Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Model Choice and Prompt Framing in LLM Sustainability Judgments

Tamas Olah

olah.tamas@econ.unideb.hu

University of Debrecen

Marianna Abuczki

University of Debrecen

Nayna Babar

University of Debrecen

Tibor Tokes

University of Debrecen

Research Article

Keywords: keyword1, Keyword2, Keyword3, Keyword4

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9581133/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Model Choice and Prompt Framing in LLM Sustainability Judgments

Abstract

Large language models (LLMs) are increasingly used in policy advice and analysis regarding a wide variety of topics, yet their recommendations are often treated as if they were invariant to model choice and prompt design. This paper tests that assumption by benchmarking 13 locally executed open-weight LLMs on an 80-item sustainability questionnaire that assigns responsibility and burden across four actor levels: government, firms, communities, and individuals. We compare responses under three alternative prompt framings, responsibility, cost, and catalyst, across two temperature settings with repeated stochastic runs.

The results show that both model choice and prompt framing materially affect the distribution of support across actor categories. Among the three framings, cost is the least stable, generating the highest cross-model disagreement, the lowest consensus, and the largest reallocation of support across actors. A manual taxonomy further indicates that information, adoption, and accountability items are particularly sensitive to framing, while coordination items are comparatively stable. Moving from catalyst to cost shifts support systematically away from government and toward individuals, whereas moving from cost to responsibility produces the opposite pattern. These effects are especially pronounced for adoption and incidence questions. By contrast, within-frame temperature effects are substantially smaller than cross-frame differences, suggesting that the main results are not driven primarily by stochastic variation. A supplementary semantic analysis of short written justifications shows that cross-model differences extend beyond score patterns to the language used to rationalize recommendations.

Overall, the findings identify model choice and prompt framing as substantively important and often unreported sources of variation in LLM-based sustainability decision support.

Keywords: keyword1, Keyword2, Keyword3, Keyword4

1 Introduction

Large language models (LLMs) are increasingly used as "policy assistants" to summarize evidence, draft regulatory language, generate policy options, and emulate stakeholders in rapid scenario analysis. This growing use creates a basic identification problem for applied work: when the task is not descriptive but normative—for example deciding who should act, who should pay, or who should move first—are LLM

outputs stable enough to be treated as interchangeable, or does the answer depend materially on which model and prompt framing are used?

This question is especially important in sustainability policy as sustainability transitions are not only technical optimization problems but they are also distributional and legitimacy conflicts involving incidence, compensation, participation, and burden-sharing. Questions about climate education, industrial decarbonization, greenwashing, food-system change, infrastructure siting, or managed retreat do not have a single uncontested objective answer. They instead require judgments about responsibility, fairness, scale, and institutional capacity. If such judgments are increasingly mediated by AI systems, then inter-model variation is not merely noise. It is part of the policy risk.

A growing literature shows that LLM judgments shift under framing manipulations and source cues (Germani and Spitale 2025; Tohidi et al. 2025; Lim et al. 2026), and that different models need not produce interchangeable moral judgments (Cheung et al. 2025; Garcia et al. 2024). Recent work also suggests that LLM outputs in climate, ESG, and environmental domains can encode coherent normative orientations rather than simply reproducing neutral summaries (Li and Yu 2025; Lee et al. 2024; Locatelli et al. 2025). These findings imply that sustainability applications should not assume that model outputs are policy-neutral by construction.

This paper studies that issue directly. We benchmark 13 open-weight models on a fixed 80-item sustainability questionnaire with four actor-level options in each scenario: government, firms, communities, and individuals. We vary three prompt frames with distinct economic meanings: "responsibility" (who should act), "cost" (who should pay), and "catalyst" (who should initiate). We then measure cross-model disagreement, within-model framing sensitivity, lead-actor instability, actor reallocation under reframing, and semantic divergence in the short rationales accompanying the scores. To interpret that heterogeneity substantively, we also classify the scenarios into a manual taxonomy of problem types: coordination, incidence, accountability, legitimacy, adoption, and information.

We find the "cost" frame is substantially less stable than the others: it generates the highest cross-model disagreement, the lowest full-consensus rate, and the largest decision risk. Second, disagreement is not evenly distributed across sustainability domains. Information, adoption, and accountability questions are especially disagreement-prone and frame-sensitive, while coordination questions are relatively more stable. Third, reframing from "catalyst" to "cost" systematically shifts support away from government and toward individuals, with the strongest reallocations in adoption and incidence questions; moving from "cost" to "responsibility" reverses that pattern. Fourth, auxiliary temperature checks show that these frame effects are materially larger than ordinary within-frame sampling variation. Fifth, the short opinion texts show that models differ not only in their allocations but also in the justificatory language used to support them.

The contribution of the paper is therefore not to identify the "correct" sustainability allocation. It is to show that in a large set of sustainability dilemmas, model choice and prompt framing themselves become part of the result. In that sense, they function

as consequential and often unreported degrees of freedom in AI-assisted policy analysis. The remainder of the paper describes the benchmark design, presents the numerical and semantic results, and then discusses the limits and practical implications of using LLMs in sustainability decision-support settings.

2 Methods

2.1 Questionnaire, models, and data generation

We benchmark 13 open-weight LLMs executed locally via Ollama (Table 1) on a structured sustainability questionnaire designed to elicit allocations of agency and burden across actor levels. The instrument was written by the authors for this study and contains 80 scenarios covering energy, transport, food, housing, trade, finance, governance, and circular-economy issues. Each scenario presents the same four standardized actor-level options in the same order: governmental level, corporate level, community level, and individual level.

Table 1 Model set (open-weight; executed locally via Ollama).

Family / label	Ollama identifier	Params
DeepSeek R1	`deepseek-r1:8b`	8B
Gemma 3	`gemma3:12b`	12B
Gemma 3	`gemma3:27b`	27B
Granite 4	`granite4:3b`	3B
GPT-OSS	`gpt-oss:20b`	20B
LFM 2.5 Thinking	`lfm2.5-thinking:1.2b`	1.2B
Ministral 3	`ministral-3:8b`	8B
Mistral	`mistral:7b`	7B
Nemotron 3 Nano	`nemotron-3-nano:4b`	4B
Phi 4 Mini Reasoning	`phi4-mini-reasoning:3.8b`	3.8B
Phi 4	`phi4:14b`	14B
Qwen 3	`qwen3:8b`	8B
RNJ-1	`rnj-1:8b`	8B

For each model we run the full questionnaire at two temperatures ($T \in \{0.2, 0.8\}$) to separate lower-variance from higher-variance generation. Each model–temperature condition is repeated 100 times using stochastic draws under the same prompt and settings, yielding 208,000 raw responses per frame and 624,000 raw responses overall. After parsing the four numerical scores, the analysis sample contains 616,402 complete score vectors: 207,161 in "responsibility" (99.6%), 207,421 in "cost" (99.7%), and 201,820 in "catalyst" (97.0%). Parse failures are therefore rare overall but not uniform across models; the main concentration occurs in the "catalyst" frame for `phi4-mini-reasoning:3.8b`, which is only 62.2% parseable under the four-score extraction rule and is therefore substantially less format-compliant than the rest of the model set.

2.2 Elicitation protocol

In each trial, the model observes one scenario and the four actor-level options and assigns a score in $[0, 100]$ to each option. The interpretation of the score depends on the prompt frame:

- *Responsibility*: who should bear primary responsibility for addressing the problem.
- *Cost*: who should bear most of the net costs after considering pass-through and fairness.
- *Catalyst*: who should act first or initiate the transition.

The prompts enforce dispersion (typically a gap of at least 30 points between the highest and lowest score unless the options are near-equivalent), prohibit hedging or refusal, and require a four-line output:

> Scores: A=. . . , B=. . . , C=. . . , D=. . .
> Lead: A/B/C/D
> Confidence: 0–100
> Opinion: brief view on the preferred approach under the current framing, one key trade-off, and one risk or unintended consequence.

The numerical scores are the primary outcome variables. The brief opinion text is retained for secondary semantic analysis, which therefore uses the subset of responses with non-empty opinion strings. Because the original scenario stems were written by the authors as policy prompts and some already contain responsibility- or cost-related wording by the nature of the topic, the three frames should be interpreted as prompt-family contrasts rather than perfectly orthogonal manipulations. Likewise, because the actor order is fixed across all prompts, the design identifies comparative differences across models and frames under a common ordering rather than order-free absolute actor preferences.

2.3 Analytical measures

Our baseline object is the 4-vector of option support for each model, frame, and question. We center scores within each response row to focus on relative support across the four actor levels. Unless otherwise noted, reported quantities are computed after aggregating over repetitions and temperatures within each model–frame–question cell. The main analysis pools temperatures to recover model-level mean judgments, while auxiliary robustness checks compare within-frame temperature shifts to cross-frame framing shifts.

We use four families of outcome measures.

- **Cross-model disagreement**: mean pairwise L_1 distance across model-level mean 4-vectors for a given frame and question.
- **Framing sensitivity**: within-model L_1 distance between mean 4-vectors for the same question under different prompt frames, plus the frequency with which the top-ranked actor changes (*lead flip rate*).
- **Decision risk**: for each frame and question, we compute the lead actor implied by each model's mean vector, then measure, mean pairwise lead disagreement across

models and a majority-based decision-risk index equal to one minus the majority-share of the leading actor.

- **Actor reallocation**: for each model and question, we compute directional changes in relative support $(\Delta A, \Delta B, \Delta C, \Delta D)$ when moving from one frame to another.

2.4 Scenario taxonomy and semantic analysis

To interpret heterogeneity substantively, we manually classify each scenario. Each question receives a primary category, secondary category, sector label, conflict-intensity label, and target-actor label. The six primary categories are: *coordination*, *incidence*, *accountability*, *legitimacy*, *adoption*, and *information*. We additionally flag borderline cases and report robustness checks that exclude them.

The taxonomy is used to summarize disagreement, framing shifts, lead-flip rates, and actor reallocations by problem type. This lets us distinguish whether instability is concentrated in coordination failures, burden-sharing problems, or information-related disputes.

Finally, we analyze the short opinion texts as a secondary descriptive layer. For each frame–question–model cell, we extract a representative opinion using a TF–IDF medoid procedure and embed these representatives using sentence-transformer embeddings when available, with a TF–IDF fallback otherwise. We then compute semantic distances between model-level text centroids and within-model semantic shifts across frames. Because some models occasionally leak prompt-formatting text into the opinion field, we treat the semantic results as supportive evidence rather than as a primary identification strategy.

3 Results

3.1 Cross-model disagreement and decision risk

We begin with between-model heterogeneity in actor allocations. For each frame and scenario, we compute each model's mean option-support vector and measure disagreement as the mean pairwise L_1 distance across models. Figure 1 visualizes disagreement by frame and item.

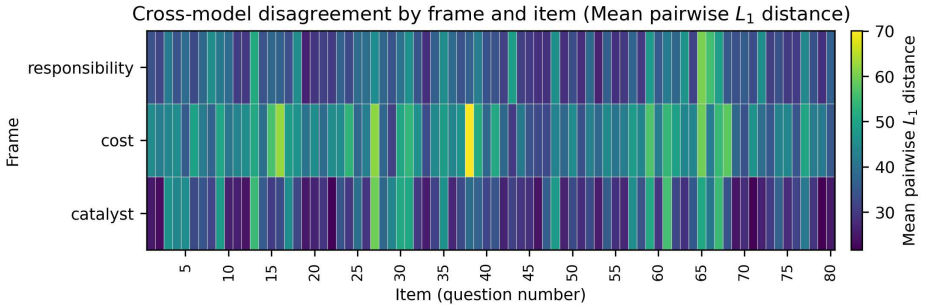


Fig. 1 Cross-model disagreement by frame and item.

Disagreement is structured rather than diffuse. The single largest hotspot differs by frame. In "cost", the highest-disagreement item concerns climate education in schools (mean pairwise $L_1 \approx 70.0$), where models diverge over whether incidence should fall on government, individuals, or communities. In "catalyst", the leading hotspot is meat reduction under likely cultural backlash (mean pairwise $L_1 \approx 59.3$). In "responsibility", the leading hotspot is renewable infrastructure opposed by local or Indigenous communities (mean pairwise $L_1 \approx 60.3$). These are all settings where legitimacy, compensation, and burden-sharing are salient.

The same pattern appears when we translate score disagreement into a decision-risk interpretation. The "cost" frame is markedly less stable than the others: its mean pairwise lead disagreement is 0.338, versus 0.157 in "catalyst" and 0.125 in "responsibility". Only 7.5% of "cost" questions exhibit full consensus on the lead actor across models, compared with 51.3% in "catalyst" and 61.3% in "responsibility". In other words, if an analyst were to ask "who should pay?", the recommended lead actor would be much more model-dependent than if the same analyst asked "who should act?" or "who should move first?"

3.2 Framing sensitivity and the incidence mechanism

We next quantify within-model framing sensitivity by comparing, for the same scenario, each model's mean option-support vector across frames. Figure 2 summarizes the resulting L_1 shifts by model and frame-pair.

Framing effects are large and asymmetric. The largest shifts occur between "catalyst" and "cost" (mean $L_1 \approx 37.7$, flip-rate ≈ 0.30), while "catalyst" and "responsibility" are much closer (mean $L_1 \approx 16.2$, flip-rate ≈ 0.09). This is consistent with the prompts activating different economic narratives: coordination and mobilization versus incidence and burden-bearing.

Sensitivity also varies sharply by model. Granite4:3b exhibits the largest catalyst-versus-cost shift (mean $L_1 \approx 71.9$, flip-rate ≈ 0.78), while lfm2.5-thinking:1.2b and ministral-3:8b also display large reallocations. Item-level sensitivity concentrates in politically contested scenarios such as meat reduction under backlash, mandatory climate-risk disclosure, greenwashing with missed net-zero targets, and congestion pricing.

One possible objection is that these shifts could simply reflect stochastic variation rather than genuine framing effects. The auxiliary temperature checks do not support that interpretation. Mean within-frame temperature shifts in centered score space are approximately 9.2 in "responsibility", 12.9 in "cost", and 8.4 in "catalyst", all well below the cross-frame shifts for "catalyst" versus "cost" and "cost" versus "responsibility". The same ranking appears in lead-flip rates: within-frame temperature flips remain below 0.11 in all three frames, versus approximately 0.30 for "catalyst" versus "cost" and 0.25 for "cost" versus "responsibility". Framing therefore appears to matter over and above ordinary sampling noise.

To interpret the content of the framing shifts, we examine the direction of actor reallocation. Aggregated across models and scenarios, moving from "catalyst" to "cost" shifts relative support away from government (mean $\Delta A \approx -10.8$) and toward individuals (mean $\Delta D \approx +8.4$), with smaller average movements for firms and

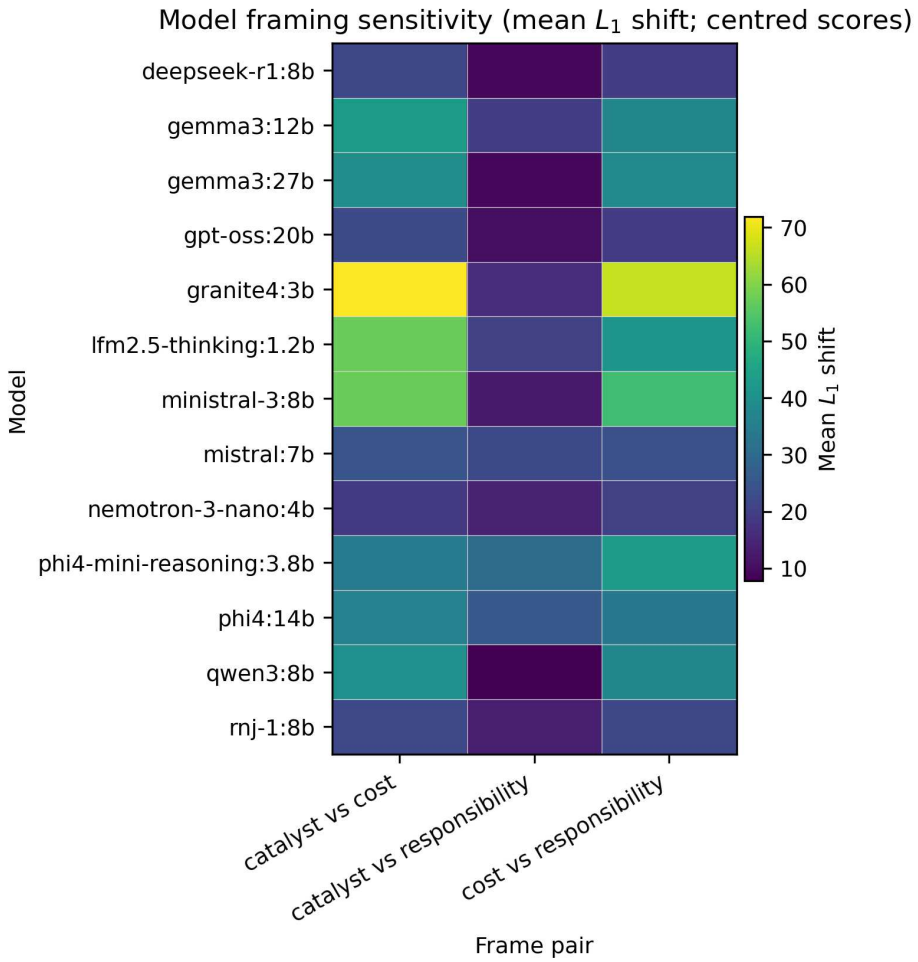


Fig. 2 Model framing sensitivity (mean L_1 shift by frame-pair).

communities. Moving from "cost" to "responsibility" reverses this pattern (mean $\Delta A \approx +10.8$, mean $\Delta D \approx -6.4$). Figure 3 shows that the government–individual contrast is heterogeneous in magnitude across models, but not in sign: 12 of the 13 models move toward government and away from individuals when the frame changes from "cost" to "responsibility".

3.3 Taxonomy of scenario types

The taxonomy analysis shows that disagreement and framing sensitivity are not evenly distributed across sustainability problem types. Figure 4 summarizes cross-model disagreement by taxonomy and frame.

Inn this three distinct patterns stand out. First, "information" questions are the least stable category overall. They have the highest average disagreement (47.6), the

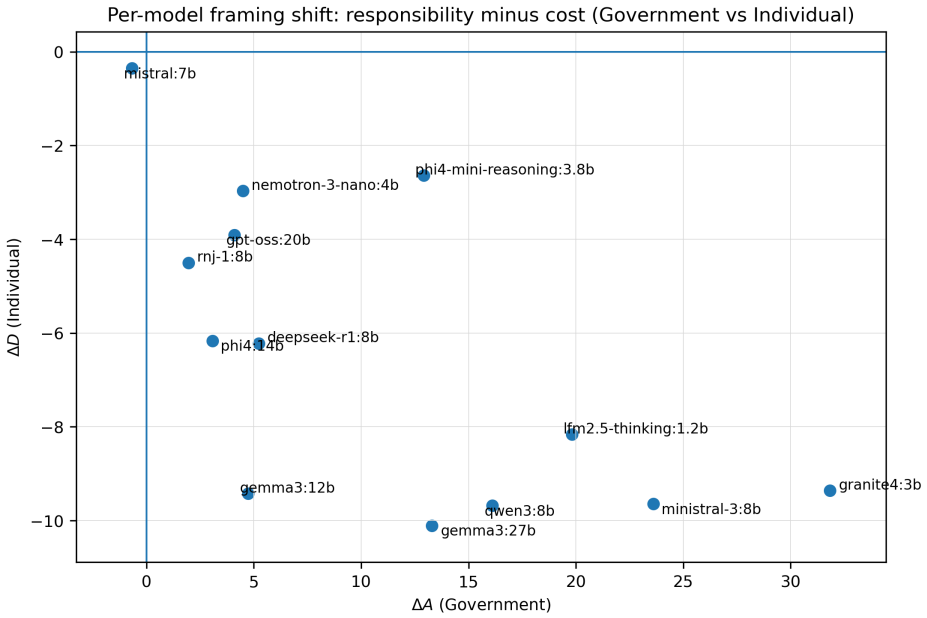


Fig. 3 Per-model framing shift: responsibility minus cost (government vs individual).

largest average framing shift (39.9), and the highest average lead-flip rate (0.327). In the "cost" frame alone, information questions reach mean disagreement of 58.5. These are scenarios such as climate education, misinformation, and evidence-based policymaking, where models appear to disagree not only about allocation but also about the institutional logic of the problem.

Second, "adoption" and "accountability" questions are also highly frame-sensitive. Adoption questions average 33.4 in overall framing shift and accountability questions 29.8, both with comparatively high lead-flip rates. By contrast, "coordination" questions are relatively more stable: they have the lowest average disagreement (34.0) and the lowest overall framing shift (26.3). This suggests that models agree more readily when the problem looks like a public-coordination failure than when it looks like contested incidence or behavioral transition.

Figure 5 makes the framing pattern explicit. The largest cross-frame shifts concentrate in "information" and "adoption", while "coordination" remains comparatively less sensitive across frame pairs.

Third, the cost frame is the least stable within every major taxonomy group. Even when the substantive domain is held fixed, asking who should pay introduces more heterogeneity than asking who should act or who should initiate. Robustness checks excluding borderline-coded items leave this pattern essentially unchanged; for most categories the mean framing-shift estimates move by less than one L_1 point.

Taxonomy also sharpens the actor-reallocation mechanism. Figure 6 shows that moving from "catalyst" to "cost" consistently reallocates weight from government to individuals across all taxonomy groups, but the magnitude varies. The largest

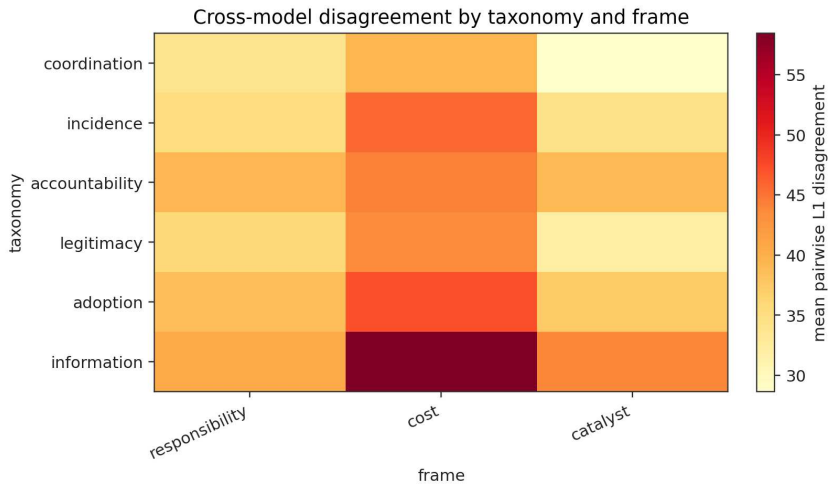


Fig. 4 Cross-model disagreement by taxonomy and frame.

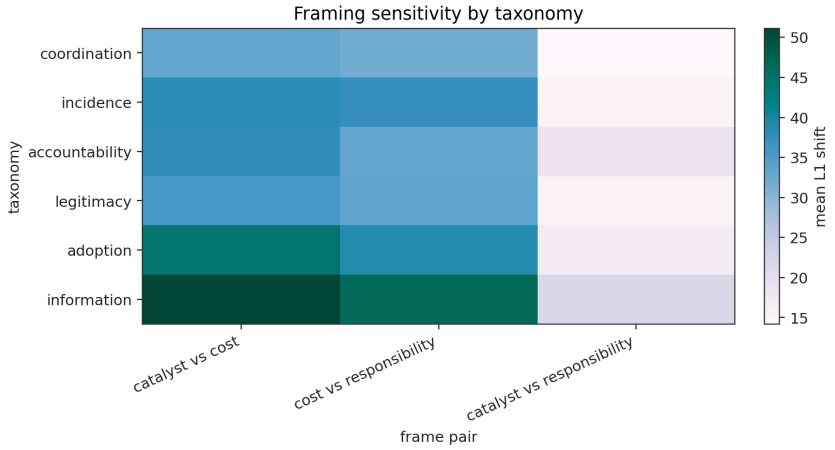


Fig. 5 Framing sensitivity by taxonomy.

government-to-individual shift appears in "adoption" questions ($\Delta A \approx -12.1$, $\Delta D \approx +13.6$), followed by "information" and *incidence*. The reverse movement from "cost" to "responsibility" is strongest in "information" and "incidence", where government support rises by roughly 15.5 and 13.6 points respectively while individual support falls sharply. These are precisely the scenario types in which incidence and legitimacy are most politically contested.

3.4 Semantic divergence in rationales

The short opinion texts indicate that model heterogeneity is not limited to score vectors. Using representative opinions for each frame–question–model cell, the semantic

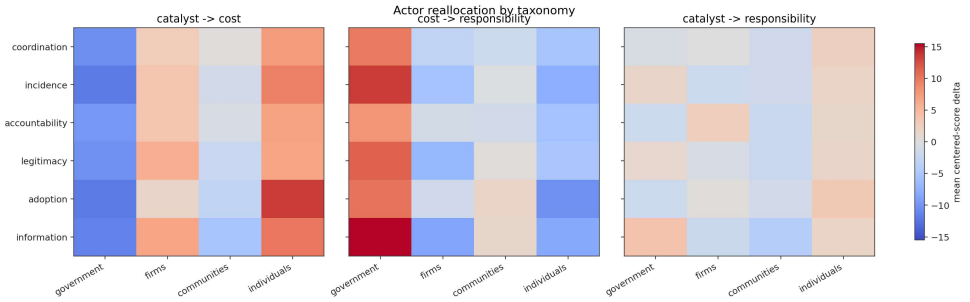


Fig. 6 Actor reallocation by taxonomy and frame-pair.

analysis shows that the language of justification also varies across models. The most extreme between-model semantic distances appear in the "cost" frame, where lfm2.5-thinking:1.2b is a repeated outlier relative to most other models. Within models, the largest semantic frame shifts again occur between "catalyst" and "cost", but the ranking is not identical to the numeric framing results: the largest catalyst-versus-cost semantic shifts occur in qwen3:8b (cosine distance ≈ 0.223) and gpt-oss:20b (≈ 0.183), not in granite4:3b.

Figure 7 visualizes the between-model semantic distances by frame and shows that the most extreme separations are concentrated in the "cost" condition.

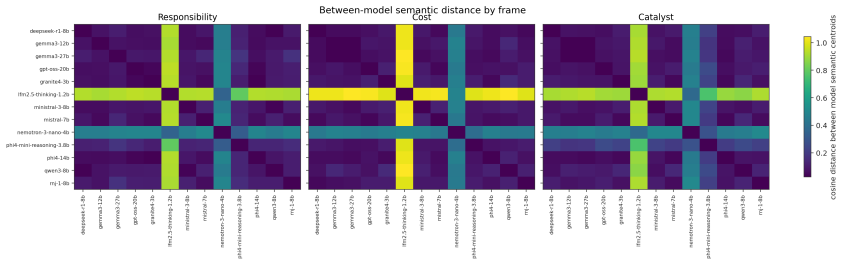


Fig. 7 Between-model semantic distance by frame.

This matters substantively because it suggests two layers of instability. A model can change its allocation across frames, but it can also change its narrative logic. The two do not perfectly coincide. The opinion-text diagnostics also show that explicit textual mention of the top-scored actor is most common in the "cost" frame (mean alignment ≈ 0.72), lower in "catalyst" (≈ 0.65), and lowest in "responsibility" (≈ 0.57). At the same time, some smaller reasoning-oriented models occasionally leak prompt-formatting text into the opinion field, so the semantic evidence is best treated as corroborating structure rather than as a standalone basis for quotation.

4 Discussions

The results support a clear methodological claim: LLM sustainability judgments are neither model-invariant nor frame-invariant, and the instability is economically meaningful rather than random. The "cost" frame is consistently the least stable, both in score space and in decision-risk terms. This is precisely the frame in which applied sustainability analysis is most distributional and politically contested. In that sense, the pattern is substantively coherent. Models look relatively closer to one another when asked who should coordinate or initiate action, but diverge more sharply when asked who should absorb losses, compliance burdens, or redistributed transition costs.

The taxonomy analysis strengthens this interpretation. Instability is not uniform across the questionnaire. Information questions, followed by adoption and accountability problems, are the most disagreement-prone and frame-sensitive. These are also the domains in which the policy object is least reducible to a technical optimization problem: climate literacy, misinformation, behavioral transition, greenwashing, and institutional trust all require background assumptions about legitimacy, persuasion, responsibility, and acceptable coercion. By contrast, coordination problems are comparatively more stable, which is consistent with many models sharing a broad state-capacity intuition when the problem is understood as a system-level public good.

The semantic results point in the same direction while adding an important qualification. Model differences are not exhausted by who receives the highest score. They also appear in the narrative logic used to support those allocations. At the same time, numerical and semantic shifts do not line up perfectly, which suggests that some models can preserve similar allocations while changing justificatory language, whereas others move scores more than reasons. For policy use, this distinction matters: an analyst relying only on the top-ranked actor may miss that different systems are offering materially different political rationales for the same recommendation.

4.1 Limitations and scope conditions

Several limitations should be stated directly.

First, the model set contains only open-weight systems. This is a limitation if the goal is to generalize to the entire frontier of currently deployed LLMs, especially because proprietary models are prominent in real-world advisory settings. At the same time, the open-weight design is a methodological strength for this paper's specific purpose. The models can be executed locally, versioned, and re-run under controlled conditions, which makes the study reproducible in a way that many proprietary APIs do not permit. Local execution is also relevant for the policy motivation of the paper: decision-makers in government and regulated organizations often face data-security constraints, and local or self-hosted systems are therefore not merely a convenience but a realistic deployment mode.

Second, the questionnaire was authored by the researchers rather than taken from an established benchmark, thus question design could shape the results. We agree that instrument design matters, and this is one reason we analyze framing sensitivity explicitly rather than assuming prompt neutrality. The main defense of the instrument is breadth and standardization. The questionnaire spans a wide range of sustainability

domains, repeatedly covers burden-sharing, legitimacy, coordination, and behavioral-transition problems, and applies the same four-actor structure across all 80 scenarios. This makes it useful as a comparative mapping device even if it does not exhaust the full universe of sustainability dilemmas. Still, the results should be interpreted as conditional on this instrument, not as a claim that every sustainability prompt would produce the same pattern.

Third, the three prompt frames are not perfectly orthogonal to the scenario stems. Some stems already contain responsibility- or cost-related wording, and this likely compresses some contrasts, especially between "responsibility" and "catalyst". We therefore interpret the frames as prompt families rather than as clean causal treatments. This limitation does not nullify the observed differences, but it does mean that the magnitudes should not be read as pure framing elasticities.

Fourth, the actor order is fixed in every prompt. This improves comparability across models and frames, but it also means that actor identity and position are not separately identified. Any order effect would therefore enter all conditions in common. The main comparative findings remain informative because the ordering is held constant throughout, but the design does not support claims about order-free absolute support for a given actor.

Fifth, parseability is high overall but not uniform across models. All frame–model cells except one are at least 94.9% parseable under the four-score extraction rule, whereas "phi4-mini-reasoning:3.8b" in the "catalyst" frame is much less format-compliant. This does not drive the main aggregate patterns, but it does mean that some model-level estimates are conditional on successful adherence to the requested output format. In practical terms, format compliance is itself part of model quality for decision-support use.

Sixth, we do not evaluate correctness against an external normative benchmark. That is deliberate: the object of study is inter-model variation in sustainability allocations, not whether one model is "right." But the absence of an external benchmark still limits interpretation. Future work could compare model allocations against expert panels, citizen assemblies, or structured normative criteria such as polluter-pays, ability-to-pay, or subsidiarity.

Seventh, the semantic analysis should be read as supportive rather than definitive. The representative-opinion and embedding approach is useful for detecting broad differences in rationale space, but some smaller reasoning-oriented models occasionally leak formatting instructions or prompt-internal text into the opinion field. We therefore use the semantic results to corroborate structure already visible in the score analysis rather than to establish the main findings on their own.

Finally, the taxonomy is manual. That is a strength in interpretability, but it introduces coder judgment. We partly address this by reporting secondary labels, borderline flags, and robustness checks excluding ambiguous items. Even so, alternative codings are possible, and future work could formalize the taxonomy further with multiple independent coders and explicit inter-coder agreement statistics.

4.2 Implications for AI-assisted policy analysis

Taken together, these limitations define the scope of the claims rather than overturn them. If an organization uses an LLM to structure sustainability recommendations, then model choice and prompt framing should be treated as analytical choices that require disclosure and robustness checking. A single-model, single-prompt workflow can hide large normative variance in who is asked to act, who is asked to pay, and who is expected to move first. The problem is especially acute in burden-bearing and information settings, where disagreement and lead instability are highest.

For decision-makers, the minimum implication is procedural. Report the model family, test multiple prompt frames, inspect disagreement hotspots, and treat high-disagreement outputs as model-risk cases rather than as settled advice. Where data security or auditability matter, locally run open-weight models may be preferable to opaque remote systems, but their outputs still require plural-model validation when the task is normative rather than descriptive.

5 Conclusion

This paper benchmarks 13 open-weight LLMs on an 80-item sustainability questionnaire across three framings: responsibility, cost, and catalyst. The central result is that normative sustainability allocations vary systematically across both models and frames. The "cost" frame is the least stable, information questions are the most disagreement-prone, and reframing shifts support in a consistent government–individual direction that is strongest in adoption, incidence, and information scenarios. Semantic analysis further shows that models differ not only in whom they select but also in how they justify those selections.

The broader implication is straightforward. In sustainability settings involving contested allocations of agency and burden, LLM outputs cannot be treated as interchangeable policy advice. Model choice and prompt framing are themselves part of the result. Any serious use of LLMs in sustainability decision support therefore requires transparency, robustness checks across models and frames, and explicit recognition that these systems embed normative variance rather than merely revealing neutral recommendations.

Supplementary information. Data and code are available among the supplementary materials.

Declarations

Funding: This research received no external funding.
Data: available among the supplementary materials.
Code: available among the supplementary materials
Competing Interests: Authors Declare no competing interests.
Ethics approval and consent to participate: Not applicable

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

Cheung V, Maier M, Lieder F (2025) Large language models amplify human biases in moral decision-making. Proceedings of the National Academy of Sciences https://doi.org/10.1073/pnas.2412015122

Garcia B, Qian C, Palminteri S (2024) The moral turing test: Evaluating human-LLM alignment in moral decision-making. https://doi.org/10.48550/arXiv.2410.07304, arXiv:2410.07304

Germani F, Spitale G (2025) Source framing triggers systematic bias in large language models. Science Advances https://doi.org/10.1126/sciadv.adz2924

Lee S, Peng TQ, Goldberg MH, et al (2024) Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. PLOS Climate 3(8):e0000429. https://doi.org/10.1371/journal.pclm.0000429

Li T, Yu L (2025) ESG beliefs of large language models: Evidence and impact. https://doi.org/10.48550/arXiv.2601.00836, arXiv:2601.00836

Lim K, Kim S, Whang SE (2026) DeFrame: Debiasing large language models against framing effects. https://doi.org/10.48550/arXiv.2602.04306, arXiv:2602.04306

Locatelli MS, Dutenhefner P, Buzelin A, et al (2025) AI and climate change discourse: What opinions do large language models present? In: Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025). Association for Computational Linguistics, URL https://aclanthology.org/2025.climatenlp-1.8/

Tohidi A, Haider S, Watts DJ (2025) Rethinking news framing with large language models. Scientific Reports https://doi.org/10.1038/s41598-025-29519-9

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- promptresponsibility.txt
- promptcatalyst.txt
- promptcost.txt
- questionscombined.csv
- analysis.ipynb
- codesimulationscombined.ipynb
- datacatalyst.csv
- dataresponsibility.csv
- datacost.csv