AI and Climate Change Discourse: What Opinions Do Large Language Models Present?

Marcelo Sartori Locatelli, Pedro Robles Dutenhefner, Arthur Buzelin,
Pedro Alzamora, Yan Aquino, Pedro Bento, Samira Malaquias,
Victoria Estanislau, Caio Santana, Lucas Dayrell,
Marisa Affonso Vasconcelos, Wagner Meira Jr., Virgilio Almeida
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{locatellimarcelo, arthurbuzelin, pedro.loures, yanaquino, pedro.bento,
samiramalaquias, victoria.estanislau, caiosantana, lucasdayrell,
marisavasconcelos, meira, virgilio}@dcc.ufmg.br, pedroroblesduten@ufmg.br

Abstract

Large Language Models (LLMs) are increasingly used in applications that shape public discourse, yet little is known about whether they reflect distinct opinions on global issues like climate change. This study compares climate change-related responses from multiple LLMs with human opinions collected through the People's Climate Vote 2024 survey (UNDP – United Nations Development Programme and Oxford, 2024). We compare country and LLM's answer probability distributions and apply Exploratory Factor Analysis (EFA) to identify latent opinion dimensions. Our findings reveal that while LLM responses do not exhibit significant biases toward specific demographic groups, they encompass a wide range of opinions, sometimes diverging markedly from the majority human perspective.

1 Introduction

Climate change is one of the most pressing global challenges of our time, shaping policy decisions, influencing public behavior, and driving scientific inquiry (Lahsen and Ribot, 2022). Public opinion plays a critical role in guiding both governmental decisions and societal responses, making its assessment indispensable for understanding the support and resistance that can influence policy effectiveness and climate action. In this context, surveys serve as fundamental tools, providing valuable insights into the diverse perspectives shaping climate discourse and enabling policymakers to craft more responsive and effective climate strategies (Shi et al., 2015). With major summits like G20 and COP30 approaching in 2025, where sustainability and climate change will be central topics (Wonneberger et al., 2020; Lochner et al., 2024), gauging public sentiment is crucial to inform discussions, anticipate challenges, and align policies with public expectations.

Recently, as artificial intelligence technologies advance, LLMs become key players in public opinion formation and information dissemination, as their integration to major search engines – such as Google and Bing – continues to expand (Costello et al., 2024). AI-generated responses frequently precede traditional search results, many of which are also algorithmically curated (Dai et al., 2023). By shaping public discourse, reflecting societal perspectives, and anticipating emerging trends (Yakura et al., 2024; Faruk, 2024), LLMs play a crucial role in how information is accessed and interpreted. Given their widespread reach, critically examining the biases they introduce and reinforce is essential.

Understanding how these models portray critical topics is not merely a technical concern, but a critical factor in assessing their impact on public perception and societal narratives (Wan et al., 2023; Motoki et al., 2024). Researchers caution that, due to LLMs being predominantly trained on data from Western and high-income countries, these models may inherently amplify the perspectives of these regions while also reflecting and perpetuating biases related to race and gender. This can lead to an oversimplification of complex societal issues (Atari et al., 2023; Cheng et al., 2023).

Therefore, assessing the alignment of LLMs in climate-related contexts is crucial. Evaluating their tendencies and biases helps determine their influence on climate narratives and broader societal and political discourses. Comparing their outputs with human opinions across different countries can provide valuable insights into how these models engage with climate discourse (Lee et al., 2024a).

In this study, we aim to examine the perspectives that large language models adopt when generating climate change-related responses. In particular, we assess which opinions their outputs reflect. Since different LLMs are trained on diverse datasets, rely on different algorithms, and are subject to distinct biases (Feng et al., 2023), discrepancies in the information they provide are expected. To address

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 113–125
July 31, 2025 ©2025 Association for Computational Linguistics

these concerns, we define the following research questions:

RQ1: To what extent do LLM responses align with different countries and geopolitical groups in climate change surveys?

RQ2: How does prompting LLMs to adopt a given citizenship influence their alignment with human responses?

RQ3: How do LLMs respond to climate-related questions, and what factors influence these responses?

We use responses from the People's Climate Vote 2024 survey (UNDP – United Nations Development Programme and Oxford, 2024), covering 77 countries, as a benchmark to evaluate eight LLMs, including both open-source and proprietary models from diverse companies and regions. The survey consists of closed-ended questions with predefined choices, which we present to the LLMs, instructing them to select the corresponding alternatives. This approach enables us to analyze token probability distributions and measure how closely their outputs align with human responses. Figure 1 illustrates the process of obtaining and evaluating LLM and human responses, highlighting the comparison and analysis framework.

Our findings reveal that, while LLMs do not exhibit systematic biases toward specific geopolitical or demographic groups, their responses often diverge significantly from majority human opinions. In particular, we found that LLMs generally express greater concern about climate change, especially regarding future risks and long-term policy commitments, than the average human respondent. However, their alignment with human perspectives on immediate climate actions varies, with some models displaying notable discrepancies. Additionally, prompting LLMs to adopt a national identity sometimes reduces divergence, but the effect is inconsistent across countries and models. These results highlight the distinct role that LLMs play in shaping climate discourse and underscore the need for careful evaluation of their potential biases and influence on public narratives.

2 Related Work

Understanding the opinions held by large language models (LLMs) has become a key area of study. Santurkar et al. (2023) proposes a framework to evaluate LLM alignment with public opinion, finding significant misalignment with U.S. views, especially in models fine-tuned with human feedback. Similarly, Durmus et al. (2024) compares LLM-generated survey responses with data from the World Values and the PEW Surveys, revealing stronger alignment with opinions in Western and South American countries. They also note that LLMs tend to assign disproportionately high probabilities to single responses, in contrast to the more diverse distributions seen in human responses.

Numerous studies have examined LLM biases across critical topics, like gender (Kotek et al., 2023), cultural perspectives (Naous et al., 2024), standardized tests (Locatelli et al., 2024), and political alignment (Motoki et al., 2024). Recent research has focused on how LLMs simulate public opinion on climate change, with studies like Wan et al. (2023) highlighting misrepresentation of demographic diversity and potential harms such as identity essentialization. Jansen et al. (2023) and Demszky et al. (2023) emphasize that LLMs are not yet reliable substitutes for human survey respondents, often misrepresenting demographic diversity. Additionally, Lee et al. (2024b) investigates social desirability response bias (SDR) in LLMs, finding limited bias with models maintaining consistent responses across varying demographic prompts.

Regarding climate change, Lee et al. (2024a) finds that GPT-based models reflect liberal, higher-income, and highly educated views, but struggle to represent beliefs of non-Hispanic Black Americans. Expanding beyond the U.S., Qu and Wang (2024) identifies regional disparities and biases based on demographic factors and ideological stances.

Our work extends on prior research by analyzing a broader set of LLMs and expanding the geographical scope of climate change simulations. We assess how these models align with human opinions and uncover which point of view they are propagating.

3 Survey Dataset

The survey used in this study is the Peoples' Climate Vote 2024 (UNDP – United Nations Development Programme and Oxford, 2024), the world's largest standalone public opinion survey on climate change. This edition introduced 15 questions organized into three main themes: (1) the direct effects of climate change on daily life, (2) how climate change is being addressed in the participant's coun-

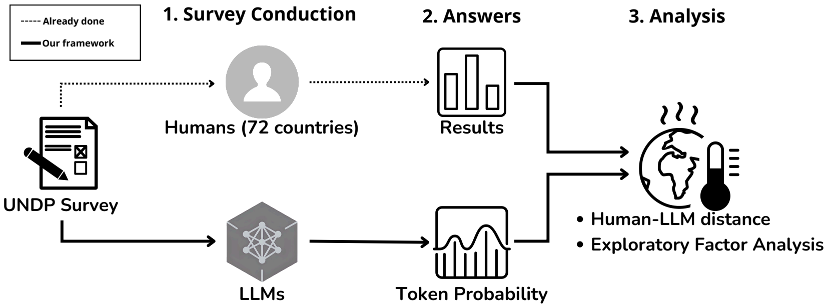

Figure 1: Diagram summarizing the proposed methodology for obtaining and evaluating responses.

try, and (3) preferences for future policy actions[1].

Administered by GeoPoll using Computer Assisted Telephone Interviewing (CATI) and Random Digit Dialing (RDD) methodologies, the survey was conducted in 87 languages, enabling the participation of a broad spectrum of demographic groups. Sample sizes per country typically ranged from 900 to 1,500 respondents, yielding in a total of 73,765 completed interviews from 1.9 million calls across 77 different countries. However, responses from only 72 countries and a global summary were available in the survey dataset.

The dataset provides a structured representation of survey responses, including the distribution of human responses for each alternative across all questions. Each entry contains the full question text, multiple-choice options, and respondents' demographic attributes, like age and education level.

4 Methodology

We evaluate LLM responses by submitting each survey question and its predefined answer choices, prompting the model to select a single-letter response. This allows us to extract log probabilities for each option, which we normalize into a probability distribution for comparison with human responses. Our analysis includes three key components: (i) measuring distributional distance using Jensen-Shannon divergence to compare model outputs with public opinion across countries; (ii) conducting Exploratory Factor Analysis to identify underlying factors influencing responses; and (iii) performing sentence embedding analysis to examine whether LLMs favor answer choices semantically closer to the question in the embedding space.

4.1 Selection of Large Language Models

To ensure a representative analysis, we selected both open-source and proprietary LLMs from diverse companies and countries to assess their alignment with human opinions across different regions. We included GPT-4o as a state-of-the-art LLM, DeepSeek and Qwen as Asian models, LLaMA, Phi, and Grok as U.S.A. representatives, and Mistral as a European counterpart. Open-source models were executed in local machines, while proprietary models were accessed via API.

4.2 Prompts for Multiple-Choice Questions

The prompting strategies in this study simulate real-world scenarios. We employed a zero-shot approach, allowing models to leverage their natural language and contextual understanding to handle unfamiliar questions. Each prompt consists of an instruction explicitly requesting the model to respond with a single letter corresponding to the selected answer, followed by the question and its predefined answer choices. All prompts were written in English, matching the language used in the survey. The prompt used in this study can be found in the Appendix C.

Consistent with current literature (Argyle et al., 2023), we set the models temperature to 0.7 to balance deterministic responses with moderate variability. Additionally, we imposed a strict token limit of 1 to ensure that only a single token—the model's answer – was generated. This setup enabled us to extract log probabilities or logits for the predicted token directly.

To obtain the probability distribution of the model's responses, we first extract the logits for all tokens in the vocabulary, which represent the unnormalized scores assigned to each token. From the logits, we select only those corresponding to the predefined answer choices ("A", "B", "C", "D",

[1] https://peoplesclimate.vote/about

etc.). Additionally, we apply a *strip* process to remove leading and trailing whitespace from tokens, ensuring that variations like "␣A" and "A" are treated identically.

To convert the selected logits into probabilities, we apply the softmax function, which normalizes the values into a probability distribution where all probabilities range between 0 and 1 and sum to 1. The softmax function is defined as:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$

where z_i represents the logit for answer choice i, and N is the total number of answer choices.

4.3 Measuring Distances Between Human And LLM Responses

Following prior literature (Locatelli et al., 2024), we use Jensen-Shannon distance as the primary metric to quantify the differences between human responses and those generated by LLMs. Applying a base-2 logarithm, this metric is bounded within the interval $[0, 1]$, enhancing its interpretability.

Let Q be a set of size N representing a collection of survey questions, and let A_q denote the set of possible answers for each question q. Since we extract the log probabilities for the tokens corresponding to the answer choices, we define the following for LLMs:

$$P_m(a|q), \forall a \in A_q, q \in Q, m \in M$$

where m refers to a specific model in our study, and $P_m(a|q)$ is the probability of model m answering question q with alternative a.

Analogously, for human responses, we also have the probability for each possible answer, which is available in the survey data. Thus, we define:

$$P_H(a|q), \forall a \in A_q, q \in Q$$

where $P_H(a|q)$ represents the probability of humans answering question q with alternative a, and this probability distribution is available for each country in the survey.

The distance between human responses and a model m is then calculated as the mean of the Jensen-Shannon distances across all questions:

$$Distance(m, H) = \frac{1}{N} \sum_{q=1}^{N} JS(P_m(A_q|q), P_H(A_q|q))$$

A larger distance indicates a greater divergence between the model and human distributions, while smaller values suggest stronger alignment.

4.4 Evaluating Question-Level Contributions to Global Alignment

To assess the structure of alignment between LLM-generated responses and human opinions at a more granular level, we applied the DISTATIS method (Abdi et al., 2005) to the distance matrix derived from each individual survey question. This approach allows us to combine multiple distance matrices into a shared structure, assigning a weight to each question based on its contribution to the global similarity pattern. Higher weights indicate that a question's distance matrix not only aligns more closely with the overall trend, but also contributes more significantly to shaping the shared structure. In contrast, lower weights suggest that a question's distance relationships deviate more from the common pattern, exerting a smaller influence on the global alignment. We leverage this analysis to evaluate the extent to which each individual question influenced the overall alignment between LLM-generated responses and human opinions, allowing us to identify which questions deviate from the global behavior.

4.5 Exploring Latent Factors in Climate Change Opinions

To understand the underlying structure of climate change opinions, we employ Exploratory Factor Analysis (EFA), a widely used statistical technique in social sciences (Teo, 2014). EFA identifies latent factors that explain patterns of correlation among observed variables, assuming that responses to individual items are influenced by these underlying dimensions. By analyzing response patterns, EFA reveals the structure of opinions in a dataset, reducing complexity while preserving key relationships.

In the context of a climate change survey, these latent factors group countries with similar response distributions on related questions. By interpreting these factors, we gain insights into the values of citizens of different countries and the opinions that large language models might generate.

For the EFA, we first construct a matrix based on our observations. Since most survey questions have ordinal answers, we assigned a value to each alternative ranging from 1 to $|A_q|$, where $|A_q|$ denotes the number of alternatives for question q. Next, we calculated the weighted average score for each

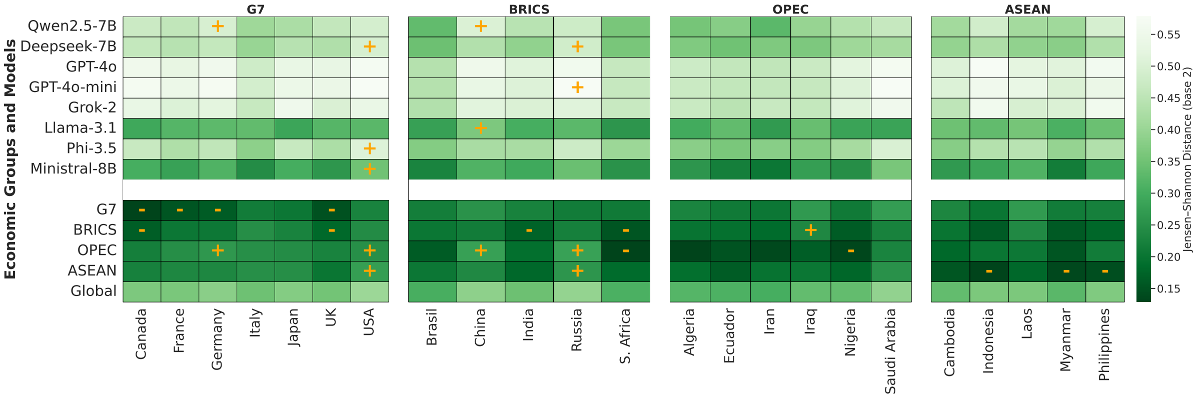

Figure 2: Distance between each country and group/LLM. Distances marked with "-" are lower than the mean distance for that row, while those marked with "+" are significantly higher at significance level 0.05. Note that the country responses are removed from its group's distribution when calculating the distance between the two. A full version including all countries is available in Appendix B.

observation (country or LLM) across question alternatives. This results in an 81x14 matrix, where each row represents a country or LLM, and each column corresponds to the average score for a given question. To simplify the process, we exclude one question whose alternatives were not ordinal. Applying EFA to this matrix reveals latent factors that capture climate-related opinion patterns, providing a more interpretable representation of potential alignments or divergences between LLMs and human respondents across regions.

We use the Factor software (Lorenzo-Seva and Ferrando, 2006), applying unweighted least squares as the optimization method and *promin* rotation to maximize factor simplicity. To determine the optimal number of factors, we use parallel analysis, comparing eigenvalues from our observations with those from Monte-Carlo simulated random data (Allen, 2017).

5 Results

In this section, we present the results of our proposed methodology, addressing the research questions (RQs) posed earlier. The results are organized as follows: Section 5.1 analyzes the distances between probabilities distributions of LLM and human responses, Section 5.2 investigates the effect of conditioning the LLM to be more similar to specific countries, Section 5.3 delves into the characteristics of individual questions, and, finally, Section 5.4 explores the alignment between LLM-generates opinions and human values.

5.1 Assessing LLM Alignment with Regional and Geopolitical Groups

To assess how closely LLM responses align with different human populations, we analyze the distances between models, geopolitical groups, and individual countries. Figure 2 presents these distances, where the probability distribution for each geopolitical organization group is obtained by averaging the distributions of all countries within that group, excluding the country of interest. This grouping approach enhances visualization and interpretability. For instance, when calculating the distance between G7 and the United States, U.S. responses are excluded from the G7 distribution, allowing us to assess whether LLMs align more closely with specific regions or geopolitical groups.

We selected the G7[2], BRICS[3], OPEC[4], and ASEAN[5] as representative geopolitical groups, given their diverse economic and political perspectives. These groups provide a broader context for evaluating alignment patterns. The distributions for each group were obtained by averaging the answers from each of its members.

When comparing LLMs responses to human responses, we find no clear evidence of alignment with any specific group. If LLMs strongly aligned with a group, we would expect significantly lower distances compared to the average for countries in that group. Instead, the distances remain relatively high, suggesting that LLMs do not show a sys-

[2] Canada, France, Germany, Italy, Japan, the United Kingdom and the United States.

[3] Brazil, Russia, India, China and South Africa.

[4] Algeria, Ecuador, Iran, Iraq, Nigeria and Saudi Arabia.

[5] Cambodia, Indonesia, Laos, Myanmar and Philippines.

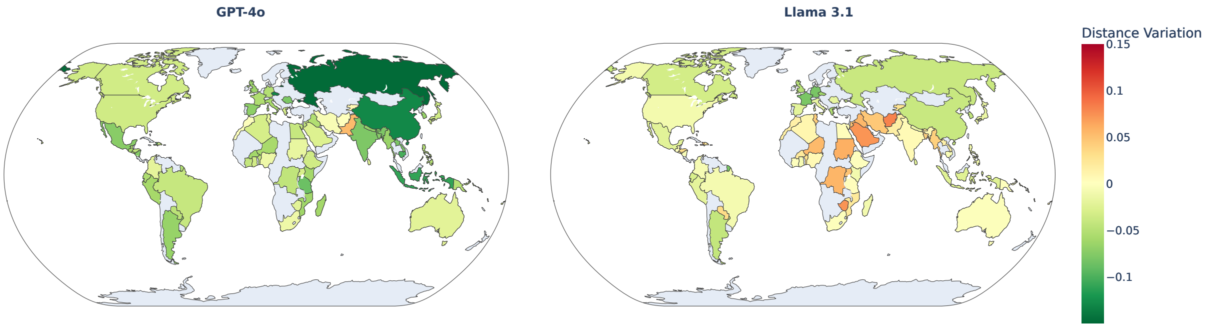

Figure 3: Variation in the distance between each surveyed country and GPT-4o/LLaMa 3.1 8B when prompting the model to respond as a citizen of that country. Positive values (red) indicate greater divergence, while negative values (green) suggest improved alignment with human responses.

tematic preference for any geopolitical or regional group.

This is further corroborated by the fact that the standard deviation in similarity between LLM and human responses across individual survey questions is relatively high ($\sigma > 0.12$ for the distance between LLMs and their closest country responses), indicating inconsistency in model predictions when compared to human populations. Additionally, the minimum distance between an LLM and its closest country is higher than the distance between two geographically or culturally similar countries (median for LLMs = 0.34, median for countries = 0.13).

These findings challenge the assumption that LLMs may be biased towards certain populations, such as Western or developed countries. In the context of climate change, our analysis provide no strong evidence of such biases. Instead, the results suggest LLM-generated response distributions do not closely resemble human distributions in general. Nevertheless, some models generate responses that significantly diverge from those of the populations they might be expected to represent. For example, the responses generated by Chinese LLM Qwen2.5 differ notably from those provided by Chinese citizens.

5.2 LLMs as Virtual Citizens: Can LLMs Adapt to Country-Specific Beliefs?

Since large language models do not inherently produce responses that align with the answer distributions of any specific country, we explored whether prompting techniques could encourage more human-like responses. To test this, we instruct the LLM to act as that countries' (country X) citizen (see Appendix C).

We then measure the impact of this intervention by comparing the distance between the model's new responses and those of country X. Figure 3 shows the change in distance before and after applying this prompt, referred to as *distance variation*. This variation is computed as:

$$\Delta_{\text{dist}}(m, H_X) = \text{dist}(H_X, m_X) - \text{dist}(H_X, m),$$

where H_X represents the human response distribution for country X, m denotes the default LLM response distribution, and m_X corresponds the LLM's response when prompted to act as a citizen of country X.

A positive $\Delta_{\text{dist}}(m, H_X)$ means the customized prompt increased the distance to human responses, whereas a negative value suggests better alignment. This analysis is limited to GPT-4o and LLaMa 3.1 8B Instruct for brevity.

Our findings reveal that, for both models – particularly GPT-4o – assigning a national identity for the LLM to mimic often reduced the distance to the target country. However, in certain cases (e.g. Pakistan), the intervention failed to bring the model's responses closer to human distributions. In some instances, it even increased the divergence, suggesting that the effectiveness of this approach varies depending on the model and the country.

Moreover, LLaMa 3.1 8B failed to reduce its distance to several African, Middle Eastern, and South Asian countries. This may derive from biases in the model's training data, as well as its relatively small number of parameters. The representation of multilingual content in the training corpus, estimated at around 8% (Grattafiori et al., 2024), could have contributed to weaker alignment with regional human responses. Additionally, its reduced model capacity may limit its ability to capture complex cultural nuances.

These results suggest that prompting LLMs to mimic a nation's citizen can sometimes improve

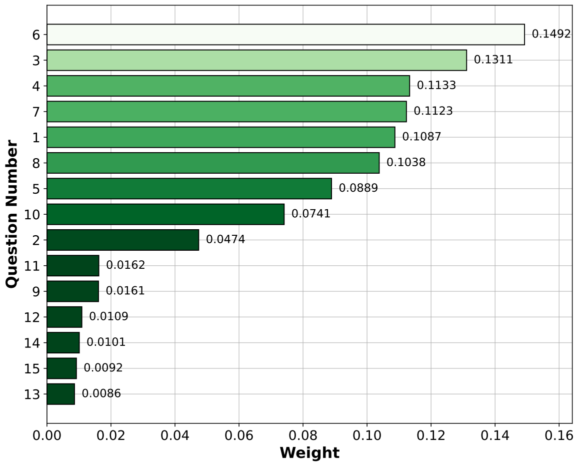

Figure 4: Weights assigned to each survey question using the DISTATIS method.

Model	F1	F2	F3
Deepseek 7b-chat	5.06	3.99	5.19
GPT-4o	4.84	5.63	4.36
GPT-4o-mini	4.91	4.18	4.05
Grok-2	4.77	7.73	5.87
LLaMa-3.1 8B-Instruct	6.12	7.10	8.53
Phi-3.5-mini-Instruct	4.99	4.82	6.03
Ministral 8B-Instruct	5.38	6.84	6.36
Qwen2.5 7B-Instruct	4.88	6.42	4.44

Table 1: Factor scores for each tested LLM. Cells highlighted in red represent values in the top 10%, while those in green represent the bottom 10%, including the countries. Due to the scale we adopt for the answers, a lower value on a factor indicates that the model is more concerned with that aspect of climate change.

alignment with human responses, but the effect is inconsistent across models and regions. Differences in model architecture, training data, and parameter count likely contribute to variations, while increased divergence in certain countries highlights the risks of misrepresentation. This underscores the importance of careful evaluation when using LLMs to simulate national public opinion.

5.3 Question-Level Contributions to LLM-Human Alignment

The results reveal substantial variation in how different survey questions contribute to the overall similarity structure. As shown in Figure 4, question 6, that addresses governmental effectiveness in climate action and question 3, related to concerns for future generations exhibit the highest contributions, suggesting that LLM responses on these topics align more consistently with the overall distance between humans and LLMs.

In contrast, question 15, related to international cooperation and question 13, about educational efforts present the lowest contributions, which suggest that regarding these topics they present a different answer pattern from the one presented by the global similarity pattern. This can be confirmed by looking at the distance matrix of these two questions and noting that it describes a much smaller distance between models and countries, specifically, the tested models seem to be express opinion much closer to the developing responses than on other questions.

5.4 Exploring the Opinions of LLMs on Climate Change

In the previous sections, we found that LLM answer distributions, even when prompted to simulate responses as citizens of specific countries, had very inconsistent alignment with those of human groups. This suggests that the models do not exhibit a strong bias towards any national perspective on climate change issues. However, this analysis alone does not reveal the underlying opinions the models may be expressing. To address this, we now turn to an Exploratory Factor Analysis (EFA) to better understand the models' perspectives on climate change.

Three factors were identified as significant in our analysis (Kaiser-Meyer-Olkin (KMO) test = 0.788, 69.4% explained variance), suggesting that our data is suitable for factor analysis. The full factor loadings are available in Appendix A. By examining the associations between factors and individual survey questions, we found that each factor aligned with one of the main themes of the survey presented in Section 3. Since these themes emerged from question groupings, we defined the factor labels *a posteriori* as:

F1: Future Actions: Concerns about long-term climate policies and commitments.

F2: Present Actions: Focus on immediate efforts and measures to address climate change.

F3: Climate Change and Daily Life: The perceived impact of climate change on everyday life and personal experiences.

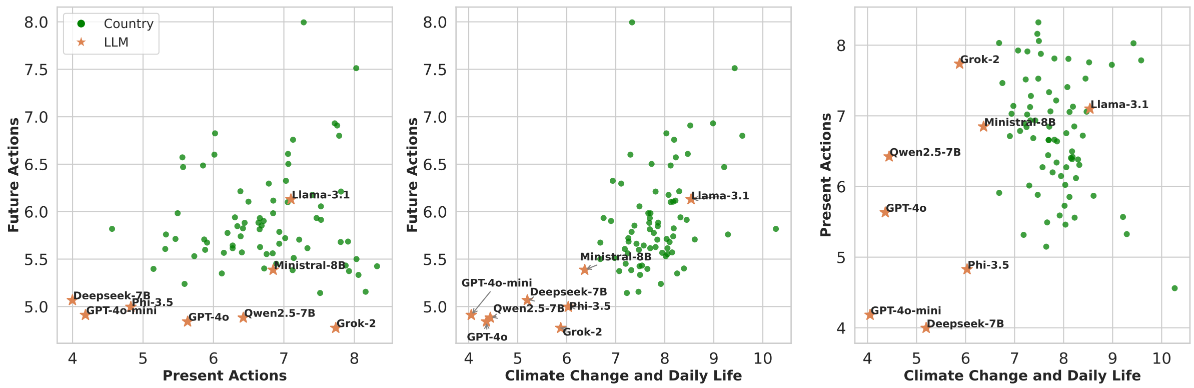

Figure 5: Scatter plots comparing factor scores for countries and LLMs. (a) Future Actions vs. Present Actions: LLMs consistently express concern for future actions, but show variation on present actions. (b) Future Actions vs. Climate Change and Daily Life: LLMs exhibit diverse positions, with LLaMa showing lower climate anxiety than most countries. (c) Present Actions vs. Climate Change and Daily Life: LLM responses differ significantly from human populations, highlighting their distinct perspectives.

Using the weighted sum method (DiStefano et al., 2019), we calculated factor scores for each country and LLM, reflecting their responses to each question. Table 1 presents the scores for LLMs analyzed. A lower score on a given factor suggests that the model is more likely to provide favourable responses to questions related to that factor.

Among the models examined, the GPT-4 family stood out as the most likely to acknowledge the impact of climate change and the importance of government actions across all factors, followed by Phi. In general, we found that LLMs expressed more concern about climate change's effects on daily lives (F3) and future actions (F1) than the average human from most countries. This was not the case for present actions (F2), where LLMs factor scores, except for Phi and GPT-4, aligned more closely with human responses. Notably, Mistral and LLaMA showed the most divergent responses: both models tended to provide more negative assessments regarding present and future actions, but differed on their stance towards F3–LLaMA being more negative than most countries, and Mistral more positive, aligning with other LLMs.

Having analyzed the performance of the models relative to each other, we now compare their responses to human answers. Figure 5 show the positions of the LLMs relative to the countries on these factors. Most models are clear outliers in relation to the factor values, positioning themselves relatively far from the countries' distributions. Even the models that are not clear outliers – LLaMa and Mistral – appear on the border of the cloud of countries, suggesting that the opinions they generate may differ significantly from those of most countries.

In practice, this highlights how unusual the answer distribution from LLMs are when compared to humans, especially when considering the combination of factors. Although some of the concerns of the large language models, in the form of factor scores, individually may approach the opinions of some countries, when assessing all three factors, we notice that the generated response distributions are inconsistent with existing countries.

6 Conclusion

As large language models gain widespread use, understanding the nature of the opinions they generate is crucial, particularly in sensitive areas like climate change. Our analysis of responses from eight LLMs compared with human answers from the People's Climate Vote 2024 survey, reveals that LLMs generally express greater concern about climate change than average human, with their responses differing significantly from human groups.

Furthermore, the higher levels of concern observed in LLM responses may be linked to various stages of model training, though the lack of transparency in training data complicates the identification of specific causes. Future research could explore the impact of these factors on LLM-generated opinions.

It is still unclear whether LLMs should mimic the public opinion or the expert opinion on a given topic. In this study, we focus solely on the first, finding that there is currently little alignment between model generated and people's response on climate change. Nevertheless, future work should explore the latter, as it can be argued that this technology should be used to gently steer people's opinions towards the scientific consent on pressing world problems.

7 Implications

As the use of LLMs as substitutes for human participants in surveys becomes increasingly debated (Jansen et al., 2023), it is crucial to be aware of the limitations these models have when representing diverse groups. As our analysis shows, the answers distributions generated by these models are considerably different from those of humans, and mitigation techniques such as prompting the model to adopt the role of a specific demographic group can only go so far, potentially without risking representational harms.

Another point to consider is that even between LLMs, their answer distributions may vary greatly, and, in some cases, this can lead them to express different views on specific issues. For instance, the degree to which each model values **F2:Present Actions** is significantly different, with LLama-3.1 and Grok-2 showing much higher scores when compared to GPT-4o-mini and Deepseek-7B.

As an user, it is hard to know which kind of bias or point of view an LLM may display a priori and one may be influenced without even realizing. With the trend in decreasing information in LLM model cards, especially in sections related to bias and limitations (Liang et al., 2024), and the sheer number of different models, it is hard to know what kind of information one may receive when interacting with a LLM-powered application. Large language model providers should be encouraged to provide accurate and transparent documentation that can inform the end users of the expected outputs of their products.

Limitations

In our study, we aim to represent a diverse range of cultures by examining the countries available in the Peoples' Climate Vote 2024 survey. However, this focus on countries means we do not account for within-country demographic variations. LLM responses may align closely with specific age, education, gender, religion or other demographic groups, which we leave for future work to explore. For the model selection, we analyze eight widely-used models from diverse companies and countries of origin. However, other state-of-art models, such as Deepseek-V3 and Claude-3, or models tailored for specific languages, could provide valuable insights. Additionally, versions of models used in the study with more parameters, such as LLaMa 3.1 405B Instruct, may offer further improvements. Finally, while we assess model opinions using controlled prompts and survey questions, our findings may not fully reflect the responses these models would generate in real-world applications.

Acknowledgements

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

H. Abdi, A.J. O'Toole, D. Valentin, and B. Edelman. 2005. Distatis: The analysis of multiple distance matrices. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 42–42.

Mike Allen. 2017. *The SAGE encyclopedia of communication research methods*. SAGE publications.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. 2023. Llms may dominate information access: Neural retrievers are biased towards llm-generated texts. *arXiv preprint arXiv:2310.20501*.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Christine DiStefano, Min Zhu, and Diana Mindrila. 2019. Understanding and using factor scores: Considerations for the applied researcher. *Practical assessment, research, and evaluation*, 14(1):20.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Tanjim Bin Faruk. 2024. Evaluating the performance of large language models in scientific claim detection and classification. *arXiv preprint arXiv:2412.16486*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Myanna Lahsen and Jesse Ribot. 2022. Politics of attributing extreme events and disasters to climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 13(1):e750.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024a. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):1–14.

Sanguk Lee, Kai-Qi Yang, Tai-Quan Peng, Ruth Heo, and Hui Liu. 2024b. Exploring social desirability response bias in large language models: Evidence from gpt-4 simulations. *arXiv preprint arXiv:2410.15442*.

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic analysis of 32,111 ai model cards characterizes documentation practice in ai. *Nature Machine Intelligence*, 6(7):744–753.

Marcelo Sartori Locatelli, Matheus Prado Miranda, Igor Joaquim da Silva Costa, Matheus Torres Prates, Victor Thomé, Mateus Zaparoli Monteiro, Tomas Lacerda, Adriana Pagano, Eduardo Rios Neto, Wagner Meira Jr., and Virgilio Almeida. 2024. Examining the behavior of llm architectures within the framework of standardized national exams in brazil. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):879–890.

Jakob H Lochner, Annika Stechemesser, and Leonie Wenz. 2024. Climate summits and protests have a strong impact on climate change media coverage in germany. *Communications Earth & Environment*, 5(1):279.

Urbano Lorenzo-Seva and Pere J Ferrando. 2006. Factor: A computer program to fit the exploratory factor analysis model. *Behavior research methods*, 38(1):88–91.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Jing Shi, Vivianne HM Visschers, and Michael Siegrist. 2015. Public perception of climate change: The importance of knowledge and cultural worldviews. *Risk Analysis*, 35(12):2183–2201.

Timothy Teo. 2014. *Handbook of quantitative methods for educational research*. Springer Science & Business Media.

UNDP – United Nations Development Programme and University of Oxford. 2024. Peoples' climate vote 2024. *UNDP (United Nations Development Programme)*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Anke Wonneberger, Marjolein H. C. Meijers, and Andreas R. T. Schuck. 2020. Shifting public engagement: How media coverage of climate change conferences affects climate change audience segments. *Public Understanding of Science*, 29(2):176–193.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

A Factor Loadings for Survey Questions

Table 2 shows the factor loadings for each question. The absolute value of the factor loading indicates how related that question is to the factor. For example, the question "Should your country strengthen or weaken its commitments to address climate change?" is highly associated with the factor **(F1) Future Actions**.

B Complete Jensen-Shannon Distances

Figure 6 shows the mean Jensen-Shannon distance between all surveyed countries and the studied LLMs. Note that the addition of the extra countries adds little information: they are generally further away from the LLMs when compared to the economical/geopolitical groups.

C System Prompt

The following system prompt was used to standardize the responses generated by the LLMs:

> "You will receive a question. You **MUST** respond with only one letter. The possible answers will be presented as follows: A: answer, B: answer, C: answer, etc. You should respond **ONLY** with the letter corresponding to the correct alternative according to you. Do not provide explanations, additional text, or repeat the answer—just the letter."

This prompt ensured that all models produced structured and comparable outputs, facilitating a consistent evaluation of their alignment with human responses.

We append the following instruction to the prompt in order to conduct the analysis proposed in section 5.2, where country X stands for any country we wish the LLM to mimic:

> "You must answer the following question as if you were a typical citizen of **{country X}**, reflecting the general opinions, beliefs, and cultural perspectives of people from that nation."

D Question-Level Contributions Distances

In this section, we present the distances between each country and its group/LLM for Question 6: "*How well is your country addressing climate change?*". This analysis helps to understand how each country perceives its own efforts in addressing climate change relative to others. The results are shown in Figure 7.

E Semantic Proximity in LLM Responses

Large Language Models rely on internal text representations to generate responses. This raises the question of whether their answer choices are influenced by semantic proximity in the embedding space. To explore this, we analyze if LLMs tend to favor answer choices closer to the questions in a pre-trained sentence embedding space.

For this analysis, we use a pre-trained sentence embedding model to encode both survey questions and answer choices into a shared embedding space. Specifically, we adopt the SentenceTransformer (Reimers and Gurevych, 2019), a bidirectional, encoder-only transformer model. Each question is encoded as a single vector, and each answer choice is separately encoded into the same space.

To assess whether LLMs are more likely to select answer choices semantically closer to the question in embedding space, we computed the correlation between the distance of each answer choice to the question and its selection probability. Figure 8 presents these correlations for the studied models.

The results indicate a clear negative correlation across all LLMs, with values ranging from approximately between -0.30 to -0.55. This suggests that the closer an answer choice is to the question in embedding space, the more likely the model is to select it. While the strength of this effect varies across models, the consistent trend implies that semantic proximity plays a significant role in shaping LLM predictions.

ID	Question Text	(F1) Future Actions	(F2) Present Actions	(F3) Climate Change and Daily Life
1	How often do you think about climate change?	0.097	0.333	0.732
2	Compared with last year, are you more or less worried about climate change?	-0.018	-0.123	0.806
3	How worried are you about the effects of climate change on the next generation?	-0.053	-0.049	0.840
4	Thinking about extreme weather events - such as, droughts, flooding, storms, and extreme heat or cold - was your community's experience this year...	-0.032	-0.027	0.743
5	How much has climate change affected any big decisions for your family, such as where to live or work, or what to buy?	-0.024	0.423	0.548
6	How well is your country addressing climate change?	0.020	0.830	0.010
7	How well are big businesses addressing climate change?	-0.052	0.944	-0.028
8	In your country, who do you think has had the most impact addressing climate change?	N/A	N/A	N/A
9	Should your country strengthen or weaken its commitments to address climate change?	0.758	0.088	0.050
10	How quickly should your country replace coal, oil, and gas with renewable energy, such as power from the wind or sun?	0.281	0.130	0.335
11	How much should your country protect and restore nature, for example, by planting trees or protecting wildlife?	0.835	-0.121	-0.116
12	When it comes to protecting people at risk from extreme weather events, such as storms or extreme heat, should your country provide...	0.932	-0.057	-0.094
13	Should countries work together on climate change even if they disagree on other issues, such as trade or security?	0.537	-0.045	0.197
14	Should rich countries give more or less help to poorer countries to address climate change?	0.824	0.076	0.037
15	Should schools in your country do more or less to teach about climate change?	0.838	0.011	0.010

Table 2: Factor loadings for each survey question. Factors with an absolute value greater than 0.3 are highlighted for easier interpretation. Question 8 was not included in the EFA as its answers are not ordinal, resulting in no factor loadings.

This finding has important implications for how LLMs respond to climate-related survey questions. Survey design typically aims to capture nuanced opinions, but an overreliance on semantic proximity may introduce biases in response selection. If LLMs prioritize answers that semantically closer to the question, they may systematically favor certain perspectives rather than reflecting a broader range of human responses. Moreover, the linguistic style of the question, such as word choice and phrasing, could reinforce these biases, influencing the model's response selection. In climate discourse, for example, questions often contrast immediate versus long-term actions or individual versus governmental responsibility, leading models to disproportionately select semantically aligned answers.

The variation in correlation strength across models also suggests that architecture and training data influence how semantic similarity impacts response selection. Models with stronger correlations might be more susceptible to this effect, limiting their ability to represent a balanced spectrum of climate opinions. This highlights the need to understand the internal biases of LLMs, particularly when using them to simulate public sentiment or inform policy decisions.

Overall, these results provide an initial insight into how sentence embeddings influence LLM decision-making, potentially introducing systematic patterns in response selection. While these findings shed light on the role of semantic alignment in model outputs, further research is needed to deepen this analysis and develop strategies to mitigate such biases. This is particularly crucial to ensure that LLM-generated responses in climate-related surveys and discussion are are scientifically grounded and not unduly influenced by embedding or linguistic biases.

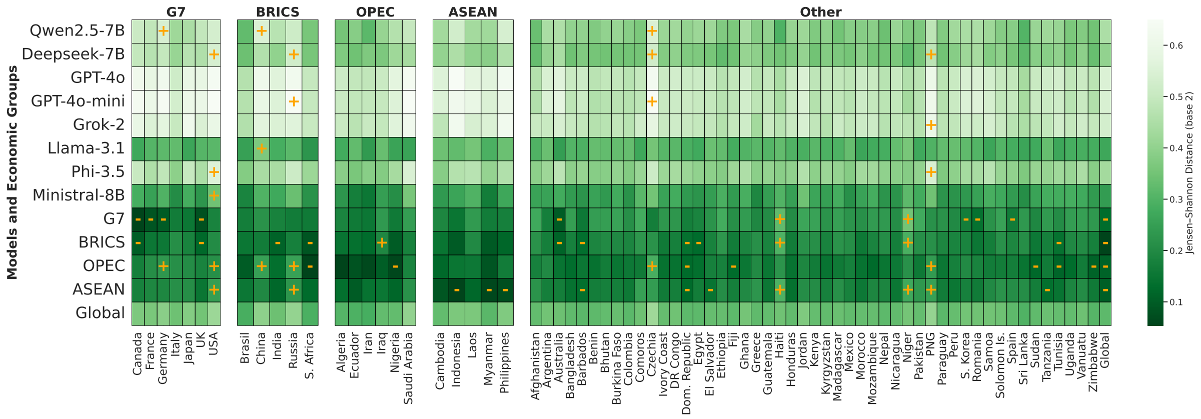

Figure 6: Distance between each country and group/LLM. Distances marked with "-" are significantly lower than the row, while those marked with "+" are significantly higher at the 0.05 significance level. Note that the country responses are removed from their group's distribution when calculating the distance.

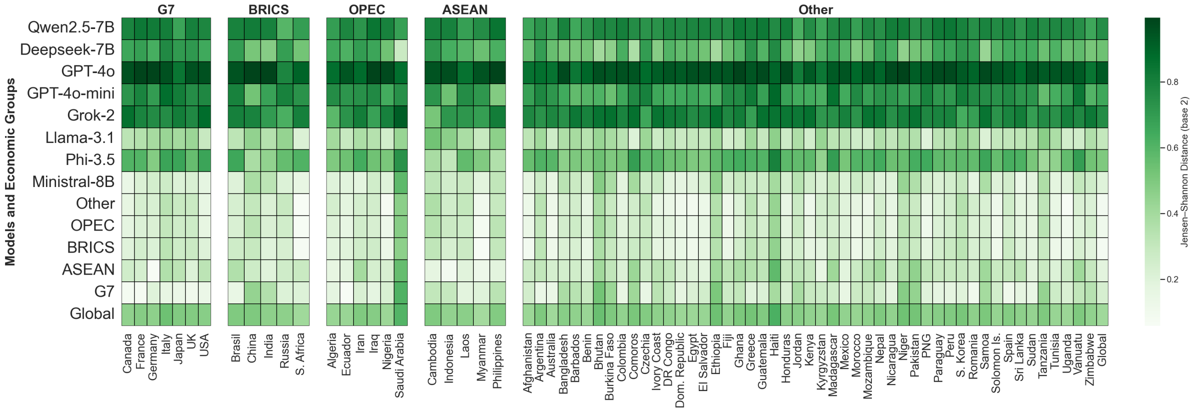

Figure 7: Distance between each country and group/LLM for Question 6.

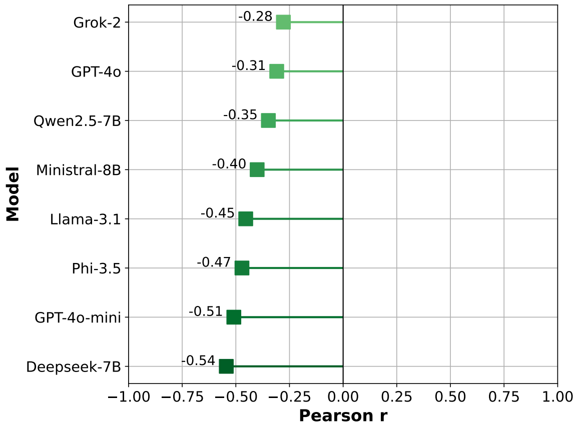

Figure 8: Correlation between the distance of answer choices from the question in embedding space and their selection probability.