The Accuracy, Robustness, and Readability of LLM-Generated Sustainability-Related Word Definitions

Anonymous Author

Abstract

A common language with shared standard definitions is essential for effective climate conversations. However, there is concern that LLMs may misrepresent and/or diversify climate-related terms. We compare 305 official IPCC glossary definitions with those generated by OpenAI's GPT-4o-mini and investigate their adherence, robustness, and readability using a combination of SBERT sentence embeddings and statistical measures. The LLM definitions received average adherence and robustness scores of 0.58 ± 0.15 and 0.96 ± 0.02, respectively. Both sustainability-related terminologies remain challenging to read, with model-generated definitions varying mainly among words with multiple or ambiguous definitions. Thus, the results highlight the potential of LLMs to support environmental discourse while emphasizing the need to align model outputs with established terminology for clarity and consistency. .

1 Introduction

Large language models (LLMs) have proven effective in a range of tasks, such as analyzing climate-related texts (Callaghan et al., 2021) and explaining sustainability reports (Ni et al., 2023). However, as citizens and politicians turn to LLMs for information and inspiration, there is concern that these probabilistic models fail to consistently convey the specificity and accuracy required to discuss climate change. For example, agreeing to a standard set of definitions is essential to achieve common ground in the climate debate (Rev, 2007). However, streamlining language around climate is already challenging. For instance, Con (2017) showed that among 114 different definitions for "circular economy," most failed to convey all nuances of the concept. Thus, this can lead to inconsistencies in research and policy-making.

To address this issue, the Interdisciplinary Panel on Climate Change (IPCC) and the United Nations (UN) maintain the online glossaries IPCC Glossary (IPCC, 2019a,b, 2018), and UNTERM (United Nations , UN). Although LLMs have access to these repositories during training, they are not constrained to them during inference. Therefore, LLMs could further diversify and confuse these terms. As more people rely on LLMs, it is of special interest to study how LLM-generated explanations adhere to the official definitions, how robust the completions are, and what lessons we should keep in mind when using these models at ever higher levels of climate discourse. Motivated by this, we analyze the adherence, robustness, and readability of word definitions generated by GPT-4o-mini compared to official IPCC definitions.

2 Related Work

Pham et al. (2024) showed that word definitions of English words given by OpenAI LLM agree well with three popular English dictionaries. However, current LLM performance is mainly dependent on prompt engineering. Atil et al. (2024) examined LLM stability and showed that LLM variation arises even given the same input and parameters, depends on the task, and is not normally distributed.

Studies show that sustainability literature can be complex to read (Smeuninx et al., 2020). This complexity challenges the accessibility and transparency of sustainability debates and reporting. Studies spanning the sustainability to medical domains use LLMs to simplify these texts and make them interactive (Ni et al., 2023; Yao et al., 2024).

3 Methodology

We present a framework for assessing the adherence and robustness of LLM sustainability word definitions. Specifically, given a term, we let OpenAI's GPT-4o-mini generate five definitions using five prompt templates (25 completions per term). Then, we use SBERT sentence embeddings to compute the sentence similarity between the official and generated definitions (adherence), as well as the similarity between the generated definitions for a given term and prompt template (robustness). Thus, we define adherence and robustness for each term as follows

$$\text{adherence} = \frac{1}{n}\sum_{k=1}^{n} \text{sim}(D, M_k)$$

$$\text{robustness} = \frac{1}{\text{comb(n)}}\sum_{p=1}^{n}\sum_{q=k+1}^{n} \text{sim}(M_p, M_q)$$

where D is the IPCC glossary definition, M_k is the k'th model definition completion across all prompts, comb(n) the number of unique pairwise combinations using n terms, and sim(A, B) the cosine distance between the SBERT sentence embeddings of texts A and B. Intuitively, adherence measures how similar model completions are to glossary definitions, while robustness measures the consistency of model completions. Finally, we used GPT-4o-mini to classify the IPCC definition and one model definition completion for each term into zero or more of the categories "environmental," "social," and "economic" for comparison.

Dataset collection: We use Selenium WebBrowser to scrape all terms and definitions from the IPCC glossary website as of December, 2024. In total, the glossary contained 911 terms. We limit the terms to those mentioned in the IPCC 2022 Special Report on Climate Change and Land Annex I Glossary (IPCC, 2019c), for a total of 305 terms. Finally, we use only the first sentence of each definition and replace all cross-references (such as "See Pathways") with the cited term.

Prompts: We prompt ChatGPT with "Write 4 versions of asking 'Define "sustainability" in one sentence.'" resulting in the following list of 5 prompt templates:

- Define "[TERM]" in one sentence.
- How would you define "[TERM]" in a single sentence?
- Can you describe "[TERM]" in just one sentence?
- What is your one-sentence definition of "[TERM]"?
- In one sentence, what does "[TERM]" mean to you?

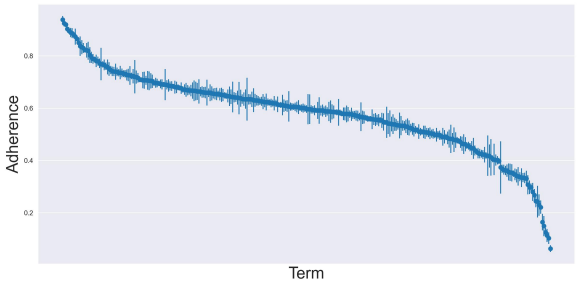

Figure 1: Distribution of SBERT adherence scores between LLM and official IPCC word definitions.

Readability analysis: We use the Python library Readability and compute five readability metrics (Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid (Kincaid et al., 1975), Gunning-Fog (Gunning, 1952), Dale-Chall (Dale and Chall, 1948), and SMOG (McLaughlin, 1969)) for the official definitions and model completions, respectively. The metrics require at least 100 words and are not directly applicable to single sentences. Therefore, we use bootstrapping with 10,000 iterations to create longer text samples by sampling 50 random definitions with replacement and assessing the Readability of these excerpts.

4 Experimental Results

4.1 Adherence

The average SBERT similarity scores between all terms and their corresponding official IPCC definitions are shown in Figure 1. There were 305 terms in total, with an average adherence of 0.58 ± 0.15 (min 0.06, max 0.94). The 10 terms with the highest and lowest adherence scores are shown in Figure 2. Table 4.1 shows the holistic category count distributions between the official definitions and the model completions. Analysis of the number of times the model completion missed a category found in the official definition shows that the model missed the economic, social, and environmental categories 25, 18, and 32 times, respectively.

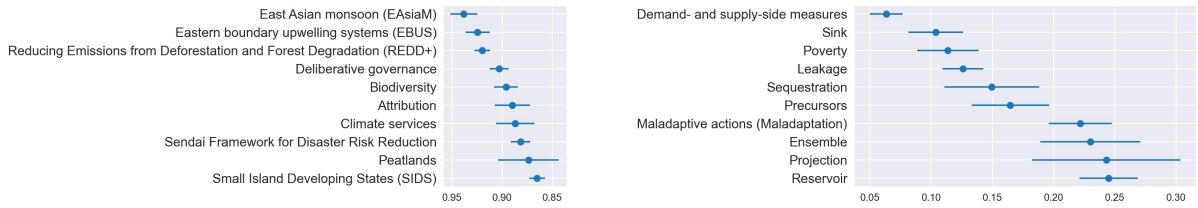

Figure 2: Top (left) and bottom (right) 10 terms with the highest and lowest semantic similarity between LLM-generated and official IPCC definitions.

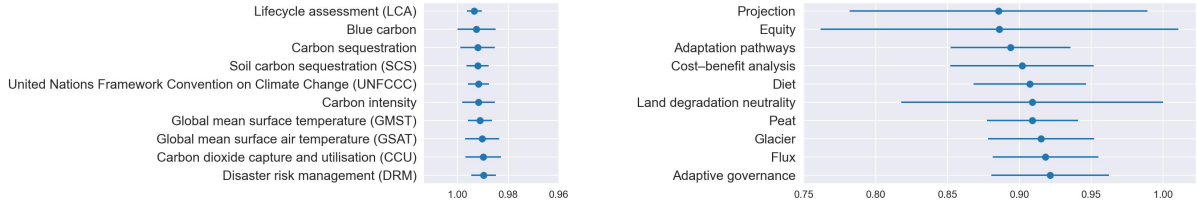

Figure 3: Top (left) and bottom (right) 10 terms with the highest and lowest semantic similarity between model completions over multiple iterations and prompts.

Category	Completions	Definitions
None	7.9%	6.0%
Environmental	46.5%	52.1%
Social	22.2%	21.5%
Economic	23.3%	20.4%

Table 1: Category classification of IPCC and model definitions.

4.2 Robustness

Figure 3 shows the 10 terms with the highest and lowest robustness scores. The average SBERT similarity score between all pairwise comparisons of term completions was 0.96 ± 0.02 (min 0.89, max 0.99), with no statistical difference between the prompt templates.

4.3 Readability

Table 4.3 shows the readability scores for five standard readability metrics. The "Syllable Score" is computed as the average total syllables to definition length. For Flesch, lower scores are more difficult to read, while for Flesh-Kincaid, Gunning-Fog, Dale-Chall, and SMOG, higher scores are more complex. The scores indicate that both IPCC- and model-generated definitions are at the reading level of college graduates, with the generated definitions being slightly more complex.

Readability	Completions	Definitions
Sentence Length	34.3 ± 51.5	30.2 ± 295.5
Syllable Score	1.8 ± 0.0	1.6 ± 0.1
Flesch	7.5 ± 7.6	22.6 ± 12.5
Flesch-Kincaid	19.4 ± 0.2	16.3 ± 0.7
Gunning-Fog	22.4 ± 0.3	19.6 ± 0.8
Dale-Chall	12.5 ± 0.0	11.7 ± 0.1
SMOG	19.8 ± 0.3	17.8 ± 0.7

Table 2: Standard readability metrics for definitions generated by the LLM and the official IPCC scores.

5 Discussion

The adherence scores indicate that the LLM generally captures the core semantic meaning of official definitions. However, there is considerable variation in similarity scores across terms. As expected, terms with multiple meanings, particularly those whose most common form is not environmentally related, tend to have lower adherence scores. Pairwise text classification comparisons of official and model-generated definitions further corroborate this trend, where the environmental category emerges as the one most frequently missed by the model. These results suggest that the model does not, by default, always include the environmental perspective, highlighting a potential gap in its ability to portray domain-specific language accurately.

Regarding robustness, the five prompt templates

Term	Definition	Define "[TERM]"...	In one sentence...
Carbon sequestration	The process of storing carbon in a carbon pool.	Carbon sequestration is the process of capturing and storing atmospheric carbon dioxide (CO2) to mitigate climate change by reducing the amount of CO2 in the atmosphere.	Carbon sequestration refers to the process of capturing and storing atmospheric carbon dioxide to mitigate climate change and reduce global warming.
Sink	Any process, activity or mechanism which removes a greenhouse gas, an aerosol or a precursor of a greenhouse gas from the atmosphere.	A sink is a fixed basin used for holding water, typically equipped with a drain, used for washing hands, dishes, or other items.	"Sink" signifies the act of descending or becoming submerged, often associated with the loss of buoyancy or support.
Equity	The principle of being fair and impartial, and a basis for understanding how the impacts and responses to climate change, including costs and benefits, are distributed in and by society in more or less equal ways.	Equity refers to the concept of fairness and justice in the allocation of resources, opportunities, and treatment, often aimed at addressing imbalances and ensuring that everyone has access to the same rights and benefits.	Equity means creating fair opportunities and access to resources for all individuals, regardless of their background or circumstances, to ensure that everyone can achieve their potential.

Table 3: Case studies of select terms with their official definitions and model completions using various prompt templates.

tested did not result in significant variations in generated model definitions. However, there was a notable variability in terms of vaguer definitions. For instance, "equity" displayed many definitions, reflecting its complex and multi-faceted meanings. This ambiguity aligns with discussions in recent sustainability reports, such as the UN's 2024 Emissions Gap Report, which dedicates an entire section to discuss different equity models (Environment, 2024). Thus, the robustness score can help target terms needing further standardization.

Regarding readability, both the IPCC and model definitions scored poorly across all five readability metrics, consistent with previous studies indicating that sustainability texts are complex and inaccessible to most readers. Future work could explore ways to improve accessibility by using LLMs to simplify language without compromising accuracy, potentially incorporating official glossaries as a port of an in-context learning approach. One challenge will be balancing simplicity with accuracy. Adherence scores could offer a helpful framework for evaluating and refining these model outputs since they do not rely on exact sentence matching but instead on semantic meaning. Studies across more models and languages would further inform how LLMs represent sustainability.

6 Conclusion

This study provides a comprehensive framework for assessing the adherence, robustness, and readability of LLM-generated definitions of sustainability terms compared to official glossaries. While the LLM captures the semantic meaning of most terms, there is significant variation, particularly for terms with multiple meanings or ambiguous definitions. In addition, IPCC and model definitions show low readability, highlighting the need for further work to simplify sustainability-related language without sacrificing accuracy. These findings highlight the potential of LLMs to support the environmental conversation but also underscore the importance of carefully aligning model outputs with established terminology to ensure clarity and consistency.

References

2007. https://doi.org/10.1016/j.jclepro.2006.12.006 Review of sustainability terms and their definitions. *Journal of Cleaner Production*, 15(18):1875–1885. Publisher: Elsevier.

2017. https://doi.org/10.1016/j.resconrec.2017.09.005 Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, Conservation and Recycling*, 127:221–232. Publisher: Elsevier.

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. https://doi.org/10.48550/arXiv.2408.04667 LLM Stability: A detailed analysis with some surprises. ArXiv:2408.04667 [cs] version: 1.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R. Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J. Brecha, Michael Hegarty, Chelsea Jones, Kaylin Lee, Agathe Lucas, Nicole van Maanen, Inga Menke, Peter Pfleiderer, Burcu Yesil, and Jan C. Minx. 2021. https://doi.org/10.1038/s41558-021-01168-6 Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, 11(11):966–972.

Edgar Dale and Jeanne S. Chall. 1948. http://www.jstor.org/stable/1473169 A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

UN Environment. 2024. https://www.unep.org/resources/emissions-gap-report-2024 Emissions gap report 2024.

Rudolf Flesch. 1948. https://doi.org/10.1037/h0057532 A new readability yardstick. *The Journal of Applied Psychology*, 32(3):221–233.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

IPCC. 2018. Annex i: Glossary. In J.B.R. Matthews, editor, *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. IPCC. In Press.

IPCC. 2019a. Annex i: Glossary. In N.M. Weyer, editor, *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. IPCC. In Press.

IPCC. 2019b. Annex i: Glossary. In R. van Diemen, editor, *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. IPCC. In press.

IPCC. 2019c. https://doi.org/10.1017/9781009157988.010 Annex i: Glossary. In R. van Diemen, editor, *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. Cambridge University Press. In Press.

J Kincaid, Robert Fishburne, L Richard, Brad Rogers, and Chissom. 1975. *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel 1-1-1975*. Institute for Simulation and Training.

G. H. McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. https://doi.org/10.48550/arXiv.2307.15770 CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. ArXiv:2307.15770 [cs].

Bach Pham, JuiHsuan Wong, Samuel Kim, Yunting Yin, and Steven Skiena. 2024. https://doi.org/10.48550/arXiv.2311.06362 Word Definitions from Large Language Models. ArXiv:2311.06362 [cs].

Nils Smeuninx, Bernard De Clerck, and Walter Aerts. 2020. https://doi.org/10.1177/2329488416675456 Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP. *International Journal of Business Communication*, 57(1):52–85. Publisher: SAGE Publications Inc.

Department for General Assembly United Nations (UN) and Conference Management. 2024. https://unterm.un.org/unterm2/en/ [link].

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, README annotation team, and Hong Yu. 2024. https://doi.org/10.48550/arXiv.2312.15561 README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP. ArXiv:2312.15561 [cs].