On Large Language Models as Data Sources for Policy Deliberation on Climate Change and Sustainability

Rachel Bina
rfbina@gmail.com

Kha Luong
khaluong@wharton.upenn.edu

Shrey Mehta
shreym@sas.upenn.edu

Daphne Pang
daphne.y.pang@gmail.com

Mingjun (Kristen) Xie
mingjunx@seas.upenn.edu

Christine Chou
chou@gms.ndhu.edu.tw

Steven O. Kimbrough   (and the Penn Climate Decisions Lab)
sok@upenn.edu

March 24, 2025

Democracy begins with conversation.

—John Dewey at his 90th birthday party[1]

Only that which has no history can be defined.

—Frederick Nietzsche *On the Genealogy of Morals*

The definition of the problem, rather than its solution, will be the scarce resource in the future.

—Esther Dyson (1990s)

Abstract

We pose the research question, *Can LLMs provide credible evaluation scores, suitable for constructing starter MCDM models that support commencing deliberation regarding climate and sustainability policies?* In this exploratory study we

i. Identify a number of interesting policy alternatives that are actively considered by local governments in the United States (and indeed around the world).

ii. Identify a number of quality-of-life indicators as apt evaluation criteria for these policies.

iii. Use GPT-4 to obtain evaluation scores for the policies on multiple criteria.

iv. Use the TOPSIS MCDM method to rank the policies based on the obtained evaluation scores.

v. Evaluate the quality and validity of the resulting ensemble of scores by comparing the TOPSIS-based policy rankings with those obtained by an informed assessment exercise.

We find that GPT-4 is in rough agreement with the policy rankings of our informed assessment exercise. Hence, we conclude (always provisionally and assuming a modest level of vetting) that GPT-4 can be used as a credible input, even starting point, for subsequent deliberation processes on climate and sustainability policies.

Keywords: large language models, LLM, MCDM, climate change, sustainability, biodiversity

---

[1]Lamont, C. (Ed.). (1959). Dialog on John Dewey. Horizon Press. Page 88.

Contents

1 Context

This study occurs in the context of a larger project that is directed at, among other things, two design goals. The first of these goals is about providing deliberation tools for supporting decision making on climate and sustainability policies.

Design Goal 1 (tools). *Construct tools to support deliberation on climate and sustainability policies. These tools may come in any of several forms, including knowledge of methods and practices, as well as databases, document corpora, and software. The list of potential forms is open-ended.*

The second motivating design goal is more specific. It constitutes a way of achieving instances that meet the first goal. The second goal is about MCDM (multi-criteria decision-making) models. These constitute a broad class of modeling methods that are designed and intended to facilitate decision-making in the presence of two or more conflicting goals and objectives. This kind of situation is rife in everyday life, for example, when we make purchasing decisions that have to trade off cost and quality. Policy decisions in climate change and sustainability are more complex, typically having more than two goals or evaluation criteria that have to be considered. The goals are typically in some degree of conflict and themselves complex.

MCDM models, while long established and used, are under-employed. This is in significant part due to the fact that developing a model requires specialized expertise, which may not be available or affordable in particular cases, as well as rather intense and laborious feedback from decision makers, who are busy and often skeptical of yet-to-be-seen tools (Keeney, 1992; Keeney and Raiffa, 1993). This occasions our second design goal.

Design Goal 2 (starter MCDM models). *Deliver to climate and sustainability policy stakeholders MCDM models ("starter MCDM models") that are prima facie acceptable, comprehensible, and suitable for follow on deliberation, revision, and evolution. These initial models should require minimal input from stakeholders, while affording revision in light of new information and stakeholder views when available.*

We foresee being able to achieve Design Goal 2 and thereby being able to afford processes in which technical analysts develop *prima facie* valid (alias "not stupid") starter MCDM models that may be used in public deliberations. Under extreme time constraints, such models would be usable

without substantial revision for decision making. (This is not to say that the model's assessments are simply accepted; rather, the assessments are subject to deliberation with the information available.) In the more usual case, the model is revised and evolved in light of new information and public comments. In this way, an ever-improving, presumably valid knowledge base contributes to ongoing discussions and is available whenever decisions have to be made.

With these goals before us, we hypothesize that large language models (LLMs), such as GPT-4, can be instrumental in creating starter MCDM models that meet Design Goal 2.

Research Question. *Can LLMs provide credible evaluation scores, suitable for constructing starter MCDM models that support commencing deliberation regarding climate and sustainability policies?*

This is all quite aspirational. Certainly, no single study can meet the design goals and dispositively answer the research question. This paper reports on an exploratory study that bears on the research question in the context of the two design goals.

The upshot of our study is to give a positive answer to the research question. We turn now to essential framing and setup for what is to follow.

$$* * *$$

Table 1 is a schema or template for what we shall alternatively call an ACS (<u>a</u>lternatives, evaluation <u>c</u>riteria, <u>s</u>cores) or P (<u>p</u>erformance) table. Such tables are essential data objects when comparing multiple alternatives (the a_i policy alternatives) on multiple dimensions of value (the c_j evaluation criteria). The entries $s_{i,j}$ in the table represent the evaluation scores for their associated a_i and c_j alternatives and evaluation criteria: $s_{i,j}$ is the evaluation score for alternative a_i on criterion c_j.

Policy Alternatives	Evaluation Criteria					
	c_1	c_2	$\ldots$	c_j	$\ldots$	c_n
a_1	$s_{1,1}$	$s_{1,2}$	$\ldots$	$s_{1,j}$	$\ldots$	$s_{1,n}$
a_2	$s_{2,1}$	$s_{2,2}$	$\ldots$	$s_{2,j}$	$\ldots$	$s_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	$s_{i,1}$	$s_{i,2}$	$\ldots$	$s_{i,j}$	$\ldots$	$s_{i,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_m	$s_{m,1}$	$s_{m,2}$	$\ldots$	$s_{m,j}$	$\ldots$	$s_{m,n}$

Table 1: General form—schema or template—for an ACS (alias P) table.

Given a complete performance table, P, decision makers have something definite and useful to work with in deliberating which alternatives to select for implementation. Of course, additional information may be useful and desired. Even so, an ACS (P) table is a fundamental requirement and starting point for serious deliberation. We emphasize *starting point* because modeling and the deliberation it supports should above all be seen as dynamic processes, halted only by practical considerations. Our concern here is with methods for arriving at data and models that are useful because they can contribute to supervening processes. Not final results, but results positioned for getting better results.[2]

This raises the question of how ACS tables of usefully high quality can and should be constructed to support decision making in specific contexts. We are especially concerned in this study with

---

[2]And that can do well enough if necessary.

climate and sustainability policies and with prospects for reducing the workload of scoring P tables. Specifically, this study addresses the paper's Research Question, page 4, §1.

In what follows, we

a. Identify a number of interesting policy alternatives that are actively considered by local governments in the United States (and indeed around the world).

b. Identify a number of quality-of-life indicators as apt evaluation criteria for these policies.

c. Use ChatGPT-4 to obtain scores ($s_{i,j}$ values in Table 1) to complete an ACS (P) table.

d. Evaluate the quality and validity of the resulting P table by comparing its policy rankings with those obtained by an informed assessment exercise using human judgments.

The report is organized into several sections. §2, "Background," extends this introductory section and provides a broader and deeper account of our multiple criteria decision making (MCDM) approach to climate and sustainability policy deliberation. This section may be skipped on first reading.

§3, "Review of the literature," focuses on recent work exploring the validity of responses obtained from large language models, especially GPT. Frequent findings that large language models "hallucinate"—give responses in terms of non–existent entities or are wildly wrong in other ways—are certainly worrisome in the context of our research question. While ultimately we are content to treat large language models as black boxes for our purpose of scoring ACS tables, awareness of systematic failures of these models is certainly relevant to any overall answer to our research question.

§4, "GPT-4 Instructions," describes the process and format used to pose the queries to GPT-4 that produced the scores in our GPT-4 ACS table. This section is supplemented by two appendices, C and D. Appendix C, "Reference Policy Set for the Study," presents and describes the collection of climate and sustainability policies—in terms of the ACS table, the alternatives—we considered in our study. Appendix D, "Reference Evaluation Criteria for the Study," presents and describes nine quality-of-life evaluation criteria used for scoring the alternative policies.

§5, "GPT-4 ACS table," presents and discusses the ACS table resulting from the queries posed regarding performance (on quality of life, mitigation, adaption) to GPT-4.

§6, "MCDM results," presents and discusses application of MCDM modeling to the GPT-4 ACS table. This results in a preference ranking of the policy alternatives under consideration.

§7, "Comparison with human scoring," works with an ACS table covering a superset of the policies in the GPT ACS table. We employed an *informed assessment* process in which members of the Climate Decisions Lab were given background information on the the policy alternatives and the evaluation criteria, then asked to report a score for the corresponding cell in the ACS table. Each cell was scored by multiple people. These scores were averaged to produce the final reported scores. This resulted in which we call the *informed assessment ACS table* or IA ACS table. We applied MCDM modeling as we did for the GPT-4 ACS table to obtain a (modeled) preference ranking of the policy alternatives. We then discuss the ranking and compare it to the ranking obtained from the GPT-4 ACS table. We find substantial agreement between the two rankings. This leads us to conclude—always provisionally—that LLMs can be used for starter or baseline models for policy deliberation, presumably as a prelude to collecting and incorporating additional data from more credible sources.

§8, "Conclusion," concludes the paper with a summary and discussion of the significance of our findings and methods, and suggests directions for future investigation.

2 Background

The "triple planetary crises" of pollution, climate change, and biodiversity loss, present a decision-making challenge—locally as well as globally—that is fraught and enormously complex (Environment, 2021). There are well more than 100 triple planetary crisis plans, programs, and policies (alias "climate and sustainability policies," "policies") known to be available to local governments (counties and municipalities) to address these crises.[3,4] Each policy (type or kind) in turn is associated with multiple program configurations and subject to many distinct local jurisdictions. Complicating matters, in no place can the available policies all be implemented at one time. At best, a portfolio of policies can be funded and managed at a particular time and place and for a particular planning cycle. Which policies should be implemented and in what order? What does implementation actually look like in detail? Who is involved in the decision-making process? Fundamentally, how are policies to be compared and evaluated for purposes of decision making? These questions present enduring challenges, if only because of the computational complexity involved.[5] This last question—posing the *policy evaluation problem* for climate and sustainability policies—is the main focus of the present study.

There are broadly three kinds of justifications used for policy evaluation and adoption. The first is financial advantage. If there is profit to be made or money to be saved with the policy, if a cost-benefit analysis is favorable, this is *ceteris paribus* justification for the policy. Climate adaptation policies are often justified in this way. For example, fire protection policies and building code updates may warrant their expense by prospects of cost avoidance in the future.[6] They serve as a sort of insurance protection against storm damage. Altruism is the second kind of justification. If the cost-benefit analysis of a policy is judged not to favor its adoption, the policy may be undertaken nevertheless because of its overall resulting social good. For example, a policy that reduces greenhouse gas (GHG) emissions or that materially assists the disadvantaged but is costly may be undertaken in the name of climate change on broadly moral grounds. Some individuals behave this way by installing uneconomic rooftop solar PV, for example. They act nobly, altruistically, and prosocially.

Altruism, however, is a vulnerable and unreliable basis for public policy; it is unlikely to be sustainable. Hence, there remains the problem of finding justification, if justification can be had, for seemingly uneconomic but still valuable policies.

The third kind of justification occurs when the policy in question fails or comes out uncertain on a strict cost–benefit basis but is judged to have sufficient co-benefits to overcome a cost–benefit shortcoming (Boyd et al., 2022; Creutzig et al., 2022; Dagnachew and Hof, 2022; Finn and Brockway, 2023; Karlsson et al., 2020; Sharifi, 2021). An example of such a policy might be a ban on single-use plastic bags (yielding co-benefits of reduced litter in the streets and reduced landfill material) or a ban on gasoline-powered leaf blowers (yielding co-benefits of cleaner air, reduced ecological damage, and reduced noise).

The present study centers upon this third kind of co-benefit-based warrant for climate and sustainability policies, and within it on factors that contribute to quality-of-life (alias well-being). Certain aspects of climate and sustainability plans, such as costs, emissions, degree of protection,

---

[3]More than 1500 are identified in (Stechemesser et al., 2024). However, these are tokens of policy types. The paper categorizes these instances into 2–3 score types of policies. Moreover, these are national-level policies, whereas we focus on locally administered and managed policies.

[4]We have assembled more than 100 policies in the spreadsheet file LocalPoliciesDatabase.xlsx, which may be found at https://tinyurl.com/yyfydfnp.

[5]With only ten policies under consideration there are about 16 million possible portfolios to consider.

[6]As we see by recent events in Los Angeles. A similar point can be made about storm water management policies, both for wet and dry climates.

etc., are technical by nature and best evaluated by climate scientists, city planners, engineers, and area experts in general, or obtained through careful assessment of the literature. Other aspects, particularly those focused on how a policy could impact well-being and quality of life in the community, perhaps should not be assessed by policy makers except through consultation with representatives of the general public. These *well-being* aspects of climate and sustainability policies are our main, but not exclusive, concern in this study.

Our study presumes that climate and sustainability policies are usefully—or even necessarily—compared across multiple evaluation criteria and that these criteria are inevitably in conflict, even limiting our attention to quality-of-life and well-being criteria. No policy is always and everywhere the best, and trade-offs are inevitable.

Given this, we further assume the value of constructing multiple criteria decision making (MCDM) models for comparative assessment of climate and sustainability policies. Briefly, MCDM models,[7] whichever of many particular methods are employed, rely fundamentally on a common data structure. We call this the ACS (<u>a</u>ltnatives, evaluation <u>c</u>riteria, and <u>s</u>cores) table, or alternatively the P (<u>p</u>erformance) table. To construct an MCDM model one must at minimum

A. Identify a consideration set of alternatives, $a_i \in \mathcal{A}$, among which choice is to be made.

C. Identify multiple evaluation criteria, $c_j \in \mathcal{C}$, with which to evaluate the alternatives.

S. Evaluate each alternative i on every evaluation criterion j to obtain a score, $s_{i,j}$.

Once an ACS table is constructed, any of several dozen MCDM methods can be applied to gain insight for decision making. Step C, scoring, while challenging, is perhaps the step most plausibly aided by automation and information retrieval on large corpora. The immediate purpose of this study can be framed as addressing the following research question.

Research Question (Repeated from §1, page 4). *Can LLMs provide credible evaluation scores, suitable for constructing starter MCDM models that support commencing deliberation regarding climate and sustainability policies?*

Our study is necessarily exploratory. Having framed the policy evaluation problem as usefully addressed as an MCDM (multiple criteria decision making) problem, in what follows we:

1. Select a consideration set of policies that are salient for local decision makers in America and, we believe, well beyond (Appendix C).

2. Identify nine quality-of-life indicators plus two additional criteria pertinent to climate and sustainability policies (Appendices D and E).

3. Use ChatGPT to assess each policy in the policy consideration set on each evaluation criterion.

4. Apply the TOPSIS MCDM method to the resulting ACS table thereby ranking the policies on the assessed evaluation criteria.

5. Discuss critically how good and how credible the ChatGPT answers are by comparing the results to the results of an informed assessment exercise performed by members of the Climate Decisions Lab.

---

[7]Some example sources among many others: (Alvarez et al., 2021; Arrow and Raynaud, 1986; Ishizaka and Nemery, 2013; Kleinmuntz, 2007; Wu and Tiao, 2018).

2.1 Further points arising

1. The ACS/P table structure is simple and easily understood:

    i. Rows correspond to policy <u>a</u>lternatives. A refers to the collection of alternatives, a_i to a member of that collection.

    ii. Columns correspond to evaluation criteria. C refers to the collection of evaluation <u>c</u>riteria, c_j to a member of that collection.

    iii. ACS table entries represent numerical <u>s</u>cores for the corresponding alternatives and evaluation criteria. The collection of such scores is denoted S and $s_{i,j}$ is an element of S. Ancillary tables or data structures may record non-numeric information pertaining to scores, such as textual comments.

2. The ACS/P table format does not formally assume a linear or additive overall evaluation function on the alternatives. This is, however, an assumption often made when comparing scored alternatives with each other.[8]

3. The $s_{i,j}$'s may be obtained from three main sources:

    (a) Published credible accounts, e.g., from peer reviewed literature. These sources are thin for local policies and well-being.

    (b) Subject matter experts. These sources are also thin for local policies and well-being.

    (c) Stakeholder deliberations and informed assessment exercises. The present study relies on informed assessment for the $s_{i,j}$ scores in the ACS/P table. The informed assessors were well-informed (but not expert) members of the research team, principally graduate and undergraduate students and a few faculty. Informed assessors are in principle always available. What has to be kept in mind is that they may be greatly mistaken or subtly biased. They may do their best, but that will not be perfect. This is an inevitable source of biased and inaccurate data, but it is the best available. We proceed on the basis of a ***draft and revise*** deliberation philosophy.

Our P table for the informed assessment exercise is given in Table 2 on page 19.[9] The scores were arrived at by a group process of *informed assessment*. Assessors received instruction on the policies, and assessments were averaged across two or more assessors after discussion.

3   Review of the Literature

We propose to explore whether large language models, and GPT-4 in particular, can effectively and reliably supply scores in ACS tables for climate and sustainability policies. It has not escaped notice that the 'safety' of GPT responses is often problematic, in that false and even fanciful ("hallucinogenic") responses often occur. This section reviews a sampling of the literature that investigates this issue, which is obviously important in the context of our proposed use of GPT.

---

[8]Given our purpose of overviewing an approach, we defer the technical questions of additivity, preferential independence, etc. to another venue.

[9]The ACS table and the P table are equivalent. When we include the row labels with the P table we call it the ACS table.

3.1 Huang et al., 2024

Huang et al. (2024) is a comprehensive survey of the state-of-the-art of hallucination research on large language models (LLMs), such as LLaMA, Claude, Gemini, and GPT-4. The latter being the chosen tool for our study.

> [H]allucinations in conventional natural language generation (NLG) tasks have been extensively studied ..., with hallucinations defined as generated content that is either nonsensical or unfaithful to the provided source content. These hallucinations are categorized into two types: intrinsic hallucination, where the generated output contradicts the source content, and extrinsic hallucination, where the generated output cannot be verified from the source.

The study categorizes

> hallucination into two primary types: *factuality hallucination* and *faithfulness hallucination*. Factuality hallucination emphasizes the discrepancy between generated content and verifiable real-world facts, typically manifesting as factual inconsistencies. Conversely, faithfulness hallucination captures the divergence of generated content from user input or the lack of self-consistency within the generated content. This category is further subdivided into instruction inconsistency, where the content deviates from the user's original instruction; context inconsistency, highlighting discrepancies from the provided context; and logical inconsistency, highlighting internal contradictions within the content. This categorization refines our understanding of hallucinations in LLMs, aligning it closely with their contemporary usage.

The extent and thoroughness of the survey is considerable. See Figure 1. For present purposes, the main findings may be summarized as follows.

1. Hallucination in LLMs is indeed a pervasive and troubling problem.

2. Factuality hallucinations are more readily detectable and remediable than faithfulness hallucinations. Note: our proposed use of LLMs, for filling in ACS tables, is exposed to both kinds of hallucination.

3. A variety of detection and remediation techniques have been proposed and have been found to be effective to some extent in particular circumstances. There is, however, as yet no normative and definite (exactly correct) general procedure for doing this.

4. The picture that emerges is one of indefinite probing of LLM responses, drawing from an evolving pool of methods, to be halted by pragmatic considerations.

5. In reinforcement of the previous point, we observe that there is very little discussion of Type I versus Type II errors. Presumably, reducing hallucinations (Type I errors) also reduces the scope of returned claims (Type II errors). But see page 24 of Huang et al. (2024) on the "inherent trade-off between diversity and context attribution." Type I–Type II dilemmas are likely inherent in any scheme to reduce hallucination. This too militates in favor of an evolutionary, indefinite probing approach to using results from LLMs, at least in our context.
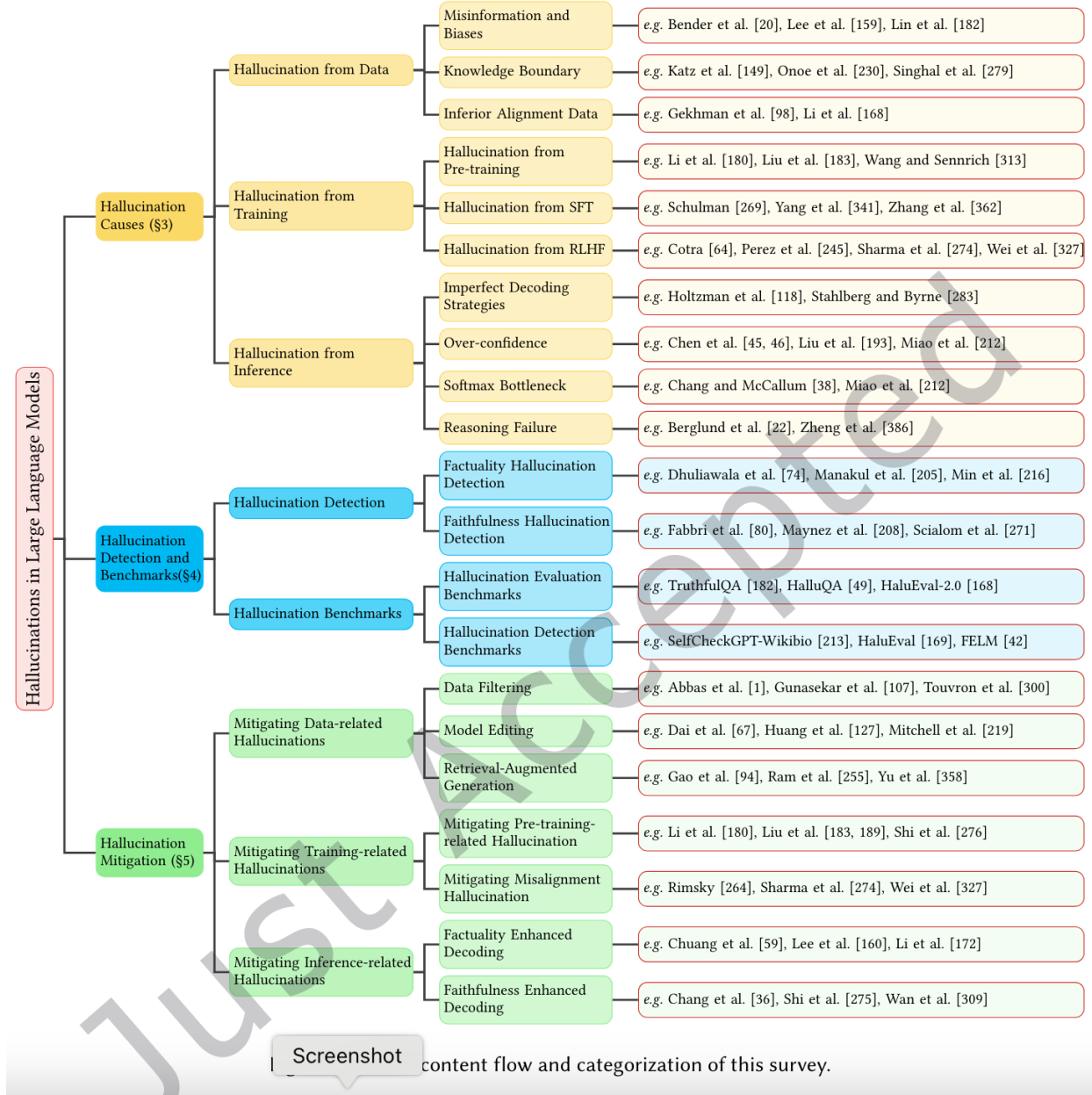
Figure 1: The main content flow and categorization of the survey Huang et al. (2024).

3.2   OpenAI, 2024

OpenAI et al. (2024) is a recent extensive study of hallucination in GPT-4 by a large group at OpenAI. The study documents the considerable accomplishments of GPT-4, for example its scoring the in the top 10% on bar exams, among many other challenging tests. These are summarized in Table 1 of (OpenAI et al., 2024). The study goes on to examine in detail failure rates in different contexts. These results are reproduced from the paper in our Figure 2. Although none of the nine categories examined are identical with ours, it would be prudent and sensible to presume similar performances.

Internal factual eval by category
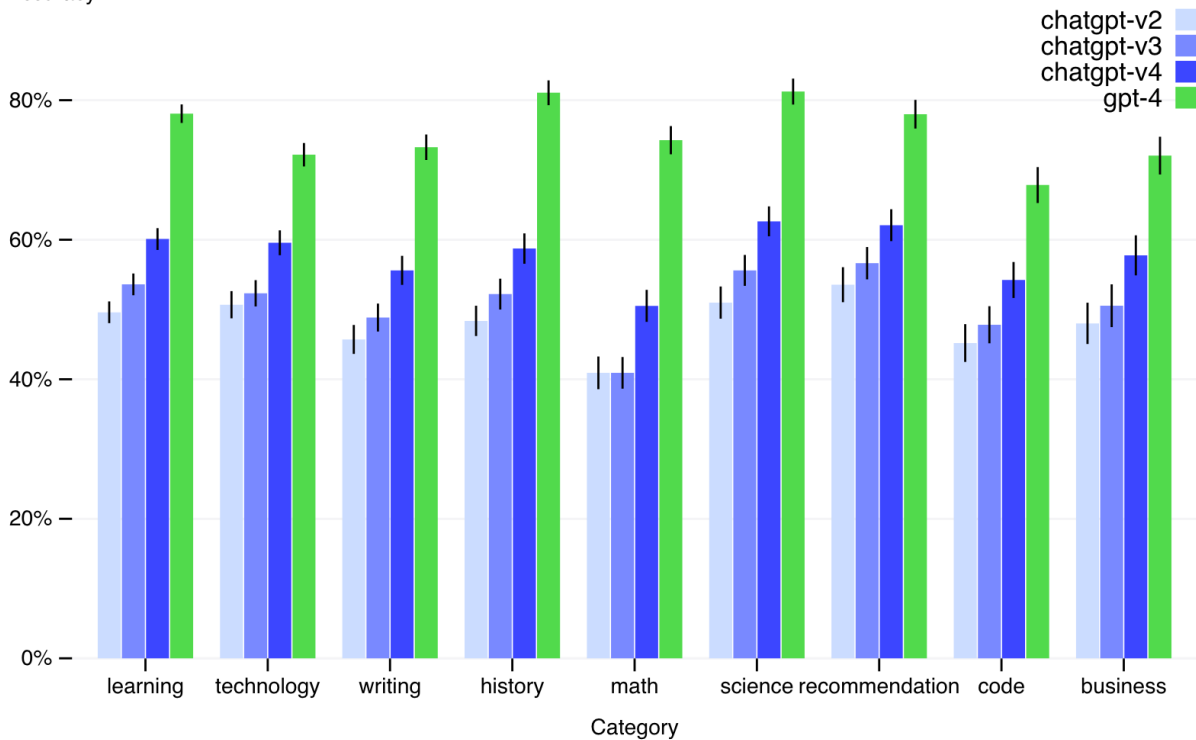
Figure 6. Performance of GPT-4 on nine internal adversarially-designed factuality evaluations. Accuracy is shown on the y-axis, higher is better. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval. We compare GPT-4 to three earlier versions of ChatGPT [64] based on GPT-3.5; GPT-4 improves on the latest GPT-3.5 model by 19 percentage points, with significant gains across all topics.

Figure 2: From (OpenAI et al., 2024).

The written summary/comment on these results is surely well-taken:

> Despite its capabilities, GPT-4 has similar limitations as earlier GPT models. Most importantly, it still is not fully reliable (it "hallucinates" facts and makes reasoning errors). Great care should be taken when using language model outputs, particularly in high-stakes contexts, with the exact protocol (such as human review, grounding with additional context, or avoiding high-stakes uses altogether) matching the needs of specific applications. (OpenAI et al., 2024, page 10)

3.3 Also of note

3.3.1 Chelli et al.

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. Journal of Medical Internet Research, 26(1), e53164. `https://www.jmir.org/2024/1/e53164`

Comparative Analysis

a. Objectives: The aim of the study is to assess the performance of LLMs such as ChatGPT and Bard (subsequently rebranded Gemini) to produce references in the context of scientific writing.

b. Results: In total, 11 systematic reviews across 4 fields yielded 33 prompts to LLMs (3 LLMs×11 reviews), with 471 references analyzed. Precision rates for GPT-3.5, GPT-4, and Bard were 9.4% (13/139), 13.4% (16/119), and 0% (0/104) respectively (P¡.001). Recall rates were 11.9% (13/109) for GPT-3.5 and 13.7% (15/109) for GPT-4, with Bard failing to retrieve any relevant papers (P¡.001). Hallucination rates stood at 39.6% (55/139) for GPT-3.5, 28.6% (34/119) for GPT-4, and 91.4% (95/104) for Bard (P<.001).

These dispiriting findings stand in contrast to the comparatively upbeat results reported in (OpenAI et al., 2024). Unfortunately, the tasks here are likely more similar to our proposed uses than are the tasks in (OpenAI et al., 2024), which probe for established facts.

3.3.2 Li et al., HaluEval, 2023

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models (arXiv:2305.11747). arXiv. `https://doi.org/10.48550/arXiv.2305.11747`

The paper introduces HaluEval, a large-scale benchmark for evaluating hallucination recognition in Large Language Models (LLMs). The benchmark consists of 35,000 samples, including 5,000 human-annotated ChatGPT responses to general queries and 30,000 automatically generated task-specific examples for question answering, knowledge-grounded dialogue, and text summarization. The authors propose a two-stage framework for generating hallucinated samples: sampling-then-filtering. They find that ChatGPT generates hallucinated content in about 19.5% of responses, primarily by fabricating unverifiable information. Experiments show that existing LLMs struggle to recognize hallucinations, with even ChatGPT achieving only 62.59% accuracy in question answering.The study also explores strategies to improve hallucination recognition, such as providing external knowledge and adding reasoning steps.

3.3.3 Dahl et al.

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis, 16(1), 64–93. https://doi.org/10.1093/jla/laae003 arXiv:2401.01301 [cs] http://arxiv.org/abs/2401.01301

a. PDF Access: https://arxiv.org/pdf/2401.01301

b. Objective and Results: This paper develops a typology of legal hallucinations, providing a conceptual framework for future research in this area. Second, it finds that legal hallucinations are alarmingly prevalent, occurring between 58% of the time with ChatGPT 4 and 88% with Llama 2, when these models are asked specific, verifiable questions about random federal court cases. Third, it illustrates that LLMs often fail to correct a user's incorrect legal assumptions in a contra-factual question setup. Fourth, this paper provides evidence that LLMs cannot always predict, or do not always know, when they are producing legal hallucinations.

3.4 Upshot

1. Hallucination problems for LLMs are amply documented and well-established.

2. This summary and comment on these problems for GPT-4 is surely well-taken in general for LLMs:

> Despite its capabilities, GPT-4 has similar limitations as earlier GPT models. Most importantly, it still is not fully reliable (it "hallucinates" facts and makes reasoning errors). Great care should be taken when using language model outputs, particularly in high-stakes contexts, with the exact protocol (such as human review, grounding with additional context, or avoiding high-stakes uses altogether) matching the needs of specific applications. (OpenAI et al., 2024, page 10)

3. Progress has been made in mitigation techniques. These hold promise of materially increasing confidence in results, but on a case-by-case basis.

4. Together, these findings warrant provisional use of LLMs for scoring in ACS tables, in conjunction with vetting results and continued, evolving mitigation.

3.5   Aside: ChatGPT-o1's epistemic status

For what it is worth:

Request posed:

> Please comment on the epistemic status of your replies to questions. How warranted are they? How reliable? You will surely want to qualify your answer by type of question, and by subject matter. (2024-09-27)

Answered with

> My responses aim to be accurate and reliable based on the information available up to my last update in October 2023. The warrant and reliability of my replies can vary depending on the type of question and subject matter:
>
> - **Factual Questions:** For well-established facts in areas like history, science, and general knowledge, my answers are generally dependable. However, I may lack information on events or developments that occurred after my last update.
>
> - **Specialized or Technical Subjects:** In fields that require deep expertise—such as advanced sciences, niche technologies, or specialized academic topics—my knowledge is as accurate as the data I've been trained on. While I strive for precision, there may be gaps or outdated information due to the rapidly evolving nature of these fields.
>
> - **Subjective or Opinion-Based Questions:** For queries involving opinions, ethics, or personal perspectives, I provide balanced viewpoints based on a wide range of sources. These responses are meant to inform rather than persuade and should not be taken as definitive answers.
>
> - **Creative or Open-Ended Queries:** When addressing creative tasks like writing prompts, brainstorming ideas, or exploring hypotheticals, my replies are designed to inspire and engage. These are generated based on patterns and examples in the data and may not reflect unique or groundbreaking ideas.
>
> Overall, while I strive to offer accurate and helpful information across various topics, it's important to consult authoritative sources or professionals for critical decisions or specialized knowledge.

A fair assessment, if not as specific, detailed, and incisive as one might hope.

4   GPT-4 instructions

In querying GPT and LLMs generally, we wish to ask for an evaluation of a specific policy on a specific evaluation criterion. For this, we need a standard name and description of the policy, as well as a name and description of the evaluation criterion. Then we assemble them with a template into a standardized query for ChatGPT.

Here is the template used for all queries with minor modifications:

> Consider first a sustainability policy of <**policy name**>, practiced at the county or municipal of government. <**policy description.**> Consider second, a policy evaluation criterion: <**evaluation criterion name**>. <**evaluation criterion description.**> How would you rate the <**policy name**> policy on a 1 to 10 scale for its capacity to do well on <**evaluation criterion name**> considerations?

See the "Policy Descriptions" chapter of the supplemental materials for the 'official' policy names and descriptions used for prompting ChatGPT with this template.

See Appendix E and the "Quality of Life" chapter of the supplemental material for the 'official' quality-of-life names and descriptions used for prompting ChatGPT with this template. See the "Criteria in Addition to QoL" chapter of the supplemental for the 'official' names and descriptions of other evaluation criteria to be used for prompting ChatGPT. These are for adaptation and mitigation.

See the "Basic query template for LLM querying" section of the Supplementary Material for further information.

Records of the queries posed to and responses given by GPT-4 can be found in Climate Decisions Lab (sok)/Research Projects/LLMs and Policy MCDM/.

5   GPT-4 ACS table

The ACS table shown in Figure 3 is the result of consolidating the ChatGPT-4 queries constructed to fit the template of the previous section.

	B	C	D	E	F	G	H	I	J	K	L	M	N
	ID	Policy Name	Mitigation	Adaptation	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
	2	personal transportation	8	6	7.5	8.5	8.5	7.5	8.5	8.5	8.5	8	8
	3	food rescue programs	5.5	7.5	6.5	5.5	9	2.5	8	7	6	7	4
	4	energy efficiency programs f	9	7	8.5	8.5	9	8	8	7	8	7.5	4.5
	5	complete streets policies	7	8	8	7	9.5	9	9.75	9	9	9.5	9
	6	implementing nature based	9	8	9	7	9	9	8	6	8	7	9
	7	safe streets programs	7	8	9	7	9	8	9	8	9	8	8
	9	green roofs	5	9	7.5	5	9	8.5	8	4.5	8	7	7.5
	13	electrification of municipal v	10	7	9	6	10	8	6	6	6	7	6
	14	residential electric cooking	6	4	8	7	8	7	6	4	6	7	3
	17	resilience hubs	6	9	8	6	9	5	9	7	8	7	7
	20	congestion charges	8	6	8	7	6	7	6	6	7	8	9
	21	plastic bag ban	4	5	8	5	9	9	8	5	7	7	7
	22	gasoline-powered leafblowe	4.5	6	7	7	8	7	7	6	6	6	6

Figure 3: Thirteen polices and their evaluations. File: CurrentDatabase.xlsx, at `https://tinyurl.com/26azw7pp`.

How valid or reliable this ACS table is can only be determined after sufficient probing, *pace* the discussion of mitigation in §3. We note in this regard that we retained all of the ChatGPT responses from which these scores were obtained and the responses uniformly contained information that explained (at least to a degree) why the numerical score was given as it was. For example, we posed the following question in seeking a score for cell D12.

> Leaf blower bans
>
> Consider first a sustainability policy of Gasoline-Powered Leaf Blowers Ban, practiced at the county or municipal of government. Some municipalities are banning or phasing out gas-powered lawn mowers and leaf blowers in favor of electric alternatives to reduce noise, pollution, and greenhouse gas emissions. Consider second, a policy evaluation criterion: adaptation to climate change. Adaptation in the context of climate change refers to the process of adjusting systems (natural, social, or economic) to minimize the negative impacts of climate changes while maximizing potential benefits. It involves understanding the current and projected impacts of climate change and developing strategies to cope with them, ensuring resilience and sustainability. How would you rate the Gasoline-Powered Leaf Blowers Ban policy on a 1 to 10 scale for its capacity to do well on adaptation considerations?

Here is ChatGPT-4's response:

> To evaluate the Gasoline-Powered Leaf Blowers Ban on its capacity to address adaptation to climate change, it is essential to analyze its direct and indirect contributions to resilience and sustainability in the face of climate change. Here's how it aligns:
>
> Strengths
>
> Reduction of Local Air Pollution: Gasoline-powered leaf blowers contribute to smog, particulate matter, and other pollutants that exacerbate respiratory and cardiovascular

issues, especially in heat-stressed urban areas. Reducing this pollution supports public health adaptation. Mitigation Synergy: While primarily a mitigation policy (reducing greenhouse gas emissions), transitioning to electric alternatives fosters cleaner air and reduces heat-trapping emissions, indirectly aiding adaptation strategies like improved air quality in urban heat islands. Encouragement of Sustainable Practices: Switching to electric alternatives aligns with broader sustainability goals, fostering community awareness and behavior change that supports adaptive mindsets. Weaknesses Indirect Link to Adaptation: The policy's primary focus is on mitigation rather than direct adaptation. While it reduces emissions and pollution, it doesn't explicitly address climate resilience, such as urban cooling, stormwater management, or ecosystem restoration. Equity Concerns: Electric alternatives may impose higher upfront costs, potentially creating barriers for lower-income households or small landscapers. Unequal access to resources can hinder overall community resilience. Energy Source Dependence: If electricity for alternatives comes from fossil fuels, the benefits might be partially offset, reducing the effectiveness of the policy in fostering a fully adaptive and sustainable energy system.

Rating: 6/10

The ban contributes to adaptation indirectly through improved air quality and reduced emissions, which are important for building community resilience to climate impacts. However, its primary focus on mitigation and potential equity challenges limit its full effectiveness as a climate adaptation policy.

This is upon review a prima facie sensible and valid answer. Compare it to ChatGPT's answer for climate mitigation:

To rate the Gasoline-Powered Leaf Blowers Ban policy on its capacity to contribute to greenhouse gas mitigation, we need to evaluate its impact in terms of emission reduction compared to other available strategies. Here's a structured breakdown:

Strengths:

Reduction in Fossil Fuel Use: Gasoline-powered leaf blowers emit greenhouse gases, including CO2 and unburned hydrocarbons. Banning them directly cuts these emissions.

Transition to Electric Alternatives: Electric leaf blowers are typically powered by a grid increasingly transitioning to renewable energy sources, resulting in lower lifecycle emissions. Localized Impact: Although small-scale, this policy directly impacts individual behavior and equipment usage within the municipality, contributing to community-wide reductions.

Limitations:

Scope of Emission Reduction: Gas-powered leaf blowers are a relatively minor source of emissions compared to sectors like transportation, energy production, and agriculture.

Dependent on Electricity Source: The mitigation impact depends on the local energy grid's reliance on renewable energy versus fossil fuels. If the grid is coal or gas-heavy, the net mitigation benefit could be lower.

Behavioral Changes Required: The effectiveness depends on the adoption of electric alternatives and proper disposal of old gas-powered equipment to avoid additional environmental impact.

Relative Impact:

While this policy addresses a specific source of emissions and aligns with broader sustainability goals, its mitigation potential is modest compared to other strategies such as transitioning to renewable energy or improving energy efficiency in buildings and transportation. However, it serves as an important symbolic and incremental step toward broader climate action.

Rating:

On a 1 to 10 scale, where 1 is negligible impact and 10 is transformative climate mitigation, the Gasoline-Powered Leaf Blowers Ban policy could be rated at 4 to 5. This reflects its moderate capacity to mitigate emissions, acknowledging both its localized benefits and its relatively limited scale of impact.

With the understanding that GPT is grading on an absolute scale of some sort, we find that on balance the generated explanations redound to the credibility of the scores produced. The scores can be taken as *prima facie* valid, subject revision under necessary subsequent probing.

6   MCDM results

We applied the TOPSIS MCDM method with equal weights on all 11 evaluation criteria (columns C–M in Figure 3). Figure 4 shows the results for three different subsets of the criteria. The column labeled C+All contains TOPSIS scores taking into account all 11 criteria. The column labeled C+Qs contains TOPSIS scores taking into account only the nine quality-of-life criteria (in columns E–M in Figure 3). Finally, the column labeled C+MA contains TOPSIS scores taking into account only the two criteria *not* about quality-of-life (in columns C–D in Figure 3), viz., mitigation and adaptation.

Points arising:

1. Figure 4 is sorted on column C+All in descending (reduced preference) order

2. Column C+Qs is also in descending order, except for policies 6 and 2, indicating very strong agreement between the net evaluations of the quality-of-life criteria and these criteria augmented by the mitigation and adaptation criteria.

3. The C+MA column is quite different from the other two columns in its ranking of the policies. Policy 6, for example, is judged the best, while it is third in C+All and fourth in C+Qs. Policy 13 is even more divergent.

4. Assuming that the ACS table scores are provisionally valid, we find that the rankings in Figure 4 are sufficiently sensible to serve as a point of departure for subsequent deliberations, accompanied by investigation and fine-tuning.

C+				C+ALL	C+Qs	C+MA
ID		Policy Name				
	5	complete streets policies		0.782	0.820	0.637
	7	safe streets programs		0.739	0.764	0.637
	2	personal transportation		0.670	0.703	0.531
	6	implementing nature based solutions		0.669	0.647	0.816
	4	energy efficiency programs for buildings		0.639	0.627	0.703
	17	resilience hubs		0.566	0.553	0.613
	13	electrification of municipal vehicles		0.525	0.487	0.745
	9	green roofs		0.505	0.493	0.549
	20	congestion charges		0.482	0.473	0.531
	21	plastic bag ban		0.439	0.489	0.135
	22	gasoline-powered leafblower ban		0.360	0.377	0.272
	3	food rescue programs		0.359	0.333	0.479
	14	residential electric cooking		0.330	0.349	0.217

Figure 4: TOPSIS results. File: CurrentDatabase.xlsx, at `https://tinyurl.com/26azw7pp`.

7 Comparison with human scoring

Recall that Table 5 (Appendix B, page 24) constitutes our master reference policy list for a series of MCDM studies on climate and sustainability policies.

We conducted an informed assessment exercise to score 21 of the 23 reference list policies on our nine quality-of-life evaluation criteria. In this exercise, members of the Climate Decisions Lab were given information about the several policies under consideration and asked to reflect on them. Individuals scored the policies on the nine quality-of-life criteria, using a 0–to–5, with 0 indicating irrelevance to or even negative impact on quality-of-life. Then the individual scores were averaged across multiple assessments. When multiple scorers were present, they had opportunity to discuss their scores and revise them. Table 2 contains the results of this exercise.

We then applied nine distinct rules or methods for decision making under uncertainty and/or MCDM to this table. The results are shown in Table 3.

ID	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9
0	2.5	1.5	3.5	1.0	0.0	0.0	1.0	1.0	2.0
1	4.0	3.0	4.0	1.5	2.0	0.0	3.5	3.0	3.5
2	4.5	4.0	5.0	3.0	4.0	3.5	4.0	3.5	4.5
3	4.0	2.0	4.5	1.0	4.5	4.0	3.5	2.5	3.5
4	3.0	4.0	3.0	1.0	1.0	0.0	3.0	5.0	4.0
5	5.0	3.0	5.0	4.0	5.0	5.0	4.5	4.5	4.0
6	3.0	2.0	4.0	4.0	3.0	1.0	5.0	3.5	4.0
7	5.0	4.0	4.5	3.5	4.5	5.0	4.0	3.0	5.0
8	5.0	4.0	3.5	5.0	3.0	3.0	3.0	2.5	5.0
9	3.5	2.0	3.5	4.5	3.0	1.0	3.5	1.5	3.5
10	4.0	5.0	4.5	3.0	4.0	5.0	3.0	3.0	4.5
11	1.0	0.0	4.0	0.0	1.0	0.0	4.0	3.0	3.0
12	4.0	4.0	3.5	3.0	4.0	5.0	3.0	4.5	4.0
13	2.0	1.0	3.0	1.0	1.0	0.0	3.5	3.0	4.0
14	4.0	4.2	2.0	4.0	2.0	2.0	2.1	2.3	3.0
15	4.5	4.0	4.0	1.0	2.5	0.0	2.0	3.0	3.5
16	5.0	3.0	5.0	3.5	4.5	4.5	3.5	4.0	4.5
17	5.0	3.0	5.0	0.0	3.0	4.0	3.0	1.5	3.0
18	5.0	4.0	5.0	3.0	3.0	3.0	1.5	2.5	4.0
19	4.0	2.5	3.5	3.0	3.5	3.0	3.5	2.0	3.0
20	3.0	0.0	3.0	2.0	2.0	2.0	2.0	3.0	2.5

Table 2: P (performance, aka ACS) table for well-being data. Row IDs are keyed to the policies in Table 5. Column headings are the evaluation criteria, discussed in Appendix E, for well-being.

ID	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
0	4.5	2.0	4.5	1.5	1.0	6.6	12.5	1.5	1.0
1	4.5	4.5	4.5	8.5	6.0	12.9	16.5	9.0	5.0
2	18.5	16.0	18.5	17.0	17.0	18.9	4.0	20.0	17.0
3	10.0	9.0	10.0	13.5	12.5	15.5	12.5	11.0	11.0
4	4.5	16.0	4.5	8.5	9.0	12.6	20.5	5.0	7.0
5	18.5	16.0	18.5	20.0	21.0	21.0	4.0	18.0	21.0
6	10.0	16.0	10.0	13.5	12.5	15.5	16.5	13.0	13.0
7	18.5	16.0	18.5	20.0	20.0	20.2	4.0	21.0	20.0
8	15.0	16.0	15.0	13.5	15.0	17.8	9.0	17.0	15.0
9	10.0	9.0	10.0	13.5	8.0	13.6	12.5	10.0	9.0
10	18.5	16.0	18.5	17.0	18.0	18.9	4.0	19.0	16.0
11	4.5	4.5	4.5	1.5	3.0	8.4	16.5	1.5	2.0
12	18.5	16.0	18.5	17.0	16.0	18.4	4.0	15.0	18.0
13	4.5	4.5	4.5	3.5	4.0	9.7	16.5	3.0	3.0
14	13.5	7.0	13.5	5.0	5.0	13.4	8.0	6.0	10.0
15	4.5	9.0	4.5	8.5	7.0	12.9	19.0	8.0	6.0
16	18.5	16.0	18.5	20.0	19.0	19.7	4.0	16.0	19.0
17	4.5	16.0	4.5	8.5	11.0	14.4	20.5	7.0	8.0
18	12.0	16.0	12.0	8.5	14.0	16.3	12.5	14.0	14.0
19	13.5	4.5	13.5	8.5	10.0	14.7	4.0	12.0	12.0
20	4.5	1.0	4.5	3.5	2.0	10.2	10.0	4.0	4.0

Table 3: E (evaluations) table. Decision rule scores for the P (performance or ACS) table. Data within columns are ranked (but based on scaled values in SAW, PROMETHEE, and TOPSIS). Produced by mcdm_library_climate_wellbeing.ipynb. D_1: maximin returns D_2: maximax returns D_3: minimax regret D_4: maximize medians D_5: lengths D_6: SAW D_7: Hurwicz D_8: PROMETHEE D_9: TOPSIS.

Finally, we aggregated the scores (for the policies) in the rows of Table 3 using three methods: the Borda count (row sums), the median, and the averaged rank median. The results are given in Table 4. The table is sorted by Borda count scores, but we note the strong agreement across the three measures.

	Policy Name	V_1 Borda (Row Sums)	V_2 Simple Median	V_3 Averaged Rank Median
7	Safe Streets programs	158.21	20.00	20.00
5	Complete Streets Policy	158.00	18.50	18.50
16	Active mobility pathways (alias micro-mobility)	150.69	18.50	18.50
2	Personal transportation	146.90	17.00	17.00
10	Public transit	145.90	18.00	17.50
12	Transit Oriented Design	141.38	17.00	16.00
8	Planting and nurturing of shade trees	133.35	15.00	15.00
6	Nature based climate and sustainability solutions	119.99	13.00	13.00
18	Facility amenities	119.28	14.00	14.00
3	Food rescue programs	104.99	11.00	11.00
9	Green roofs	95.65	10.00	10.00
17	Resilience hubs	94.44	8.50	8.50
19	Community gardens	92.70	12.00	11.00
4	Energy efficiency programs for buildings	87.60	8.50	7.00
14	Residential electric cooking	81.44	8.00	8.00
15	Electrification of grounds care equipment	79.36	8.00	7.00
1	Residential space heating and cooling with heat pumps	71.36	6.00	6.00
13	Electrify Municipal Vehicles	53.21	4.50	4.00
11	Municipal composting	46.40	4.50	3.00
20	Congestion charges	43.74	4.00	4.00
0	Residential hot water heating with heat pumps	35.06	2.00	1.50

Table 4: A (aggregation) table. Three robust aggregations of the E table.

8 Conclusion

8.1 Comparing the GPT-4 and informed assessment results

Comparing the Borda count results in Table 4 with the TOPSIS results in Table 4 we see the following.

1. There is good but hardly exact agreement between the GPT-4 rankings (under TOPSIS) and the Borda count rankings with TOPSIS as a component in essential agreement.

2. Specifically, the GPT-4 ranking for the policies evaluated is 5, 7, 2, 6, 4, 17, 13, 9, 20, 21, 22, 3, 14.

3. Informed assessment ranking of these policies (where evaluated) is 7, 5, 2, 6, 3, 9, 17, 4, 14, 13, 20. (Policies 21 and 22 were not evaluated by the informed assessment exercise. The TOPSIS alone ranking is 5, 7, 2, 6, 3, 14, 9, 17, 4, 20, 13; see Table 3.)

4. Focusing on the eleven policies in common, the top four are

   (a) GPT TOPSIS: 5, 7, 2, 6
   (b) Informed assessment overall: 7, 5, 2, 6
   (c) Informed assessment TOPSIS: 5, 7, 2, 6

5. The bottom four are

   (a) GPT TOPSIS: 9, 20, 3, 14
   (b) Informed assessment overall: 4, 14, 13, 20
   (c) Informed assessment TOPSIS: 17, 4, 20, 13

   Again, decent agreement.

   In sum, recall our research question:

Research Question. *Can large language models and GPT in particular provide credible scores in ACS tables for climate and sustainability policies?*

In this exploratory study we find that GPT-4 is in rough agreement with the policy rankings of informed assessment. Hence, we conclude (always provisionally and assuming a modest level of vetting) that GPT-4 can be used as a credible input, even starting point, for subsequent deliberation processes on climate and sustainability policies.

8.2 Discussion

Recall our goals and research question.

Design Goal 1 (tools). *Construct tools to support deliberation on climate and sustainability policies. These tools may come in any of several forms, including knowledge of methods and practices, as well as databases, document corpora, and software. The list of potential forms is open-ended.*

Design Goal 2 (starter MCDM models). *Deliver to climate and sustainability policy stakeholders MCDM models ("starter MCDM models") that are prima facie acceptable, comprehensible, and suitable for follow on deliberation, revision, and evolution. These initial models should require minimal input from stakeholders, while affording revision in light of new information and stakeholder views when available.*

Research Question. *Can LLMs provide credible evaluation scores, suitable for constructing starter MCDM models that support commencing deliberation regarding climate and sustainability policies?*

As noted above, we find a positive answer to our research question. Using two approaches, one based on populating ACS tables using informed judgments and one using ChatGPT, we produced two *prima facie* MCDM starter models for preferential ranking of a number of policies. We found, moreover, that the two approaches are in reasonable agreement with each other. The upshot of this is that we arguably have a Design Goal 2 success example, and with that some positive steps toward meeting Design Goal 1. Put most briefly, we have demonstrated use of LLMs for construction of a starter MCDM model for policy evaluation, and we can expect such starter models will facilitate policy deliberation on climate change and sustainability. To this end we also have produced artifacts useful for deliberation and decision making:

1. Goodness rankings of climate and sustainability policies, taking into account quality-of-life, climate mitigation, and climate adaptation.

2. ACS tables of scores upon which the goodness rankings can be produced using various methods.

3. A spreadsheet representation of the ACS tables and the TOPSIS MCDM method applied to them, affording thereby model analysis and revision in light of new information.

* * *

We emphasize our ***draft and revise*** philosophy for modeling and deliberation (§2). There are no real alternatives to it when, as in the present case of climate and sustainability policies, we have a dynamic environment in which the availability and credibility of relevant data is constantly in flux, is forever being contested, and is subject to continual need for revision in light of developments. Exacerbating the data problems, modeling and analyses based on them are concomitantly subject to revision. In short, we have the Nietzscheian condition of historicity and with it a fundamental barrier to definitional fixation.

How to proceed? The findings of this study suggest and support an approach that seeks a credible starting point (draft), followed by continual exploration, deliberation, and information collection, leading to revision and a new draft. The process is then repeated indefinitely. Decision are taken when necessary, based on the admittedly developing draft and assessment at the time of its reliability.

This process description should, we think, meet with general ascent for all inquiry. The upshot of this study concerns the specific, particular, and concrete matter of climate and sustainability policy making. We find that LLMs and informed assessment exercises applied in MCDM modeling plausibly serve as apt points of departure for *prima facie* bases for decision making and for follow on deliberation and investigation. These tools do in the present case and can more generally be used with confidence to produce first draft decision models and to support continued investigation.

A   Supplementary Material

The *Climate and Sustainability Policies Sourcebook* and other documents referenced in this report can be found at `https://tinyurl.com/2yfe88st`.

B   Master reference policy list

Table 5 constitutes our master reference policy list for a series of MCDM studies on climate and sustainability policies. This set was curated by the Climate Decisions Lab from the master database of climate and sustainability policies. These are assembled in the spreadsheet file LocalPolicies-Database.xlsx, which may be found at `https://tinyurl.com/yyfydfnp`. The basis for this subjective curation (by many people) was the combined likelihood of being considered by municipalities, significance for mitigation of or adaptation to climate change, and effect on quality-of-life (alias well-being).

ID	Policy
0	Residential hot water heating with heat pumps
1	Residential space heating and cooling with heat pumps
2	Personal transportation (EVs, scooters, e-bikes, etc.; supporting infrastructure such as charging stations, extra parking, etc.)
3	Food rescue programs
4	Energy efficiency programs for all kinds of buildings, e.g., commercial, industrial, non-profit, and residential
5	Complete Streets Policy
6	Nature based climate and sustainability solutions (Includes storm water management, bio-swales, wetlands, green planted areas for storm buffering, community forests (alias tiny forests), etc.)
7	Safe streets programs (Includes safe routes to schools for active mobility students.)
8	Planting and nurturing of shade trees
9	Green roofs, for non-residential as well as residential buildings
10	Public transit
11	Municipal composting
12	Transit Oriented Design
13	Electrification of municipal vehicles
14	Residential electric cooking
15	Electrification of grounds care equipment (grass cutting, hedge trimming, leaf blowing, etc.)
16	Active mobility pathways (alias micro-mobility). (For walking, biking, etc. Facilitating both recreational use and utilitarian use, e.g., access to service sites such as schools, shopping, restaurants, personal services.)
17	Resilience hubs
18	Facility amenities (bike racks at bus stops, shelters at bus stops and other destinations, etc.)
19	Community gardens
20	Congestion charges
21	Plastic bag ban
22	Gasoline-powered leaf blower ban

Table 5: Policy support consideration set, 23 in all. ID: index, unique identifier. Policy: name.

C Reference Policy Set for the Study

We essay to evaluate and compare policies *available to municipalities* on the basis of their contributions to quality of life and well-being. Our ultimate purpose is to identify for decision makers and the general public policy opportunities (and threats) to well-being. With such information to hand, policies that fail a standard cos-benefit analysis may on balance be deemed advisable. Moreover, policies that fare well on cost-benefit assessment, may be productively illuminated on the basis of quality of life.

We conducted Internet searches to assemble a list of more than 120 climate and sustainability policies that are in fact considered if not implemented by local governments (counties and municipalities in the USA). These are assembled in the spreadsheet file LocalPoliciesDatabase.xlsx, which may be found at `https://tinyurl.com/yyfydfnp`. Drawing on that database, the study team selected a subset of apparently promising policies to be targets of this study. The list is as follows:

1. personal transportation

2. food rescue programs

3. energy efficiency programs for buildings

4. complete streets policies

5. implementing nature based solutions

6. safe streets programs

7. green roofs

8. electrification of municipal vehicles

9. residential electric cooking

10. plastic bag ban

11. gasoline-powered leaf blower ban

12. resilience hubs

13. congestion charges

D Reference Evaluation Criteria for the Study

These are (climate) mitigation and adaptation, (see "Sustainability Policies Sourcebook" in the Supplementary Materials), plus the nine quality-of-life criteria described in Appendix E.

Mitigation prompt:

In the context of climate change, mitigation refers to efforts to reduce or prevent the emission of greenhouse gases (GHGs) into the atmosphere, thereby minimizing the pace and impact of global warming. The goal of mitigation is to address the root causes of climate change by reducing emissions or enhancing "sinks" (natural or artificial processes that absorb more carbon than they emit, such as forests, oceans, or carbon capture technologies). Mitigation strategies can include: Transitioning to Renewable Energy: Replacing fossil fuels with renewable energy sources like wind,

solar, and hydroelectric power to reduce carbon emissions. Energy Efficiency Improvements: Enhancing the efficiency of appliances, buildings, and industrial processes to use less energy. Carbon Sequestration: Capturing and storing CO_2 through techniques like reforestation, soil management, or carbon capture and storage (CCS) technologies. Sustainable Transportation: Shifting to electric vehicles, public transportation, and other low-emission modes of transportation. Changes in Agriculture and Land Use: Adopting sustainable farming practices and protecting or restoring forests and wetlands to maintain or increase natural carbon sinks.

Adaptation prompt:

Adaptation in the context of climate change refers to the process of adjusting systems (natural, social, or economic) to minimize the negative impacts of climate changes while maximizing potential benefits. It involves understanding the current and projected impacts of climate change and developing strategies to cope with them, ensuring resilience and sustainability.

E   Quality-of-life evaluation criteria

As supplemental material for this study, the study team (alias Climate Decisions Lab at the University of Pennsylvania) developed a report entitled *Climate and Sustainability Policies Sourcebook*. The *Sourcebook* may be found at: `https://tinyurl.com/2cw757ew`. We will draw upon it multiple times in this report. Typically, we excerpt the *sourcebook* for the sake of brevity in this report.

The study team undertook iterative literature reviews, brainstorming, and critical discussions to arrive at a list of nine quality of life criteria we found to be relevant to climate and sustainability policies. The list is as follows.

E.1   Q-1: Health, safety, and hygiene.

E.1.1   Characterization

Q_1: Health, safety, and hygiene. In the main, health considerations arise with (indoor or outdoor) air pollution, and amenities for exercise, active mobility, and recreation; safety mainly comes up with avoiding accidents, e.g., during transit; hygiene issues arise in multiple ways, perhaps most often in the design of parks and other public spaces. Both physical and mental health fall under this criterion.

E.1.2   Version for prompting

Health, safety, and hygiene. In the main, health considerations arise with (indoor or outdoor) air pollution, and amenities for exercise, active mobility, and recreation; safety mainly comes up with avoiding accidents, e.g., during transit; hygiene issues arise in multiple ways, perhaps most often in the design of parks and other public spaces. Both physical and mental health fall under this criterion.

E.2   Q-2: Time, attention and convenience.

E.2.1   Characterization

Q_2: Time, attention and convenience. Policies that economize on time or relieve people of having to maintain an item or that in some way relieve people of inconveniences score highly. Policies that may increase discretionary time for people should score highly. Conversely, policies

that may reduce discretionary time should receive very low scores (Goodin et al., 2008). Discretionary time may be characterized as follows:

> The time beyond that necessary to attend to necessary functions is yours to use as you please. That is what we will call 'discretionary time'. That is how much 'temporal autonomy' you possess. (Goodin et al., 2008, page 5)

To illustrate, some respondents find electric vehicles very attractive on this criterion, typically because home charging and reduced maintenance for electric vehicles frees up time that had to be spent on internal combustion engine (ICE) vehicles for refueling and scheduled maintenance. Other respondents report to the contrary. Typically this is because they lack access to convenient charging and so would need to commit more time to refueling an electric vehicle than an ICE vehicle.

To take another example, a policy that reduced commuting times would perforce be increasing discretionary time for the commuters affected.

E.2.2 Version for prompting

Time, attention and convenience. Policies that economize on time or relieve people of having to maintain an item or that in some way relieve people of inconveniences score highly. Policies that may increase time for people to use as they please should score highly. Conversely, policies that may reduce freely disposable time should receive very low scores.

E.3 Q-3: Moral considerations.

E.3.1 Characterization

Q_3: Moral considerations. This dimension of evaluation is meant to record broadly moral or ethical aspects of the policy in question. These ethical considerations may constitute a reason why the policy should be implemented, or they may enter as a consequence of implementing the policy. The single criterion links with the flourishing moral foundations literature (empirical study of ethics and morality), compressing several dimensions into one. See Adger et al. (2017); Haidt (2012); Culiberg et al. (2023). We were surprised when these considerations arose spontaneously and fairly frequently in a number of semi-structured interviews we conducted, both in Taiwan (described above) and in the U.S. For this reason, we include moral considerations in our well-being criteria.

E.3.2 Version for prompting

Moral considerations. This dimension of evaluation is meant to record broadly moral or ethical aspects of the policy in question. These ethical considerations may constitute a reason why the policy should be implemented, or they may enter as a consequence of implementing the policy. The single criterion links with the flourishing moral foundations literature (empirical study of ethics and morality), compressing several dimensions into one.

E.4 Q-4: Aesthetics

E.4.1 Characterization

Q_4: Aesthetics. Policies score favorably if they result in improvements in how aesthetically pleasing the local environs are. Aesthetic concerns were often mentioned in interviews and the online survey comments.

E.4.2 Version for prompting

Aesthetics. Policies should be scored favorably if they result in improvements in how aesthetically pleasing the local environs are.

E.5 Q_5: Community strengthening.

E.5.1 Characterization

Q_5: Community strengthening. Policies score favorably if their implementation would tend to build stronger, more cohesive communities, fostering a sense of togetherness. This could be expected to happen, for example, with policies that afford activities and interactions in and among the public. We are assuming the policy is well or at least adequately implemented.

E.5.2 Version for prompting

Community strengthening. Policies score favorably if their implementation would tend to build stronger, more cohesive communities, fostering a sense of togetherness. This could be expected to happen, for example, with policies that afford activities and interactions in and among the public. It should be assumed that the policy is well or at least adequately implemented.

E.6 Q_6: Access to service points.

E.6.1 Characterization

Q_6: Access to service points. Service points include just about any location that would be a destination for a trip, such as grocery stores, schools, places of work, pharmacies, libraries, and retail outlets generally. A policy scores well to the extent that it improves ease of access to services or is itself a service to which access is valuable.

E.6.2 Version for prompting

Access to service points. Service points include any location that would be a destination for a trip, such as grocery stores, schools, places of work, pharmacies, libraries, and retail outlets. A policy scores well to the extent that it improves ease of access to services or is itself a service to which access is valuable.

E.7 Q_7: Affordance of capabilities.

E.7.1 Characterization

Q_7: Affordance of capabilities. Policies that score well make positive contributions to the exercise and development of human capabilities, including physical (exercise), mental, and social capa-

bilities. Prototypically, this can be done through education, learning, knowledge acquisition, and enriching, uplifting, broadly educative experiences.

E.7.2 Version for prompting

Affordance of capabilities. Policies that score well make positive contributions to the exercise and development of human capabilities, including physical (such as exercise and athletics), mental, and social capabilities. This is typically done through education, learning, knowledge acquisition, and generally through enriching, uplifting, broadly educative experiences.

E.8 Q_8: Economy

E.8.1 Characterization

Q_8: Economy. Policies that score well on economy make positive contributions to saving money, or they stimulate employment, or they attract investment, or they tend to increase tax revenues (without raising tax rates), and so on.

E.8.2 Version for prompting

Economy. Policies that score well on economy make positive contributions to saving money, or they stimulate employment, or they attract investment, or they tend to increase tax revenues (without raising tax rates), and so on.

E.9 Q_9: Improving and creating destinations.

E.9.1 Characterization

Q_9: Improving and creating available destinations. Q_6 is about improving access to existing destinations. Q_9 is about improving the collection of destinations worth accessing. Policies that score well make positive contributions to creating or improving destinations or to making certain destinations more valuable or available to the public. For example, a nature-based storm water management policy may lead to the creation or improvement of a natural setting that would be suitable for recreational purposes at least much of the time. The destinations attribute is distinct from the accessibility attribute (access to service points, Q_6). Accessibility is about easing travel to a fixed destination. The destinations attribute is about making it more attractive to bother traveling to a place.

E.9.2 Version for prompting

Improving and creating available destinations is about enhancing the collection of destinations worth accessing. Policies that score well make positive contributions to creating or improving destinations or to making certain destinations more valuable or available to the public. For example, a nature-based storm water management policy may lead to the creation or improvement of a natural setting that would be suitable for recreational purposes at least much of the time. To take another example, measures to improve safety and pleasantness of pedestrian access may improve the attractiveness of a retail corridor.

References

Adger, W. N., Butler, C., and Walker-Springett, K. (2017). Moral reasoning in adaptation to climate change. *Environmental Politics*, 26(3):371–390.

Alvarez, P. A., Ishizaka, A., and Martínez, L. (2021). Multiple-criteria decision-making sorting methods: A survey. *Expert Systems with Applications*, 183:115368.

Arrow, K. J. and Raynaud, H. (1986). *Social Choice and Multicriterion Decision-Making*. The MIT Press, Cambridge, Massachusetts.

Boyd, D., Pathak, M., van Diemen, R., and Skea, J. (2022). Mitigation co-benefits of climate change adaptation: A case-study analysis of eight cities. *Sustainable Cities and Society*, 77:103563.

Creutzig, F., Niamir, L., Bai, X., Callaghan, M., Cullen, J., Díaz-José, J., Figueroa, M., Grubler, A., Lamb, W. F., Leip, A., Masanet, E., Mata, É., Mattauch, L., Minx, J. C., Mirasgedis, S., Mulugetta, Y., Nugroho, S. B., Pathak, M., Perkins, P., Roy, J., de la Rue du Can, S., Saheb, Y., Some, S., Steg, L., Steinberger, J., and Ürge-Vorsatz, D. (2022). Demand-side solutions to climate change mitigation consistent with high levels of well-being. *Nature Climate Change*, 12(1):36–46. Number: 1 Publisher: Nature Publishing Group.

Culiberg, B., Cho, H., Koklic, M. K., and Zabkar, V. (2023). The Role of Moral Foundations, Anticipated Guilt and Personal Responsibility in Predicting Anti-consumption for Environmental Reasons. *Journal of Business Ethics*, 182:465–481.

Dagnachew, A. G. and Hof, A. F. (2022). Climate change mitigation and SDGs: modelling the regional potential of promising mitigation measures and assessing their impact on other SDGs. *Journal of Integrative Environmental Sciences*, 19(1):289–314. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/1943815X.2022.2146137.

Environment, U. N. (2021). Making Peace With Nature; UNEP - UN Environment Programme. Section: publications.

Finn, O. and Brockway, P. E. (2023). Much broader than health: Surveying the diverse co-benefits of energy demand reduction in Europe. *Energy Research & Social Science*, 95:102890. https://www.sciencedirect.com/science/article/pii/S2214629622003930.

Goodin, R. E., Rice, J. M., Parpo, A., and Erikson, L. (2008). *Discretionary Time: A New Measure of Freedom*. Cambridge University Press, Cambridge, UK.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon, New York.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* https://dl.acm.org/doi/10.1145/3703155.

Ishizaka, A. and Nemery, P. (2013). *Multi-Criteria Decision Analysis: Methods and Software*. John Wiley & Sons, Inc., New York.

Karlsson, M., Alfredsson, E., and Westling, N. (2020). Climate policy co-benefits: a review. *Climate Policy*, 20(3):292–316. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14693062.2020.1724070.

Keeney, R. L. (1992). *Value-Focused Thinking: A Path to Creative Decisionmaking.* Harvard University Press, Cambridge, MA.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with multiple objectives : preferences and value tradeoffs.* Cambridge University Press, Cambridge, UK.

Kleinmuntz, D. N. (2007). Resource Allocation Decisions. In Edwards, W., Miles, M., and von Winterfeldt, D., editors, *Advances in Decision Analysis.* Cambridge University Press.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R.,

Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). GPT-4 Technical Report. arXiv:2303.08774 [cs] https://arxiv.org/pdf/2303.08774.

Sharifi, A. (2021). Co-benefits and synergies between urban climate change mitigation and adaptation measures: A literature review. *Science of The Total Environment*, 750:141642.

Stechemesser, A., Koch, N., Mark, E., Dilger, E., Klösel, P., Menicacci, L., Nachtigall, D., Pretis, F., Ritter, N., Schwarz, M., Vossen, H., and Wenzel, A. (2024). Climate policies that achieved major emission reductions: Global evidence from two decades. *Science*, 385(6711):884–892. Publisher: American Association for the Advancement of Science https://www.science.org/doi/10.1126/science.adl6547.

Wu, J.-Z. and Tiao, P.-J. (2018). A validation scheme for intelligent and effective multiple criteria decision-making. *Applied Soft Computing*, 68:866–872.