



Testing the use of a large language model (LLM) for performing data quality assessment

Steven MacMaster¹ · Julie Sinistore²

Received: 6 February 2024 / Accepted: 7 November 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Purpose The purpose of this study is to evaluate the capability of large language models (LLMs) to perform data quality assessment on background data with respect to a study scenario.

Methods LLMs generate coherent and contextually relevant text in response to prompts. Using a chat interface and prompting the model in a conversational style, OpenAI's DaVinci model was prompted to perform a data quality assessment of background data against a study scenario using a revised Pedigree Matrix. The model performed DQA across the temporal, geographic, and technology coverage indicators. The outputs were evaluated for correctness in reasoning as well as the final scoring. Prompts underwent several iterations in some cases to improve the correctness.

Results The model was able to provide correct reasoning and scores for 100% of scenarios across the temporal coverage, 91% for geographic coverage, and 73% for technology coverage. The technology coverage test was repeated for two scenarios where contextual clues in the prompt were obscured to evaluate if the model was indexing its training data to produce scores. Both additional technology coverage scenarios had 100% success in reasoning and scoring. Basic iteration and refinement of the prompts was successful in improving the correctness of outputs.

Conclusions The results of this research indicate that LLMs are capable of performing data quality assessment across the temporal, geographic, and technology coverage indicators. Outsourcing DQA to artificial intelligence (A.I.) can eliminate practitioner's biases and reduce liability for practitioners. The success of an off-the-shelf model and improvement in results from little prompt refinement indicates that a LLM may successfully be fine-tuned for the purpose of performing DQA.

Keywords Artificial intelligence · Large language model · DQA · Automation

1 Introduction

The impending effects of climate change have driven growth in the use of life cycle assessment (LCA) up and down the value chain to quantify environmental impacts at scale. Assessments of GHGs in particular are a critical prerequisite to achieving the rapid decarbonization of products, production systems, and activities.

As the need for LCA grows, so does the challenge of scaling to meet demand. Conventional laborious methods constrain practitioners in the amount of quality work they are able to produce. Rising demand for LCA drives costs up, making LCA less accessible.

LCA practitioners must become more efficient, and artificial intelligence (A.I.) has the capability to enhance practitioner's workflows and address their major challenges, as it has done in other industries. Circa Technologies is applying A.I. to various practitioner workflows, and major challenges such as mapping background data to product system models have been the focus of A.I. applications by other companies in the past although this frontier remains largely unexplored..

Though the task of mapping background data to product system models demands significant time from practitioners, and is therefore a valuable target for optimization, it is

Communicated by Daniela Sica.

✉ Steven MacMaster
steven@circa.ai

¹ Circa Technologies, Inc., Boston, USA

² WSP USA Inc., New York, USA

underpinned by the need for efficient data quality analysis (DQA) alongside it.

ISO standard 14044 defines “data quality” as “characteristics of data that relate to their ability to satisfy stated requirements” (ISO 14044:2006 2016). As LCA relies heavily on background data from external sources, evaluating the quality of the data used in the life cycle inventory (LCI) is crucial to the integrity of a study. The process of evaluating the quality of data is known as data quality assessment (DQA).

As LCA becomes more prevalent and LCIs become more complex, the need for efficient DQA grows as well. Data quality is determined to be a “critical problem” affecting all four phases of LCA by Reap et al. (2008). The inability to effectively scale the DQA process has led to the suggestion by Edelen and Ingwersen (2018) that Data Quality Scores should be aggregated as “...it becomes impractical to examine all processes directly with the intention of drawing conclusions on data quality.”

Compounding the problem of scale, DQA has also been difficult to standardize because “quality” is effectively a subject of a practitioner’s judgment, and the same standard may not apply across products or industries. One primary method for evaluating Data Quality, called the “Pedigree Matrix,” is most widely used across the field of LCA. The Pedigree Matrix has been adapted many times, and although there has been initiative to improve the reproducibility of DQA with tools such as the Pedigree Matrix, it is still challenging to standardize value choices across the DQA process. Reproducibility of DQA scores is a challenge that has been stressed in an evaluation of competing DQA methods by Edelen and Ingwersen (2018) and researched by Wiedema (1998).

A.I. has the potential to rapidly and accurately perform DQA between product system data and background data at scale and consistently. A literature review indicates that A.I. has not previously been applied to perform DQA, or any other value-choice-based LCA tasks. This may be for several reasons; A.I. technology has only recently become accessible at a consumer level, prohibiting research into its applications. The few explorations into A.I. applications for LCA have been focused on other areas, such as data mining (de Jesus et al 2021), extrapolating/predicting life cycle inventories, and estimating impacts (Algren et al. 2021).

LCA practices that require value choices are the most difficult to automate due to the subjectivity of the task, especially within data quality assessment, where “The contextual aspects of data quality often require practitioners to make subjective decisions on the relevance of certain data quality indicators” (Edelen and Ingwersen 2018). Context-dependent tasks, however, become prime

candidates for automation by large language models, which unlike a static algorithm, “...displayed the ability to learn more novel tasks when given examples in context” (Tamkin et al. 2021).

There are currently few applications of LLMs for the purpose of scoring and evaluating subjective scenarios. The existing body of work is primarily focused on evaluating the performance of LLM outputs using LLMs (Liu et al. 2023), including correlation to human ratings (Chiang and Lee 2023).

In this paper, findings are shared from an experiment in which OpenAI’s DaVinci model is used to perform data quality assessment on sample LCI data. The findings show that DQA can be accurately performed by A.I. including making value choices.

2 Methods

2.1 Research approach

In developing approaches to performing DQA using A.I., it was decided to use a large language model (LLM). LLMs are a category of machine learning models that process human language. LLMs are easier to interface with because they can be engaged in natural language. Off-the-shelf products such as OpenAI’s DaVinci model make LLMs accessible for research purposes.

LLMs are utilized by providing a prompt and receiving a response in natural language, imitating a conversational style. The LLM was prompted to follow the DQA workflow established by the US Environmental Protection Agency’s Data Quality Assessment Guidance (Edelen and Ingwersen 2016) in which Data Quality Goals are established based on the study criteria and compared with the background data against several indicators. This approach was chosen because it allows each step of the process to be supervised and evaluated for error. Being that the ultimate goal of Circa Technologies’ research is to evaluate A.I. capabilities for enhancing LCA workflows, closely imitating the role of a human practitioner is crucial for evaluating the effectiveness of the LLM against human performance. In short, using a LLM to follow the DQA workflow was the best approximation for the performance of a human practitioner with the capabilities of A.I.

2.2 Choosing indicators

Data quality is composed of several different dimensions, called “indicators,” and the data must be evaluated against each of them.

Three quality indicators were chosen for study in this research:

1. Time-related (temporal) coverage
2. Geographical coverage
3. Technology coverage

Several data quality indicators have been excluded from the scope of this test, such as methodological or process indicators. A review of various DQA matrices found that these three flow indicators are the most commonly found among different matrices, as they are the most crucial and ubiquitous criteria for evaluating data quality. Excluding these indicators narrowed the scope of this research to indicators that have more straightforward success criteria before attempting more complex indicators.

The indicators were researched in the order listed above because each successive indicator introduces a new complexity. The temporal indicator is relatively straightforward, where the model must only assign a score based on a single dimension: the range between the given dates. The geographic indicator introduces a second dimension, where each location is assigned a resolution, and then a score is given based on the difference between each resolution. The technology indicator is the most complex and least empirical, constituting a result more akin to a value choice.

The scoring matrices for the indicators used in this research are based on a modification of the Pedigree Matrix

by the EPA that clarifies ambiguities in the scoring criteria (Edelen, A. and W. Ingwersen 2016).

2.3 Scoring matrices

The EPA DQA guidance provides definitions and scoring matrices for the three chosen indicators, as shown below:

- Time-related (temporal) coverage

Age of data and the minimum length of time over which data should be collected (Table 1). Degree to which the dataset reflects the specific conditions of the system being considered regarding the time/age of the data, and including background datasets, if any, i.e., of the given year (and, if applicable, of intra-annual or intra-daily differences).

Relevance of the primary data collected to the period of study.

Relevance of the vintage of secondary data to the period of study.

- Geographical coverage

Geographical area from which data for unit processes should be collected to satisfy the goal of the study. Degree to which the dataset reflects the true population of interest regarding geography, including background datasets, if any, i.e., of the given location/site, region, country, market, continent, etc. (Table 2).

Table 1 Temporal coverage scoring matrix, reproduced from Edelen, A. and W. Ingwersen (2016)

Indicator	1	2	3	4	5 (default)
Temporal correlation	Less than 3 years of difference	Less than 6 years of difference	Less than 10 years of difference	Less than 15 years of difference	Age of data unknown or more than 15 years

Table 2 Geographic resolution scoring matrix, reproduced from Edelen, A. and W. Ingwersen (2016)

Indicator	1	2	3	4	5 (default)
Geographic correlation	Data from the same resolution and same area of study	Within one level of resolution and a related area of study	Within two levels of resolution and a related area of study	Outside of two levels of resolution but a relate study area	From a different or unknown area of study

Resolution	A	B	C	D	E	F	G
Name	Global	Continental	Sub-region	National	(Province/state/region)	(County/city)	(Site specific)
Example	World	North America	North America	USA	Ohio	Hamilton	26 W Martin Luther King Dr

- Technology coverage

Specific technology or technology mix. Degree to which the dataset reflects the true population of interest regarding technology, including for included background datasets, if any, i.e., of the technological characteristics including operating conditions.

EPA guidance suggests dividing the technology coverage into four dimensions: process design, operating conditions, material quality, and process scale (Table 3).

1. Process design—the flow of materials and energy through the designated system boundaries. Flow diagram should be included.
2. Operating conditions—site-specific parameters such as temperature, pressure or flow rates, and production rate. If process design and production rates are the same, the operating conditions may also be assumed to be the same unless otherwise documented.
3. Material quality—materials as well as material quality (composition)
4. Process scale—production scale vs. bench scale

1. Determine conditions of the study (establish data quality goals)

The model shall request conditions of the study and establish data quality goals (DQGs). DQA is the comparison of the data collected against the DQGs, which describe the ideal representativeness for the project.

The conditions of the study needed to evaluate the selected indicators are illustrated in Table 4.

2. Determine conditions of the data

The model shall request the user's data which is being compared against the DQGs (same as above, if available).

3. Evaluation

The model shall evaluate the submitted data against the DQGs and score the data according to the indicator scoring matrices above on all 3 indicators.

4. Report

The model shall report the score.

Between indicators and prompt updates, a new chat thread was begun to avoid confusion with previous indicator scoring criteria. Between scenarios within the same indicator and starting prompt, the model was given the prompt “Let's repeat, please” to restart the dialogue exchange instructed.

2.4 Workflow

The scoring matrices provided by the EPA were translated into a basic workflow that was used to establish prompts and dialogue exchanges with the large language model.

2.5 Starting prompts

LLMs are a type of artificial intelligence model that enables computers to understand, interpret, and generate human language. By training a neural network on large

Table 3 Technology coverage scoring matrix, reproduced from Edelen, A. and W. Ingwersen (2016)

Indicator	1	2	3	4	5 (default)
Technological correlation	All technology categories are equivalent	Three of the technology categories are equivalent	Two of the technology categories are equivalent	One of the technology categories is equivalent	None of the technology categories are equivalent

Table 4 Data quality evaluation criteria by indicator

Temporal coverage
Time or time range in which the study takes place
Geographic coverage
Location where the data is intended to represent
Technology coverage
The flow of materials and energy through the designated system boundaries. Flow diagram should be included
Site-specific parameters such as temperature, pressure or flow rates, and production rate
Materials as well as material quality (composition)
Scale of production (commercial scale vs. bench scale)

datasets of text, it learns the patterns, structures, and semantics of language. It can then generate coherent and contextually relevant text in response to prompts.

The LLM is prompted within a chat interface in a conversational style. An operator enters a text prompt such as a topic, statement, or question in natural language. The model then produces text responses to the prompt.

The workflow stated above was translated into prompts for the LLM to correctly perform the assessment. The final versions of the prompts for each indicator are below:

Temporal coverage I would like you to evaluate the quality of a datapoint that will be used in a study. I'd like you to score the datapoint based on criteria I provide. First, ask me for the date of the study. I will respond with a year.

Next, ask me for the date of the datapoint. I will respond with a year.

Subtract the dates from one another. If the dates are less than 3 years apart, say 'The score is 1'. If the dates are less than 6 years apart, say 'The score is 2', if the dates are less than 10 years apart, say 'The score is 3', if the dates are less than 15 years apart, say 'The score is 4', and if the dates are 15 years apart or greater, or one of the input dates is unknown, say 'The score is 5'.

Geographic coverage I would like you to evaluate the quality of a datapoint that will be used in a study. You will do this by comparing the resolution of the location of the datapoint against the resolution of the location of the study. The goal is not to achieve any particular resolution or level of detail. The goal is only to determine the resolutions, then compare them.

First, ask me for the location of the study. I will respond with a location.

Next, ask me for the location of the datapoint. I will respond with a location.

Determine the level of resolution of the location of the study and location of the datapoint. The levels of resolution are a scale, as follows:

A is Global, aka "GLO", worldwide, or any resolution more general than continental

B is Continental

C is Sub-region

D is National

E is Province/State/Region

F is County/City

G is Site-specific, e.g., an address

For example, resolution B and resolution F would be 4 levels of resolution from one another.

Compare the resolution of the locations of the study and datapoint. If they are the same resolution, say 'The score is 1'.

If they are within one level of resolution, say 'The score is 2'.

If they are within two levels of resolution, say 'The score is 3'.

If they are outside two levels of resolution, say 'The score is 4'.

If the location of the data is unknown, say 'The score is 5'.

Technology coverage A I would like you to evaluate the quality of a datapoint that will be used in a study. You will do this by comparing the relevance of the technical details of the datapoint against the technical details of the study scenario. The goal is not to achieve any particular level of detail. The goal is only to determine the relevance by comparison.

First, ask me for the technical details of the study scenario. I will respond with details that may include the process design, operating conditions, material quality, and process scale.

Next, ask me for the technical details of the datapoint. I will respond with details that may include process design, operating conditions, material quality, and process scale.

Determine the relevance of the datapoint to the study scenario by comparing the detail's process design, operating conditions, material quality, and process scale. These are defined as follows:

Process Design- The flow of materials and energy through the designated system boundaries.

Operating Conditions- Site specific parameters such as temperature, pressure or flow rates, and production rate. If process design and production rates are the same, the operating conditions may also be assumed to be the same unless otherwise noted.

Material Quality- Materials as well as material quality, material composition.

Process Scale- e.g. Production scale vs. bench scale.

Compare the details of the datapoint and the study scenario. If all 4 technological categories are equivalent, say 'The score is 1'. If all 3 technological categories are equivalent, say 'The score is 2'. If all 2 technological categories are equivalent, say 'The score is 3'. If all 1 technological category is equivalent, say 'The score is 4'. If none of the technological categories are equivalent, say 'The score is 5'.

The process of prompt refinement is described below the results, and the previous versions of prompts are included in the [Appendix](#).

3 Results

The LLM was given first the sample study scenario, then provided the sample datapoint, according to the prompt exchanges instructed. The sample study scenarios and datapoints used in this research are fictitious yet realistic examples chosen. The decision to use plausible yet fictitious data allowed the difficulty of each scenario to be strategically increased with each iteration. The first two columns contain the content of the sample study scenario and background datapoints respectively.

Because it was asked to describe the reasoning or logic for each scenario, the LLM was evaluated for correctness within both the process it described and the final score it produced. The “Process” column includes a checkmark if the reasoning provided by the model was correct, and an X if it was incorrect. The final score was evaluated separately, likewise with a checkmark or X. In some cases, it was not possible to objectively approve or reject a result—such scenarios are marked with a dash “–”.

The last column, titled “Notes” includes information for scenarios that had an aberration or error, for example noting an erroneous score. Content in quotation marks is quoted directly from the LLM, meaning it was generated by the model as part of its prompt responses and is included

to disclose the nature of the error; content in quotes in the Notes column is not the author’s interpretation.

3.1 Temporal coverage

The temporal coverage test underwent a single iteration because the LLM was successful on the initial attempt, as seen in Table 5. This is presumably because of the simplicity of the temporal indicator, as described above in the “Choosing indicators” section and provided an encouraging outcome for the ensuing tests.

3.2 Geographic coverage

The geographic coverage test underwent four iterations of prompt refinement. The results in Table 6 are the results of the last test and reflect the improvements to the prompt from the previous tests. The prompt refinement process is described in further detail below, and the results of the previous iterations are available in the [Appendix](#).

3.3 Technology coverage A

The correctness of the technology indicator results, relying more on judgment, are disputable. i.e., a practitioner may or may not agree with the “judgment” of the model in assigning a quality score, which underpins the subjective nature of the practice. If a result could be deemed “acceptable,” even if disputable, it was considered correct because the model achieved its aim of imitating a practitioner’s judgment, which itself may be disputed. The scenarios and results are outlined in Table 7.

Table 5 Temporal coverage test scenarios and results

Test	Study scenario	Background data	Process	Calculation
1.0	2023	2020	✓	✓
1.1	2023	1970	✓	✓
1.2	2023	2020	✓	✓
1.3	2023	2016	✓	✓
1.4	2018	2012	✓	✓

Table 6 Geographic coverage test scenarios and results

Test	Study scenario	Background data	Process	Calculation	Notes
5.0	USA	USA	✓	✓	
5.1	1 Spring St. Portland Maine	Portland Maine	✓	✓	
5.2	Ohio	Global	✓	✓	
5.3	Boston, MA	USA	✓	✓	
5.4	Austin, Texas	United States of America	✓	✓	
5.5	London, UK	Europe	✓	✓	
5.6	California	Global	✓	✓	
5.7	South America	Global	✓	X	Correctly determined the level of resolution, incorrect answer: “Therefore, the score is 4, as the resolutions of the two locations are outside two levels of resolution.”
5.8	1600 Pennsylvania Avenue, Washington D.C	Washington, D.C	✓	✓	
5.9	Global	Global	✓	✓	
5.91	Global	worldwide	✓	✓	

Table 7 Technology coverage test scenarios and results

Test	Study scenario	Background data	Process	Calculation	Notes
6.0	Land-clearing debris batch digester system, 950 bph, 400 kg/h Land-clearing debris makeup: 3% leaves and grass 7% bark 10% pallets 42% brush and mixed logs 38% logs under 20' in diameter	Land-clearing debris batch digester system, 950 bph, 400 kg/h Land-clearing debris makeup: 3% leaves and grass 7% bark 10% pallets 42% brush and mixed logs 38% logs under 20" in diameter	✓	✓	Replied to the datapoint conditions: "It seems that you have provided the technical details of the study scenario again. Could you please check and provide me the technical details of the datapoint that needs to be evaluated for relevance to the study scenario?" Responded: "Those are the details of the datapoint, please proceed with the evaluation"
6.1	Tetra Pak TBA/19 filling machine The machine is set up in a controlled environment to ensure sterility and cleanliness. The packaging material is fed into the machine, where it is sterilized with hydrogen peroxide before being formed into a package. The tetra pak is composed of a laminate of polyethylene, aluminum, and paperboard, and processes 9000 packages per hour	9000 packages/h Tetra Pak TBA/19 filling machine Sterilized, formed into package Composition: polyethylene, aluminum, and paperboard	X	X	"The process scale of the datapoint is not equivalent to the study scenario, as the datapoint only provides information on the production rate, while the study scenario involves a specific filling machine." Both scenarios include the same filling machine information. Otherwise the assessment was correct
6.2	Tetra Pak TBA/19 filling machine The machine is set up in a controlled environment to ensure sterility and cleanliness. The packaging material is fed into the machine, where it is sterilized with hydrogen peroxide before being formed into a package The tetra pak is composed of a laminate of polyethylene, aluminum, and paperboard, and processes 9000 packages per hour	9000 packages/h Tetra Pak TBA/19 filling machine Sterilized, formed into package Composition: polyethylene, aluminum, and paperboard	✓	✓	
6.3	B&W Biomass boiler: feed auger, furnace, fly ash conveyor and storage silo, flue gas scrubber 16% bark residue 84% wood chip residue 0.086 tonne steam/h	16% bark residue 84% wood chip residue	✓	✓	
6.4	Carding Machine C10 Condensed Cotton 42 kg/h	A Trützschler DK 803 carding machine	✓	-	"The process design of the datapoint partially matches that of the study scenario, as both involve carding machines." This may be disputed

Table 7 (continued)

Test	Study scenario	Background data	Process	Calculation	Notes
6.5	Newspaper making machine ZDM CO Newsprint pulp, Corn Starch 0.9 tonne/h	Newspaper making machine ZDM CO 1 tonne/h	✓	✓	"The process design of the datapoint partially matches that of the study scenario, as both involve newspaper making machines. The process scale of the datapoint is very similar to that of the study scenario, as both involve the production of around 1 tonne/hr of newsprint paper. However, the operating conditions and material quality of the datapoint have not been provided, and it is not possible to determine if they match the study scenario. Since 2 technological categories (process design and process scale) are equivalent, the score for the relevance of the datapoint to the study scenario would be 3. Therefore, I would say that the datapoint has some relevance to the study scenario, but more information is needed to determine its full relevance."
6.6	Land-clearing debris batch digester system 950 bph 400 kg/h Land-clearing debris: 3% leaves and grass 7% bark 10% pallets 42% brush and mixed logs 38% logs under 20" in diameter	Morbark 1300B Tub Grinder The debris is loaded onto the machine via a conveyor belt, which feeds it into the tub chamber where it is ground up by a spinning hammer rotor. The ground-up material is then discharged onto a separate conveyor belt for further processing or transport. The debris is composed of 80% tree trunks and limbs, and 20% other woody material such as brush and leaves. 150 cubic yards per hour	✓	✓	
6.7	Tetra pak manufacturing cardboard, polyethylene, paperboard, PE, Film 13,030 cartons/month	Tetra Pak TBA/19 filling machine The machine is set up in a controlled environment to ensure sterility and cleanliness. The packaging material is fed into the machine, where it is sterilized with hydrogen peroxide before being formed into a package. The tetra pak is composed of a laminate of polyethylene, aluminum, and paperboard, 9000 packages per hour	✓	✓	

Table 7 (continued)

Test	Study scenario	Background data	Process	Calculation	Notes
6.8	Newspaper making machine ZDM CO Newsprint pulp, Corn Starch 0.9 tonne/h	Fourdrinier machine A mixture of wood pulp, water, and chemicals is fed onto a continuously moving wire mesh screen, where water is drained from the mixture and the fibers begin to bond together. The paper is then passed through a series of press rollers to remove more water and improve the smoothness of the paper. The paper is then dried using a series of heated rollers, before being wound onto a large roll. 70 °F, 50% humidity wood pulp mixed with water and chemicals, such as fillers and sizing agents, to improve the strength and printability of the paper. 200–400 m per minute	✓	✓	
6.9	B&W Biomass boiler: feed auger, furnace, fly ash conveyer and storage silo, flue gas scrubber 16% bark residue 84% wood chip residue 0.086 tonne steam/h	Hurst Biomass Boiler 30,000 lbs/h of wood chips The boiler operates at a temperature of around 800 °F (427 °C) and pressure of 150 psi (10.3 bar) 70:30 ratio of wood chips to bark produces 100,000 lbs/h of steam	✓	✓	
6.91	Carding Machine C10 Condensed Cotton 42 kg/h	A Trützschler DK 803 carding machine Condensed Cotton 150 kg/h of cotton fiber	✓	X	“However, the operating conditions and process scale of the datapoint are different than that of the study scenario. The study scenario processes 42 kg/hour, while the datapoint processes 150 kg/hour of cotton fiber. This difference in production rate implies that the operating conditions for the two scenarios could be different.” Scored 2 even though it acknowledged that only 2 dimensions are comparable

3.4 Technology coverage B

The technology indicator test was repeated for two scenarios where either the machine model name or machinery description was omitted, seen in Table 8.

For example, “A Trützschler DK 803 carding machine” → “Trützschler DK 803”

The purpose of these modified scenarios was to evaluate if the LLM was assessing the technology based off of the description in the provided scenario, or if it was indexing its training data for additional context about the technology. Before running the scenarios, the machinery model numbers were searched online to verify that context about the machine is available on the internet. Results indicated that the LLM indexed information outside of the scenario to develop the context needed to perform the DQA.

4 Conclusions

4.1 Prompt refinement

The prompt adjustments were based on the nature of the errors that would appear in the results. The refinements included emphasizing certain instructions, clarifying the nature of the instructions, and reformatting the prompt for better clarity.

For example, adding “The goal is not to achieve any particular resolution, only to compare the resolution given.” to the instructions corrected a tendency of the LLM to assign incorrect scores.

When the model produced incorrect results, the incorrect results were generally biased towards the later scenarios. LLMs use a generative process to create text. They produce results one character, also known as token, at a time and predict the

next most likely token based on the tokens previously generated as well as its training data. This is analogous to making a photocopy of an already photocopied image—through each additional copy, aberrations in the image become compounded. Because the model was relying on the previous context to output results, as the previous context became unreliable, the results would become increasingly incorrect. This can be rectified by fine-tuning the model to prevent initial errors and restarting the prompt between scenarios to avoid compounding errors. To use the photocopier analogy, this means copying the original image each time instead of the previously made copy.

4.2 Limitations

As stated in the Section 4, the scenarios provided to the model are realistic yet fictitious examples. The use of fictitious scenarios ensured consistency across test scenarios and allowed the complexity of each example to be escalated with each scenario. However, applications of LLMs for data quality scoring should be tested on actual data quality scenarios before use. Future research could elaborate on this experiment by introducing actual data quality scoring scenarios and perhaps comparing against human decisions.

The technology coverage B scenario demonstrated that the LLM was able to score scenarios with limited context based on its reference data. However, this experiment did not evaluate how the model performs for scenarios where the model was not able to fill in such gaps with its training data. Generalization, the ability of the model to perform for scenarios which it has not been trained on, is a potential challenge area. This experiment broadly suggests that the model handles generalization well because the model was not trained on any of the scenarios presented to it, as the model was not trained for DQA at all; however, the potential uncertainty introduced by generalization limitations should be considered for future applications.

Table 8 Technology indicator test repeated without descriptions

Test	Study scenario	Background data	Process	Calculation	Notes
7.0	9000 packages/h Tetra Pak filling machine Sterilized, formed into package Composition: polyethylene, aluminum, and paperboard	Land-clearing debris batch digester system, 950 bph, 400 kg/h Land-clearing debris makeup: 3% leaves and grass 7% bark 10% pallets 42% brush and mixed logs 38% logs under 20" in diameter	✓	✓	“Process Scale: The process scale is not explicitly mentioned in either the study scenario or the datapoint. Therefore, we cannot determine whether the scale is the same or different.” This is debatable; a practitioner may consider the throughput to represent scale
7.1	Carding machine Condensed Cotton 150 kg/h of cotton fiber	9000 packages/h Tetra Pak TBA/19 filling machine Sterilized, formed into package Composition: polyethylene, aluminum, and paperboard	✓	✓	“Process Design: The study scenario involves a carding machine, while the datapoint also mentions a Trützschler DK 803 machine. Since both are carding machines, the process design is equivalent.” Here, the model understands that both are carding machines

This experiment highlights another limitation of LLMs: their sensitivity to prompt structure. As evidenced by the prompt iterations described under Section 4.1, even small changes to the prompt structure and content can radically alter the correctness of results. It has encouraged for future work to engage in additional prompt refinement and consider the effect of prompt sensitivity on the reproducibility of results.

4.3 Implementation

Given the model performance and limitations, it is the author's recommendation that LLMs be used for data quality scoring or other value choices with dedicated applications such as specifically trained models and/or a potential commercially available tool. The guardrails introduced by purpose-built systems reduce the risk from prompt sensitivity.

While LLMs provide scale benefits to data quality scoring process, as the technology is still nascent, automated data quality scoring should occur under practitioner supervision.

The authors envision that this technology will not replace scoring by human practitioners, but instead practitioners may use LLM-generated scores to check their own for bias, mitigating liability and improving consistency of scoring across the field. In the future, practitioners may confer with an AI agent when executing value choices as a scalable way of acquiring a “second opinion.”

4.4 Takeaways

Large language models are capable of imitating the decision-making process that a practitioner would perform when performing DQA on background data.

Outsourcing tasks that require value-choices to LLMs offers the benefit of removing liability from the practitioner, as the model can apply consistent decision-making based on its training data. This can avoid practitioner's biases influencing value-choices and other subjective decisions, reducing risks brought by human bias.

The ability of the OpenAI DaVinci model (ChatGPT) to consistently perform DQA without any specialized background training and limited prompt refinement suggests that there is strong potential for A.I. to augment LCA practitioner workflows by automating the DQA process.

Future research could build upon the findings of this study by using a dedicated model trained on larger datasets of background data and example LCIs. Language models could be trained on specific information that would aid assessments, such as machinery specifications. Fine-tuning small language models could permit results more fit-for-purpose for specific tasks. Additional research should explore using real examples from a study LCI instead of fictitious scenarios.

The tests in this study also explore common cases, such as traditional forms of technology, and additional research could investigate how the A.I. responds to edge cases and less common examples of technology and geographic descriptions.

This research serves as an indicator that it is worth exploring LLM applications for DQA further. The application of A.I. to LCA, both for data quality assessment and other applications, could allow for more rapid and consistent exercise of value choices within LCA.

Appendix

Geographic coverage prompt iterations

As aforementioned, the geographic coverage indicator required several iterations of prompt refinement. The last and most successful prompt version is included in the results. Below, the prompts for the three previous iterations are included for the purpose of evaluating how the prompt was refined based on errors in the results.

Prompt 1

I would like you to evaluate the quality of a datapoint that will be used in a study. I'd like you to score the datapoint by evaluating the resolution of a location. First, ask me for the location of the study. I will respond with a location.

Next, ask me for the location of the datapoint. I will respond with a location.

You will determine the level of resolution of the location of the study and location of the datapoint. The levels of resolution are as follows: A is Global, B is Continental, C is Sub-region, D is National, E is Province/State/Region, F is County/City, and G is Site-specific, e.g. an address. You will then compare the resolution of the locations of the study and datapoint. If they are the same resolution, say ‘The score is 1’. If they are within one level of resolution, say ‘The score is 2’. If they are within two levels of resolution, say ‘The score is 3’. If they are outside two levels of resolution, say ‘The score is 4’. If the location of the data is unknown, say ‘The score is 5’..

Prompt 2

I would like you to score the quality of a datapoint that will be used in a study. You will do this by evaluating the resolution of the location of the datapoint. The goal is not to achieve any particular resolution, only to compare the resolution given.

First, ask me for the location of the study. I will respond with a location.

Next, ask me for the location of the datapoint. I will respond with a location.

Determine the level of resolution of the location of the study and location of the datapoint. The levels of resolution are as follows:

A is Global, worldwide, or any resolution more general than continental

B is Continental

C is Sub-region

D is National

E is Province/State/Region

F is County/City

G is Site-specific, e.g. an address

Compare the resolution of the locations of the study and datapoint. If they are the same resolution, say ‘The score is 1’. If they are within one level of resolution, say ‘The score is 2’. If they are within two levels of resolution, say ‘The score is 3’. If they are outside two levels of resolution, say ‘The score is 4’. If the location of the data is unknown, say ‘The score is 5’.

In all responses, explain the rationale.

Prompt 3

I would like you to evaluate the quality of a datapoint that will be used in a study. You will do this by comparing the resolution of the location of the datapoint against the resolution of the location of the study. The goal is not to achieve any particular resolution, only to determine the resolutions, then compare them.

First, ask me for the location of the study. I will respond with a location.

Next, ask me for the location of the datapoint. I will respond with a location.

Determine the level of resolution of the location of the study and location of the datapoint.

The levels of resolution are as follows:

A is Global, aka ‘GLO’, worldwide, or any resolution more general than continental

B is Continental

C is Sub-region

D is National

E is Province/State/Region

F is County/City

G is Site-specific, e.g. an address

Compare the resolution of the locations of the study and datapoint. If they are the same resolution, say ‘The score is 1’.

If they are within one level of resolution, say ‘The score is 2’.

If they are within two levels of resolution, say ‘The score is 3’.

If they are outside two levels of resolution, say ‘The score is 4’.

If the location of the data is unknown, say ‘The score is 5’.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

- Algren M, Fisher W, Landis AE (2021) Chapter 8 - machine learning in life cycle assessment, Data Science Applied to Sustainability Analysis. Elsevier, pp 167–190, ISBN 9780128179765. <https://doi.org/10.1016/B978-0-12-817976-5.00009-7>
- Chiang C-H, Lee H (2023) A closer look into using large language models for automatic evaluation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8928–8942, Singapore. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.599>
- de Jesus JO et al (2021) Integration of artificial intelligence and life cycle assessment methods. IOP Conf Ser: Mater Sci Eng. <https://doi.org/10.1088/1757-899X/1196/1/012028>
- Edelen A, Ingwersen W (2016) Guidance on data quality assessment for life cycle inventory data. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-16/096
- Edelen A, Ingwersen WW (2018) The creation, management, and use of data quality information for life cycle assessment. Int J Life Cycle Assess 23(4):759–772. <https://doi.org/10.1007/s11367-017-1348-1>
- ISO 14040:2006 (2016) Environmental management — life cycle assessment — requirements and guidelines
- Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023) G-EVAL: NLG Evaluation using GPT-4 with better human alignment. <https://doi.org/10.48550/arXiv.2303.16634>
- Reap J, Roman F, Duncan S et al (2008) A survey of unresolved problems in life cycle assessment. Int J Life Cycle Assess 13:374–388. <https://doi.org/10.1007/s11367-008-0009-9>
- Tamkin A, Brundage M, Clark J, Ganguli D (2021) Understanding the Capabilities, limitations, and societal impact of large language models. <https://doi.org/10.48550/arXiv.2102.02503>
- Weidema BP (1998) Multi-user test of the data quality matrix for product life cycle inventory data. Int J LCA 3:259–265. <https://doi.org/10.1007/BF02979832>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.