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

Alice Heiman
Stanford University
aheiman@stanford.edu

Abstract

A common language with standardized definitions is crucial for effective climate discussions. However, concerns exist about LLMs misrepresenting climate terms. We compared 300 official IPCC glossary definitions with those generated by GPT-4o-mini, Llama3.1 8B, and Mistral 7B, analyzing adherence, robustness, and readability using SBERT sentence embeddings. The LLMs scored an average adherence of $0.57 - 0.59 \pm 0.15$, and their definitions proved harder to read than the originals. Model-generated definitions vary mainly among words with multiple or ambiguous definitions, showing the potential to highlight terms that need standardization. The results show how LLMs could 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 (Peter Glavič, 2007). However, streamlining language around climate is already challenging. For instance, Julian Kirchherr (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 (UN, 2024a). 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 one closed-source and two open-source models compared to official IPCC definitions.

2 Related Work

Pham et al. (2024) showed that word definitions of English words given by OpenAI LLMs 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 even the same input and parameters can result in variation, which is task-dependent and not normally distributed.

Studies show that sustainability literature can be complex to read (Smeuninx et al., 2020; Barkemeyer et al., 2016). 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 an LLM generate five definitions for each of the 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:

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

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

where D is the IPCC glossary definition, M_k is the k'th model definition completion across all prompts, cmb(n) the number of unique pairwise combinations using n terms, and sim(A, B) the cosine distance between the SBERT sentence embeddings of the texts A and B. Intuitively, adherence measures how similar model completions are to glossary definitions, while robustness measures the consistency of model completions.

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 with an overlap in the IPCC 2022 Special Report on Climate Change and Land Annex I Glossary (IPCC, 2022), and get a subset of 300 terms. Finally, we use only the first sentence of each definition and replace all cross-references (such as "See Pathways") with the cited term.

Models: We use three different models in the experiments. We use GPT-4o-mini as our closed source model, and Meta-Llama-3.1-8B-Instruct (Meta, 2024) and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) as our open source models. We use the default parameter settings for all models.

Prompts: We prompt ChatGPT with "Write 4 versions of asking 'Define "[TERM]" 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?

Readability analysis: We use the Python library Readability (Py-Readbility-Metrics, 2019) to compute the two readability metrics Flesch-Kincaid (Kincaid et al., 1975) and Gunning-Fog (Gunning, 1952) for the official definitions and model completions, respectively. Higher Flesh-Kincaid and Gunning-Fog scores indicate more complex material. The metrics require at least 100 words and are not directly applicable to single sentences. Therefore, we use bootstrapping with 1,000 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. The terms vary greatly, ranging from an adherence score of 0.06 to 0.94. Table 1 shows that all three models received similar results, with average adherence scores of $0.57 - 0.59 \pm 0.15$. The terms with the highest and lowest adherence scores are shown in Table 2. Notably, there is a significant overlap between models, with the term "East Asian monsoon (EAsiaM)" scoring highest and "Demand- and supply-side measures" scoring lowest.

4.2 Robustness

Table 1 includes the robustness scores across all term completions. The average robustness falls between $0.96 - 1.00 \pm 0.02$ (min 0.89, max 1.00), with no statistical difference between the prompt templates. Some terms produce notable variations, however, in definitions across prompt templates, as listed in Table 3. For instance, GPT-4o-mini's

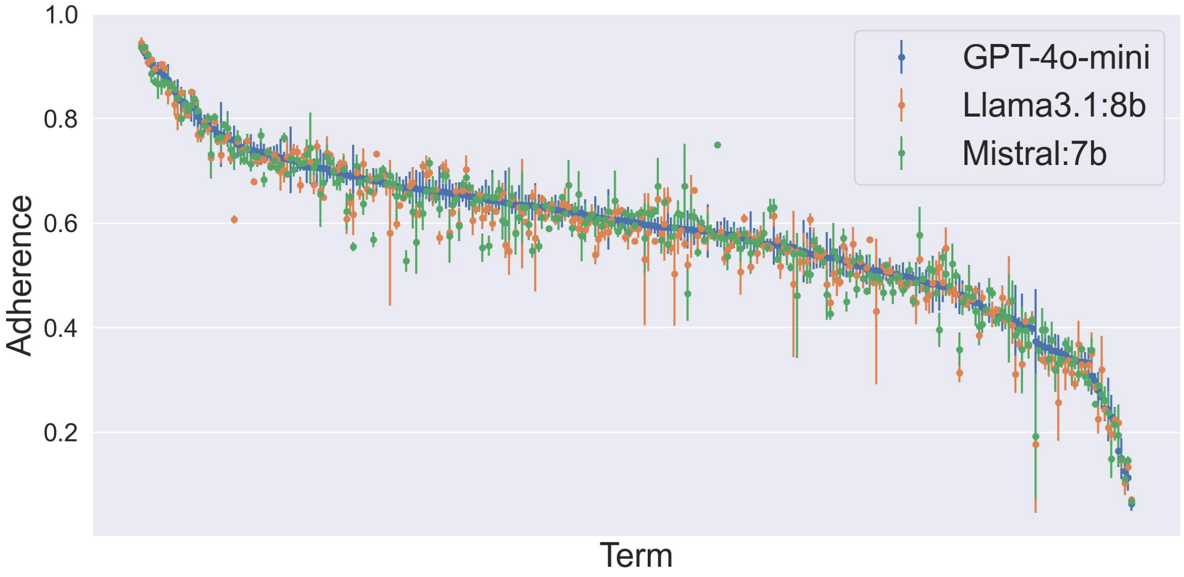

Figure 1: Distribution of SBERT adherence scores between LLM and official IPCC word definitions.

Model	**Adherence**	**Robustness**	**Num Words**	**Gunning Fog**	**Flesch-Kincaid**
GPT-4o-mini	0.59 ± 0.15	0.96 ± 0.02	34.3 ± 51.5	22.4 ± 0.3	19.4 ± 0.2
Llama 3.1 8B	0.57 ± 0.15	1.00 ± 0.01	39.7 ± 61.4	22.9 ± 0.3	19.9 ± 0.2
Mistral 7B	0.58 ± 0.15	1.00 ± 0.00	33.6 ± 69.5	20.8 ± 0.3	18.1 ± 0.2
Definitions	-	-	30.2 ± 295.5	19.7 ± 0.8	16.3 ± 0.7

Table 1: Adherence, robustness, and readability scores for various LLMs.

definition of “Projection” spanned the psychological (“Projection is a psychological defense mechanism...”), mathematical (“Projection is the process of transferring an image, shape, or data representation...”), and environmental (“Projection” refers to the process of estimating or forecasting future events”) topics. This is to be expected, however, since the prompt did not constrain the model to a particular context. On the other hand, prompting without context gives a hint into potential ambiguities when adapting terms such as “Equity”, “Exposure”, and “Adaptation pathways” into the climate debate.

4.3 Readability

Table 1 shows the definitions’ average lengths and readability scores. The scores indicate that both IPCC- and model-generated definitions are at the reading level of college graduates. Nevertheless, the IPCC definitions are significantly less complex according to both readability metrics and use fewer words than all model-generated definitions.

4.4 Ablation Case Studies

We perform three additional ablation studies using Llama3.1 8B, using the following prompts:

- **IPCC**: ‘Define “[TERM]” in one sentence. Adhere to the official Intergovernmental Panel on Climate Change (IPCC) glossary without citing it.’
- **Readable**: ‘Define “[TERM]” in one sentence. You must also make the definition understandable by a 10-year old.’
- **IPCC+Readable**: ‘Define “[TERM]” in one sentence. Adhere to the official Intergovernmental Panel on Climate Change (IPCC) glossary without citing it. You must also make the definition understandable by a 10-year old.’

Table 4 shows the adherence and readability scores using the ablation prompt templates. Notably, the adherence score remains roughly unchanged using the IPCC-specific prompt. Instead, the readability prompt seems to have a greater effect, decreasing the Flesch-Kincaid score from 19.9 ± 0.2 to

Model	Highest Adherence Terms	Lowest Adherence Terms
GPT-4o-mini	1. East Asian monsoon (EAsiaM) 2. Eastern boundary upwelling systems (EBUS) 3. Reducing Emissions from Deforestation and Forest Degradation (REDD+)	1. Demand- and supply-side measures 2. Poverty 3. Leakage
Llama3.1:8b	1. East Asian monsoon (EAsiaM) 2. Eastern boundary upwelling systems (EBUS) 3. Deliberative governance	1. Demand- and supply-side measures 2. Leakage 3. Poverty
Mistral:7b	1. Eastern boundary upwelling systems (EBUS) 2. East Asian monsoon (EAsiaM) 3. Reducing Emissions from Deforestation and Forest Degradation (REDD+)	1. Demand- and supply-side measures 2. Leakage 3. Poverty

Table 2: Terms with the highest and lowest adherence scores between generated and official definitions.

Model	Lowest Robustness Scores
GPT-4o-mini	1. Projection 2. Equity 3. Adaptation pathways
Llama3.1:8b	1. Exposure 2. Glacier 3. Forest
Mistral:7b	1. Sea ice 2. Global mean surface air temperature (GSAT) 3. Ensemble

Table 3: Terms with the lowest robustness score between the generated and official definitions.

16.4 ± 0.02. Although the prompt specified language for a 10-year hold, the Flesh-Kincaid score still corresponds to a college reading level. The relatively high score may partly be explained by the increased sentence length in the LLM's attempt to elaborate and explain parts of the concepts. Table 5 shows case studies for the term "Radiative Forcing" for the official IPCC definition and ablations comparing the definitions generated from different prompts.

5 Discussion

The adherence scores suggest that all LLMs generally capture the core semantic meanings of official definitions. Intriguingly, all LLMs achieved similar average adherence scores and had many common outlier terms. This similarity may be due to the models being trained using similar methods and on roughly the same training data. Notably, the adherence score did not significantly improve when we explicitly prompted the model for IPCC definitions. These results imply that providing a climate context may not automatically align language models for a given terminology group. The models do not have perfect recall of definitions; instead, they operate based on probability distributions. Therefore, it is advisable to include the exact definitions in the prompts or LLM systems to ensure they are readily available for reference.

Regarding robustness, the five prompt templates tested did not result in significant variations in generated model definitions. However, there was a notable variability among several terms. As anticipated, the terms with lower robustness scores tend to have multiple meanings, such as "Projection", "Exposure", and "Equity". 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 (UN, 2024b). Thus, the robustness score can help target terms needing further standardization. However, we must also note that robustness is very dependent on the temperature settings of the models. In this paper, we use the default temperature for the models. However, model parameters play a significant role in the consistency and variability of model outputs. These variations could impact how the model presents terms to different users across time.

In terms of readability, both the IPCC and model definitions scored poorly across both readability metrics. This finding is consistent with previous studies, which suggest that sustainability texts are inaccessible to most readers. Notably, all model completions consistently received more complex readability scores than the already

Ablation	**Adherence**	**Num Words**	**Gunning Fog**	**Flesch-Kincaid**
LLama+IPCC	0.60 ± 0.14	38.1 ± 45.1	22.5 ± 0.3	19.4 ± 0.2
Llama+Readable	0.55 ± 0.16	43.4 ± 58.6	19.2 ± 0.2	16.4 ± 0.2
Llama+IPCC+Readable	0.58 ± 0.14	42.6 ± 47.9	20.2 ± 0.2	17.3 ± 0.1

Table 4: Adherence and readability scores for different ablation studies using Llama3.1 8B.

Ablation	**Radiative Forcing Definition**
Official Definition	The change in the net, downward minus upward, radiative flux (expressed in W m–2) due to a change in an external driver of climate change, such as a change in the concentration of carbon dioxide (CO2), the concentration of volcanic aerosols or in the output of the Sun.
Llama+IPCC	Radiative forcing is the change in the balance between the incoming and outgoing radiation at the top of the atmosphere, usually expressed in watts per square meter (W/m²), resulting from a specific perturbation, such as the increase in greenhouse gases.
Llama+Readable	Radiative forcing is the amount of extra energy that the Earth's atmosphere receives from the sun, usually because of something like pollution or a change in the atmosphere, which can make the planet warmer or cooler.
LLama+IPCC-Readable	Radiative forcing is the change in the amount of energy that the Earth's atmosphere receives from the sun, usually caused by human activities or natural changes, which can make the planet warmer or cooler.

Table 5: Case Study: Ablation study using LLama 3.1 8B to define "Radiative Forcing" using three different prompting strategies. "IPCC" explicitly asks for a definition in line with the official definition, "Readable" for an easily understandable description, and "IPCC+Readable" combines the two.

intricate official definitions. This discrepancy may partly be attributed to the longer model responses. Moreover, increasing the readability proved difficult. Although the model used more straightforward terminology, prompting for readability made the model more verbose. Additionally, the readability metrics were not initially designed for single sentences, suggesting that using multiple sentences may yield a more representative assessment.

Future work could explore ways to improve accessibility by using LLMs to simplify language without compromising accuracy and incorporating relevant official glossaries as part 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 rely not on exact sentence matching but 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 LLMs capture 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. Moreover, the case studies show the difficulty in retrieving official definitions even using explicit prompting, indicating the need to include official definitions directly in the prompt. 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.

Acknowledgments

I want to express my sincere gratitude to Professor Chris Piech for his continued support and encouragement. I also appreciate Maty Bohacek and Anjali Sreenivas's thoughtful comments and invaluable advice.

References

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

Ralf Barkemeyer, Suraje Dessai, Beatriz Monge-Sanz, Barbara Gabriella Renzi, and Giulio Napolitano.

2016. Linguistic analysis of IPCC summaries for policymakers and associated coverage. *Nature Climate Change*, 6(3):311–316. Publisher: Nature Publishing Group.

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. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, 11(11):966–972.

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

IPCC. 2018. Annex i: Glossary. In V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield, editors, *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 H.-O. Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, and N.M. Weyer, editors, *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. IPCC. In Press.

IPCC. 2019b. Annex i: Glossary. In P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.-O. Pörtner, D.C. Roberts, P. Zhai, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. Portugal Pereira, P. Vyas, E. Huntley, K. Kissick, M. Belkacemi, and J. Malley, editors, *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. 2022. *Climate Change and Land: IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems*. Cambridge University Press.

Albert Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Singh Devendra, Diego Chaplot, Florian De Las Casas, Gianna Bressand, Guillaume Lengyel, Lucile Lample, Renard Saulnier, Marie-Anne Lavaud, Pierre Lachaux, Teven Stock, Thibaut Le Scao, Thomas Lavril, Timothée Wang, William Lacroix, and Sayed. 2023. *Mistral 7B*. arXiv.

Marko Hekkert Julian Kirchherr, Denise Reike. 2017. Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, Conservation and Recycling*, 127:221–232. Publisher: Elsevier.

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.

Meta. 2024. *The Llama 3 Herd of Models*. arXiv.

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. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. ArXiv:2307.15770 [cs].

Rebeka Lukman Peter Glavič. 2007. Review of sustainability terms and their definitions. *Journal of Cleaner Production*, 15(18):1875–1885. Publisher: Elsevier.

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

Py-Readbility-Metrics. 2019. [link].

Nils Smeuninx, Bernard De Clerck, and Walter Aerts. 2020. 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.

UN. 2024a. [link].

UN. 2024b. Emissions gap report 2024.

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