IV Congresso Brasileiro em Engenharia de Sistemas em Processos (PSE BR 2026)
03 a 05 de Agosto de 2026, Campina Grande, PB, Brasil

Use of Generative AI Tools for Evaluating Qualitative Indicators in Sustainability Analyses

Gabriel de Figueiredo da Costa[a,*], André Ferreira Young[a,b], Raquel Massad Cavalcante[a]

[a]*Federal University of Rio de Janeiro - UFRJ, School of Chemistry, Rio de Janeiro-RJ, Brasil*
[b]*Federal Fluminense University - UFF, Department of Chemical and Petroleum Engineering, Niterói-RJ, Brasil*
**gabriel.de.figueiredo@eq.ufrj.br*

ABSTRACT

Generative Artificial Intelligence models (GenAI) and Large Language Models (LLMs) are becoming increasingly accessible, allowing for aid in everyday life and also in specialized applications such as educational, industrial and engineering tasks. Parallel to that, quali-quantitative and multi-criteria methods have been used for a holistic assessment of engineering processes, specially regarding their sustainability, requiring a specialized, unbiased and inherently subjective analysis of such processes. In this work, we have submitted to five commercial LLM models (in the standard version, and in the "advanced" version) and to a panel of five human specialists the task of assessing the alkaline homogeneous transesterification of soybean oil using methanol, to produce biodiesel, regarding the 12 Principles of Green Chemistry. The preliminary results of our work point to the fact that, in the present state of the technology, AI models do not replace, but can possibly aid, in the assessment of sustainability of chemical processes made by specialized researchers. In this sense, LLMs can help by enumerating and compiling pros and cons for each criterion, which can then be assessed by specialists in terms of relevance, possibly influencing the scores.

Keywords: large language models, generative artificial intelligence, principles of green chemistry.

1. Introduction

Generative Artificial Intelligence models (GenAI) and Large Language Models (LLMs) are becoming increasingly accessible, allowing for aid in everyday life and also in specialized applications such as educational, industrial and engineering tasks (Chakraborty et al., 2025). Parallel to that, qualitative, quantitative and multi-criteria methods have been used for a holistic assessment of engineering processes, specially regarding their sustainability (Sorkhab et al., 2026).

The use of such methods provide different points of view in the assessment of a chemical process, delving into technical, economical, social and environmental themes, as well as regulatory aspects and safety issues, amongst others. Therefore, the application of sustainability assessments require a specialized, unbiased and inherently subjective analysis of such processes. The specialists who are involved in the assessment should not be biased by each other, and should be able to provide their ideas without limitations or boundaries, so that every possibility can be further detailed and applied (or not, if it proves to be unfeasible) regarding the improvement of the process.

Therefore, the objective of this work was to assess the ability of GenAI and LLMs in aiding specialists regarding the multi-criteria assessment of a chemical process (in this case, a biodiesel production plant), in order to analyze whether these models are able to provide new insights for the specialists, or to provide final results for the sustainability assessment which are comparable to the results expected from a panel of specialists.

2. Methodology

In this work, we have tested the application of five commercial LLM models (in the standard version, and in the "advanced" version) in order to assess the alkaline homogeneous transesterification of soybean oil using methanol, to produce biodiesel, regarding the 12 Principles of Green Chemistry (Anastas and Warner, 2000). The same process description was presented to five specialized human researchers, who answered a form regarding the same analysis.

2.1. Principles of Green Chemistry

Realização:

The 12 Principles of Green Chemistry have been defined by Anastas and Warner (2000) as the main aspects that must been taken into account in the design step of chemical facilities, so that the final process can minimize and mitigate the damage to the environment from the start, in contrast to such mitigation being taken into account in later stages of the development. When such efforts are present since the beginning of the chemical process design, specialists can make better decisions and choose better equipment, with a significantly lower cost (e.g. in purification steps and waste management) than in the case that such decisions are made after the core process is already defined.

The 12 Principles of Green Chemistry are the following, as defined by Anastas and Warner (2000) and Araújo *et al.* (2015): 1) Prevention of Waste; 2) Atom Economy; 3) Less Hazardous Chemical Syntheses; 4) Designing Safer Chemicals; 5) Safer Solvents and Auxiliaries; 6) Energy Efficiency; 7) Use of Renewable Feedstocks; 8) Reduce Derivatives; 9) Catalysis; 10) Design for Degradation; 11) Real-time Analysis for Pollution Prevention; 12) Inherently Safer Chemistry for Accident Prevention.

2.2. Definition of the case study

The first step of the work was the definition of the case study to be assessed by the LLMs and human researchers, as well as the prompt for LLMs and instructions to human researchers. For that, the following simplified description and instruction was provided:

> *"Soybean oil is reacted with methanol (6:1 molar ratio) in the presence of NaOH in order to generate biodiesel. Then, the product is washed with water to remove the glycerol, which is neutralized with HCl and distilled. Assess the process regarding the 12 Principles of Green Chemistry, with scores from 1 to 10 for each criterion, and justify".*

The prompt was defined in such a simplified fashion so that the process can be assessed in a conceptual manner, relying on the assumption of typical parameters, when not specified, so that both AI models and human researchers could fill the gaps when needed, considering similar processes in the industry. The use of a simplified definition also seeks to represent the first steps in the design of a chemical process, when engineers are collecting the first impressions on the process, and starting to define the basal aspects of the chemical plant, without further details for the complete facility (however, this is the step with more opportunity for changes and optimization).

2.3. AI models selected for the study

The standard versions of the following LLMs were used: ChatGPT (version GPT-5.4), DeepSeek (version 3.2), Grok (version 4.20 Beta), Claude (version 4.6 Opus) and Gemini (version 3.1); although Copilot is also mentioned in most works regarding LLMs, it was not used in this work due to the fact that it applies a previous ChatGPT engine (version GPT-5.2).

The "pro" versions and/or "advanced settings" for the same LLMs were used: ChatGPT Pro (version GPT-5.4 with "thinking mode" on), DeepSeek (version 3.2 with "specialist mode", "deep thinking mode", and "smart research" on), SuperGrok (version 4.20 Beta with "Expert" mode on), Claude Pro (version 4.6 Opus with "extended thinking" mode on) and Gemini Pro (version 3.1 with "thinking mode" on). It is worth noting that DeepSeek does not provide a Pro version, besides the advanced settings which are available for all users; and that all of the other models require a payment for the use of the respective "pro" versions.

2.4. Human researchers selected for the study

Regarding the answers from the human researchers, we have chosen 5 specialists to assess the base process, based on their knowledge about biodiesel production and/or sustainability assessments:

A. Specialist who is familiar with the Principles of Green Chemistry and with the biodiesel production process
B. Specialist who is familiar with the biodiesel production process, but not with the Principles of Green Chemistry
C. Specialist who is familiar with the Principles of Green Chemistry, but not with the biodiesel production process
D. Specialist who is familiar with sustainability analysis, but not with this particular methodology or process
E. Specialist who is familiar with chemical processes, but not with this particular methodology or process

The different levels of familiarity with the Principles of Green Chemistry methodology, and with the biodiesel production process via the methylic alkaline route with NaOH as a catalyst, were chosen in order to increase the variability of opinions and points of view across different assessments. A Google Forms questionnaire[1] was applied in order to collect their assessments. The results reported by LLMs and researchers were then compared, in order to see how reliable these generative AI tools would be for the task.

3. Results and Discussion

Based on the results from the five standard AI models, the average overall sustainability score for the case study process was 6.52 (varying from 5.42 to 7.25 between the LLMs). The "advanced" versions of the same LLMs resulted in an average overall sustainability score of 6.23 (varying from 5.75 to 6.83 between the LLMs). Regarding the assessment by the five human researchers, the average overall sustainability score for the case study process was 6.27 (varying from 4.50 to 7.08 between researchers). It is possible to observe that the average scores are similar over the 3 groups of assessments.

When data from the LLMs was compared with the results from human researchers, we found out that the deviation between human and AI assessments was up to 20% for each criterion, which was comparable to the deviation between humans and between AIs regarding the same criterion. The comparison between the standard and the "pro/advanced" versions of each LLM resulted in minimum changes, ranging from -13% to +9% for each model.

3.1. Results for the standard versions of AI models

The use of the standard version of the AI models resulted in a relatively narrow range of scores for the 12 Principles of Green Chemistry, as shown in Figure 1.

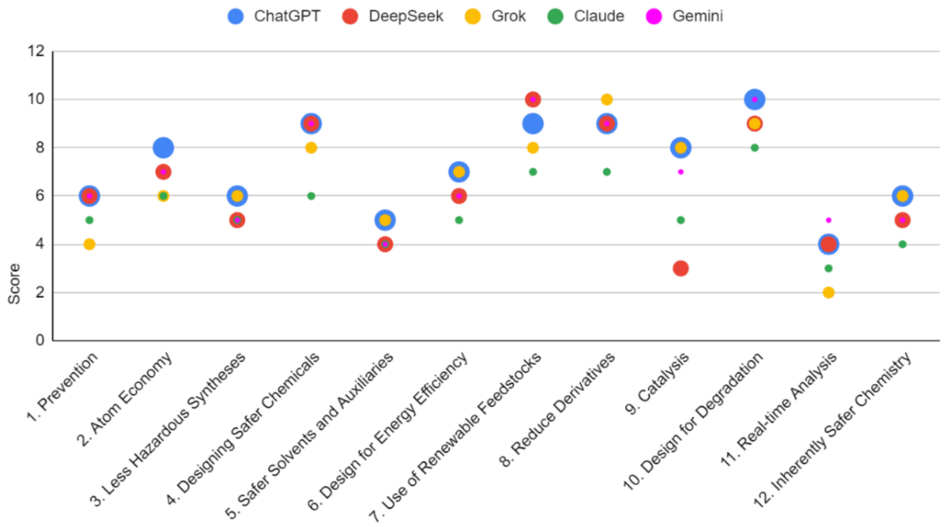

Figure 1. Results for the standard versions of AI models

Apart from the "9. Catalysis" criterion, with a variation of 5 points between the lowest and highest scores, the other 11 criteria presented a variation of 1 to 3 points between LLMs. All of the models highlighted that, despite being used in catalytic amounts, NaOH is consumed in side-reactions or neutralized after the reaction, not being recoverable as a catalyst; however, the score penalties provided by each LLM regarding this fact were considerably different.

3.2. Results for the "advanced" versions of AI models

The use of the "advanced" version of the AI models resulted in a somewhat similar range of scores for the 12 Principles of Green Chemistry when compared to their respective standard versions, as shown in Figure 2.

[1] Link for the questionnaire: https://docs.google.com/forms/d/e/1FAIpQLSf0d4F-JXCd4omjLxxr2ljA7RJ4Fjt8t85X8xeziSpgJbkYKA/viewform

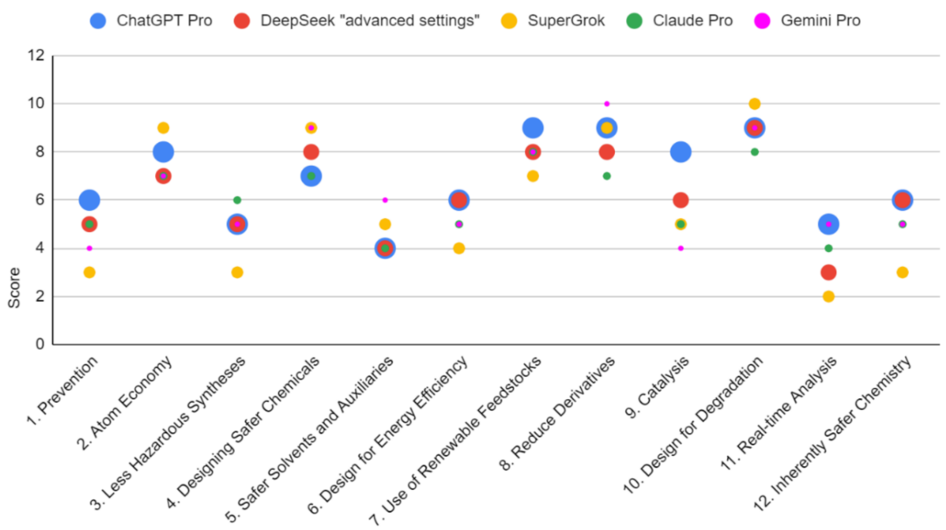

Figure 2. Results for the “advanced” versions of AI models

Once again, the “9. Catalysis” criterion presented the most important variation between LLMs, with 4 points of difference between the lowest and highest scores, while other criteria presented 2 to 3 points of difference. The reason for that was the same as before: the penalties for NaOH not being reusable were different between models. There has not been observed any improvements in assessment quality when “pro” versions and “advanced settings” for AI models were used, in comparison to the standard (and free) versions of the same models. The justifications for each criterion were roughly the same for each LLM in the “standard” and “advanced” versions, which also confirms the lack of significative improvement when the latter were used.

3.3. Results for human researchers

The results for the sustainability analysis performed by the 5 human researchers presented a broad distribution of scores, as can be observed in Figure 3.

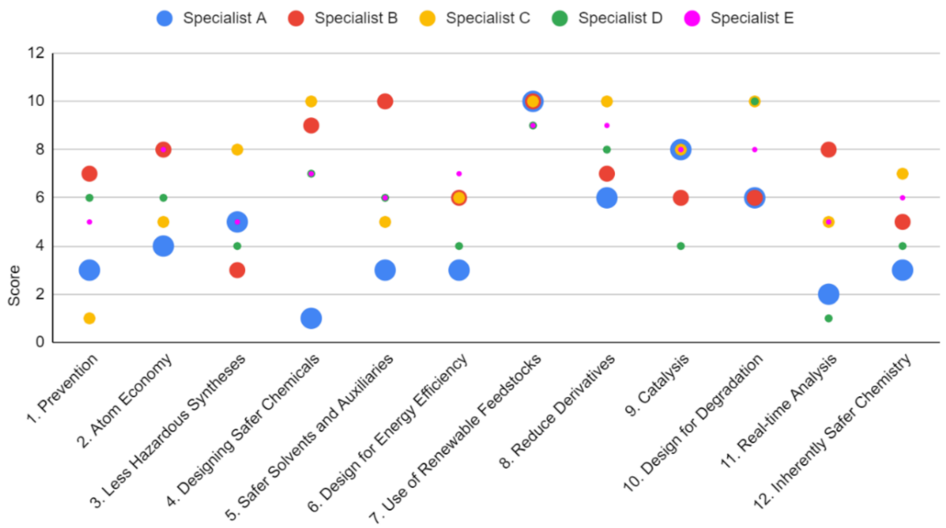

Figure 3. Results for human researchers

The criteria with a higher scattering of scores were “4. Designing Safer Chemicals” (with a difference of 9 points between the lowest and highest scores), “5. Safer Solvents and Auxiliaries” (with a difference of 7 points between the lowest and highest scores) and "1. Prevention" (with a difference of 6 points between the lowest and highest scores). All of the other criteria presented a difference of 1 to 5 points between the lowest and highest scores.

Regarding the criterion "4. Designing Safer Chemicals", Specialist "A" justifies that the process uses NaOH and methanol, which are not environmentally acceptable and are toxic to humans (resulting in a score of 1); while Specialist "C" observes that no toxic products are formed, already considering that methanol is recycled and NaOH is neutralized (resulting in a score of 10). When the criterion "5. Safer Solvents and Auxiliaries" is taken into account, Specialist "B" observes that the only solvent which is used is water (resulting in a score of 10), and only in the separation step, while Specialist "A" observes that an acid is used as an auxiliary substance to neutralize NaOH (resulting in a score of 3). Regarding the criterion "1. Prevention", Specialist "C" considers that no waste prevention efforts are being made (resulting in a score of 1), while Specialist "B" considers that the waste which is generated in the process can still be managed (resulting in a score of 7).

This shows that, even considering the same process, the differences between the subjective analysis carried out by specialists can result in different scores for the same criteria. Some specialists can focus on certain substances and their characteristics, or focus on a different "part" of the criterion, while others already take into account the starting and final forms of each compound that leaves the process (not focusing on intermediary streams). There are also some situations where there are several "pros" and "cons" for the same criterion, and some specialists can choose to apply higher penalties for the "cons" when defining their respective scores.

This highlights the diversity of points of view and the richness of depth that arise when such an assessment is carried out by a diverse set of human researchers, in a panel of specialists, which is precisely the objective of such quali-quantitative and multi-criteria assessments.

3.4. Comparison between assessments and insights for each criterion

Overall, it can be observed that LLM assessments and human assessments were similar, as observed in Figure 4, which presents the comparison between the ranges of results of the assessments carried out by the standard and "advanced" versions of the AI models, as well as the results for human researchers.

Figure 4. Comparison of results

It can be observed that the minimum scores for human researchers were lower than the minimum scores for LLMs in 8 of the 12 criteria, equal in 3, and higher in only 1, which shows that human researchers were more rigid in their assessments, and have taken into account different pros and cons, and/or higher weights for the different pros and cons which were considered. The maximum score for criterion "4. Designing Safer Chemicals" was significantly lower for human researchers, due to a more detailed observation of unsafe and toxic solvents and streams within the process. On the other hand, the maximum score for criterion "5. Use of Renewable Feedstocks" was higher for human researchers, due to the focus on vegetable oil being the main feedstock for the process, and considering that unreacted methanol would be recycled.

When the justifications for the scores given by LLMs are taken into account, it is possible to notice that the insights brought by each model are roughly the same for each criterion. This is probably due to the same data-sets, texts, documents, books, etc., being used by the models, which resulted in a similar "thinking sequence" that provided similar insights. The collective answers from the panel of specialists was able to arrive at the same insights in an independent fashion, while also adding new perspectives into account for some of the criteria.

4. Conclusions

The results of our work point to the fact that, in the present state of the technology, AI models do not replace, but can possibly aid, in the assessment of sustainability of chemical processes made by specialized researchers. In this sense, LLMs can help by enumerating and compiling pros and cons for each criterion, which can then be assessed by specialists in terms of relevance, possibly influencing the scores.

However, the assessment performed by LLMs were significantly standardized (probably due to the same data-sets being used for the training of the models), with a really low difference in most scores, which shows that Generative AI still faces a deficiency in providing answers with a broad diversity of points of view or a high variability in the assessment procedures.

Besides that, we concluded that, for the present task, the use of "advanced settings" or "pro" versions of different LLMs was not able to improve the assessment results, neither the diversity of insights provided, when compared to the standard (and free) versions of the same LLMs.

Acknowledgements: Financial support of the Human Resources Program of the Brazilian National Agency of Petroleum, Natural Gas and Biofuels - PRH-ANP, through the program PRH17/UFRJ (FAPESP grant number 2025/02818-8). The use of LLMs and GenAI was carried out solely regarding the object of the research, but not for aiding in writing of the present article.

References

Anastas, Paul T, and John C Warner, Green Chemistry: Theory and Practice (Oxford, 2000; online edn, Oxford Academic, 31 Oct. 2023), https://doi.org/10.1093/oso/9780198506980.001.0001

Araújo, O.; Medeiros, J.L.; Yokoyama, L.; and Morgado, C. Metrics for Sustainability Analysis of Post-Combustion Abatement of CO_2 Emissions: Microalgae Mediated Routes and CCS (Carbon Capture and Storage). Energy 92 (2015) 556–568, doi:10.1016/j.energy.2015.03.116

Arijit Chakraborty, Naz Pinar Taskiran, Rishab Kottooru, Vipul Mann, Venkat Venkatasubramanian, Building hybrid AI models in chemical engineering: A tutorial review, Computers & Chemical Engineering, Volume 201, 2025, 109236, ISSN 0098-1354, https://doi.org/10.1016/j.compchemeng.2025.109236

Mohammadali Babanejadi Sorkhab, Archishman Bose, Jerry D. Murphy, Richard O'Shea, A critical review of the application of multi-criteria decision-making in the sustainability assessment of biorefineries, Biomass and Bioenergy, Volume 210, 2026, 109036, ISSN 0961-9534, https://doi.org/10.1016/j.biombioe.2026.109036