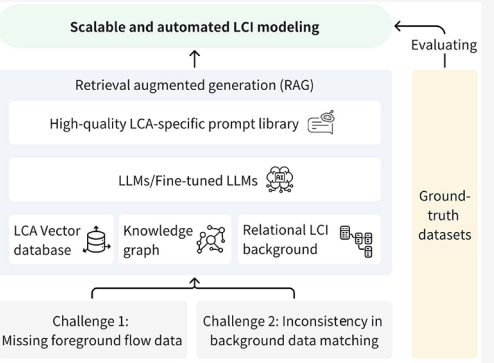
Qingshi Tu,* Jing Guo, Nan Li, Jianchuan Qi, and Ming Xu

Cite This: *Environ. Sci. Technol.* 2024, 58, 19595−19603

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The accuracy of life cycle assessment (LCA) studies is often questioned due to the two grand challenges of life cycle inventory (LCI) modeling: (1) missing foreground flow data and (2) inconsistency in background data matching. Traditional mechanistic methods (e.g., process simulation) and existing machine learning (ML) methods (e.g., similarity-based selection methods) are inadequate due to their limitations in scalability and generalizability. The large language models (LLMs) are well-positioned to address these challenges, given the massive and diverse knowledge learned through the pretraining step. Incorporating LLMs into LCI modeling can lead to the automation of inventory data curation from diverse data sources and to the implementation of a multimodal analytical capacity. In this article, we delineated the mechanisms and advantages of LLMs to addressing these two grand challenges. We also discussed the future research to enhance the use of LLMs for LCI modeling, which includes the key areas such as improving retrieval augmented generation (RAG), integration with knowledge graphs, developing prompt engineering strategies, and fine-tuning pretrained LLMs for LCI-specific tasks. The findings from our study serve as a foundation for future research on scalable and automated LCI modeling methods that can provide more appropriate data for LCA calculations.



KEYWORDS: *life cycle assessment, life cycle inventory, missing data, background data mapping, large language models, automation, scalability*

INTRODUCTION

Life cycle assessment (LCA) is a systematic method to estimate the environmental impacts of a product system from its entire life cycle. Conducting LCA for emerging technologies remains a challenge, primarily due to the data gap in life cycle inventory (LCI).[1,2] Currently, there is a significant disparity between the swift advancements in technology in areas like biochemicals and biomaterials and the availability of high-quality inventory data for evaluating their environmental impacts.[3,4] This discrepancy hinders the assessment and enhancement of the sustainability of new technologies, especially in their developmental stage.[5−7] For instance, deep eutectic solvents (DES) are increasingly investigated within new technological applications to decrease energy use in the production of lignocellulosic materials. The difficulty in evaluating the environmental effects of these nascent technologies arises from the absence of LCA studies for many DES compounds (e.g., choline chloride, p-Toluenesulfonic acid).[8]

An LCI model consists of a foreground inventory data set and an associated background inventory data set. Currently, there is a remarkable difficulty in transforming the information on real-world activities (e.g., electricity use, waste treatment) into data for LCI modeling, particularly for emerging technologies, due to the challenges in information collection and processing. Specifically, there are two grand challenges: (1) creating missing foreground flow data and (2) background data matching.[9] The former refers to the situations where key foreground data (e.g., heat consumption of a chemical reaction) is not available. Traditionally, for creating missing foreground data, LCA practitioners have been relying on the "mechanistic approach" which typically requires a high level of domain knowledge. For example, process simulation using chemical engineering software (e.g., AspenPlus) is a common mechanistic approach for estimating the energy consumption, material, and waste flows of the chemical manufacturing processes.[10] Effective as it is, there is a high knowledge barrier (e.g., chemical engineering expertise and knowledge of facility-specific design) for applying the process simulation.[11]

Received: July 26, 2024
Revised: October 11, 2024
Accepted: October 11, 2024
Published: October 21, 2024

Background data matching refers to the situations where one needs to select the most suitable unit processes from a given background database (e.g., Ecoinvent[12]) to represent the respective activities that generate the foreground flow (i.e., "flow-provider" pairing). For example, LCA practitioners may face the question "what unit process should I choose to represent the production of NaOH involved in my chemical reaction?" Similar to foreground data creation, background data matching also relies heavily on the domain knowledge of the LCA practitioners, which is prone to high subjectivity and hence leading to high uncertainties in the LCA results.

The traditional methods are limited in scalability due to the low level of automation. Given the heavy reliance on domain knowledge and case-specific nature of the mechanistic approach (e.g., an AspenPlus model may only be valid for a specific configuration of a process), the generalizability of traditional methods is also limited. On the other hand, machine learning methods have the advantage of leveraging large data sources across multiple domains, as well as an improved level of automation,[13] which is an important improvement for scalability compared to the traditional methods. Consequently, multiple studies have explored the feasibility of applying machine learning to estimating the missing inventory data.[14−20] These studies are primarily based on the concept of "similarity" that is quantified by calculating the distance (e.g., Minkowski distance) between the process of interest (e.g., a unit process with missing flow data) and a given process in the pool of candidates (e.g., unit processes in a LCA database such as Ecoinvent), using their respective values in a multidimensional embedding space.[21,22] The missing data of interest is then estimated by either selecting a proxy (e.g., corresponding flow data in the process with the shortest distance) or creating a new value (e.g., average value from top K most similar processes) from the pool of candidates, based on the calculated distances.[9] Meron et al. (2020)[14] developed a similarity-based method for choosing the most suitable proxy process for an unknown unit process by using a set of attributes (e.g., regional population density, regional average waste composition index for municipal solid waste (MSW) management processes) to describe the available unit processes. These attributes create a multidimensional space in which each unit process is placed according to its attribute values. Expert knowledge is needed to select the correct attributes and their value ranges, ensuring that the Euclidean distance (determined by the attribute values) between two more similar unit processes is shorter than between two that are less alike. When the attribute values for an unknown process are given, the most suitable (i.e., closest in distance) unit process can be automatically chosen as the appropriate proxy process. The efficacy of this proxy selection approach is sensitive to the choice of the characteristics and their value ranges. Besides scalability, ML methods have also demonstrated the potential of improving generalizability. For example, by training from a sufficiently large pool of simulation data, ML models can predict the dynamics of a complex industrial system (e.g., the governing mechanism in a distillation column)[23] or serve as a surrogate model to directly output the results of an evaluation metric (e.g., cost, energy consumption).[24] Such ML models, once trained for some key processes that are commonly used in a particular industry, can be applied broadly to different analytical situations for that industry.

Despite their recent success in addressing the two grand challenges of LCI modeling, there are still significant limitations of the existing ML applications. For similarity-based ML approach, the performance is generally instable due to the variations in the factors such as the composition of training data (e.g., imbalanced representation of certain industrial processes) and the extent of missing information, which hinders its generalizability for a broad application. For example, the accuracy of Hou et al. (2018)'s method was limited even when the fraction of missing inventory data was small (e.g., 5%) in a unit process.[21] Despite an improvement in handling a larger percentage of missing inventory data compared to Hou et al. (2018), the performance of Zhao et al. (2021)'s method deteriorated considerably when the missing data to be created had an extremely small value (e.g., less than 10^{-7}).[22] For surrogate ML models trained from simulation data, it is still time-consuming to manually generate sufficient amount of training data, which restricts their scalability. Also, the approach may only be feasible for certain domains where "mass production" of synthetic data is feasible because of the availability of the established tools (e.g., AspenPlus). In addition, assumptions behind the training data collected from literature are often not consistently mentioned in the original studies (e.g., assumptions may be summarized in a table, included in a figure or are scattered throughout the text of a study). Manually examining these assumptions presented in different data modes (i.e., tables, figures, texts) significantly limits the scalability of training data validation, and hence may compromise the validity of ML applications in LCI modeling. In fact, the root causes of these limitations can be summarized as (1) lack of diverse and large-size data sources, and (2) lack of the capacity to conduct multimodal analysis to accommodate diverse formats of data.

LARGE LANGUAGE MODELS (LLMS) TO IMPROVE LCI MODELING

To address these root causes, new methodologies that can (1) avail diverse data sources through automation and (2) incorporate a multimodal analytical capacity into LCI modeling are critical. In this perspective study, we illustrate that such a paradigm shift is achievable through the implementation of the large language models (LLMs) that possess the unique advantages of handling mixed type of data, performing semantic search, and conducting reasoning.

LLMs have incorporated an enormous amount of information that covers a wide range of domains through pretraining,[25] which contributes to the remarkable potential of their applications in the research and development in many fields, such as environmental research, material science, chemistry and biology.[26−31] This indicates that LLMs have automatically availed a wide variety of domain knowledge, when applied to answer LCI modeling questions related to specific products or services. For example, Deng et al. (2023) developed an automated product carbon footprint (AutoPCF) calculation framework that utilizes LLMs as the reasoning engine to infer the plausible process (and activities within) involved in a product system, and to match activities with their respective emission factors to calculate the PCF.[32] AutoPCF has the potential to significantly reduce the cost and time required to complete a PCF calculation. Another advantage of AutoPCF is its access to a vast collection of common knowledge (that is learnt by LLMs during pretraining step), which can surpass the knowledge of the individual human

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603

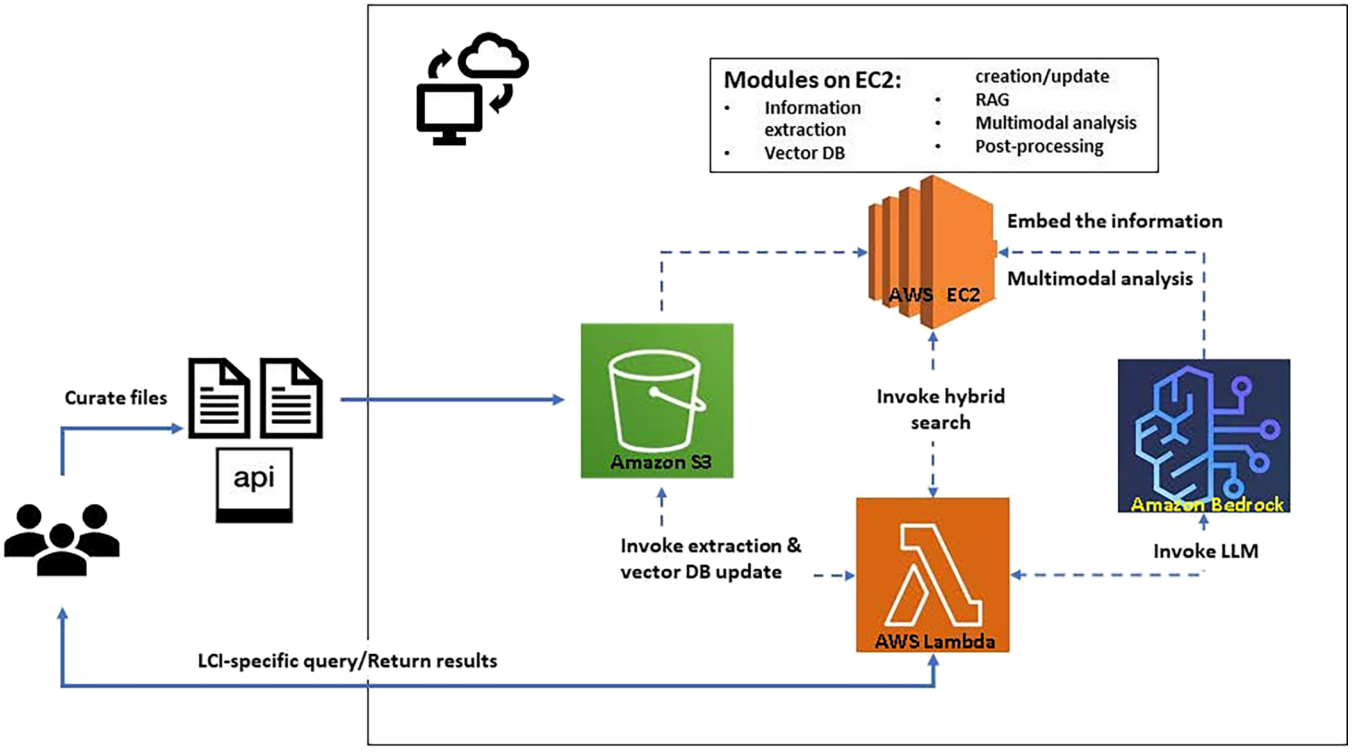


Figure 1. Illustration of LLM-assisted LCI modeling workflow (using Amazon AWS services as an example).

experts in some cases, leading to a possibly more accurate PCF calculation result. The authors reported that AutoPCF based on GPT-3.5 provided comparable results to the benchmark approach (i.e., expert-knowledge based PCF calculation) for several products. On the other hand, the generalization potential of AutoPCF is still limited by the availability and diversity of data used during the pretraining step of the LLMs, which may be enhanced by applying retrieval augmented generation (RAG) algorithms[33,34] that can introduce additional information at inference that was not originally included in the pretraining step. Balaji et al. (2023)[35] developed a SBERT (Sentence Bidirectional Encoder Representations from Transformers))-based algorithm named "Flamingo" that automates the semantic matching between a product description and the most relevant Environmental Impact Factor (EIF) from an Ecoinvent database. Flamingo leverages industry sector classification codes (e.g., Harmonized System (HS) at 4-digit level) to determine the existence of potential match, as well as to improve EIF matching precision. The algorithm demonstrates a potential to reduce the time for LCA modeling by avoiding the manual matching process. One major limitation of flamingo algorithm is recognizing relationships between closely related EIFs or product classifications, especially when broad or nonspecific descriptions are involved. Mismatches were observed when a product category contains a diverse group of items. Improving the algorithm's ability to understand these complex relationships and utilize the hierarchical structure of product classifications could mitigate these errors.

LLMs are also uniquely positioned to perform multimodal analyses to generate different formats of information for LCI modeling, by leveraging tabulated data, text and images from documents.[36] There are different approaches for text mining tasks, including rule-based methods (e.g., applying predefined linguistic rules), statistical methods (e.g., applying algorithms such as term frequency-inverse document frequency (TF-IDF) for keyword extraction), machine learning-based methods (e.g., recognizing specific entities or relationships within a sentence using trained classifiers), and traditional natural language processing (NLP) based methods (e.g., Part-of-Speech (PoS) tagging and parsing). The (probably) closest comparison is "NLP methods vs LLMs" to extract domain knowledge from literature. Traditional NLP methods involve a series of steps such as tokenization, POS tagging, syntactic parsing, and the use of handcrafted rules or statistical models to extract information. These methods often require extensive preprocessing and are heavily dependent on domain-specific knowledge and feature engineering. They can be effective for structured and well-formatted texts but may struggle with the variability and complexity found in extensive literature databases. On the other hand, LLMs offer a more flexible approach. Trained on vast amounts of data, LLMs can understand context, semantics, and nuanced language patterns without the need for explicit feature extraction or rule-based programming. They can process unstructured text directly, handling diverse formats and writing styles found in literature. This capability allows LLMs to potentially extract domain knowledge more efficiently and accurately, even from complex and heterogeneous documents like PDFs.

Several key elements to the success of an LLM-assisted method for information extraction and curation include: (1) segmentation of elements within a document (e.g., tables, text, figures in a PDF file), (2) information extraction and synthesis from each element (e.g., extract a table from a PDF file and save it as a csv file, summarize the material quantity involved in an industrial process, identify the chemical reactions of interest), (3) formulate query-ready information that allows human-machine interaction tasks (e.g., SQL query on tabulated data, prompt for reasoning), (4) LLM agents[37,38] that can fulfill different tasks (e.g., SQL query, search Internet,

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603

search database). With this automated, multimodal data extraction capability, LLMs can also be applied to create missing inventory data by fusing the information extracted from technoeconomic analysis studies. For example, by simultaneously querying the content of a process flow diagram created by AspenPlus, synthesizing semantic meaning from text related to technology description and identifying operational conditions from tables, it is possible to create a system boundary diagram and an estimation of energy, material and waste flows using the reasoning capability of LLMs. LLMs can also assist inventory data validation, by identifying the existence (or lack thereof) the description of assumptions (often in text of a study) behind values in an inventory table. Besides, LLMs can extract the meta-data of the study (e.g., geographical location, year of study) which can be used to assign values for pedigree matrix[39,40] as an indication of data quality. For example, it is possible to write a function to calculate the difference between the year of study conducted (extracted from the study) and the year of creation of the background data cited (e.g., extracted from an Ecoinvent process' data description field). One can use this difference to assign a value (e.g., if difference >15 years then assign a value of 5, indicating a "poor" data quality) for the indicator "Temporal correlation" to indicate the data quality of the unit process selected for the modeling (e.g., a unit process created more than 15 years ago will likely have a poor representativeness of the technology to be modeled in current study). This (assigning value) step can also be automated using LLM or rule-based heuristics.

In summary, LLMs can have versatile applications in LCI modeling, including multimodal information extraction and curation, system boundary completion, creating missing values of foreground inventory flows and background data matching. The substantial and diverse amount of domain knowledge in pretrained LLMs, assisted by RAG algorithms and scalable external knowledge bases, provide a potentially scalable solution for LCI modeling for many domains. A typical workflow of LLM-assisted LCI modeling may contain the following key steps: (1) curate a list of files (documents, spreadsheets, images, etc.) which contains the information on interest. This step can also be automated, if the documents are paired with digital object identifiers (DOI), by using the APIs of crossref.org and reference manage software such as Zotero. (2) Automate the information extraction using LLMs as mentioned above. (3) Prepare the knowledge base by processing the files to store the information locally or in a cloud-based vector database (e.g., pinecone.io). Steps 1−3 can be periodically executed to update the knowledge base of interest. (4) A script that is customizable for users to ask LCI-related questions, which performs RAG to retrieving relevant chunks of embedded information from the knowledge base to generate a response (e.g., identify the appropriate background data from Ecoinvent database to match a foreground inventory flow of interest). Figure 1 illustrates such a workflow using Amazon AWS services at an example. The system begins with the setup and input management using Amazon S3 to store and manage both input PDF files and subsequent outputs. This is complemented by AWS Lambda, which triggers the workflow upon new PDF uploads. The next phase involves extracting text, tables, and images from these PDFs using "unstructured" (unstructured-io.github.io) and other relevant python libraries. An EC2-hosted vector database such as Pinecone will store and manage the embeddings created by an LLM (e.g., Amazon Titan or Claude 3) from Amazon Bedrock. When a user query about inventory data and its assumption is entered, relevant information (text, table and/or image) will be retrieved from the vector database. API calls are made to LLMs to interpret the semantic meaning of the search results, as well as their relationship with respect to the query. The integration of search results and the multimodal analysis are facilitated by Lambda functions. The interpretation and the search results are processed and formatted through additional Lambda functions, before returned to the user.

INTEGRATION BETWEEN LLMS AND KNOWLEDGE GRAPHS

An LCI model can also be transformed into a knowledge graph where the nodes represent the activities (e.g., biomass production, electricity generation) and the edges are the links that define the relationship between two activities (e.g., quantity of biomass from biomass production node fed to the electricity generation node).[41] A major challenge of RAG, particularly when handling complex queries, is its limited capacity to incorporate relationship information among different sources (e.g., aggregating information from multiple documents).[42] This can lead to decreased relevance in retrieved context information for LCI modeling. Knowledge graph, with its inherent advantage of graph traversal, can improve retrieval quality significantly.[43−45] By organizing the embedding chunks as a knowledge graph, the relevant chunks can be readily retrieved through graph traversal. Besides, compared to vector similarity search, knowledge graph-assisted query can capture relationships other than topological relationship between text and question. In addition, knowledge graph can also provide LLMs with up-to-date, domain-specific knowledge given its advantage in easiness for update, which is particularly valuable for specialized tasks such as LCI modeling where domain knowledge updates frequently.

Creating such knowledge graphs entails the application of the ontologies representing both the LCA methodology and domain knowledge of interest.[46] There have been several developments of ontology-based LCA methodology. Ingwersen et al. (2015) proposed the use of Resource Description Framework (RDF) to organize the knowledge in "triples" for an LCI model (e.g., "Intermediate chemical A" → "consumes" → "electricity" during a given synthesis step of making final product B).[47] Ghose et al. (2022) developed a core set of ontology definitions for LCA (e.g.,"hasActivityType", "hasLocation") that allows a meaningful linking and storing of data under a common framework for all LCA models.[48]

Overall, the contextual understanding capacity of LLMs and structured, semantically rich information from knowledge graphs are complementary, which can lead to more accurate and relevant responses when applied together. This unification also improves data organization, allowing LLMs to efficiently access a vast array of structured information, which is crucial for the scalability of missing inventory data generation.[49] Furthermore, the structured nature of knowledge graphs helps in reducing ambiguity in language understanding, thereby enhancing the clarity and precision of the language generated by LLMs,[50] which is crucial for reducing the variations in background data mapping results during LCI modeling.

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603

DISCUSSION

Benefit for Climate Change Research. For the LCA community, LLM-assisted LCI modeling methods bring a paradigm shift in sourcing data. Not restricted by tabulated data and relational databases, more precise matching between foreground and background data can be achieved by leveraging text descriptions, knowledge graph, and many other structured or unstructured data sources. RAG provides the new capacity that allows access to the constantly growing domain knowledge, which significantly accelerates the new inventory data creation. LCA is the foundation of many climate change research work and policy decisions. When it comes to modeling emerging technologies in large systems models (e.g., IAMs), an up-to-date and comprehensive representation of their environmental impacts is critical.[51] The LLM-assisted LCI modeling methods can significantly strengthen the representation of LCA information on emerging technology by mitigating the inventory data gaps. This can lead to a significant improvement on the understanding of their potential role in mitigating climate change impacts, which accelerates the climate change research and policy making processes.

Challenges and Future Research. The following discussion aligns with a recent study by Preuss et al.[52] that proposed a wide range of possible applications of LLM to LCA. We provide a more focused discussion on data and methodological challenges with respect to LCI modeling. Creating LLM-assisted LCI modeling methods is essentially equivalent to building context-aware reasoning applications, which requires a well-aligned workflow that performs data engineering, prompt engineering, and evaluation (for both RAG and fine-tuning).

Data Challenges and Mitigation Strategies. Although data ingestion can be achieved through the established workflow provided by major databases (e.g., singlestore.com, pinecone.io), one remaining challenge is the lack of automated data curation methods (i.e., data fetching, parsing and cleaning). Avoiding the scalability issue of the current manual workflow is crucial for the curation of representative (i.e., sufficiently large size, balanced) data sets for RAG and fine-tuning LLMs. The groundtruth data for missing inventory creation should consist of inventory tables from relevant LCA studies and process simulation results (e.g., from relevant techno-economic analysis studies), as well as text descriptions that are not included in inventory tables (e.g., description of electricity consumption for a particular step of the chemical synthesis). The groundtruth data for background data mapping should consist of "flow-provider" pairs extracted from relevant LCA studies. Besides increasing the overall availability of training data, attentions should also be given to ensuring a balanced representation of all relevant categories. Methods for unstructured data extraction from documents will play an important role in automated data curation.[53] Multimodal data fusion from a diverse collection of sources, including tabulated data, figures (e.g., flowcharts from process simulations), text and knowledge graph, can be a breakthrough for LLM-assisted LCI modeling. In addition, creating a set of LCA-specific prompt templates and an example library will significantly improve the consistency and efficiency of using the LLM-assisted LCI modeling methods by LCA researchers and practitioners. A prompt template for LCI-related queries may include: (1) a proper setting of "role", "context", "system" messages, (2) instructions to extract information (e.g., zero-/few-shot prompt, chain-of-thought reasoning), (3) chain of prompts, (4) instructions to use different LCA-specific agents.

We acknowledge that data sets containing incorrect information extracted from literature (e.g., incorrect mapping between flow and provider, incorrect value of electricity use for a specific manufacturing step) can compromise the efficacy of applying LLMs to bridge the data gaps for LCI modeling. Also, completely eliminating such incorrect data using automated data curation methods is currently not feasible. Nonetheless, by applying techniques and best practices developed for the machine learning field in general—such as thorough data preprocessing and cleaning, using high-quality data sources, implementing robust, human-in-the-loop validation protocols, applying statistical noise reduction methods, and ensuring transparency and explainability—we can significantly reduce the incorrect information in the curated data sets.

Method Development Challenges and Mitigation Strategies. Determining the threshold of relevance for responses from an LLM is critical for its practical application in LCI modeling, as LLMs will not generative perfect answers even after RAG or fine-tuning.[54,55] This involves a combination of qualitative and quantitative measures. Currently, there is no consensus on what constitutes a "relevant" response in the context of LCI modeling (e.g., to what extent are the matching results of background data considered acceptable?). Factors to consider include: accuracy, completeness and specificity (e.g., is the "market for fruits and vegetables" a good background data match for "watermelon growth"?). Accordingly, quantitative metrics are needed to measure the relevance of a response. In addition to the conventional metrics such as precision, recall and F1 score, another possible solution is to adapt the widely used pedigree matrix (a semiquantitative data quality measurement method). Equally important is the design of a scoring rubric for evaluating the relevance metrics. Next, a benchmark data set of representative queries and ideal responses relevant to LCI modeling needs to be compiled. The initial threshold of relevance (based on the quantitative metrics) is set by using the benchmark data set to evaluate the LLM's baseline performance using the scoring rubric. The evaluation can be conducted by the same or another LLM through proper prompting techniques. It is crucial to analyze the evaluation outcome to identify patterns or consistent issues that indicate where the initial threshold may be too strict, too lenient, or misaligned with actual information needs. It is integral to incorporate domain expert feedback in this step. The evaluation from domain experts on the LLM's responses, particularly in edge cases or where the metrics suggest borderline relevance, can provide qualitative insights that are not captured by quantitative metrics. Next, the relevance thresholds may need to be adjusted which involves redefining what constitutes relevance or changing how different aspects of a response are weighted. Given the iterative nature of this evaluation-adjustment cycle, the Reinforcement Learning from Human Feedback (RLHF)[56] approach can be adopted to enhance the scalability of this step.

With the above-mentioned pipeline for relevance calibration, LLM-assisted LCI modeling methods can be further enhanced by one or a combination of the following approaches: domain-specific prompt engineering[57,58] (e.g., refine the prompts by incorporating insights from the relevance calibration to guide the model toward generating more relevant responses), advanced RAG,[59,60] domain-specific fine-turning,[61,62] and

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603

postprocessing enhancement (e.g., apply response ranking or filtering).

The success of RAG depends on multiple factors including the nature of query (e.g., answering complex queries spanning multiple documents), query quality and source of external information (e.g., distributed among multiple documents). The retrieved text chunks from LCA studies may not be directly useful to complete the queries such as "complete the system boundary of product A", "choose the most appropriate provider from the inventory flow B", "breakdown the environmental impacts by stages (e.g., raw acquisition, manufacturing, use phase)". There are multiple directions for future research to improve the retrieval quality/relevance, ranging from synergizing prompt engineering with RAG, adjusting the key RAG parameters such as chuck size, the top_k threshold and chuck overlap, to testing different advanced retrieval mechanisms. Another important strategy is to conduct domain-specific fine-tuning of a pretrained LLM to further boost the performance of LLM-assisted LCI modeling. LLMs are primarily trained on large data sets and leverage patterns and information present within that data. In scenarios lacking sufficient domain-specific data, the reasoning ability of LLMs may be limited. A pretrained LLM can be enhanced to generate answers specifically for LCI modeling, through algorithms such as parameter efficient fine-tuning (PEFT) algorithms such as Low Rank Adaptation (LoRA)[63] and direct preference optimization.[64] A sufficiently large and diverse "{instruction: completion}" training data set is the key to the success of such a domain-specific fine-tuning task. Such a data set can be created by domain experts with the help from a pretrained LLM (e.g., OpenAI's GPT-4 model).[65] With improved RAG and/or fine-tuning pretrained LLMs, the LLM-assisted LCI modeling methods can automate the retrieval of relevant information before conducting the quantitative modeling. These methods can also scan multiple sources to create the system boundary of a product system from scratch, as well as suggest the possible missing steps involved in a product system by using the available information as the context for LLMs to generate relevant information.

The success of LLM-assisted LCI modeling will greatly enhance the availability and quality of inventory data for a wide variety of domains, which has a profound implication for the LCA community and climate change research. It represents a potential solution to the critical need for scaling up the LCA research with the rapid growth of technologies, which is essential for policy design for promoting the sustainable development of these technologies. Given the challenges identified above, we suggest the current LLM-assisted LCI modeling methods be viewed as complementary to expert knowledge and traditional LCI modeling practices. We also expect that, with the rapid advancements in the research field of LLMs, further integration of LLMs into human LCI modeling practices to improve efficiency is achievable.

AUTHOR INFORMATION

Corresponding Author

Qingshi Tu − *Sustainable Bioeconomy Research Group, Department of Wood Science, The University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada;* orcid.org/0000-0001-7113-0564; Email: Qingshi.tu@ubc.ca

Authors

Jing Guo − *School of Environment, Tsinghua University, Beijing 100084, China;* orcid.org/0000-0003-1962-6559

Nan Li − *School of Environment, Tsinghua University, Beijing 100084, China*

Jianchuan Qi − *School of Environment, Tsinghua University, Beijing 100084, China;* orcid.org/0000-0001-7026-2442

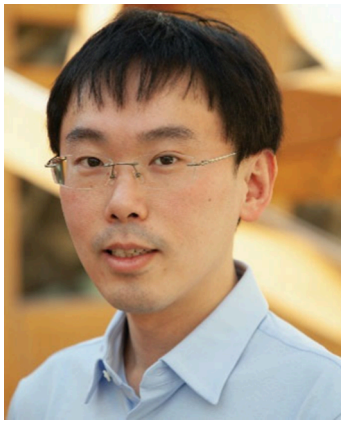
Ming Xu − *School of Environment, Tsinghua University, Beijing 100084, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c07634

Notes

The authors declare no competing financial interest.

Biographies



Dr. Qingshi Tu is an Assistant Professor of Industrial Ecology at Department of Wood Science at UBC. Dr. Tu has a strong record of life cycle assessment (LCA) and techno-economic analysis (TEA) research on a variety of topics. Dr. Tu's research focuses on 1) creating open-source databases and models for evaluating the environmental, economic, and social impacts of emerging technologies, 2) transforming knowledge into user-friendly tools and educational materials, and 3) engaging different stakeholders to collaborate on sustainable bioeconomy projects. A main focus of his recent projects is developing scalable, transparent, and automated methods to address the two grand challenges in life cycle inventory (LCI) modeling: (1) missing foreground flow data and (2) inconsistency in background data matching. Dr. Tu's team has been investigating different machine learning algorithms and implementation libraries to delineate their advantages and limitations in addressing these two grand challenges. A particular interest is in leveraging the versatility of large language models (LLMs) and enhancing their applications to the LCA domain.

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603

Dr. Jing Guo serves as an Assistant Researcher at Tsinghua University, specializing in the application of data-driven artificial intelligence methodologies to address challenges in sustainable environmental resource management. Her current research focus lies in the utilization of generative AI within environmental and ecological domains, with particular emphasis on life cycle assessment, Environmental, Social & Governance (ESG) industry benchmarking, and greenwashing detection. She has made significant contributions to open-source projects related to Large Language Models on GitHub. Leveraging this expertise, she has provided AI solutions to prominent enterprises, including Alibaba Group, demonstrating the practical applications of her research in industry settings.

Nan Li is an Associate Professor at the School of Environment, the Deputy Director of the Ecological Environment Artificial Intelligence Research Center, and the Deputy Director of the Institute for Environmental Sustainability Research, Tsinghua University. He has extensive experience in environmental systems analysis, information systems architecture and implementation, and environmental big data analysis and intelligent applications. His current research focuses on the application of generative artificial intelligence in the environmental sector. He has developed open-source lifecycle assessment database and software and Large Language Model-assisted tools as the most primary contributor (https://github.com/linancn).

Dr. Jianchuan Qi is an Assistant Researcher at Tsinghua University, specializing in Life Cycle Assessment (LCA), LCA database development, and the application of Generative AI (GenAI) in LCA. His research focuses on advancing digital intelligence-based lifecycle assessments, product eco-design, and sustainable supply chain solutions, bridging academic innovation with industrial applications. He is actively involved in the TianGong Initiative, a nonprofit, research-driven international community that promotes sustainability through open LCA and AI. In this role, he contributes to the development of methods and standards for database construction, data development tools, and AI applications, while promoting collaborative innovation with industry partners.

Ming Xu is the Chair Professor of Carbon Neutrality, Associate Dean for Research, and Director of the Center of Artificial Intelligence for Ecology and Environment in School of Environment at Tsinghua University. Prior to that, he was a Professor in the School for Environment and Sustainability and a Professor in the Department of Civil and Environmental Engineering at the University of Michigan, Ann Arbor. His research focuses on environmental systems engineering, life cycle assessment, and environmental artificial intelligence. He was awarded the Robert A. Laudise Medal from the International Society for Industrial Ecology in 2015, the US National Science Foundation Faculty Early Career Development (CAREER) Award in 2016, and the Walter L. Huber Civil Engineering Research Prize from the American Society of Civil Engineers. He was elected to Chair the 2024 Gordon Research Conference on Industrial Ecology and President of the International Society for Industrial Ecology (term 2023-2024). Currently he serves as the Editor-in-Chief of Elsevier's flagship journal in environmental management, Resources, Conservation & Recycling. He has been leading the TianGong Initiative (https://www.tiangong.earth) to develop the TianGong LCA Database and the world's first large language model application in environment and sustainability (TianGong AI).

■ ACKNOWLEDGMENTS

The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number RGPIN-2021-02841].

■ REFERENCES

(1) Tsoy, N.; Steubing, B.; van der Giesen, C.; Guinée, J. Upscaling Methods Used in Ex Ante Life Cycle Assessment of Emerging Technologies: A Review. *Int. J. Life Cycle Assess* **2020**, *25* (9), 1680−1692.

(2) Hellweg, S.; Milàì Canals, L. Emerging Approaches, Challenges and Opportunities in Life Cycle Assessment. *Science* **2014**, *344* (6188), 1109−1113.

(3) Canals, L. M. i; 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. *Journal of Industrial Ecology* **2011**, *15* (5), 707−725.

(4) Grabowski, A.; Selke, S. E. M.; Auras, R.; Patel, M. K.; Narayan, R. Life Cycle Inventory Data Quality Issues for Bioplastics Feedstocks. *Int. J. Life Cycle Assess* **2015**, *20* (5), 584−596.

(5) Bergerson, J.; Cucurachi, S.; Seager, T. P. Bringing a Life Cycle Perspective to Emerging Technology Development. *Journal of Industrial Ecology* **2020**, *24* (1), 6−10.

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603

(6) Bergerson, J. A.; Brandt, A.; Cresko, J.; Carbajales Dale, M.; MacLean, H. L.; Matthews, H. S.; McCoy, S.; McManus, M.; Miller, S. A.; Morrow, W. R.; Posen, I. D.; Seager, T.; Skone, T.; Sleep, S. Life Cycle Assessment of Emerging Technologies: Evaluation Techniques at Different Stages of Market and Technical Maturity. *Journal of Industrial Ecology* **2020**, *24* (1), 11−25.
(7) Bartolozzi, I.; Daddi, T.; Punta, C.; Fiorati, A.; Iraldo, F. Life Cycle Assessment of Emerging Environmental Technologies in the Early Stage of Development: A Case Study on Nanostructured Materials. *Journal of Industrial Ecology* **2020**, *24* (1), 101−115.
(8) Zargar, S.; Jiang, J.; Jiang, F.; Tu, Q. Isolation of Lignin-Containing Cellulose Nanocrystals: Life-Cycle Environmental Impacts and Opportunities for Improvement. *Biofuels, Bioproducts and Biorefining* **2022**, *16* (1), 68−80.
(9) Zargar, S.; Yao, Y.; Tu, Q. A Review of Inventory Modeling Methods for Missing Data in Life Cycle Assessment. *Journal of Industrial Ecology* **2022**, *26* (5), 1676−1689.
(10) Tu, Q.; Parvatker, A.; Garedew, M.; Harris, C.; Eckelman, M.; Zimmerman, J. B.; Anastas, P. T.; Lam, C. H. Electrocatalysis for Chemical and Fuel Production: Investigating Climate Change Mitigation Potential and Economic Feasibility. *Environ. Sci. Technol.* **2021**, *55* (5), 3240−3249.
(11) Parvatker, A. G.; Eckelman, M. J. Comparative Evaluation of Chemical Life Cycle Inventory Generation Methods and Implications for Life Cycle Assessment Results. *ACS Sustainable Chem. Eng.* **2019**, *7* (1), 350−367.
(12) Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B. The Ecoinvent Database Version 3 (Part I): Overview and Methodology. *Int. J. Life Cycle Assess* **2016**, *21* (9), 1218−1230.
(13) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57* (46), 17671−17689.
(14) 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* (4), 771−783.
(15) Khadem, S. A.; Bensebaa, F.; Pelletier, N. Optimized Feed-Forward Neural Networks to Address CO2-Equivalent Emissions Data Gaps - Application to Emissions Prediction for Unit Processes of Fuel Life Cycle Inventories for Canadian Provinces. *Journal of Cleaner Production* **2022**, *332*, 130053.
(16) 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), 5531.
(17) 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 Sustainable Chem. Eng.* **2020**, *8* (2), 1252−1261.
(18) Liu, Z.; Saito, R.; Guo, J.; Hirai, C.; Haga, C.; Matsui, T.; Shirakawa, H.; Tanikawa, H. Does Deep Learning Enhance the Estimation for Spatially Explicit Built Environment Stocks through Nighttime Light Data Set? Evidence from Japanese Metropolitans. *Environ. Sci. Technol.* **2023**, *57* (9), 3971−3979.
(19) Dai, T.; Jordaan, S. M.; Wemhoff, A. P. Gaussian Process Regression as a Replicable, Streamlined Approach to Inventory and Uncertainty Analysis in Life Cycle Assessment. *Environ. Sci. Technol.* **2022**, *56* (6), 3821−3829.
(20) Ghoroghi, A.; Rezgui, Y.; Petri, I.; Beach, T. Advances in Application of Machine Learning to Life Cycle Assessment: A Literature Review. *Int. J. Life Cycle Assess* **2022**, *27* (3), 433−456.
(21) 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), 5259−5267.
(22) 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), 8439−8446.
(23) Subramanian, R.; Moar, R. R.; Singh, S. White-Box Machine Learning Approaches to Identify Governing Equations for Overall Dynamics of Manufacturing Systems: A Case Study on Distillation Column. *Machine Learning with Applications* **2021**, *3*, 100014.
(24) Huntington, T.; Baral, N. R.; Yang, M.; Sundstrom, E.; Scown, C. D. Machine Learning for Surrogate Process Models of Bioproduction Pathways. *Bioresour. Technol.* **2023**, *370*, 128528.
(25) Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large Language Models: A Survey. arXiv February 20, 2024. https://doi.org/10.48550/arXiv.2402.06196. (accessed 2024−06−01).
(26) Zhu, J.-J.; Jiang, J.; Yang, M.; Ren, Z. J. ChatGPT and Environmental Research. *Environ. Sci. Technol.* **2023**, *57* (46), 17667−17670.
(27) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; D. Bocarsly, J.; M. Bran, A.; Bringuier, S.; Catherine Brinson, L.; Choudhary, K.; Circi, D.; Cox, S.; Jong, W. A. de; L. Evans, M.; Gastellu, N.; Genzling, J.; Victoria Gil, M.; K. Gupta, A.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; W. Merz, G.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Caldas Ramos, M.; Ranković, B.; G. Rodriques, S.; N. Sanders, J.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; E. Smith, B.; Herck, J. V.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Ahmad Zia, G.; Scourtas, A.; J. Schmidt, K.; Foster, I. D.; White, A.; Blaiszik, B. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digital Discovery* **2023**, *2* (5), 1233−1250.
(28) Boiko, D. A.; MacKnight, R.; Gomes, G. Emergent Autonomous Scientific Research Capabilities of Large Language Models. arXiv April 11, 2023. https://arxiv.org/abs/2304.05332. (accessed 2024−06−01).
(29) White, A. D.; M. Hocky, G.; A. Gandhi, H.; Ansari, M.; Cox, S.; P. Wellawatte, G.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Ccoa, W. J. P. Assessment of Chemistry Knowledge in Large Language Models That Generate Code. *Digital Discovery* **2023**, *2* (2), 368−376.
(30) Yu, B.; Baker, F. N.; Chen, Z.; Ning, X.; Sun, H. LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset. arXiv April 1, 2024. https://arxiv.org/abs/2402.09391 (accessed 2024−06−01).
(31) Decardi-Nelson, B.; Alshehri, A. S.; Ajagekar, A.; You, F. Generative AI and Process Systems Engineering: The next Frontier. *Comput. Chem. Eng.* **2024**, *187*, 108723.
(32) Deng, Z.; Liu, J.; Luo, B.; Yuan, C.; Yang, Q.; Xiao, L.; Zhou, W.; Liu, Z. AutoPCF: Efficient Product Carbon Footprint Accounting with Large Language Models. arXiv August 11, 2023. https://arxiv.org/abs/2308.04241 (accessed 2024−06−01).
(33) Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv April 12, 2021. https://arxiv.org/abs/2005.11401 (accessed 2024−06−01).
(34) Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv March 27, 2024. https://arxiv.org/abs/2312.10997 (accessed 2024−06−01).
(35) Balaji, B.; Vunnava, V. S. G.; 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.
(36) Perot, V.; Kang, K.; Luisier, F.; Su, G.; Sun, X.; Boppana, R. S.; Wang, Z.; Mu, J.; Zhang, H.; Hua, N. LMDX: Language Model-Based Document Information Extraction and Localization. arXiv September 19, 2023. https://arxiv.org/abs/2309.10952 (accessed 2024−06−01).
(37) Hu, Z.; Shu, T. Language Models, Agent Models, and World Models: The LAW for Machine Reasoning and Planning. arXiv December 8, 2023. https://arxiv.org/abs/2312.05230 (accessed 2024−06−01).

(38) Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; Zhang, X. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. arXiv April 18, 2024. https://arxiv.org/abs/2402.01680 (accessed 2024−06−01).
(39) Ciroth, A.; Muller, S.; Weidema, B.; Lesage, P. Empirically Based Uncertainty Factors for the Pedigree Matrix in Ecoinvent. *Int. J. Life Cycle Assess* **2016**, *21* (9), 1338−1348.
(40) Muller, S.; Lesage, P.; Ciroth, A.; Mutel, C.; Weidema, B. P.; Samson, R. The Application of the Pedigree Approach to the Distributions Foreseen in Ecoinvent V3. *Int. J. Life Cycle Assess* **2016**, *21* (9), 1327−1337.
(41) Saad, M.; Zhang, Y.; Tian, J.; Jia, J. A Graph Database for Life Cycle Inventory Using Neo4j. *Journal of Cleaner Production* **2023**, *393*, 136344.
(42) Tang, Y.; Yang, Y. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. arXiv January 27, 2024. https://arxiv.org/abs/2401.15391 (accessed 2024−06−01).
(43) Nunes, S.; Sousa, R. T.; Pesquita, C. Multi-Domain Knowledge Graph Embeddings for Gene-Disease Association Prediction. *Journal of Biomedical Semantics* **2023**, *14* (1), 11.
(44) Wang, Y.; Lipka, N.; Rossi, R. A.; Siu, A.; Zhang, R.; Derr, T. Knowledge Graph Prompting for Multi-Document Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence* **2024**, *38* (17), 19206−19214.
(45) Buehler, M. J. Accelerating Scientific Discovery with Generative Knowledge Extraction, Graph-Based Representation, and Multimodal Intelligent Graph Reasoning. arXiv March 26, 2024. http://arxiv.org/abs/2403.11996 (accessed 2024−04−30).
(46) Malek, K.; Dreger, M.; Tang, Z.; Tu, Q. Novel Data Models for Inter-Operable LCA Frameworks. arXiv May 16, 2024. http://arxiv.org/abs/2405.10235 (accessed 2024−05−17).
(47) Ingwersen, W. W.; Hawkins, T. R.; Transue, T. R.; Meyer, D. E.; Moore, G.; Kahn, E.; Arbuckle, P.; Paulsen, H.; Norris, G. A. A New Data Architecture for Advancing Life Cycle Assessment. *Int. J. Life Cycle Assess* **2015**, *20* (4), 520−526.
(48) Ghose, A.; Lissandrini, M.; Hansen, E. R.; Weidema, B. P. A Core Ontology for Modeling Life Cycle Sustainability Assessment on the Semantic Web. *Journal of Industrial Ecology* **2022**, *26* (3), 731−747.
(49) Sequeda, J.; Allemang, D.; Jacob, B. A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. arXiv November 13, 2023. https://arxiv.org/abs/2311.07509 (accessed 2024−06−01).
(50) Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv June 20, 2023. http://arxiv.org/abs/2306.08302 (accessed 2023−11−18).
(51) Pauliuk, S.; Arvesen, A.; Stadler, K.; Hertwich, E. G. Industrial Ecology in Integrated Assessment Models. *Nature Clim Change* **2017**, *7* (1), 13−20.
(52) Preuss, N.; Alshehri, A. S.; You, F. Large Language Models for Life Cycle Assessments: Opportunities, Challenges, and Risks. *Journal of Cleaner Production* **2024**, *466*, 142824.
(53) Peng, J.; Gao, J.; Tong, X.; Guo, J.; Yang, H.; Qi, J.; Li, R.; Li, N.; Xu, M. *Advanced Unstructured Data Processing for ESG Reports: A Methodology for Structured Transformation and Enhanced Analysis*. arXiv.org. https://arxiv.org/abs/2401.02992v1 (accessed 2024−01−08).
(54) Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv May 12, 2024. https://arxiv.org/abs/2405.07437 (accessed 2024−06−01).
(55) Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv September 26, 2023. https://arxiv.org/abs/2309.15217 (accessed 2024−06−01).
(56) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; Lowe, R. Training Language Models to Follow Instructions with Human Feedback. arXiv March 4, 2022.https://arxiv.org/abs/2203.02155 (accessed 2024−06−01).
(57) Ahmed, A.; Zeng, X.; Xi, R.; Hou, M.; Shah, S. A. MED-Prompt: A Novel Prompt Engineering Framework for Medicine Prediction on Free-Text Clinical Notes. *Journal of King Saud University - Computer and Information Sciences* **2024**, *36* (2), 101933.
(58) Ribary, M.; Krause, P.; Orban, M.; Vaccari, E.; Wood, T. Prompt Engineering and Provision of Context in Domain Specific Use of GPT. In *Legal Knowledge and Information Systems*; IOS Press, 2023; pp 305−310. DOI: 10.3233/FAIA230979.
(59) Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv December 20, 2022. https://arxiv.org/abs/2212.10496 (accessed 2024−06−01).
(60) Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; Ling, Z.-H. Corrective Retrieval Augmented Generation. arXiv February 16, 2024. https://arxiv.org/abs/2401.15884 (accessed 2024−06−01).
(61) Xie, T.; Wan, Y.; Huang, W.; Yin, Z.; Liu, Y.; Wang, S.; Linghu, Q.; Kit, C.; Grazian, C.; Zhang, W.; Razzak, I.; Hoex, B. DARWIN Series: Domain Specific Large Language Models for Natural Science. arXiv August 24, 2023.https://arxiv.org/abs/2308.13565 (accessed 2024−06−01).
(62) Jeong, C. Fine-Tuning and Utilization Methods of Domain-Specific LLMs. arXiv January 24, 2024. https://arxiv.org/abs/2401.02981 (accessed 2024−06−01).
(63) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv October 16, 2021. https://arxiv.org/abs/2106.09685 (accessed 2024−06−01).
(64) Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; Finn, C. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. arXiv December 13, 2023. https://arxiv.org/abs/2305.18290 (accessed 2024−06−01).
(65) Yu, Y.; Zhuang, Y.; Zhang, J.; Meng, Y.; Ratner, A.; Krishna, R.; Shen, J.; Zhang, C. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. arXiv October 17, 2023. https://arxiv.org/abs/2306.15895 (accessed 2024−06−01).

https://doi.org/10.1021/acs.est.4c07634
Environ. Sci. Technol. 2024, 58, 19595−19603