On the Alignment of Large Language Models with Global Human Opinion

Yang Liu[1], **Masahiro Kaneko**[2], **Chenhui Chu**[1]

[1]Kyoto University
[2]MBZUAI
yangliu@nlp.ist.i.kyoto-u.ac.jp, Masahiro.Kaneko@mbzuai.ac.ae, chu@i.kyoto-u.ac.jp

Abstract

Today's large language models (LLMs) are capable of supporting multilingual scenarios, allowing users to interact with LLMs in their native languages. When LLMs respond to subjective questions posed by users, they are expected to align with the views of specific demographic groups or historical periods, shaped by the language in which the user interacts with the model. Existing studies mainly focus on researching the opinions represented by LLMs among demographic groups in the United States or a few countries, lacking worldwide country samples and studies on human opinions in different historical periods, as well as lacking discussion on using language to steer LLMs. Moreover, they also overlook the potential influence of prompt language on the alignment of LLMs' opinions. In this study, our goal is to fill these gaps. To this end, we create an evaluation framework based on the World Values Survey (WVS) to systematically assess the alignment of LLMs with human opinions across different countries, languages, and historical periods around the world. We find that LLMs appropriately or over-align the opinions with only a few countries while under-aligning the opinions with most countries. Furthermore, changing the language of the prompt to match the language used in the questionnaire can effectively steer LLMs to align with the opinions of the corresponding country more effectively than existing steering methods. At the same time, LLMs are more aligned with the opinions of the contemporary population. To our knowledge, our study is the first comprehensive investigation of the topic of opinion alignment in LLMs across global, language, and temporal dimensions.

Code —
https://github.com/ku-nlp/global-opinion-alignment

1 Introduction

Large language models (LLMs) have become crucial to everyday decision-making and assistance (Achiam et al. 2023; Bubeck et al. 2023; Bommasani et al. 2021). Once LLMs are deployed as products, they are inevitably asked to answer subjective questions, not just objective ones (Ouyang et al. 2022; Santurkar et al. 2023; Meister, Guestrin, and Hashimoto 2025). From an alignment perspective, next-token prediction during pretraining gives LLMs a statistical prior over human opinions (Radford et al. 2019; Brown et al. 2020), while subsequent alignment stages, such as instruction tuning (Wei et al. 2021; Sanh et al. 2021) and preference optimization (Christiano et al. 2017; Ouyang et al. 2022), shape this prior toward the opinions in their preference datasets, which are typically annotated by annotators from various countries, speaking diverse languages, and across different ages (Mason and Suri 2012).

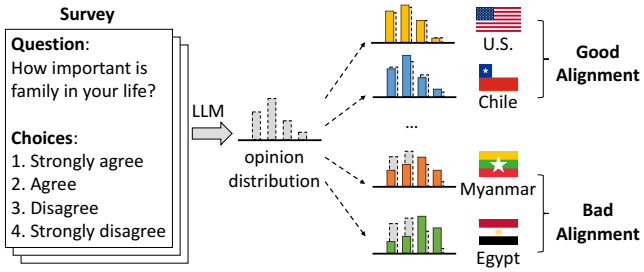


Figure 1: An example of our main idea. When asking the LLM a subjective question, is the LLM's opinion distribution closer to specific countries' opinion distributions?

Existing study mainly focuses on exploring opinion alignment between LLMs and demographic groups in the United States (Santurkar et al. 2023). Durmus et al. (2023) expand Santurkar et al. (2023)'s study to explore the alignment between the LLM Claude and other countries, and they further find that translating the questions into Russian, Chinese, and Turkish fails to improve the alignment between the LLM and the language speakers. A crucial challenge to previous studies is that the LLMs' opinion distribution is expressed as the next token log probabilities from the model, which suffers from a mis-calibration issue and requires the LLMs to be able to output logits (Meister, Guestrin, and Hashimoto 2025). Therefore, in this study, we adopt the verbalized distribution method to investigate the alignment between LLMs and human opinions. We expand to investigate seven state-of-the-art LLMs and eight languages covering 10 countries to validate the effectiveness of language steering. In addition, we innovatively explore the alignment of LLMs and human opinions in the temporal dimension. Therefore, our study covers the alignment of LLMs with human opinions in three dimensions: **global, languages, and temporal**.

The key question we focus on is illustrated in Figure 1:

when asking the LLMs a subjective question, is the LLM's opinion distribution closer to specific countries' opinion distributions? If an LLM systematically reflects the opinions of specific countries, this may lead to unexpected country dominance or misalignment in international applications (Bender et al. 2021; AlKhamissi et al. 2024; Sukiennik et al. 2025). Additionally, analyzing the opinions reflected by the LLMs can help improve the transparency and accountability of the LLMs, allowing developers to better assess and mitigate potential geopolitical or ethical risks. Finally, this study can provide a foundation for building more human opinion adaptive or neutral AI systems that respect diverse views in the globally interconnected world.

The World Values Survey (WVS; Haerpfer et al. 2022), a comprehensive global research program, systematically collects human opinions on a wide range of topics, including politics, religion, and ethics. The standardized questionnaire design of the WVS allows for consistent comparisons of opinions across countries, languages, and historical periods. In addition, the survey covers highly subjective and value-oriented questions, which align well with the types of prompts used to probe LLMs' opinions. By using WVS as the benchmark, in this study, we focus on the following research questions: **RQ1**: Do LLMs appropriately align with the opinions of different countries? **RQ2**: Can language steer LLMs to express opinions closer to language speakers? **RQ3**: Which human historical period's opinions are reflected in LLMs' responses?

Our work makes four main contributions: *1)* We systematically investigate human opinion alignment in global, language, and temporal dimensions with LLMs; *2)* We reveal that current LLMs tend to align with the opinions of English- or Spanish-speaking countries such as the United States, Canada, and Chile, but under-align with the opinions of information-regulated or non-Latin-speaking countries such as Myanmar and Egypt; *3)* We employ language steering on eight languages, which is effective in over 82% of the 240 cases (3 baselines[1] $\times$ 8 models $\times$ 10 countries) in our experiments; *4)* We evaluate the alignment of LLMs with human opinions in the temporal dimension, revealing that LLMs are most aligned with contemporary human opinions.

2 The World Values Survey

In this section, we explain why WVS is suitable for our study from three dimensions: global, language, and temporal. Further details are provided in the supplementary materials.

Global Dimension The WVS covers highly subjective and value-oriented questions, which align well with the types of prompts used to probe LLMs' opinions. The latest wave covers survey data from 66 countries and regions, enabling us to compare the alignment of LLMs with different countries' human opinions. In addition, each country has more than 1,000 human questionnaire results, which ensures that the survey data reflects real human behavior to some extent. These features make it possible for us to study RQ1.

[1]The three baselines include no steering, persona steering, and few-shot steering.

Language Dimension The WVS questionnaire was translated into all languages which serve as the first language for 15% (or more) of the population. For countries that participated in previous waves of the WVS, translation for the replicated items is used from the previous wave questionnaires to minimize bias during overtime comparison of answer distributions.[2] According to our statistics, the latest wave provided more than 50 language translations, ensuring that the questionnaire covers languages supported by LLMs and providing linguistic diversity for RQ2.

Temporal Dimension Since 1981, the WVS has conducted seven waves of globally representative surveys. The WVS maintains functional consistency across most questions in all waves, allowing us to study long-term trends of alignment between LLMs and human opinions across different waves, providing temporal data support for our RQ3.

3 Experiment Settings

In this section, we introduce the experiment settings. More details such as the question filter rules and the complete country list (§3.1), the full prompts (§3.3), and more details of the distribution expression methods (§3.4) are provided in the supplementary materials.

3.1 Dataset

In RQ1, we conduct a question filter process on the English version of wave 7, which includes survey data from 66 countries, to filter out questions that are not suitable for this study. For example, questions that require actual experience, objectivity, etc. After this filter process, we get 144 opinion-related questions. In RQ2, we used the same 144 questions and collected them from other language versions of the questionnaire. In RQ3, to ensure a certain number of common questions, we chose to analyze WVS's data from wave 5 (2005-2009), wave 6 (2010-2014), and wave 7 (2017-2022).[3] Each wave corresponds to a historical questionnaire version. We organized each English version from wave 5 to 7 and then selected questions that are common in all three waves. We obtained 75 opinion-related questions that are shared across all three waves.

3.2 Models

We experiment with seven instruction-tuned multilingual LLMs: Aya-23-35B (Aryabumi et al. 2024), Llama3-70B-Instruct (AI@Meta 2024), Qwen2.5-72B-Instruct (Team 2024), GPT-3.5-Turbo (`gpt-3.5-turbo-0125`; OpenAI 2023), GPT-4 (`gpt-4-0613`; Achiam et al. 2023), GPT-5 (`gpt-5`; OpenAI 2025), DeepSeek-V3 (`DeepSeek-V3-0324`; Liu et al. 2024), and DeepSeek-R1 (`DeepSeek-R1-0528`; Guo et al. 2025). For open-weight community models, we select the largest publicly released weights in each series. For commercial LLMs, we include OpenAI's GPT-3.5-Turbo, GPT-4, and GPT-5 (closed-source models), as well as DeepSeek-V3 and

[2]https://www.worldvaluessurvey.org/wvs.jsp

[3]Our statistical result indicates that, when wave 4 is taken into account, there are only 48 common questions.

DeepSeek-R1, noting that DeepSeek releases open-weight models under a permissive license.

3.3 Prompts

As the instruction fine-tuned LLMs use "task instructions + examples" as input during training, it is easier for them to follow the requirements of the prompt and output content in the required format. We refer to Meister, Guestrin, and Hashimoto (2025) to control the LLMs' output distribution in JSON format (e.g., {`1: 31%, 2: 4%, 3: 30%, 4: 35%`}) by providing five few-shot examples.

3.4 Distribution Expression Methods

LLMs' Opinion Distribution Being able to accurately express LLMs' opinion distributions is a precondition of our study. Meister, Guestrin, and Hashimoto (2025)'s work compares three distribution expression methods: *model log-probabilities*, *sequence of tokens*, and *verbalized distribution*. Their analysis reveals that *verbalized distribution* outperforms the other two methods. Therefore, we use *verbalized distribution* to represent the opinion distribution $D_{\mathcal{M}}(q)$ of the LLM $\mathcal{M}$ on answering question q.

Human Opinion Distribution We calculate human opinion distributions from the statistical data of the WVS, which contain more than 1,000 pieces of human data for each surveyed country. The distribution of the country $c \in \mathcal{C}$ answers to the question $q \in \mathcal{Q}$ with $|\mathcal{N}|$ options can be denoted as $D_c(q) = \{D_{c,n} | n \in \mathcal{N}\}$. Specifically, $D_{c,n}$ is as follows:

$$D_{c,n} = \frac{|\mathcal{P}_c^{(n)}|}{\sum_{i \in \mathcal{N}} |\mathcal{P}_c^{(i)}|} \tag{1}$$

where $|\mathcal{P}_c^{(n)}|$ denotes the number of people in the country c who choose option n when answering question q.

3.5 Alignment Metrics

Measuring the alignment between the LLM's opinion distribution $D_{\mathcal{M}}(q)$ and human opinion distribution $D_c(q)$ requires considering the order of the options in these two distributions. Moreover, it is desirable for the metrics to range from 0 to 1. In this paper, we use the *alignment* proposed by Santurkar et al. (2023) as our metric. This metric is used to measure the alignment score between distributions $D_{\mathcal{M}}$ and D_c. The formal definition of *alignment* is as follows:

$$\mathcal{A}(D_{\mathcal{M}}, D_c; \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left(1 - \frac{\mathrm{WD}(D_{\mathcal{M}}(q), D_c(q))}{|\mathcal{N}| - 1}\right) \tag{2}$$

where $\mathcal{Q}$ is the question set and $|\mathcal{N}|$ is the number of options of the question q. The denominator $|\mathcal{N}| - 1$ serves to normalize the Wasserstein distance (WD)[4] so that the value of the metric falls between 0 and 1. The metric's value of 1 indicates that the two distributions are perfectly matched. In addition to *alignment*, we also discuss the normalized mean absolute error and the Jensen-Shannon divergence as alignment metrics in the supplementary materials.

[4]https://en.wikipedia.org/wiki/Wasserstein_metric

4 Global Opinion Alignment (RQ1)

In this section, we calculate the alignment score between LLMs' opinion distributions and human opinion distributions in wave 7. Then, we discuss the alignment of LLMs with human opinion at three levels: *1)* **Model level**: Which LLM most matches the human opinion distribution? *2)* **Country level**: Which countries' opinions are aligned by LLMs? *3)* **Alignment difference level**: If we take the average human opinion distribution as a possible goal, how do LLMs and this average distribution differ in terms of alignment with a country?

4.1 Model Level

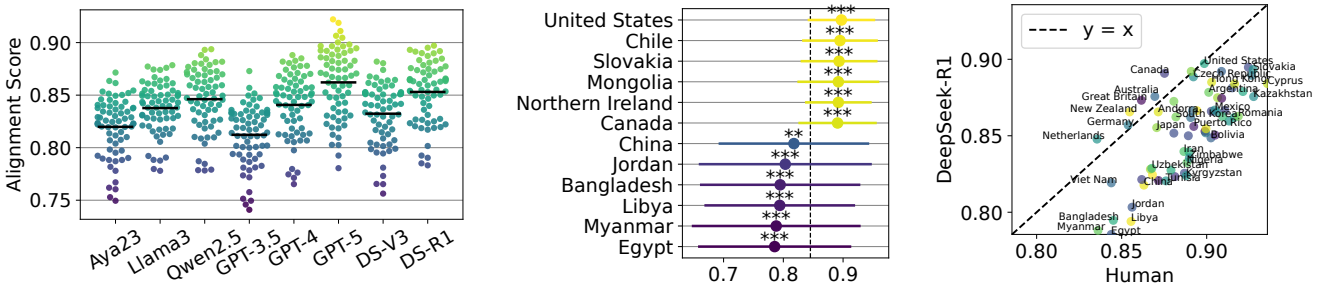
As shown in Figure 2(a), we demonstrate the overall alignment scores of LLMs' opinion distributions with human opinion distributions. The experimental results show that GPT-5 achieves the highest average alignment score (0.8621). Meanwhile, the average alignment scores of Llama3, Qwen2.5, GPT-4, DeepSeek-V3, and DeepSeek-R1 closely follow those of GPT-5, also demonstrating high human opinion alignment performance. In contrast, Aya-23-35B and GPT-3.5-Turbo, as earlier generations of instruction fine-tuning LLMs, have lower average alignment scores and a number of countries with low tail scores; the relative lack of multilingual long-tail coverage may be part of the reason. In addition, the alignment scores of each LLM differ significantly across countries, indicating that no single LLM can simultaneously align with all countries' opinions.

4.2 Country Level

Next, we move our attention to the question of *which countries LLMs are more aligned with*. Figure 2(b) shows the five countries' opinion distributions that are most and least aligned with DeepSeek-R1' opinion distributions.[5] We can see that DeepSeek-R1 has the highest average alignment scores with countries such as the United States and Chile (0.8972 and 0.8948), and relatively low scores with countries such as Egypt and Myanmar (0.7855 and 0.7881). Notably, the standard deviation of the alignment scores in the high-alignment countries is significantly lower than that in the low-alignment countries, indicating that DeepSeek-R1 is both **approximate** and **robust** in the high-alignment countries. In contrast, low-alignment countries suffer from both systematic misalignment and high standard deviations. This may be due to the widespread presence of English and Spanish content,[6] which makes the LLMs highly homogeneous with the dominant views of the United States, Canada, and Chile societies, while low-aligned countries tend to be in non-Latin-speaking environments, which results in a severe scarcity of their political and cultural alignments in the training set. In addition, the alignment with Slovakia and Mongolia is surprisingly high, which implies that the LLM may have learned indirect social signals to Slovakia and Mongolia by borrowing from regional media or cross-language interconnected corpora, suggesting that estimating cultural

[5]Results for all LLMs and the complete list of countries are provided in the supplementary materials.

[6]https://en.wikipedia.org/wiki/Languages_used_on_the_Internet



(a) The alignment score of LLMs with each country. (b) The first and last 6 countries or regions. (c) DeepSeek-R1 vs. Human

Figure 2: The overall results of **RQ1**. (a) The alignment score of LLMs with each country, where each point represents a country and the black line represents the average alignment score of all countries. "DS" is the abbreviation for "DeepSeek." (b) The first and last 6 countries ranked by their alignment scores with DeepSeek-R1's opinion distributions, the point represents the average alignment score of the country on all questions and the line represents the standard deviation. *** denotes p-value $<$ 0.001 (t-test). (c) The relationship of different countries to DeepSeek-R1's opinion distribution and the average human opinion distribution alignment scores. For visibility, we hide some countries. See the supplementary materials for the full version.

alignment purely in terms of corpus size is still insufficient. It is worth noting that although DeepSeek-R1 supports Chinese well, its opinion distribution is not highly aligned with China. In fact, as we show in the supplementary materials, DeepSeek-R1 exhibits a similar pattern to other LLMs (such as GPT-4) in terms of country's opinion alignment.

4.3 Alignment Difference Level

It is impossible to align LLMs with all countries at the same time due to the diversity of cultures. LLMs aligning with the average human opinion distribution is a plausible solution, and we hypothesize that it is the ideal goal. Then, we take the average human opinion distribution as the baseline to investigate how well LLMs fit the different countries' opinions. Figure 2(c) illustrates the relationship of different countries to the DeepSeek-R1's opinion distribution and the average human opinion distribution alignment scores. Consider $y = x$ as the reference line: points that fall on the line indicate that the LLM **appropriately align** with the country's opinions, points that fall above the line indicate that the LLM **over-align** with the country's opinions, and points that fall below the line indicate **under-alignment**. We can see that DeepSeek-R1 most appropriately aligns with the United States' opinions, as well as over-aligns with some developed countries' opinions. It is worth noting that although DeepSeek-R1 is relatively close to Chile, Slovakia, and Mongolia's opinion distributions (Figure 2(b)), when compared to the average human opinion distribution, DeepSeek-R1 under-aligns with these countries even more (Figure 2(c)). Overall, DeepSeek-R1 appropriately aligns or over-aligns with the views of a few countries, while under-aligning with the views of most countries, especially developing countries such as Myanmar, Egypt, and Libya.

5 Steerability of Language (RQ2)

Baselines Steerability refers to the ability of LLMs to adjust and align with the opinion of a target demographic group (Meister, Guestrin, and Hashimoto 2025). Existing steering methods involve prepending additional context to the prompt describing the group we want the model to emulate. We consider the following steering methods as our baselines.

- **Persona Steering** operates by inserting a concise persona description into the original task prompt, prompting the LLM to first "locate" itself as a member of a specific group and thereby explicitly emulate that group's opinion orientation. In this paper, we follow Santurkar et al. (2023) and (Meister, Guestrin, and Hashimoto 2025), in which the LLM is instructed to pretend to be a member of the target country.
- **Few Shot Steering** (Meister, Guestrin, and Hashimoto 2025) provides the persona steering setting while also giving five in-context examples of the real group opinion distributions and asking the LLM to emulate the group's responses. This setting is representative when we already have the opinion distributions of the target group.

Method Language affects the way we think (Turner 2000), that is, speakers of different languages end up with different conceptual structures, and these differences can affect their worldview (Plebe and Vivian 2015). However, it is unclear how language affects the views of LLMs. In this section, we validate the **language steering**, which enables an LLM to represent the opinions of speakers of language l by changing the language of the prompt from English to non-English language l. Specifically, for the task instructions, we translate them using GPT-4 and manually check the accuracy of the translated versions; for the in-context examples, we obtain the translations from the corresponding translated versions of the WVS questionnaires. Unlike previous work (Durmus et al. 2023) that only evaluates three languages, uses log probabilities of one LLM, and validates the language steering itself, we evaluate eight languages, use the verbalized distribution of seven LLMs, and validate the effect of combining the language steering with other steering methods.

Method	Aya23	Llama3	Qwen2.5	GPT-3.5	GPT-4	GPT-5	DS-V3	DS-R1	AVG.
				China					
No Steering (En.)	0.8038	0.8235	0.8273	0.7816	0.8152	0.8348	0.7952	0.8174	0.8124
+*Language Steering* (Zh.)	**0.8103**	**0.8308**	**0.8439**	**0.7976**	**0.8526***	**0.8601**	**0.8471****	**0.8498***	**0.8365*****
Persona Steering (En.)	0.8154	0.8439	0.8316	0.7846	0.8247	0.8953***	0.8278*	0.8625**	0.8357***
+*Language Steering* (Zh.)	**0.8216**	**0.8553***	**0.8629***	**0.8088**	**0.8718*****	**0.8979*****	**0.8732*****	**0.8828*****	**0.8593*****
Few-shot Steering (En.)	0.8115	**0.8666****	0.8596	0.8149	0.8629**	0.8996***	0.8775***	0.8884***	0.8601***
+*Language Steering* (Zh.)	**0.8469****	0.8652**	**0.8838*****	**0.8345****	**0.8961*****	**0.9041*****	**0.8884*****	**0.8976*****	**0.8771*****
				Germany					
No Steering (En.)	0.8175	0.8441	0.8541	0.7993	0.8468	**0.8826**	0.8280	0.8543	0.8408
+*Language Steering* (De.)	**0.8445***	**0.8494**	**0.8815****	**0.8140**	**0.8765****	0.8801	**0.8689****	**0.8715**	**0.8608****
Persona Steering (En.)	0.8081	**0.8510**	0.8596	0.8193	0.8736*	0.9129***	0.8731***	0.8879***	0.8607***
+*Language Steering* (De.)	**0.8482***	0.8507	**0.8904*****	**0.8250**	**0.8938*****	**0.9130*****	**0.8905*****	**0.9053*****	**0.8771*****
Few-shot Steering (En.)	0.8345	0.8348	0.8619	0.8352*	0.8670	0.8983	0.8709***	0.8581	0.8576***
+*Language Steering* (De.)	**0.8661*****	**0.8683**	**0.8916*****	**0.8476****	**0.9097*****	**0.9219*****	**0.9072*****	**0.9103*****	**0.8903*****

Table 1: Culture representation scores for China and Germany under different steering methods for LLMs. The content in "()" is to indicate the language being used, where "En." denotes English, "Zh." denotes Chinese, and "De." denotes German. In all cases, the language of our inputs (task instruction, few-shot examples, and question) always keeps the same. Moreover, the significance is assessed using the t-test: * denotes p-value < 0.05, ** denotes p-value < 0.01, and *** denotes p-value < 0.001.

Because WVS provides a multilingual translated version of the survey questionnaire for countries that use multiple languages (e.g., Singapore uses Chinese, English, and Malay; respondents may see the questionnaire in both English and Malay), the survey results from these countries may not reflect the opinions of single-language speakers. Therefore, in order to quantitatively analyze the effect of language steering, we must avoid respondents using multilingual questionnaires. The languages and countries we choose to investigate must satisfy the following conditions: *1)* the country must be surveyed by WVS using a single language; and *2)* the language must be supported by the LLM. After our selection, our experiments cover Spanish, Chinese, Japanese, Korean, German, Russian, Vietnamese, and Portuguese. As WVS covers survey data from 12 Spanish-speaking countries, we select three countries, Argentina, Chile, and Uruguay, to validate language steering. For other languages, we select one country for each language.

Effectiveness of Steering Table 1 shows the alignment scores for China and Germany under different steering methods for LLMs.[7] We show these two countries (or languages) due to their representation of non-Latin (Chinese) and Latin writing systems (German), respectively, forming a striking contrast in terms of language families and verifying the universality of the steering methods. We can see that the baseline methods persona steering and few-shot steering show a certain steering ability on all countries compared to no steering. In general, few-shot steering exhibits higher steering ability than persona steering, and our findings are consistent with existing studies (Meister, Guestrin, and Hashimoto 2025; Studdiford et al. 2025). Moreover, "few-shot + language steering" achieves the highest alignment scores on both countries, which indicates the effectiveness of language in steering LLMs to emulate the opinions of language speakers. Although persona steering can be inaccurate (Meister, Guestrin, and Hashimoto 2025), "persona + language steering" also yields promising steering ability on China and Germany. In addition, among the LLMs tested, GPT-5 exhibits the strongest response to the composite strategy, reflecting its greater ability to understand the countries we tested. Overall, language steering is effective in steering LLMs to emulate country-specific opinions. This result implicates that while chasing higher benchmark scores may suggest using a language other than the target one (Kaneko, Aji, and Baldwin 2025), for opinion-oriented tasks, sticking to the target language can yield better alignment. Therefore, whenever we consider deploying other languages, we must pay careful attention to value alignment and similar factors.

[7]The results of other languages are available in the supplementary materials.

6 Temporal Opinion Alignment (RQ3)

Human cognition has a social historical character (Berger and Luckmann 2016), which means that human cognition is not static; it is constantly being reshaped as historical conditions evolve (Vygotsky 2012; Mannheim 2013). Therefore, it is important to know whether LLMs reflect contemporary human opinions. In this study, we suppose that *it is desirable for LLMs to reflect contemporary human opinions*. Next, we will show the representativeness of LLMs with human opinion distributions in the temporal dimension.

Method As shown in the results of Figure 2(c), the LLM is under-aligned with most countries' opinions. We therefore consider which countries are well aligned by the LLM for comparison. Formally, we use the following equation to filter the countries for comparison:

$$\mathcal{C}^* = \{\, c \in \mathcal{C} \mid \left| \mathcal{A}_{\mathcal{M}}^{(c)} - \mathcal{A}_{avg}^{(c)} \right| < \tau \,\}. \tag{3}$$

where $\mathcal{A}_{\mathcal{M}}^{(c)}$ denotes the alignment score between the LLM's opinion distribution and that of country c. Likewise, $\mathcal{A}_{avg}^{(c)}$

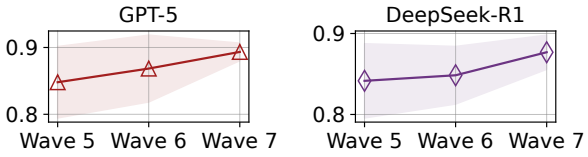


Figure 3: The trend of the average alignment scores of GPT-5 and DeepSeek-R1 with countries filtered using Eq. (3) across waves. The shaded area indicates the standard deviation of the alignment scores at the current wave.

is the alignment score between the average human opinion distribution and that of country c. The margin τ is the threshold, which we set to 0.02 to ensure that we can filter out at least five countries that are well aligned by LLMs. We refer readers to the supplementary materials for the filtering procedures, the filtered countries, and the experimental results.

Results Figure 3 illustrates the trend of the average alignment scores of GPT-5 and DeepSeek-R1 across waves. We can see that LLMs most align with the latest wave of human opinions. One reason may be due to the fact that most of the human feedback in the fine-tuning of LLMs comes from contemporary annotators, whose ethical and social orientations are more in line with recent public opinions. The models are rewarded for emulating these behaviors. Another reason may be due to the development of mobile internet and social media, which provide more accessible data than in the past. In addition, the standard deviation reflects an LLM's coverage degree of the represented opinions. A lower standard deviation indicates a better coverage degree of the opinions. Therefore, LLMs achieve better coverage of human opinions in the latest wave and poorer coverage of human opinions in earlier waves. This further indicates that the alignment of LLMs is influenced by contemporary annotators' human feedback. In contrast, it implies that there may be challenges for LLMs to align to earlier human opinions. The reason is that it is more difficult to access earlier customized human feedback data.

7 Discussion

7.1 Alignment of Human Opinions across Countries

To understand the alignment between different countries' opinions, we show the alignment scores between countries and countries in Figure 4. Similar to Figure 2(b), we plot the first and last 6 countries ranked by their alignment scores with DeepSeek-R1's opinion distributions. For comparison, we also show the alignment scores between the countries and DeepSeek-R1 in the results. More results can be found in the supplementary materials. The alignment scores of opinion distributions among the first 6 ranking countries are generally high, indicating significant similarity in their views. In contrast, the alignment scores among the last 6 ranking countries show relative differences. For example, Jordan, Libya, and Egypt have high alignment scores

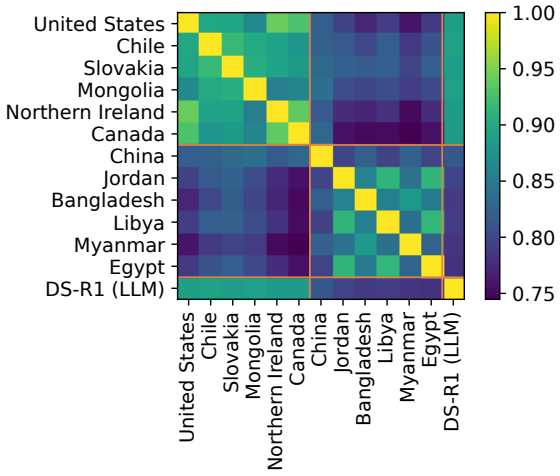


Figure 4: Alignment scores among different countries. For comparison, we also show the alignment scores between the countries and DeepSeek-R1 (DS-R1).

with each other, while Bangladesh, China, and Myanmar have low alignment scores with other countries. Meanwhile, the alignment scores between the first and last 6 ranking countries are generally low. Furthermore, by calculating the alignment scores between DeepSeek-R1 and each country, we find that despite some heterogeneity in alignment scores among the first 6 ranking countries, DeepSeek-R1's alignment scores with these countries are surprisingly consistent. This pattern suggests that the model captures an "average" cross-country view rather than aligning specifically to individual countries, revealing a lack of country-specific sensitivity in the alignment of opinion distributions.

7.2 Internal Consistency of LLMs

In order to investigate whether LLMs hold self-contradictory opinions when answering questions, we conduct a closer look at the internal consistency of LLMs. Concretely, we manually select four questions about gender fairness (agree or disagree that men are superior to women), four questions about atheism (believe or disbelieve in theism), and two questions about democracy (important or not important that democracy).[8] Questions with the same topic reflect the same logic of opinion judgment. In our experiments, we consider the options "Strong agree" and "Agree," which both express positive attitudes, to be consistent opinions. For options related to degree (e.g., from "1. Not at all important" to "10. Absolutely important"), we group options 1 to 5 and options 6 to 10 as consistent opinion groups, respectively.

We use the internal consistency rate, which is the ratio of answers within the same opinion group, as the evaluation measure. If the answers to a set of questions are [1, 1, 2, 1], the internal consistency rate is 75%. The experimental results show that on the topic of gender fairness, all the LLMs remain at a 100% internal consistency rate. However,

[8] A complete list is provided in the supplementary materials.

Models	Order	#few-shot	Prob.
Aya23	0.9660	0.9893	0.9815
Llama3	0.9927	0.9964	0.9933
Qwen2.5	0.9928	0.9928	0.9973
GPT-3.5	0.9852	0.9810	0.9736
GPT-4	0.9925	0.9783	0.9902
GPT-5	0.9973	0.9968	0.9980
DeepSeek-V3	0.9895	0.9805	0.9881
DeepSeek-R1	0.9885	0.9861	0.9924

Table 2: Pearson's r between the default and sensitivity test results. Here, the default result means the alignment scores between LLMs' opinion distributions and those of different countries under default prompts. The sensitivity test result means the alignment scores between LLMs' opinion distributions and those of different countries under sensitivity test prompts. **Order** means shuffle the input question's options, **#few-shot** means limit the number of the few-shot examples from five to three, and **Prob.** means change the probability distribution of the few-shot examples. *Note*: The p-values for all results are less than 0.001.

GPT-3.5-Turbo shows only 50% internal consistency on the topics of atheism and democracy, in contrast to other LLMs, which maintain full consistency. This suggests that, relative to other LLMs, GPT-3.5-Turbo not only poorly aligns with human opinions (§4), but also lacks internal consistency.

7.3 Sensitivity of LLMs

We acknowledge that current LLMs are very sensitive to the prompt format (Zhuo et al. 2024). Inspired by Santurkar et al. (2023), we test the prompt sensitivity of the LLMs used in our experiments against the following factors:

- The order in which the question options are presented to the model;
- The few-shot examples (the number of the few-shot examples and the probability distribution of the few-shot examples).

We employ the Pearson's r of the alignment scores of different sensitivity test prompts for evaluating prompt sensitivity. As shown in Table 2, despite changing prompts, the results output by LLMs still maintain a significantly high Pearson's r, and the p-value is less than 0.001. Moreover, the supplementary materials provide more results.

8 Related Work

Human Opinion and Culture. Culture can be defined as the pattern of thinking, feeling, and reacting, distinguishing human groups (Wallace and Fogelson 1961; Shweder et al. 2007; Cao et al. 2023). Human opinions are the beliefs, preferences, and judgments held by individuals, and are deeply influenced by cultural backgrounds (Peterson 2003; Markus and Kitayama 2014). Culture provides people with a shared framework through which they interpret experiences, internalize norms, and express their opinions (Geertz 2017). It is challenging to align the opinions of human societies with different cultural backgrounds (Hall 1976).

Culture in LLMs. Previous studies have shown that LLMs exhibit cultural challenges similar to those found in human society (Li et al. 2024). Cecilia Liu et al. (2024) pointed out that there is a cultural gap when using LLMs, especially when dealing with content translated from other languages. After that, Masoud et al. (2025) demonstrated that GPT-4 exhibits remarkable capabilities in adapting to these cultural gaps. Furthermore, Wang et al. (2024) showed that LLMs have cultural dominance issues because they are mainly trained using English data. In addition, recent study (Meister, Guestrin, and Hashimoto 2025) has challenged the existing use of LLMs to emulate opinion distributions, proving that LLMs are more accurate when describing opinion distributions. Therefore, one technical difference from existing studies is that we deployed this method of describing opinion distribution (known as *verbalized distribution*).

Opinion of LLMs. LLMs often respond to subjective questions in a way that implicitly favors certain human groups. Early study (Santurkar et al. 2023) quantified the differences between LLMs' predicted distributions of response options and human distributions based on surveys. Meister, Guestrin, and Hashimoto (2025) expanded the assessment scope to multiple demographic dimensions and compared different distribution expressions (model log-probability, sequence of tokens, and verbalized distribution). Building on this research direction, this study expands the scope from a few countries to a truly worldwide country sample and explicitly introduces the temporal dimension.

Steering Methods. To enable LLMs to emulate specific groups, previous studies (Meister, Guestrin, and Hashimoto 2025) have explored persona and few-shot steering. However, persona steering is not always accurate, leading to undesirable effects such as stereotypes and exacerbated polarization (Perez et al. 2023; Cheng, Durmus, and Jurafsky 2023; Wang, Morgenstern, and Dickerson 2025). We further expanded on these two methods and proposed language steering: changing the language of the prompt can significantly improve alignment with the opinion distribution of the corresponding country, indicating that language itself, as an alignment cue, can drive opinion alignment.

9 Conclusion

In this paper, we propose a framework to investigate the global human opinion alignment of LLMs using the WVS questionnaires and data. We manually filter the 259 questions in the WVS questionnaires across multiple language versions and historical periods to build our dataset. Through experimenting with LLMs on our data, our main findings are as follows: *1)* LLMs only appropriately align with a few countries, while under-aligning with most countries. *2)* Language can significantly steer LLMs to emulate country-specific opinions. *3)* LLMs most align with the contemporary human opinions. Our study also discusses various aspects of LLMs, such as internal consistency and sensitivity of LLMs. This paper will provide insights for studies related to the value alignment of LLMs.

Acknowledgments

We thank the reviewers for their valuable feedback. This work was supported by JST BOOST, Grant Number JPMJBS2407.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 Model Card. https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3. Accessed: 2025-12-04.

AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12404–12422. Bangkok, Thailand: Association for Computational Linguistics.

Aryabumi, V.; Dang, J.; Talupuru, D.; Dash, S.; Cairuz, D.; Lin, H.; Venkitesh, B.; Smith, M.; Campos, J. A.; Tan, Y. C.; et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Berger, P.; and Luckmann, T. 2016. The social construction of reality. In *Social theory re-wired*, 110–122. Routledge.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cao, Y.; Zhou, L.; Lee, S.; Cabello, L.; Chen, M.; and Hershcovich, D. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In Dev, S.; Prabhakaran, V.; Adelani, D. I.; Hovy, D.; and Benotti, L., eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 53–67. Dubrovnik, Croatia: Association for Computational Linguistics.

Cecilia Liu, C.; Koto, F.; Baldwin, T.; and Gurevych, I. 2024. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2016–2039. Mexico City, Mexico: Association for Computational Linguistics.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1504–1532. Toronto, Canada: Association for Computational Linguistics.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Geertz, C. 2017. *The interpretation of cultures*. Basic books.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; and Puranen, B. 2022. World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0. Dataset.

Hall, E. T. 1976. *Beyond culture*. Anchor.

Kaneko, M.; Aji, A. F.; and Baldwin, T. 2025. Balanced Multi-Factor In-Context Learning for Multilingual Large Language Models. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 20096–20115. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Li, C.; Chen, M.; Wang, J.; Sitaram, S.; and Xie, X. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37: 84799–84838.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Mannheim, K. 2013. *Ideology and utopia*. Routledge.

Markus, H. R.; and Kitayama, S. 2014. Culture and the self: Implications for cognition, emotion, and motivation. In *College student development and academic life*, 264–293. Routledge.

Mason, W.; and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1): 1–23.

Masoud, R.; Liu, Z.; Ferianc, M.; Treleaven, P. C.; and Rodrigues, M. R. 2025. Cultural Alignment in Large Language

Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 8474–8503. Abu Dhabi, UAE: Association for Computational Linguistics.

Meister, N.; Guestrin, C.; and Hashimoto, T. 2025. Benchmarking Distributional Alignment of Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 24–49. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

OpenAI. 2023. GPT-3.5 Turbo. https://platform.openai.com/docs/models/gpt-3.5-turbo. Accessed: 2025-12-04.

OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5. Accessed: 2025-12-04.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; Jones, A.; Chen, A.; Mann, B.; Israel, B.; Seethor, B.; McKinnon, C.; Olah, C.; Yan, D.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Khundadze, G.; Kernion, J.; Landis, J.; Kerr, J.; Mueller, J.; Hyun, J.; Landau, J.; Ndousse, K.; Goldberg, L.; Lovitt, L.; Lucas, M.; Sellitto, M.; Zhang, M.; Kingsland, N.; Elhage, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Rausch, O.; Larson, R.; McCandlish, S.; Johnston, S.; Kravec, S.; El Showk, S.; Lanham, T.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Clark, J.; Bowman, S. R.; Askell, A.; Grosse, R.; Hernandez, D.; Ganguli, D.; Hubinger, E.; Schiefer, N.; and Kaplan, J. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434. Toronto, Canada: Association for Computational Linguistics.

Peterson, M. F. 2003. Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations.

Plebe, A.; and Vivian, M. 2015. When language shapes perception. *Rivista Italiana di Filosofia del Linguaggio*, 9(2).

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Shweder, R. A.; Goodnow, J. J.; Hatano, G.; LeVine, R. A.; Markus, H. R.; and Miller, P. J. 2007. The cultural psychology of development: One mind, many mentalities. *Handbook of child psychology*, 1.

Studdiford, Z.; Rogers, T. T.; Suresh, S.; and Mukherjee, K. 2025. Evaluating Steering Techniques using Human Similarity Judgments. *arXiv preprint arXiv:2505.19333*.

Sukiennik, N.; Gao, C.; Xu, F.; and Li, Y. 2025. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Turner, J. H. 2000. *On the Origins of Human Emotions*. Redwood City: Stanford University Press. ISBN 9780804764360.

Vygotsky, L. S. 2012. *Thought and language*, volume 29. MIT press.

Wallace, A. F.; and Fogelson, R. D. 1961. Culture and personality. *Biennial review of anthropology*, 42–78.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 1–12.

Wang, W.; Jiao, W.; Huang, J.; Dai, R.; Huang, J.-t.; Tu, Z.; and Lyu, M. 2024. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6349–6384. Bangkok, Thailand: Association for Computational Linguistics.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. *arXiv preprint arXiv:2410.12405*.