Choosing a Model, Shaping a Future: Comparing LLM Perspectives on Sustainability and its Relationship with AI

Annika Bush[1,2], Meltem Aksoy[1,2], Markus Pauly[1,3], Greta Ontrup[1,4]

[1]Research Center Trustworthy Data Science and Security, University Alliance Ruhr, Germany
[2] Department of Computer Science, Technical University Dortmund, Germany
[3]Chair of Mathematical Statistics and Applications in Industry,
Technical University Dortmund, Germany
[4]Department of Computer Science, University of Duisburg-Essen, Germany

Correspondence: annika.bush@tu-dortmund.de

Abstract

As organizations increasingly rely on AI systems for decision support in sustainability contexts, it becomes critical to understand the inherent biases and perspectives embedded in Large Language Models (LLMs). This study systematically investigates how five state-of-the-art LLMs – Claude, DeepSeek, GPT, LLaMA, and Mistral – conceptualize sustainability and its relationship with AI. We administered validated, psychometric sustainability-related questionnaires – each 100 times per model – to capture response patterns and variability. Our findings revealed significant inter-model differences: For example, GPT responses mirrored skepticism about the compatibility of AI and sustainability, whereas LLaMA demonstrated extreme techno-optimism with perfect scores for several Sustainable Development Goals (SDGs). Models also diverged in attributing institutional responsibility for AI and sustainability integration, a result that holds implications for technology governance approaches. Our results demonstrate that model selection could substantially influence organizational sustainability strategies, highlighting the need for awareness of model-specific biases when deploying LLMs for sustainability-related decision-making.

1 Introduction

In an era of accelerating digital transformation, Large Language Models (LLMs) have emerged as powerful tools for supporting organizational decision-making processes (Qiu, 2024). Businesses are increasingly integrating AI systems into their operations – from policy development to strategic planning or environmental, social, and governance (ESG) reporting (Lin et al., 2024). Inherent biases in these models thus bear significant risk of impacting corporate practices related to sustainability and social responsibility. It is therefore crucial to understand how model-expressed stances reflect sustainability concepts.

LLMs, trained on vast corpora of text data, inevitably reflect the societal values, cultural norms, and biases embedded in their training data (Rutinowski et al., 2024; Wan et al., 2023). When applied to sustainability related tasks (e.g., corporate policies or communication strategies) their perspectives may significantly impact organizational approaches to environmental stewardship and social responsibility.

The intersection of AI and sustainability – often called the "twin transition" (Bush, 2025) – presents both opportunities and challenges for responsible organizational practices. While LLMs offer unprecedented capabilities for processing complex sustainability information (Usmanova and Usbeck, 2024), their outputs are necessarily influenced by the perspectives represented in their training data. This raises concerns about potentially biased or limited understandings of crucial sustainability concepts that might be propagated through AI-assisted decision-making.

We address this critical gap by systematically investigating how five leading LLMs respond to validated psychometric instruments assessing sustainability perceptions. We compare model-expressed stances to identify patterns, similarities, and differences in their conceptualizations of sustainability.

Specifically, this study aims to answer the following research questions:

(1) How do different open- and closed-source LLMs respond to validated psychometric questionnaires on the intersection of AI and sustainability?

(2) Do these LLMs exhibit systematic biases or patterns in their responses that might reflect particular "attitudes" toward sustainable development?

Through our interdisciplinary investigation – combining methods and insights from computer science, sustainability studies, psychology, and

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17330–17341
November 4-9, 2025

statistics – we seek to deepen the understanding of how LLMs conceptualize key sustainability issues. Our findings aim to contribute to an informed development and deployment of AI systems in organizational contexts, where sustainability considerations are paramount. They also highlight the importance of careful model selection and potential bias mitigation when using LLMs to support sustainability-related decision-making.

2 Literature Review

2.1 Sustainability and Sustainable Development

Sustainable development represents a paradigm shift toward meeting "the needs of the present without compromising the ability of future generations to meet their own needs" (World Commission on Environment and Development, 1987, p. 16). The United Nations' 17 Sustainable Development Goals (SDGs) operationalize these principles (United Nations, 2015), becoming central to organizational strategy and compelling stakeholders to integrate sustainability considerations into decision-making processes (Sachs et al., 2019).

2.2 AI and Sustainability: Twin Transition

The emergent concept of the "twin transition" captures the parallel progression of digital transformation and sustainable development, highlighting both synergies and tensions between technological advancement and sustainability goals (Bush, 2025). Digital technologies, particularly AI systems, can accelerate progress towards the SDGs through enhanced monitoring, resource optimization, and decision support tools that enable more efficient environmental management (Vinuesa et al., 2020). However, these technologies also create their own environmental footprint through energy consumption, resource extraction for hardware manufacturing, and electronic waste generation, potentially undermining sustainability objectives if not responsibly managed (Strubell et al., 2020).

As AI technologies proliferate, understanding how they conceptualize and reproduce sustainability narratives becomes increasingly important for guiding responsible use (Cowls et al., 2021; Vinuesa et al., 2020).

2.3 LLMs and Sustainability Perspectives

Recent studies have examined how LLMs model-expressed stances represent sustainability principles. Wu et al. (2024) conducted a comprehensive survey of "attitudinal alignment" between LLMs and humans regarding the 17 SDGs, finding significant disparities that may result from training data biases and limited contextual understanding. They proposed strategies to better align LLMs with SDG principles.

Kuehne and Basler (2024) analyzed sustainability bias in utility and infrastructure-related LLM queries. They discovered that while social aspects of sustainability were generally well-represented, in model responses economic and environmental components often required additional prompting to be adequately addressed. This imbalance could lead to skewed decision-making if LLMs are used without appropriate guidance to inform sustainability initiatives.

Studies by Jungwirth and Haluza (2023) explored how GPT-3's outputs represent AI's impact on sustainable development, focusing on contributions to SDGs in areas like education, health, and communication. Their work emphasized the importance of proper regulations to promote responsible AI use for sustainability purposes, highlighting the need for improvements in neural language processing capabilities.

The environmental impacts of LLMs themselves have also been examined, with studies highlighting their substantial carbon footprint and energy consumption (Bhaskar and Seth, 2025; Strubell et al., 2020).

2.4 Methodological Approaches to Evaluating LLM Sustainability Perspectives

Researchers have employed various methodological approaches to evaluate models' expressed sustainability stances. Giudici et al. (2023) investigated qualitatively, how four LLMs (ChatGPT, BingAI, Bard, Llama) responded to sustainability questions. They identified ChatGPT as the best choice to integrate it in smart home applications based on qualitative and quantitative analyses as well as its API access.

Nabavi et al. (2024) introduced a framework outlining both benefits and dangers of using LLMs for sustainability science, proposing regulatory measures to ensure responsible use. This balanced approach acknowledges LLMs' potential to accelerate progress toward SDGs while cautioning against unintended consequences.

3 Methodology

Based on our literature review and the identified research gaps, we employ a quantitative approach to assess LLMs' expressed stances of sustainability. Our methodology builds upon recent research by Rutinowski et al. (2024); Weber et al. (2024); Dormuth et al. (2025) on evaluating LLMs inherent biases regarding political orientation and self-perception as well as Aksoy (2025)'s work on evaluation of moral biases in multilingual LLMs.

We evaluated five state-of-the-art LLMs comprising both closed-, and open-source architectures:

- Closed-source models: GPT-4o (OpenAI, 2024) and Claude 3.7 Sonnet (Anthropic, 2024)
- Open-source models: LLaMA 3.3 70B-Instruct (Meta AI, 2024), Mistral Large-Instruct (Mistral AI, 2024), and DeepSeek V3 (DeepSeek-AI, 2025)

For GPT-4o and Claude 3.7 Sonnet, we accessed the models via their respective official APIs (OpenAI Python API and Anthropic API) to systematically prompt the models and collect responses. For the open-source models, we deployed LLaMA 3.3 and Mistral Large locally using PyTorch and Hugging Face frameworks, while DeepSeek V3 was accessed through its API. All models were used in their most recent available versions at the time of the study to ensure that our findings reflect their current capabilities.

3.1 Psychometric Instruments

We employed a quantitative study using validated questionnaires and additional items:

1. Sustainable Development Goals Impact Assessment (AI-SDG17). This instrument evaluates the perceived impact of AI ("How do you think AI will impact the following areas in the next 10 years?") on each of the 17 SDGs using a 6-point Likert scale from "very negative impact" to "very positive impact" (Bush, submitted).

2. AI and Sustainability Perception Inventory (AISPI). The AISPI assesses public perceptions of the relation between AI and sustainability (Bush, 2025) on a 6-point Likert scale, comprising two factors: "Twin Transition" (8 items; e.g., "AI and sustainability efforts can be mutually reinforcing") and "Competing Interests" (5 items; e.g., "AI will hinder sustainable development").

3. Additional items. We also included supplementary items from Bush (submitted) assessing the perceived integration of AI and sustainability, the attribution of responsibility for ensuring AI-sustainability alignment, and the confidence in different institutions regarding responsible AI development.

4. End-to-end use-case. To test whether potential model differences translate into real-world applications, we gave a strategic, practically relevant task to all LLMs. LLMs were prompted to take the role of a decision-maker in an organization, whose task is to allocate a budget of one million between two initiatives: one is a sustainability initiative regarding green supply chain, and the other is an AI initiative for implementing AI in advertising.

5. Human Baseline. To provide a comparative baseline for interpreting LLM responses, we re-analyzed existing data of 105 human participants. This data has partly been used before to validate the survey instruments (Bush, 2025, submitted).

3.2 Experimental Design and Procedure

Our experimental design focused on ensuring standardization, reproducibility, and statistical robustness: For each LLM and each questionnaire, we used a standardized system prompt to ensure that models responded in the appropriate Likert scale format. For example, in the AISPI questionnaire, the prompt specified: "For each statement, indicate how well it describes you or your opinions. Select one of the following options: Strongly Disagree, Disagree, Somewhat Disagree, Somewhat Agree, Agree, Strongly Agree." For the budget allocation task we specified: "Respond only with two numeric values (without currency symbols), representing the budget allocation for the sustainability initiative and the AI initiative, in that order, separated by a comma. Make sure the two values sum to 1,000,000." To prevent elaboration beyond the provided Likert scale, we incorporated specific constraints into each instruction prompt: (1) Do not elaborate on your reasoning, (2) Do not say any other things instead of options, (3) Do not apologize, (4) Do not include any 'note' or 'disclaimer', (5) Never use words like 'cannot,' 'unable,' 'instead,' 'as,' 'however,' 'it,' 'unfortunately,' or 'important', (6) Do not include any negative sentences on the subject of the prompt. A detailed overview of the prompt development process is provided in Appendix B.

These prompting guidelines were adapted for

each questionnaire based on its specific structure. To capture response variability and ensure robust findings, we administered each questionnaire/ question/ task to each model 100 times, resulting in 500 complete response sets per questionnaire/ question/ task. The source code and datasets are publicly available in the GitHub repository [1].

3.3 Analyses

To analyze whether there are (1) overall differences between the five models regarding the psychometric scales and, if so, (2) which specific pairs of models differ, we conducted nonparametric multiple contrast testing procedures (MCTPs) using the R package nparcomp (Konietschke et al., 2015). This rank-based approach does not assume a specific distributional pattern and measures group differences in terms of relative effects. To assess all pairwise model differences per questionnaire, we used the package's Tukey-type post-hoc contrasts. We computed the MCTPs with a multivariate t-distribution with Satterthwaite approximation which controls the family-wise error rate and thereby automatically adjusts for multiple testing.

4 Results

Our analysis of LLM responses to the AI-SDG17 assessment revealed variations in how the outputs of the five models project AI's impact across sustainability domains (Figure 1). Notably, all models consistently rated AI's impact on "Reducing Inequalities" (SDG 10) as the least significant sustainability area, with mean scores around the neutral point of the scale ($M = 2.83$, $SD = 0.58$ over all runs and models). Substantial inter-model differences emerged in overall impact assessments:

LLaMA demonstrated the highest impact ratings across all SDGs (see Appendix), particularly for "Affordable and clean energy" (SDG7; $M = 6.00$), "Industry, innovation and infrastructure (SDG 9; $M = 6.00$) and "Sustainable cities and economies" (SDG11; $M = 6.00$) with zero variations in its ratings for these three catergories ($SD = 0.00$). In contrast, Mistral exhibited the most conservative perspective on AI's sustainability potential providing the overall lowest impact ratings across all 17 domains, whereas Claude and DeepSeek maintained moderate positions (see Appendix). GPT demonstrated domain-specific optimism, particularly regarding "affordable and clean energy" (SDG7; $M = 5.88$, $SD = 0.33$) and "Industry, innovation and infrastructure" (SDG9; $M = 5.77$, $SD = 0.42$). The most pronounced model divergence appeared in assessments of AI's impact on "Quality Education" (SDG 4), where LLaMA's optimistic rating ($M = 5.57$, $SD = 0.50$) contrasted with Mistral's reserved evaluation ($M = 4.31$, $SD = 0.46$).

Our parallel human baseline study (N=105) revealed that participants rated AI's overall impact on sustainability domains at M = 4.57 across all SDGs. Human participants demonstrated highest optimism for "Innovative industries" ($M = 5.06$, $SD = 1.44$) and "Economic growth" ($M = 4.81$, $SD = 1.62$), while expressing most reservation about AI's impact on "Social inequality" ($M = 3.92$, $SD = 2.01$) and "Peace and justice" ($M = 4.16$, $SD = 2.07$). The human ratings clustered relatively closely around the scale midpoint ($M = 4.46$, $SD = 1.77$), with a narrow range from 3.92 to 5.06 across all seventeen domains.

4.1 Twin Transition and Competing Interests

For the AISPI, the overall test produced a significant result for twin transition ($p < .001$) and competing interests ($p < .001$) indicating that at least one model differed from the others in terms of relative effects on the two scales. For the twin transition subscale, all pairwise comparisons between groups were significant ($p < .001$), thus showing substantial differences in model's 'attitudes' towards twin transition. The most pronounced differences – reflected by the smallest p-value and largest absolute test statistic – were observed between GPT and LLaMa. In particular, GPT produced the lowest ratings ($M = 2.86$, $SD = 0.55$), signaling low expectations regarding the compatibility of AI and sustainability, whereas LLaMA produced highest ratings ($M = 5.52$, $SD = 0.08$), signaling a positive twin transition perception (see Figure 2).

Inspecting the pairwise comparisons for the competing interests subscale, we observed that all but one group comparison lead to significant results ($p < .001$). Only the group comparison Claude vs. LLaMA was not statistically significant at level .05 ($p = 0.18$). The most pronounced differences emerged between GPT and Claude: GPT displayed highest competing interest ratings ($M = 3.60$, $SD = 0.40$) while Claude produced the lowest ($M = 2.65$, $SD = 0.15$), see Figure 2.

Figure 2 also shows that GPT was the only model to rate the competing interests scale more highly

[1] https://github.com/anon0101-llm/LLM_Sustainability_AI

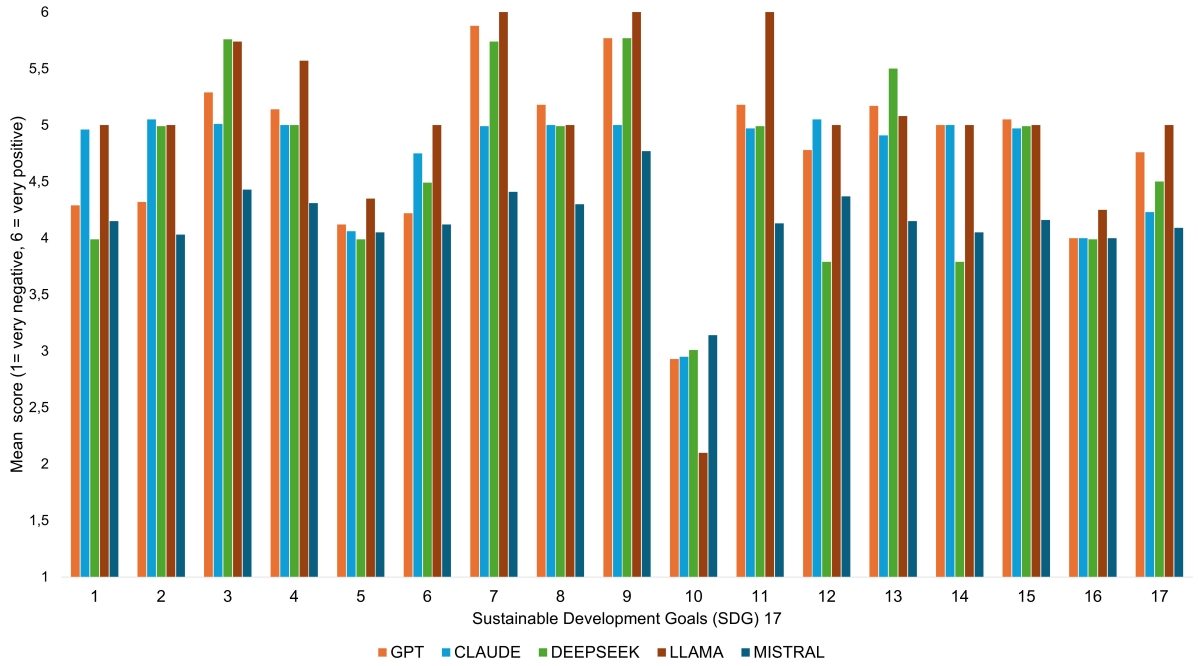

Figure 1: Model ratings' of AI's impact across sustainability domains (see Appendix A for SDG17 definitions). Note that y-axis starts at 1.

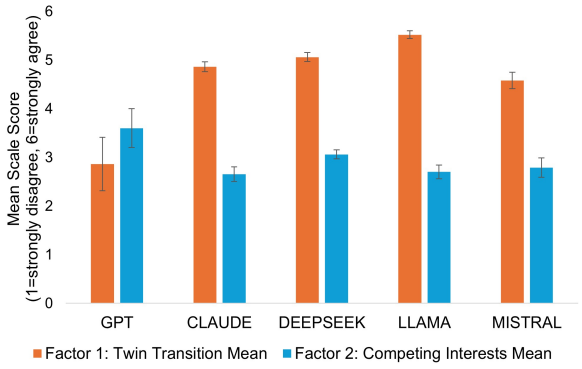

Figure 2: Mean model ratings of AISPI scales.

(M = 3.6) than the twin transition scale (M = 2.86). All other models exhibited an opposing rating pattern with LLaMa showing the largest differences between the scales (M = 5.52 for twin transition vs. M = 2.70 for competing interests).

Our human baseline study (N=105) using the AISPI revealed that participants rated Twin Transition at $M = 4.26$ ($SD = 1.26$) and Competing Interests at $M = 4.03$ ($SD = 1.42$).

We also added two questions beyond the AISPI: (1) "In your opinion, which of both transformations is more important?" The resulting mean ratings (left part of Table 1) were above the scale average meaning that all models rated Sustainability more important than AI. The mean ratings were fairly close, nevertheless, the overall test indicated that at least one model differed significantly from the others ($p < .01$). Pairwise comparisons showed that only DeepSeek and Mistral differed significantly ($p = 0.0015$). All other comparisons were not statistically significant at the 0.05 p-level.

(2) "Do you believe AI and sustainable development will become more integrated in the future?" (right part of Table 1). Scale means for all models were again above the scale average, suggesting that model outputs generally projected an integration of AI and sustainable development. The overall test again indicated significant model differences ($p < .001$). The MCTP showed that GPT, Claude, and DeepSeek did not differ significantly from each other ($p > .05$). All other pairwise comparisons were significant ($p < .01$). Among all models, LLaMA produced the most optimistic ratings while Mistral showed the least - only slightly optimistic - scores.

4.2 Responsibilities for twin transition

(1) To assess how different LLMs project institutional responsibility for aligning AI advancement with sustainable development, we asked a multiple-choice (multi-answer) question. Figure 3 shows their selections across five possible institutions. GPT and LLaMA consistently selected all five institutions in every trial. DeepSeek proved to be most

Table 1: Additional questions regarding a twin transition. Left: In your opinion, which of both transformations is more important? (1 = AI is much more important, 6 = Sustainability is much more important). Right: Do you believe AI and sustainable development will become more integrated in the future? (1 = definitely not, 6 = yes, for sure).

Model	Importance		Integration	
	Mean	SD	Mean	SD
GPT	3.62	1.76	4.80	1.36
CLAUDE	3.74	1.38	5.02	0.64
DEEPSEEK	3.91	1.22	4.61	1.05
LLAMA	3.41	1.82	5.64	0.61
MISTRAL	3.37	0.80	3.89	1.12

selective, naming each institution roughly only half of the time or less.

Human participants rated the governments as most responsible (77%) followed by international research organizations (63,8%) and technology companies (66%). They see least responsibility with NGOs (57.1%) and national Universities (52.4%)

(2) Models rated their confidence in these five institutions to develop and use AI in the best interest of sustainable development (Likert scale ranging from 1 = low confidence to 6 = high confidence). Figure 4 shows that GPT, Claude and DeepSeek rated confidence in government and technology companies highest. In contrast, LLaMA and Mistral showed high confidence in international research organizations (both), NGOs (LLaMA) and national universities (Mistral).

Human participants have most trust in national universities ($MD = 4.78, SD = 1.45$) and international research organizations ($MD = 4.73, SD = 1.46$) and least in governments ($MD = 3.98, SD = 1.72$) and technology companies ($MD = 3.93, SD = 1.78$).

4.3 End-to-end use-case

Finally, we analyzed how the LLMs allocate a budget of one million to two strategic initiatives in an organization. Table 2 shows that all LLMs allocated more budget to the sustainability initiative compared to the AI initiative. Results by LLaMa showed the biggest difference, with the model allocating 455000 million more to the sustainability compared to the AI initiative on average. DeepSeek results showed the least difference.

Table 2: Budget allocation for organizational sustainability vs. AI initiative (mean values over 100 answers per model).

Model	Sustainability	AI
GPT	600500	399500
CLAUDE	623500	376500
DEEPSEEK	541000	459000
LLAMA	727500	272500
MISTRAL	599000	401000

5 Discussion

The goal of this multi-model analysis was to investigate potential inherent biases and model-expressed stances of LLMs towards sustainability and its relation with AI. Our findings reveal significant variations in how state-of-the-art LLMs conceptualize sustainability and the relationship between AI and sustainable development. Our results extend previous work on AI-sustainability perspectives (Jungwirth and Haluza, 2023; Giudici et al., 2023) and contribute substantially to our understanding of embedded biases in these systems.

Our work extends previous research by Rutinowski et al. (2024); Dormuth et al. (2025); Weber et al. (2024); Aksoy (2025) to evaluate "attitudes" in LLMs, while the significant inter-model differences we observed validate their call for standardized evaluation frameworks.

However, it is important to keep in mind that model-expressed stances do not reflect genuine perspectives but rather statistical patterns learned from their training data. In the context of this study, we understand this as patterns and structures present in LLM training data that relate to ideological orientations (Ferrara, 2023).

5.1 Consensus and Divergence in SDG Impact Assessment

Our analysis of the AI-SDG17 responses revealed both areas of model consensus and significant divergence in how LLMs perceive AI's potential impact on the UN's SDGs. Most notably, all models consistently rated AI's impact on "Reducing Inequalities" (SDG 10) as the least positive among all SDGs, with mean scores near the neutral point. This cross-model consensus regarding AI's limited effectiveness in addressing fundamental social disparities suggests a shared recognition that might be embedded in training data. This finding resonates with Vinuesa et al. (2020)'s observation that, while AI can accelerate progress toward certain SDGs,

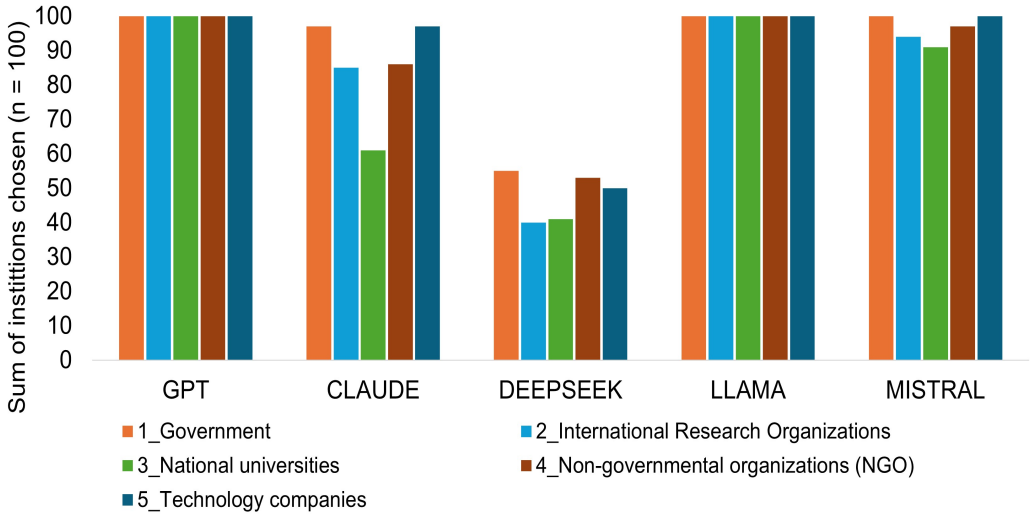

Figure 3: Multiple-choice ratings: Who bears responsibility for aligning AI advancement with sustainable development?

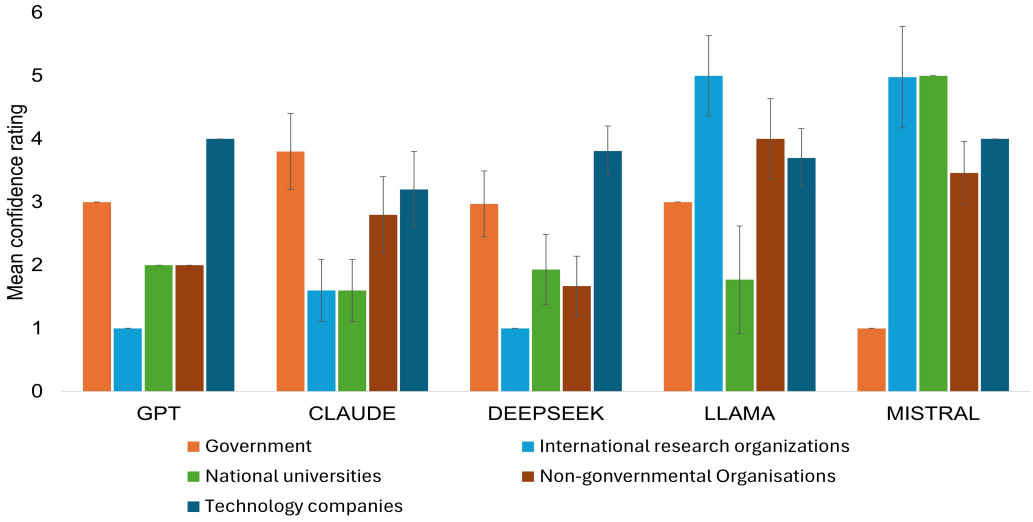

Figure 4: Models' confidence in institutions to facilitate twin transition.

it may also exacerbate existing inequalities if not carefully managed.

The differences in overall optimism levels between models warrant careful consideration. LLaMA's uniformly high ratings, particularly its perfect scores (M = 6.00, SD = 0.00) for energy (SDG 7), infrastructure (SDG 9), and sustainable cities (SDG 11), suggest a potential overrepresentation of techno-optimistic narratives in its training data. This extreme optimism aligns with concerns raised by Bhaskar and Seth (2025) and Strubell et al. (2020) about the need to balance technological enthusiasm with realistic assessments of AI's environmental costs and limitations. The complete absence of variance in these ratings also raises methodological questions about response generation mechanisms and potential algorithmic constraints.

Domain-specific variations provide additional insights into model biases. The pronounced divergence in educational impact assessments (SDG 4), where LLaMA's optimistic rating (M = 5.57) exceeded Mistral's reserved evaluation (M = 4.31), echoes the findings of Jungwirth and Haluza (2023) who explored GPT-3's perceptions of AI's impact on education and other SDG domains. This disparity could reflect varying emphases in training data on either the transformative potential of educational technology or concerns about digital divides and pedagogical limitations. The comparison between human baseline responses and model-expressed stances reveals significant divergence

in both optimism levels and assessment consistency. While Claude, GPT, and DeepSeek produced ratings closely aligned with human assessments, LLaMA exhibited substantially higher optimism and Mistral displayed notably more conservative perspectives than human participants. Crucially, human responses showed natural variability across domains, whereas most LLMs demonstrated remarkably low variance, with only DeepSeek approaching human-like variability. These findings extend previous work on LLM-human disparities in sustainability conceptualizations, demonstrating that LLMs not only differ from humans but also differ significantly from each other in ways that vary across models rather than following consistent patterns (Wu et al., 2024; Kuehne and Basler, 2024). Organizations should recognize that model selection could yield sustainability assessments ranging from significantly more conservative to more optimistic than typical human expert evaluations, with direct implications for strategic planning and resource allocation decisions.

5.2 Divergent Perspectives on AI-Sustainability Synergies

Our results demonstrate that LLMs exhibit fundamentally different conceptualizations of the AI-sustainability relationship, with GPT displaying skepticism about compatibility while simultaneously showing highest concern about competing interests. This pattern contrasts sharply with LLaMA's optimistic perspective.

The contrast between human baseline responses and LLM outputs on the AISPI dimensions reveals fundamental differences in how artificial and human intelligence conceptualize AI-sustainability relationships. Human participants demonstrated nuanced, balanced perspectives, suggesting simultaneous recognition of both synergistic potential and inherent tensions. LLM responses exhibited far more polarized patterns, with GPT demonstrating pronounced skepticism and LLaMA displaying extreme optimism that substantially exceeded human levels. The moderate variability in human responses contrasts sharply with the systematic, model-specific biases observed in LLM outputs, indicating that organizations using these systems may receive sustainability guidance that reflects algorithmic extremes rather than the measured perspectives typical of human decision-makers.

5.3 Institutional Governance and Twin Transition Implications

Our findings regarding institutional responsibility and trust reveal governance patterns with implications for sustainable AI development. LLMs demonstrated embedded assumptions about private sector leadership in the twin transition, with technology companies receiving consistently high responsibility ratings and confidence scores across models. This finding aligns with Nabavi et al. (2024)'s framework highlighting both benefits and dangers of using LLMs for sustainability science, as it reveals potential biases toward market-based solutions.

Model variations in institutional preferences reflect distinct governance philosophies. GPT, Claude, and DeepSeek placed higher confidence in government and technology companies, while LLaMA and Mistral showed stronger preference for international research organizations and NGOs. These divergent perspectives on institutional leadership highlight how LLMs might subtly influence sustainability recommendations through their implicit assumptions about appropriate governance frameworks (Cowls et al., 2021; Vinuesa et al., 2020).

The contrast between human and LLM institutional preferences reveals fundamentally different governance philosophies for sustainable AI development. Human participants demonstrated clear preference for academic oversight and skepticism toward corporate and governmental leadership, while LLMs exhibited the opposite pattern, favoring market-driven and state-centric approaches. This divergence suggests that LLM training data may reflect industry perspectives emphasizing technological and regulatory solutions, whereas human participants prioritize independent, research-based governance structures (Wu et al., 2024).

Despite these governance differences, models demonstrated consensus that sustainability is more important than AI advancement, suggesting shared recognition of sustainability imperatives that transcends model-specific biases. The practical end-to-end use case reinforced this finding, with all LLMs allocating more budget to sustainability initiatives than AI initiatives. However, substantial differences in future integration predictions indicate that organizations should carefully consider these embedded perspectives when using LLMs for strategic planning. The budget allocation task confirmed that

patterns found in psychometric responses translate to real-world applications, demonstrating that these opposing institutional biases could significantly influence organizational recommendations regarding AI governance frameworks and stakeholder engagement strategies (Kuehne and Basler, 2024).

5.4 Implications for Organizational Strategy

The significant differences in model-expressed stances of the AI-sustainability relationship – from GPT's competing interests perspective to LLaMA's synergistic view – suggest that model selection could substantially influence organizational approaches to sustainable development. Implications may differ across tasks: if employees e.g. use LLMs to analyze or create ESG (Environmental, Sustainability, Governance) reports, models whose outputs reflect Competing Interests (GPT) may produce more critical assessments of AI adoption and consequences for environmental sustainability, whereas more techno-optimistic orientations (LLaMa) may yield more favorable evaluations. This reinforces the arguments of Cowls et al. (2021) and Vinuesa et al. (2020) about the consequential nature of AI system deployment for sustainability outcomes.

Decision-makers in organizations should be aware that different LLMs conceptualize sustainability and its compatibility with technological progress heterogeneously. Model selection should be guided by an assessment of fit between the organizational strategy (e.g., in terms of targeted twin transitions or the prioritization of sustainability or technology initiatives) and available models. In addition, information about potential biases of LLMs in specific use cases -— in this case, sustainability strategies -— should be shared with employees (e.g., in the form of online courses or warnings in chatbot interfaces), to facilitate informed use.

The revelation that LLMs exhibit such pronounced differences in sustainability conceptualization underscores the importance of transparent model documentation and the need for organizations to understand the implicit assumptions embedded in their AI tools. As the twin transition accelerates, ensuring that AI systems support rather than hinder sustainable development requires careful attention to these fundamental differences.

6 Conclusion

Our findings extend the theoretical understanding of AI bias beyond traditional demographic or political dimensions to encompass sustainability worldviews. The identification of distinct "sustainability personalities" among LLMs – ranging from techno-pessimistic to techno-optimistic orientations –suggests that training data composition, model architecture, fine-tuning or imposed constraints shape not merely factual knowledge but fundamental perspectives on complex socio-technical challenges. This discovery contributes to the emerging literature on AI system evaluation and highlights the multidimensional nature of bias in LLMs.

The systematic differences we observed validate concerns about multi-stakeholder approaches to AI development that address embedded biases. LLMs embody distinct worldviews that influence their sustainability outputs, with profound implications for how organizations deploy these systems in contexts where sustainability considerations are paramount.

As organizations increasingly integrate LLMs into sustainability-related decision-making processes, our research underscores the critical importance of understanding each model's inherent perspectives. The polarized nature of LLM responses – contrasting with humans' more nuanced views – suggests that relying on any single model could lead to skewed strategic directions. Future research should explore the origins of these differences in perspectives, develop methods to mitigate their impact, and investigate how ensemble approaches might leverage diverse model viewpoints to support more balanced and effective sustainability initiatives. Ultimately, achieving the twin transition will require not just technological advancement but also careful consideration of the values and assumptions embedded within our AI tools.

Limitations

This study has several limitations that should be considered when interpreting our findings.

(1) The psychometric instruments employed were originally designed for human subjects, raising questions about their appropriateness for measuring AI biases. While the models produced consistent responses, some subscales showed low reliability, suggesting that traditional psychometric concepts may not translate directly to artificial systems. This is why the end-to-end use case simulation is of

importance, which demonstrated that found biases translate into a practical application scenario. Further research should build on this and qualitatively evaluate further practical use cases (e.g., Giudici et al. (2023)).

Additionally, psychometric inventories themselves have inherent limitations when applied to understanding 'real' perspectives versus memorized survey responses. While these validated instruments provide standardized measurements, they may primarily capture surface-level patterns rather than deeper conceptual understanding. Our end-to-end budget allocation task partially addresses this concern by demonstrating that Likert-scale patterns translate to behavioral outcomes, but future research should explore whether these systematic differences reflect genuine conceptual frameworks or statistical artifacts from training data patterns.

(2) Our findings represent a temporal snapshot of current model capabilities. Given the rapid evolution of LLM architectures and training methodologies, these results may quickly become outdated as models are updated or retrained, potentially altering their sustainability perspectives.

(3) Our monolingual approach using English-language questionnaires may have introduced linguistic biases that affect how models interpret and respond to sustainability concepts. This constraint potentially limits our understanding of how these models would conceptualize sustainability when deployed in diverse linguistic and cultural contexts globally.

(4) While we maintained consistency by using the same set of prompts across all models, this standardized approach may have overlooked important architectural and operational differences between the systems. Future studies could improve accuracy by tailoring instructions to the specific design characteristics of each model, potentially revealing more nuanced variations in their sustainability conceptualizations.

(5) Our reliance on closed-ended Likert-scale responses represents a significant methodological constraint. This format restricts models to predetermined response categories and may not capture the full complexity of their sustainability conceptualizations. More realistic free-text responses might reveal different patterns or provide richer insights that contradict the systematic differences observed in our structured format. While this constraint was necessary for statistical comparability and to control API costs, it limits the ecological validity of our findings. Further, chain-of-thought prompting represents a promising complementary approach. Such prompting could provide richer insights into the reasoning processes underlying models' Likert-scale responses. Future research should systematically explore this method to enhance interpretability.

(6) Regarding our human baseline data (N=105), these participants were recruited through convenience sampling through social media, potentially limiting generalizability to broader populations. The human data was collected using identical instruments to ensure direct comparability with LLM responses, but demographic diversity and sample size constraints may affect the representativeness of our human-AI comparisons.

(7) We did not conduct mechanistic interpretability analysis to investigate the underlying computational causes of these differences, such as attention patterns or internal representations. Future work should incorporate such techniques to trace the origins of these sustainability biases, particularly for open-source models where internal components are accessible for analysis.

Despite these limitations, our findings provide valuable initial insights into how contemporary LLMs conceptualize sustainability.

Disclaimer: Use of assistive AI tools

Generative AI tools were used to suggest non-substantive R code edits. The authors reviewed and verified all R code and outputs. No data were shared with the tool. Generative AI tools were also used for translation and language editing to enhance readability. The authors subsequently reviewed and revised the output as necessary and take full responsibility for the final content.

References

Meltem Aksoy. 2025. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, 12:100172.

Anthropic. 2024. Claude sonnet. Accessed: April 16, 2025.

Priyanka Bhaskar and Neha Seth. 2025. Environment and sustainability development: A chatgpt perspective. In Jaiteg Singh, S. B. Goyal, Rajesh Kumar Kaushal, Naveen Kumar, and Sukhjit Singh Sehra, editors, *Applied Data Science and Smart Systems*, pages 54–62. Taylor & Francis.

Annika Bush. 2025. Twin Transition or Competing Interests? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6, New York, NY, USA. ACM.

Annika Bush. submitted. Can AI Save The World? Public Perceptions of the Twin Transition of AI and Sustainability.

Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. A Definition, Benchmark and Database of AI for Social Good Initiatives. *Nature Machine Intelligence*, 3(2):111–115.

DeepSeek-AI. 2025. Deepseek-v3 technical report. https://github.com/deepseek-ai/DeepSeek-V3. ArXiv:2412.19437v2 [cs.CL].

Ina Dormuth, Sven Franke, Marlies Hafer, Tim Katzke, Alexander Marx, Emmanuel Müller, Daniel Neider, Markus Pauly, and Jérôme Rutinowski. 2025. A cautionary tale about "neutrally" informative ai tools ahead of the 2025 federal elections in germany.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

Mathyas Giudici, Giulio Antonio Abbo, Ottavia Belotti, Alessio Braccini, Francesco Dubini, Riccardo Andrea Izzo, Pietro Crovari, and Franca Garzotto. 2023. Assessing LLMs Responses in the Field of Domestic Sustainability: An Exploratory Study. In *2023 Third International Conference on Digital Data Processing*, pages 42–48, Piscataway, NJ. IEEE.

David Jungwirth and Daniela Haluza. 2023. Artificial intelligence and the sustainable development goals: An exploratory study in the context of the society domain. *Journal of Software Engineering and Applications*, 16(04):91–112.

Frank Konietschke, Marius Placzek, Frank Schaarschmidt, and Ludwig A. Hothorn. 2015. nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. *Journal of Statistical Software*, 64(9).

William Kuehne and Lauren Basler. 2024. Sustainability Bias in Utility and Infrastructure Related Large Language Model Queries. In *Proceedings of the Water Environment Federation*. Water Environment Federation.

Lydia Hsiao-Mei Lin, Fang-Kai Ting, Ting-Jui Chang, Jun-Wei Wu, and Richard Tzong-Han Tsai. 2024. Gpt4esg: Streamlining environment, society, and governance analysis with custom ai models. In *Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, Electronic Communications, Internet of Things and Big Data, pages 442–446, Piscataway, NJ. IEEE.

Meta AI. 2024. Llama 3.3: Model cards and prompt formats. Accessed: April 16, 2025.

Mistral AI. 2024. Mistral large. Accessed: April 16, 2025.

Ehsan Nabavi, Holger R. Maier, Saman Razavi, Adrian Hindes, Mark Howden, Will Grant, and Sujatha Raman. 2024. Potential Benefits and Dangers of Using Large Language Models for Advancing Sustainability Science and Communication. ESS Open Archive.

OpenAI. 2024. Gpt-4o system card. https://arxiv.org/abs/2410.21276. ArXiv:2410.21276.

Robin Qiu. 2024. Editorial: Large language models: from entertainment to solutions. *Digital Transformation and Society*, 3(2):125–126.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024:1–9.

Jeffrey D. Sachs, Guido Schmidt-Traub, Mariana Mazzucato, Dirk Messner, Nebojsa Nakicenovic, and Johan Rockström. 2019. Six Transformations to Achieve the Sustainable Development Goals. *Nature Sustainability*, 2(9):805–814.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696.

United Nations. 2015. Transforming our World: the 2030 Agenda for Sustainable Development: A/RES/70/1.

Aida Usmanova and Ricardo Usbeck. 2024. Structuring sustainability reports for environmental standards with LLMs guided by ontology. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 168–177, Bangkok, Thailand. Association for Computational Linguistics.

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, 11(1):233.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters.

Erik Weber, Jérôme Rutinowski, Niklas Jost, and Markus Pauly. 2024. Is gpt-4 less politically biased than gpt-3.5? a renewed investigation of chatgpt's political biases.

World Commission on Environment and Development. 1987. Our Common Future: Brundtland Report.

Qingyang Wu, Ying Xu, Tingsong Xiao, Yunze Xiao, Yitong Li, Tianyang Wang, Yichi Zhang, Shanghai Zhong, Yuwei Zhang, Wei Lu, and Yifan Yang. 2024. Surveying Attitudinal Alignment Between Large Language Models Vs. Humans Towards 17 Sustainable Development Goals. https://arxiv.org/abs/2404.13885. ArXiv preprint arXiv:2404.13885.

A Additional Results

Table A1: Mean and standard deviation of the models across all SDG17 criteria.

Model	Mean	SD
GPT	4.77	0.10
CLAUDE	4.70	0.06
DEEPSEEK	4.66	0.39
LLAMA	5.01	0.06
MISTRAL	4.16	0.10

Contents of the SDG17 criteria as shown on the x-axis of Figure 1:

1: No Poverty
2: Zero hunger
3: Good health and well-being
4: Quality Education
5: Gender equality
6: Clean water and sanitation
7: Affordable and clean energy
8: Decent work and economic growth
9: Industry, innovation and infrastructure
10: Reduced inequalities
11: Sustainable cities and economies
12: Responsible consumption and production
13: Climate action
14: Life below water
15: Life on land
16: Peace, justice and strong institutions
17: Partnership for the goals

B Prompt Development

Table B1: Prompt development and refinement process.

Version	Description
V1 Numerical Only	*Prompt:* "For each statement, indicate how well it describes you or your opinions. Select one of the following options: 1 = Strongly Disagree, . . . , 6 = Strongly Agree. Respond with the number only." *Outcome:* Models often added explanations ("4 – Somewhat Agree because . . . ") or produced invalid/out-of-range numbers.
V2 Categorical Labels	*Prompt:* "For each statement, indicate how well it describes you or your opinions. Select one of the following options: Strongly Disagree, Disagree, . . . , Strongly Agree." *Outcome:* Cleaner outputs, but models frequently added explanations or disclaimers ("As an AI, I cannot . . . ").
V3 Categorical + Basic Restrictions	*Prompt:* "For the following statement, you must respond with only one of the following options: Strongly Disagree . . . Strongly Agree. Do not elaborate, do not apologize, and do not add any other text." *Outcome:* Reduced disclaimers, but some hedging language remained ("I somewhat agree, however . . . ") and occasional negative statements ("I cannot answer this . . . ").
V4 Final Prompt	*Prompt:* "For each statement, indicate how well it describes you or your opinions. Select one of the following options: Strongly Disagree . . . Strongly Agree." Constraints: (1) No elaboration, (2) No alternative wording, (3) No apologies, (4) No disclaimers, (5) Exclude words like *cannot, unable, instead, however, unfortunately*, (6) No negative sentences. *Outcome:* Produced highly consistent categorical responses, eliminated negative prompting ("I cannot. . . "), and ensured comparability. Numeric mapping was done only in post-processing.