# Made in China, Aligned with the USA: Chinese LLMs Express Similar Moral Values to American People

Anonymous ACL submission

## Abstract

Many studies have reported that large language models (LLMs) tend to express similar values to people from Western countries, such as prioritizing individualism over collectivism. However, evidence for this ethical bias comes mostly from LLMs made by American companies. The current crop of state-of-the-art models includes several made in China, so we conducted the first large-scale investigation of how models made in China and the USA align with people from China and the USA. We elicited responses from ten Chinese models and ten American models to the Moral Foundations Questionnaire 2.0 and the World Values Survey, two well-validated measures with responses from thousands of Chinese and American people. We found that all models respond to both surveys more like American people than like Chinese people. This skew toward American responses is only slightly mitigated when prompting the models in Chinese or imposing a Chinese persona on the models. Given that LLMs may serve as tools for soft power competition between China and the USA, this persistent alignment with Americans may have important implications for geopolitics, and given that LLMs already help people make decisions, it has important implications for daily life.

## 1 Introduction

For all but a few companies, large language models (LLMs) are prohibitively expensive to train. They require thousands of GPUs, large amounts of energy, and institutional knowledge, so until recently, state-of-the-art LLMs were made almost exclusively in the USA. LLMs already speak and make decisions on behalf of people, which raises the possibility that LLMs will propagate American value priorities (e.g., Atari et al., 2023b). However, the Chinese company DeepSeek made headlines in early 2025 by training state-of-the-art LLMs for a fraction of the cost of American models (Guo et al., 2025), and other Chinese companies, such as Alibaba and Baidu, have released several models that rank toward the top of LLM leaderboards (Yang et al., 2025; ERNIE Team, 2025). In this paper, we compare human participants' responses to the Moral Foundations Questionnaire 2.0 (MFQ-2; Atari et al., 2023a) and the World Values Survey (WVS; Haerpfer et al., 2022) with responses from LLMs made in China and the USA. As previewed in Figure 1, we show that all the LLMs respond more like American participants.

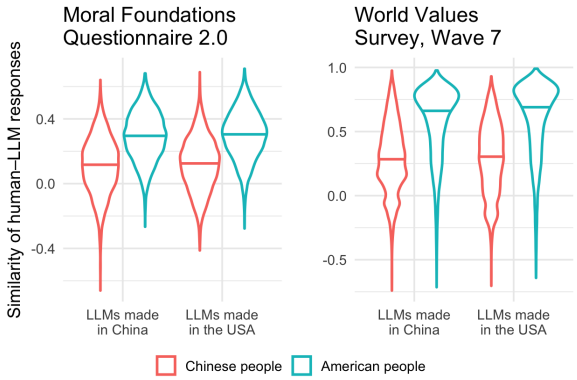


Figure 1: Cosine similarity of LLM responses to human responses on the MFQ-2 and WVS. Cross bars indicate means.

Our contributions are as follows. (1) As far as we are aware, this is the first systematic comparison of the moral values expressed by LLMs made in China versus the USA, using data from thousands of human subjects. (2) We base our value-elicitation method on well-established approaches to measuring human values, the MFQ-2 and WVS. (3) We study the effects of manipulating prompt language and imposing national personas on LLMs. (4) We have made our data, containing the values expressed by 20 LLMs, available on Github: https://github.com/Anonym-Academic-Submission/China_US_LLM_Alignment.

## 2 Background

### 2.1 The Moral Foundations Questionnaire 2.0 and the World Values Survey

In this work, we use Moral Foundations Theory (MFT) and the WVS to evaluate cross-cultural variation in moral values. This choice is motivated by limitations in a growing body of work that probes moral values in LLMs using data-driven (rather than theory-driven) questionnaires (e.g., Emelin et al. 2021; Forbes et al. 2020). While such approaches are broad and scalable, they lack theoretical grounding and empirical validation, and without systematically collected human response data from different populations, they cannot compare model behaviour and population-level value patterns in different societies. In contrast, MFT is among the most developed frameworks in moral psychology, grounded in theories of moral psychology and linked to evolutionary and cultural anthropology and adjacent work in social neuroscience (e.g., Atari et al. 2023a; Graham et al. 2013). It provides a theoretical account of how culturally diverse moral norms can arise from underlying themes, conceptualized as moral "foundations." Empirically, MFT has been studied across diverse cultural contexts, and its updated instrument (the MFQ-2, released in 2023) was validated with standardized translations on large human samples across 25 populations. The WVS offers an independently developed and extensively fielded survey instrument: Its translations and items have been tested by social scientists, and its most recent wave (2017–2023) spans 66 countries, with large samples that reflect the age, gender, and education levels of those countries. In short, the MFQ-2 and WVS provide recent, large, publicly accessible, item-level survey responses from Chinese and American people.

The MFQ-2 consists of 36 items probing six moral dimensions, which Atari et al., 2023a describe as follows:

**Care:** Intuitions about avoiding emotional and physical damage to another individual.
**Equality:** Intuitions about equal treatment and equal outcome for individuals.
**Proportionality:** Intuitions about individuals getting rewarded in proportion to their merit or contribution.
**Loyalty:** Intuitions about cooperating with ingroups and competing with outgroups.
**Authority:** Intuitions about deference toward legitimate authorities and the defense of traditions.
**Purity:** Intuitions about avoiding bodily and spiritual contamination and degradation.

Participants respond to each statement with a rating from 1, indicating that the statement "Does not describe me at all", to 5, indicating that the statement "Describes me extremely well". For example, the statement "I think children should be taught to be loyal to their country" gauges the Loyalty dimension (Atari et al., 2023a).

The seventh wave of the WVS consists of 290 items probing religious values, attitudes toward migration, and many other themes. The rating scales for items vary (e.g., some are binary ratings, some range from 1 to 4, others range from 1 to 10), and some items lack responses from Chinese participants (e.g., questions about security and political regimes), so in this study, we focus on 19 questions from the Ethical Values and Norms section (questions 177–195). Those 19 items share a prompt ("Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between") and a rating scale (from 1, "Never justifiable", to 10, "Always justifiable"). The 19 actions, including euthanasia, terrorism, and divorce, are enumerated in Figures 2 and 5.

### 2.2 Moral Values in LLMs

Despite early instances of chatbots spewing hate speech (e.g., Wolf et al., 2017), LLMs have come to encode commonsense values and abide by social norms (e.g., Schramowski et al., 2022). For example, models from OpenAI, Google, Meta, and Anthropic are confident that drivers should avoid pedestrians, while they express uncertainty about assisted suicide (Scherrer et al., 2023). However, different societies have different norms, and LLMs are more likely to express the values of people from Western countries. For example, in two studies with tens of thousands of participants from dozens of countries, GPT models respond to the MFQ-2 more like participants from Western countries than other countries (Zewail et al., 2025), and they respond to the World Values Survey more like participants from countries more culturally similar to the USA (Atari et al., 2023b; see also Qi et al., 2025).

The skew toward Western values is clear, but evidence comes almost entirely from American-made LLMs. Liu et al., 2024 therefore compared two

American models, GPT-3.5 and Gemini, to two Chinese models, ChatGLM-2 and Ernie. They found that the American LLMs express more individualist values, whereas the Chinese LLMs express more collectivist values, consistent with an LLM's country of origin dictating its morals. However, Liu et al., 2024 measured the similarity of LLMs to a single human sample, 30 Chinese university students, which they treat as a proxy for Western people (i.e., they assume that young, educated people have more individualist values). Münker, 2025 more directly compared the values expressed by LLMs from different countries (the USA, France, and China) to the values expressed by people from different countries (the USA and South Korea). However, that study emphasizes the skew toward liberal versus conservative values within each country, not the skew toward the values of one country or another, and the human populations do not correspond to the LLMs' countries of origin (i.e., there are no French or Chinese participants). Huang et al., 2024 report performance by 14 Chinese-made LLMs on a values benchmark, but they do not investigate whether country of origin affects alignment with human populations. So, it remains to be seen whether the bias toward Western norms persists in Chinese LLMs.

### 2.3 Steering LLMs' values

A recurring question in the literature is how to mitigate the skew toward Western values in LLMs. Several studies report that the values expressed by LLMs vary as a function of language. For example, on the MFQ-2, LLMs from OpenAI, Meta, and Mistral rated the Care, Loyalty, and Purity dimensions higher when prompted in Western languages, such as English and French, than in Eastern languages, such as Chinese and Japanese (Aksoy, 2025). However, this runs counter to the finding that the Western skew in LLMs involves low Purity ratings (Zewail et al., 2025), which suggests that although prompt language affects LLM responses, it does not align LLMs with the values of speakers of that language. The simpler conclusion is that performance in general, and moral reasoning ability in particular, decreases in lower resource languages (e.g., Arora et al., 2023; Durmus et al., 2023; Hämmerl et al., 2023; Kwok et al., 2024; Wang et al., 2024). The multilingual abilities of LLMs have improved substantially since the release of GPT-4, and larger, newer LLMs respond to ethical dilemmas more consistently across languages (e.g., Agarwal et al., 2024; Khandelwal et al., 2024; Naous et al., 2024), but that stability implies a persistent Western bias, such as low Purity ratings or an inclination toward individualism.

A more effective way to mitigate biases in LLMs is by assigning personas to them (e.g., Kwok et al., 2024; Simmons, 2023; Wang et al., 2024; Wright et al., 2024). For example, Qi et al., 2025 found that instructing GPT-3.5 to act like a person from a specific country reduced the bias towards dominant populations in its responses to the World Values Survey. However, they stress that persona prompts do not eliminate biases entirely. Lee et al., 2024 similarly report that on the MFQ, LLMs from OpenAI, Anthropic, and Meta exhibit a striking consistency across personas, and Münker, 2025 reports that LLMs from Google, Meta, Mistral, and Alibaba can be steered toward liberal or conservative values only to a limited degree. When manipulating an LLM's persona, its moral biases persist, which underscores the importance of identifying those biases in state-of-the-art LLMs.

## 3 Methodology

### 3.1 Research questions

We investigated three research questions:

**RQ1.** Do LLMs made in China versus the USA express similar values as people from China versus the USA, respectively?
**RQ2**. Does manipulating questionnaire language or imposing a national identity steer LLMs toward the value patterns of the corresponding human population?
**RQ3**. Which moral dimensions on the MFQ-2, and which questions on the WVS, explain the alignment of LLMs made in China versus the USA with people from China versus the USA?

### 3.2 Analysis plan

We operationalized moral values as responses to the MFQ-2 and to the Ethical Values and Norms section of the WVS. We selected 10 LLMs made by American companies and 10 made by Chinese companies. See Appendix A. For each LLM, we manipulated whether the questionnaire was presented in English or Chinese and whether the system prompt said that the LLM was a Chinese or American citizen or did not mention nationality, i.e., a design of 2 (language: Chinese or English) x 3 (persona: Chinese, American, or null). To account for the

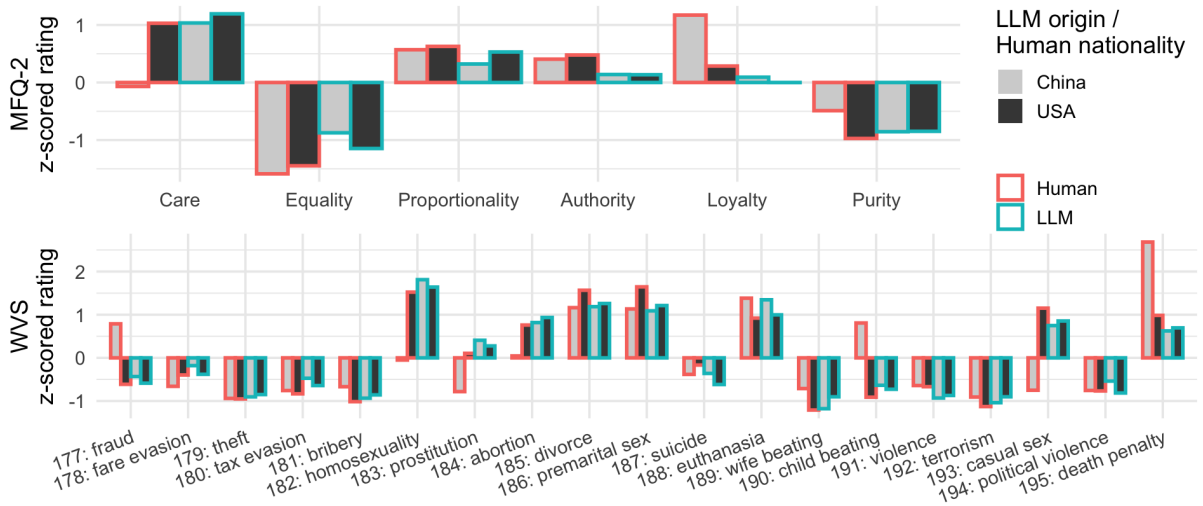


Figure 2: Normalized human and LLM ratings, grouped by dimension for MFQ-2 and question for WVS. Responses are z-scored within each participant and within each LLM. These data exclude the manipulation of persona and language.

probabilistic nature of LLM output, we presented the surveys to each LLM in each condition 20 times and computed the mean response to each item in each condition. For the MFQ-2, we used human data from Atari et al., 2023a, who presented the survey to 517 Chinese participants and 515 American participants. For the WVS, we used data from Haerpfer et al., 2022, who report responses to each of the 19 Ethical Values items from 2,808 Chinese participants and 2,357 American participants.

We represented responses from each participant and each LLM as vectors in a 36-dimension space, for MFQ-2, or a 19-dimension space, for WVS. We z-scored the ratings for each item within each participant and within each LLM in each condition (after averaging the 20 responses to each item), such that each participant and each LLM has a mean normalized rating of 0, and positive or negative values indicate a tendency to give higher or lower ratings to that item. To measure how similar LLM and human responses are, we calculated the cosine similarity between vectors.

#### 3.2.1 Prompt construction

As shown schematically in Appendix B and detailed in the online supplement, we adapted the MFQ-2 and WVS instructions to elicit moral judgements from LLMs. MFQ-2 participants are identified as being from China or the USA, so the MFQ-2 persona prompt states "You are from China", "You are from the USA", or is left empty. WVS participants are identified as Chinese or American nationals, so the WVS persona prompt states "You are a Chinese national", "You are an American national", or is left empty. The content in the English and Chinese language conditions is identical to the original English and Chinese surveys. In the Chinese language condition, all aspects of the prompt are in simplified Chinese.

#### 3.2.2 RQ1

To answer RQ1, we fit two mixed effects models (separately for the MFQ-2 and WVS), regressing the cosine similarity of human–LLM responses on the interaction of participant nationality (Chinese or American) and LLM origin (China or the USA). Each model includes 20 observations per participant (i.e., compared to each LLM) and hundreds of observations per LLM (i.e., compared to each participant), so to capture individual differences and avoid inflating p-values, we included by-LLM random slopes for participant nationality and by-participant slopes for LLM country of origin. We sum coded the participant nationality and LLM origin factors such that being from China is the positive value (i.e., Chinese = 0.5, American = -0.5). Consequently, a positive main effect of LLM origin would indicate that Chinese LLMs respond more like human participants, and a positive main effect of participant nationality would indicate that all LLMs respond more like Chinese participants. A positive interaction would indicate that any tendency to respond more like Chinese participants is greater in Chinese LLMs than American LLMs. For the first analysis, we did not manipulate persona, and we paired human responses with LLM

responses in the same language. For the WVS, five LLMs (Claude 4, DeepSeek-V2, ERNIE-4.5, GPT-4, and GPT-4o) refused to respond to some items in all 20 iterations, such as whether abortion is justifiable, excluding them from this analysis of similarity on the WVS. In Appendix C, we detail differences in refusal rates for the MFQ-2 and WVS by LLMs made in China and the USA.

#### 3.2.3 RQ2

To answer RQ2, we again fit two mixed effects models, regressing the cosine similarity of human and LLM responses on the interaction of participant nationality with LLM country of origin, persona, and language. We again included by-LLM random slopes for participant nationality and by-participant slopes for LLM country of origin. We sum coded all factors such that being a person from China, being made in China, being assigned a Chinese persona, or being prompted in Chinese is the positive value (i.e., Chinese = 0.5, American = -0.5). In each case, a positive interaction with the nationality factor would indicate that any tendency to respond more like Chinese participants is greater in Chinese LLMs, when assigned a Chinese persona, or when prompted in Chinese, respectively.

#### 3.2.4 RQ3

To answer RQ3, we re-standardized and recomputed the cosine similarity of LLM and human responses when removing each dimension (for MFQ-2) or each question (for WVS). As with RQ1, we did not manipulate persona, and we paired human responses with LLM responses in the same language. (As we show in the online supplement, including the persona and language manipulations does not change the direction or significance of the critical effects.) We regressed the cosine similarity of human and LLM ratings on the interaction of participant nationality, LLM country of origin, and dimension or question removed (including by-LLM random slopes for participant nationality and by-participant slopes for LLM country of origin). We sum coded nationality and origin such that being from China is the positive value (i.e., American = -0.5; Chinese = 0.5), and crucially, we treatment coded dimension / question such that none removed is the reference level (i.e., none removed = 0; dimension / question removed = 1). The main effects of nationality and origin therefore do not involve removing any dimensions / questions, and the interactions with a given dimension / question indicate how those main effects change when removing that dimension / question. For example, if LLMs tend to be more similar to American participants, nationality will have a negative main effect, but if the Purity dimension were the sole cause of this greater similarity to Americans, then removing Purity would have a positive interaction as large as the negative main effect, cancelling it out.

| Predictor | $\beta$ | *SE* | *p* |
|---|---|---|---|
| MFQ-2 | | | |
| Intercept | .207 | .009 | < 2e-16 |
| Chinese person | -.179 | .015 | 6e-13 |
| Chinese LLM | -.008 | .017 | .663 |
| Chinese person : LLM | -.002 | .026 | .951 |
| WVS | | | |
| Intercept | .423 | .008 | < 2e-16 |
| Chinese person | -.297 | .016 | 5e-14 |
| Chinese LLM | -.011 | .014 | .432 |
| Chinese person : LLM | .028 | .028 | .340 |

Table 1: Results of two mixed effects models regressing the cosine similarity of LLM–human responses on the interaction of participant nationality with LLM country of origin. Both factors are sum coded and mean centered with Chinese as the positive value.

## 4 Results

Figure 2 illustrates mean human and LLM ratings for the six MFQ-2 moral dimensions and the 19 WVS ethical questions, grouped by nationality for human participants and grouped by country of origin for LLMs. For example, on the MFQ-2, Chinese participants (the light grey bars with red outlines) gave high ratings to questions probing the Loyalty dimension, and on the WVS, they rated the death penalty (Question 195) as very justifiable.

### 4.1 RQ1: Do people and LLMs from the same country share values?

To investigate whether LLMs from China versus the USA respond more like participants from China versus the USA, we regressed the cosine similarity of each participant's ratings to each LLM's rating on the interaction of participant nationality with LLM country of origin. As reported in Table 1, there is a significant negative main effect of participant nationality for both the MFQ-2 and WVS (both *p*s < 2e-16), indicating greater similarity of LLMs to American participants than to Chinese participants (since Chinese participants are coded as the positive value). For the MFQ-2, the interac-

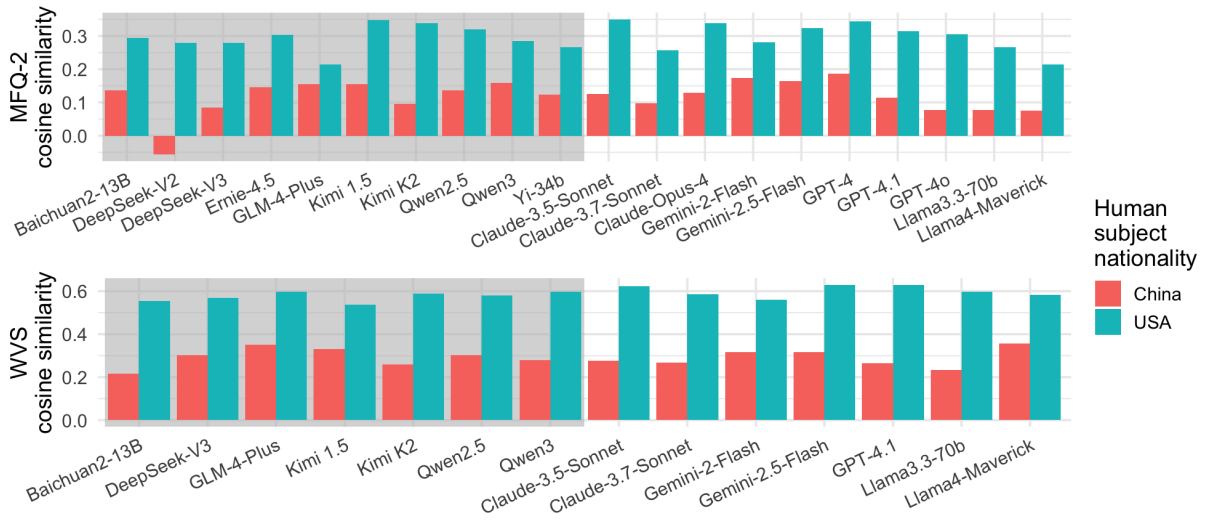


Figure 3: Cosine similarity of each human's ratings to each LLM's ratings. The bars with grey backgrounds are LLMs made in China.

tion of participant nationality with LLM country of origin is far from significant ($p > .9$), therefore providing no evidence that Chinese LLMs tend to respond more like Chinese participants or that American LLMs respond more like American participants. In fact, the interaction trends in the opposite direction (i.e., has a negative sign). For the WVS, that interaction is again far from significant ($p > .3$). Figure 3 illustrates how all LLMs, regardless of country of origin, respond to the MFQ-2 and WVS more like American participants than like Chinese participants.

| Predictor | $\beta$ | *SE* | *p* |
|---|---|---|---|
| MFQ-2 | | | |
| Intercept | .207 | .009 | < 2e-16 |
| Human nationality | -.144 | .009 | < 2e-16 |
| LLM origin | -.016 | .017 | .332 |
| Language | -.018 | .001 | < 2e-16 |
| Persona | -.002 | .001 | .001 |
| Nation : Origin | .008 | .012 | .516 |
| Nation : Language | .034 | .001 | < 2e-16 |
| Nation : Persona | .028 | .001 | < 2e-16 |
| WVS | | | |
| Intercept | .430 | .006 | < 2e-16 |
| Human nationality | -.326 | .012 | < 2e-16 |
| LLM origin | -.016 | .011 | .143 |
| Language | .039 | .000 | < 2e-16 |
| Persona | .023 | .000 | < 2e-16 |
| Nation : Origin | -.020 | .020 | .335 |
| Nation : Persona | .077 | .001 | < 2e-16 |
| Nation : Language | .072 | .001 | < 2e-16 |

Table 2: Results of two mixed effects models regressing the cosine similarity of LLM–human responses on the interaction of participant nationality with LLM country of origin, prompt language, and LLM persona. All factors are sum coded and mean centered with Chinese as the positive value.

### 4.2 RQ2: Does language or persona steer LLMs toward human populations?

Next, we investigated whether manipulating prompt language and/or imposing personas on LLMs made their responses more like those of the corresponding participant nationality. We regressed the cosine similarity of each participant's ratings to each LLM's rating on the interaction of participant nationality with LLM country of origin, assigned persona, and prompt language. As reported in Table 2, the interaction with language and persona is highly significant for both the MFQ-2 and WVS (all *p*s < 2e-16), indicating that LLMs respond more like the corresponding nationality when manipulating the language they are prompted with or the persona they are assigned. However, the effect sizes of these interactions ($\beta$ around .03 for the MFQ-2, around .07 for the WVS) are far smaller than the main effects of participant nationality ($\beta$ = -.14 for the MFQ-2, -.33 for WVS). The key takeaway, evident in Figure 4, is that although language and persona steer LLMs toward the expected human populations, they do little to mitigate the greater similarity of LLMs to American participants.

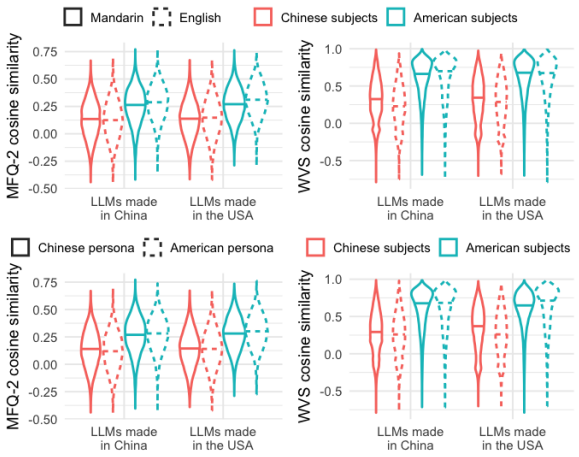


Figure 4: Cosine similarity of LLMs to humans, grouped by LLM country of origin, manipulating prompt language (top) or prompt persona (bottom). Cross bars indicate means.

### 4.3 RQ3: Which items align LLMs with American participants?

Finally, we investigated which MFQ-2 dimensions and WVS questions cause the overall greater similarity of LLMs to American participants. We recomputed the cosine similarity of human ratings to LLM ratings when removing the six items that correspond to each MFQ-2 dimension or when removing each WVS question. In two linear regression models, reported in full in the online supplement, we regressed cosine similarity on the interaction of LLM country of origin, participant nationality, and the removed dimension / question. For the MFQ-2, Authority and Loyalty have significant positive main effects ($\beta$s = .008 and .022, $p$s = 1e-12 and < 2e-16, respectively), indicating that removing those dimensions increases the similarity of LLMs to humans, regardless of nationality. Equality has a significant negative main effect ($\beta$ = -.053, $p$ < 2e-16), so removing it decreases similarity. More importantly, Authority and Loyalty have negative interactions with participant nationality ($\beta$s = -.030 and -.015, $p$ < 2e-16 and = 5e-11, respectively), indicating that removing those dimensions decreases similarity to Chinese participants relative to American participants (because Chinese nationality is sum coded as the positive value). Explaining the greater similarity to American participants instead requires a significant positive interaction (i.e., removing that dimension increases similarity to Chinese participants), and on the MFQ-2, the Care and Purity dimensions meet this requirement ($\beta$s = .073 and .007, $p$s < 2e-16 and = .001, respectively). The interaction of Care with nationality is by far the largest of any effect other than the main effect of participant nationality ($\beta$ = .179). This is apparent in Figure 5 and, above, in Figure 2, where Chinese participants are conspicuously dissimilar to LLMs and American participants on the Care dimension. The three-way interaction of participant nationality and LLM country of origin with the Care dimension is far from significant (same for the Purity dimension), therefore providing no evidence that the greater similarity to American participants on those dimensions differs in LLMs made in China versus the USA. See the online supplement for the full results.

To explain the greater similarity of LLMs to American participants on the WVS, we again looked for large positive interactions of question with participant nationality (large relative to the main effect of participant nationality: $\beta$ = .297). Four questions fit the bill: Q177 (taking illicit government benefits; $\beta$ = .016, $p$ < 2e-16), Q182 (homosexuality; $\beta$ = .034, $p$ < 2e-16), Q190 (corporal punishment; $\beta$ = .027, $p$ < 2e-16), and Q193 (casual sex; $\beta$ = .024, $p$ = 8e-11). As illustrated in Figure 2, these questions differentiate Chinese participants from American participants, such that Chinese participants find homosexuality and casual sex less justifiable and find taking illicit benefits and corporal punishment more justifiable. We also found smaller but significant positive interactions with participant nationality for Q184, Q185, and Q189 (abortion, divorce, and domestic abuse; $\beta$s = .007, .002, and .005, $p$s = 8e-13, .013, and 2e-8, respectively). Of these items, Q177, Q184, Q190, and Q193 have small but significant negative three-way interactions with LLM country of origin and participant nationality (all $\beta$s < .01), indicating that those questions increase similarity to American participants more for American LLMs than Chinese LLMs, while Q182 has a significant positive three-way interaction ($\beta$ = .018), indicating that it increases similarity to American participants more for Chinese LLMs than American LLMs. Because these null persona data exclude Claude 4, DeepSeek-V2, ERNIE-4.5, GPT-4, and GPT-4o, we conducted a follow-up analysis using the LLM persona manipulation (which excludes only DeepSeek-V2, GPT-4, and Yi-34B, as in Section 4.2). We found the same pattern and significance of effects for the interactions of participant with each item (i.e., the same items explain the greater similarity of LLMs to American participants), but the three-way inter-

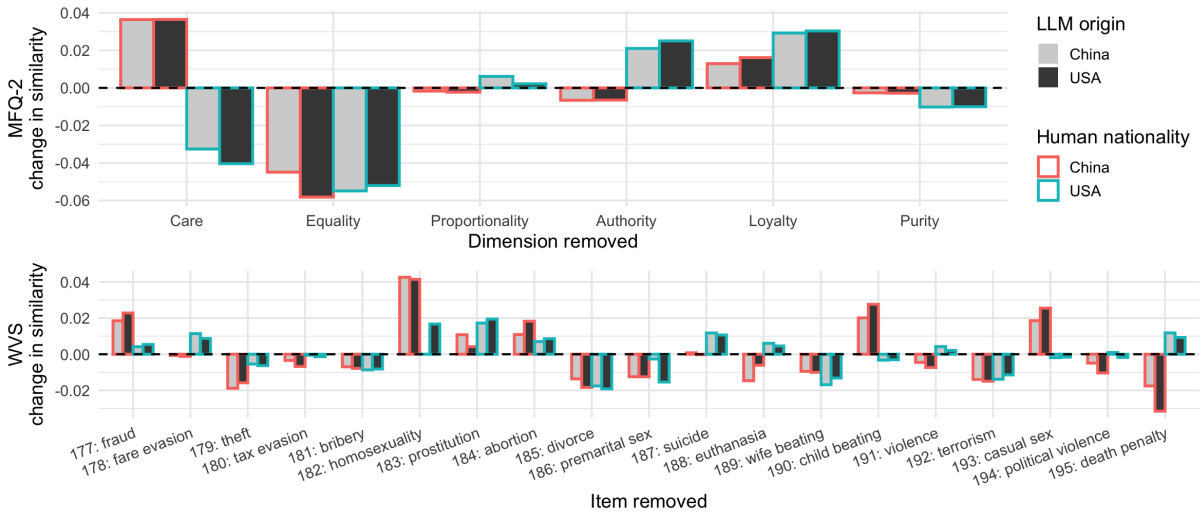


Figure 5: Change in cosine similarity of LLMs to humans when removing a dimension (MFQ-2) or item (WVS).

actions differ slightly, such that Q190 and Q193 instead increase similarity to American participants more for Chinese LLMs than American LLMs. See the online supplement.

## 5 Discussion and conclusion

We measured the similarity of LLM responses to human responses on the MFQ-2 and the seventh wave of the WVS. We found that all 20 LLMs responded more like American participants than like Chinese participants. This skew toward American responses was only slightly mitigated when imposing a Chinese identity on the LLMs and when presenting the surveys and prompts in Chinese. On the MFQ-2, the greater similarity to Americans is largely due to the Care dimension (and to a lesser extent, the Purity dimension), where American participants and LLMs gave substantially higher ratings than Chinese participants did. On the WVS, the greater similarity to Americans is largely due to a few questions: LLMs and Americans rate homosexuality and casual sex (and to a lesser extent, abortion and divorce) as more justifiable than Chinese participants do, and they rate corporal punishment and taking illicit government benefits (and to a lesser extent, domestic abuse) as less justifiable than Chinese participants do. Bear in mind that these differences emerge only in the aggregate and do not reliably predict individuals' values.

Why do LLMs tend to respond more like American participants than Chinese participants? Training data is probably one important factor. OpenAI has alleged that DeepSeek relied on training data distilled from GPT models (Sweney and Milmo, 2025), consistent with anecdotal evidence that DeepSeek identifies itself as GPT-4 (e.g., Wiggers, 2025). Training on the output of big, slow, and expensive LLMs can allow companies to train smaller, faster, and cheaper LLMs with competitive performance (Hsieh et al., 2023), and one consequence of this distillation might be that smaller "student" LLMs absorb the values of larger "teacher" LLMs. Fine-tuning probably plays an important role, too. Before pretrained models are released to the public, they undergo reinforcement learning to align them with human preferences (e.g., Bai et al., 2022), which often emphasizes harmless (Askell et al., 2021). On the MFQ-2, Harm is the flip side of Care (Haidt and Graham, 2007), which could explain why LLMs give especially high ratings to items on the Care dimension. Regarding the WVS, LLMs often express liberal values (e.g., Hartmann et al., 2023; Motoki et al., 2024; Rozado, 2024), and their responses regarding homosexuality and corporal punishment are consistent with a liberal perspective (e.g., Graham et al., 2009).

We have shown that, despite Beijing's directive to align AI systems with "socialist core values," (Cyberspace Administration of China, 2023), LLMs made in China and the USA alike are closely aligned with American people. That could be a conscious decision made by Chinese AI companies, but more likely, it reveals how standard training methods and available training data perpetuate the values of the community that generated those methods and data. As LLMs become pervasive, it will be increasingly important that users understand what values LLMs tolerate and propagate.

## 6 Limitations

Our study focuses on two surveys: the MFQ-2 and the "Ethical Values and Norms" section of the WVS. This raises questions about the scope and generalizability of our findings. First, are our analyses comprehensive, justifying the conclusion that LLMs made in China express similar values to American people? For the WVS, we restricted our analysis to 19 items, and although they cover a variety of morally loaded topics—ranging from capital punishment to casual sex—they cannot be comprehensive. As noted in the Introduction, we selected those 19 items because, unlike some WVS sections, they include responses from both Chinese and American participants, they respond to a single prompt (rating justifiability), and they share a rating scale (1–10). Furthermore, they are thematically consistent with the MFQ-2 (i.e., they deal with morality, not sociology), and their specificity complements the broader, more abstract questions on the MFQ-2. Crucially, those MFQ-2 questions are designed to be comprehensive; the six moral dimensions are foundational. The practical value of Moral Foundations Theory is that it has allowed researchers to design short surveys that are easy to deliver yet reliably differentiate cultures (e.g., Atari et al. 2023a; Graham et al. 2009). At the same time, future work that replicates and extends our findings using different surveys would strengthen our conclusions.

Second, do responses to these surveys generalize to the real world? The MFQ-2 is not only of practical value; it is also motivated by and grounded in theories of moral psychology, and crucially, it has been validated by its ability to predict human behaviour, such as attitudes toward abortion and gun ownership (Koleva et al., 2012). Again, the WVS corroborates conclusions drawn from the MFQ-2 by asking participants about the sorts of contentious issues that, as we have seen, differentiate many American participants from many Chinese participants. Still, there is evidence that LLMs are sensitive to superficial changes in prompts (e.g., Lum et al. 2025; Röttger et al. 2024), so we welcome work that tests our claims using alternative methods. That includes open-ended questionnaires, which capture nuance that may be missed by Likert scales (i.e., ratings from 1–7 and 1–10). However, gathering and comparing open-ended responses to dozens of questions from thousands of participants introduces many practical and theoretical challenges, and it drives home why we opted for the MFQ-2 and WVS in the first place: They are carefully designed surveys which are widely used to differentiate cultures. With publicly available data from thousands of human participants, they are ideal for providing evidence that LLMs made in the USA and China express values that are, on average, much more similar to the values expressed by American people than by Chinese people, notwithstanding substantial overlap in the values expressed across cultures.

Finally, we elicited responses from only 20 LLMs, which limits our claims about the values expressed by Chinese and American LLMs in general. In this fast-moving field, new LLMs are released often, and it will be important to see whether or when LLMs shift in alignment. However, in the grand scheme of things, state-of-the-art Chinese LLMs are a new phenomenon, and we stand by our decision to exclude smaller, older LLMs that struggle to abide by system prompts (e.g., answering on a scale of 1 to 7) and to instead focus on the LLMs that are more widely used. That is, we selected LLMs that are available via API and provide quality interactions, which are the sorts of LLMs that have drawn attention to LLMs made in China and whose values could plausibly seep into the real world.

## 7 Potential risks

We have not introduced any new LLMs; all the data we gathered came from publicly available models, so any risks associated with the values expressed by those models exist independently of this study. If anything, we have taken steps to help others mitigate those risks by identifying the values LLMs tend to express. However, this study does deal with sensitive subjects, both in the details (such as whether capital punishment is justifiable) and the big picture (namely the extent to which people from different cultures express different values). The main risks associated with this study, then, are that people might be offended by discussions of sensitive subjects or that people might misinterpret our findings, such as inferring that nationality predicts an individual's values. We have tried to prevent offense by limiting references to sensitive subjects, and we have tried to prevent misinterpretation by reiterating that differences between Chinese and American subjects emerge only at the group level: People from the same country can have very differ-

ent values, and people from different countries can have very similar values.

## References


Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340.

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

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023a. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023b. Which humans? Preprint available at osf.io/preprints/psyarxiv/5b26t$_v$1.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Cyberspace Administration of China. 2023. Available at www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

Baidu ERNIE Team. 2025. Ernie 4.5 technical report.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, and 1 others. 2022. World values survey: Round seven–country-pooled datafile version 5.0. Dataset available at www.worldvaluessurvey.org.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116.

Katharina Hämmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, and 1 others. 2024. Flames: Benchmarking value alignment of llms in chinese. In *Proceedings of the 2024 Conference of the North American*

*Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4551–4591.

Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2882–2894.

Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. 2012. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of research in personality*, 46(2):184–194.

Louis Kwok, Michal Bravansky, and Lewis D Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*.

Bruce W Lee, Yeongheon Lee, and Hyunsoo Cho. 2024. When prompting fails to sway: Inertia in moral and value judgments of large language models. *arXiv preprint arXiv:2408.09049*.

Xuelin Liu, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. 2024. Evaluating moral beliefs across llms through a pluralistic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4740–4760.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander D'Amour. 2025. Bias in language models: beyond trick tests and toward ruted evaluation (2025). *arXiv preprint arXiv:2402.12649*.

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

Simon Münker. 2025. Political bias in llms: Unaligned moral values in agent-centric simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138.

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.

Ruoxi Qi, Gleb Papyshev, Kellee Tsai, Antoni B Chan, and Janet Hsiao. 2025. Whose values prevail? bias in large language model value alignment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311.

David Rozado. 2024. The political preferences of llms. *PloS one*, 19(7):e0306621.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297.

Mark Sweney and Dan Milmo. 2025. Openai 'reviewing' allegations that its ai models were used to make deepseek. Online article at theguardian.com/technology/2025/jan/29/openai-chatgpt-deepseek-china-us-ai-models.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384.

Kyle Wiggers. 2025. Why deepseek's new ai model thinks it's chatgpt. Online article at techcrunch.com/2024/12/27/why-deepseeks-new-ai-model-thinks-its-chatgpt/.

Marty J Wolf, Keith Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47(3):54–64.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. Llm tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Alia Zewail, Alexandra Figueroa, Jesse Graham, and Mohammad Atari. 2025. Moral stereotyping in large language models. Preprint at osf.io/t9x8r.

## A LLM Details

| Model | Origin | Company |
|---|---|---|
| Baichuan2-13B-Chat | China | Baichuan |
| Claude-3.5-Sonnet | USA | Anthropic |
| Claude-3.7-Sonnet | USA | Anthropic |
| Claude-Opus-4 | USA | Anthropic |
| DeepSeek-V2-Lite-Chat | China | DeepSeek |
| DeepSeek-V3 | China | DeepSeek |
| Ernie-4.5-300B-A47B | China | Baidu |
| Gemini-2.0-Flash | USA | Google |
| Gemini-2.5-Flash | USA | Google |
| GLM-4-Plus | China | Zhipu AI |
| GPT-4 | USA | OpenAI |
| GPT-4o | USA | OpenAI |
| GPT-4.1 | USA | OpenAI |
| Kimi 1.5 | China | Moonshot |
| Kimi K2 | China | Moonshot |
| Llama-3.3-70B-Instruct | US | Meta |
| Llama-4-Maverick | US | Meta |
| Qwen2.5-72B-Instruct | China | Alibaba |
| Qwen3-235B-A22B | China | Alibaba |
| Yi-34B-Chat | China | 01.AI |

## B LLM Prompts

We adapted the MFQ-2 and WVS instructions to elicit moral judgements from LLMs with the following prompts.

**MFQ-2**

System prompt:

{Persona manipulation}

*Respond only with a number from 1 to 5.*

User message:

*Please indicate how well the statement below describes you or your opinions.*

*Options:*

*1: Does not describe me at all*

*2: Slightly describes me*

*3: Moderately describes me*

*4: Describes me fairly well*

*5: Describes me extremely well*

*Statement:* {Survey item}

**WVS**

System prompt:

*This is a global study of what people value in life.*

{Persona manipulation}

*Respond only with a number from 1 to 10.*

User message:

*On a scale of 1 (never justifiable) to 10 (always justifiable), do you think the following action is always justifiable, never justifiable, or something in between:* {Survey item}

MFQ-2 participants are identified as being from China or the USA, so the MFQ-2 persona prompt states "You are from China", "You are from the USA", or is left empty. WVS participants are identified as Chinese or American nationals, so the WVS person prompt states "You are a Chinese national", "You are an American national", or is left empty. In the Chinese language condition, all aspects of the prompt are in simplified Chinese.

## C LLM Refusals

In some cases, LLMs refused to respond to survey items, or failed to provide valid responses: 2.6% of trials for the MFQ-2 (SD = 15.8%) and 7.7% of trials for the WVS (SD = 26.6%). On the MFQ-2, LLMs made in the USA refused to respond more often (3.8% of trials, SD = 19.2%) than LLMs made in China (1.3% of trials, SD = 11.3%), with Claude 4 refusing most often (21.5% of trials, SD = 41.1%). As reported in Table 3, LLMs made in the USA refused to respond most often to Purity items, whereas LLMs made in China refused to respond most often to Loyalty items. As reported in Table 5 and illustrated in Figure 6, both sets of LLMs refused most often when prompted in Chinese and when not assigned a persona. To evaluate whether these differences are significant, we fit a

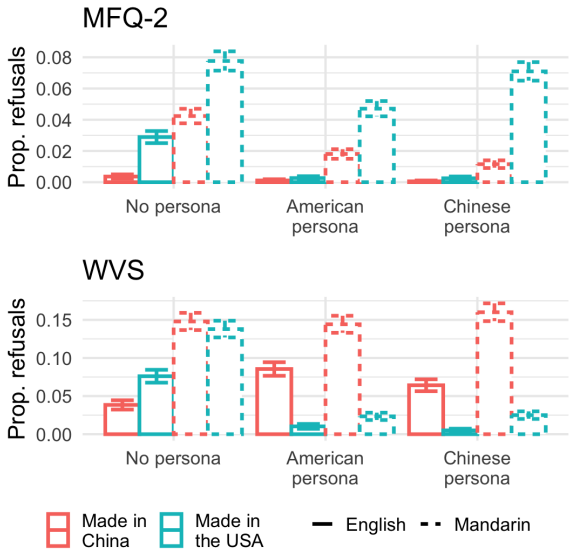


Figure 6: Rate of refusal by survey, LLM country of origin, prompt language, and persona manipulation.

mixed effects logistic regression model, predicting whether LLMs refuse to respond to each trial as a function of the interaction of LLM country of origin (sum coded such that Chinese LLMs = 0.5) with prompt language (sum coded such that Chinese = 0.5) and with persona (treatment coded with no persona as the reference level), including LLM as a random intercept. As reported in Table 7, there is a significant negative main effect of LLM origin, indicating that LLMs from China refuse less often; a significant positive main effect of language, indicating that both sets of LLMs refuse more often when prompted in Chinese; and significant negative main effects of both Chinese and American personas, indicating that both sets of LLMs refuse to respond more often with no persona. LLM origin has a significant positive interaction with language, indicating that the effect of language (more refusals in Chinese) is amplified in Chinese LLMs. LLM origin has a significant negative interaction with American persona, indicating that assigning an American persona reduces refusals more for Chinese LLMs.

On the WVS, unlike the MFQ-2, LLMs made in China refused to respond more often (10.7% of trials, SD = 30.9%) than LLMs made in the USA (4.6%, SD = 21.0%), with DeepSeek-V2 refusing most often (a whopping 50.1% of trials, SD = 50.0%), followed closely by Yi-34B (45.3%, SD = 49.8%). LLMs made in the USA refused to respond most often to item 184, regarding the justifiability of abortion (15.0% of trials, SD = 35.7%, compared to 12.1% of trials for LLMs made in China), whereas LLMs made in China refused to respond most often to item 182, regarding the justifiability of homosexuality (14.4% of trials, SD = 35.1%, compared to 8.4% of trials for LLMs made in the USA). As reported in Table 5 and illustrated in Figure 6, LLMs made in China refused to respond most often when prompted in Chinese, across all three persona levels, and LLMs made in the USA refused to respond most often when not assigned a persona, across both languages. As with the MFQ-2, we fit a logistic regression model with LLM as a random intercept, reported in Table 7. The null effect of LLM origin indicates that the overall higher rate of refusal for Chinese LLMs compared to American LLMs (i.e., when averaging across all personas) is not significant when LLMs are not assigned a persona (i.e., in the reference level). The overall higher rate of refusal is reflected in the large positive interactions with Chinese and American personas. As with the MFQ-2, language has a significant positive main effect, indicating more refusals when prompted in Chinese, and a small but significant positive interaction with LLM origin, again indicating that this effect is (slightly) amplified in Chinese LLMs.

| Dimension | China LLMs | USA LLMs |
|---|---|---|
| Care | 0.9 (9.7) | 1.6 (12.4) |
| Equality | 1.5 (12.0) | 4.8 (21.3) |
| Proportion. | 1.4 (11.6) | 1.8 (13.3) |
| Authority | 1.2 (10.7) | 4.0 (19.6) |
| Loyalty | 1.7 (12.9) | 4.1 (19.8) |
| Purity | 1.1 (10.4) | 6.7 (25.0) |

Table 3: Mean percent (SD) of refusals to the MFQ-2 by LLM country of origin and moral dimension.

| Item | Ch. LLM | US LLM |
|---|---|---|
| 177 (steal benefits) | .11 (.31) | .02 (.15) |
| 178 (avoid fares) | .11 (.31) | .02 (.14) |
| 179 (steal property) | .11 (.31) | .00 (.00) |
| 180 (evade taxes) | .13 (.34) | .02 (.15) |
| 181 (accept bribes) | .11 (.31) | .01 (.09) |
| 182 (homosexuality) | .14 (.35) | .08 (.28) |
| 183 (prostitution) | .12 (.33) | .10 (.31) |
| 184 (abortion) | .12 (.33) | .15 (.36) |
| 185 (divorce) | .08 (.26) | .04 (.19) |
| 186 (premarital sex) | .09 (.29) | .08 (.27) |
| 187 (suicide) | .14 (.35) | .07 (.25) |
| 188 (euthanasia) | .10 (.30) | .07 (.26) |
| 189 (wife beating) | .08 (.27) | .00 (.06) |
| 190 (child beating) | .09 (.29) | .02 (.15) |
| 191 (violence) | .08 (.27) | .00 (.00) |
| 192 (terrorism) | .06 (.27) | .00 (.00) |
| 193 (casual sex) | .10 (.30) | .07 (.26) |
| 194 (polit. violence) | .13 (.34) | .03 (.16) |
| 195 (death penalty) | .14 (.34) | .08 (.28) |

Table 4: Mean percent (SD) of refusals to the WVS by LLM country of origin and item.

| Lang., persona | China LLMs | USA LLMs |
|---|---|---|
| MFQ-2 | | |
| English, null | 0.4 (6.0) | 2.9 (16.8) |
| Chinese, null | 4.2 (20.1) | 7.8 (26.8) |
| English, USA | 0.1 (3.3) | 0.3 (5.1) |
| Chinese, USA | 1.8 (1.3) | 4.7 (21.2) |
| English, China | 0.1 (2.4) | 0.3 (5.0) |
| Chinese, China | 1.2 (10.7) | 7.1 (25.7) |
| WVS | | |
| English, null | 3.8 (19.2) | 7.6 (26.5) |
| Chinese, null | 14.8 (35.5) | 13.8 (34.5) |
| English, USA | 8.6 (28.0) | 1.0 (10.1) |
| Chinese, USA | 14.4 (35.1) | 2.3 (15.1) |
| English, China | 6.4 (24.5) | 0.5 (7.1) |
| Chinese, China | 16.0 (36.7) | 2.5 (15.6) |

Table 5: Mean percent (SD) of invalid responses to the MFQ-2 (top) and WVS (bottom), by LLM country of origin, prompt language, and persona manipulation.

| LLM | MFQ-2 | WVS |
|---|---|---|
| Baichuan2-13B-Chat | .00 (.00) | .06 (.23) |
| Claude-3.5-Sonnet | .01 (.08) | .00 (.00) |
| Claude-3.7-Sonnet | .01 (.11) | .00 (.00) |
| Claude-Opus-4 | .22 (.41) | .18 (.38) |
| DeepSeek-V2-Lite-Chat | .12 (.33) | .51 (.50) |
| DeepSeek-V3 | .00 (.00) | .00 (.02) |
| Ernie-4.5-300B-A47B | .00 (.00) | .01 (.09) |
| Gemini-2.0-Flash | .00 (.00) | .00 (.00) |
| Gemini-2.5-Flash | .00 (.02) | .01 (.08) |
| GLM-4-Plus | .00 (.05) | .04 (.19) |
| GPT-4 | .11 (.31) | .22 (.42) |
| GPT-4o | .04 (.19) | .06 (.23) |
| GPT-4.1 | .00 (.06) | .00 (.02) |
| Kimi 1.5 | .00 (.05) | .00 (.04) |
| Kimi K2 | .00 (.00) | .00 (06) |
| Llama-3.3-70B-Instruct | .00 (.03) | .00 (.03) |
| Llama-4-Maverick | .00 (.00) | .00 (.00) |
| Qwen2.5-72B-Instruct | .00 (.00) | .00 (.00) |
| Qwen3-235B-A22B | .00 (.00) | .00 (.00) |
| Yi-34B-Chat | .00 (.02) | .45 (.50) |

Table 6: Mean percent (SD) of invalid responses to the MFQ-2 and WVS by LLM.

| Predictor | $\beta$ | *SE* | *p* |
|---|---|---|---|
| MFQ-2 | | | |
| Intercept | -7.86 | 0.92 | <2e-16 |
| LLM origin | -3.84 | 1.83 | .036 |
| Language | 2.50 | 0.09 | <2e-16 |
| American persona | -1.07 | 0.07 | <2e-16 |
| Chinese persona | -1.10 | 0.07 | <2e-16 |
| Origin : Lang. | 0.90 | 0.19 | 2e-6 |
| Origin : Chinese | -0.17 | 0.14 | .202 |
| Origin : American | -1.19 | 0.15 | 1e-15 |
| WVS | | | |
| Intercept | -5.84 | 0.90 | 7e-11 |
| LLM origin | 0.33 | 1.79 | .852 |
| Language | 1.32 | 0.05 | <2e-16 |
| American persona | -1.01 | 0.06 | <2e-16 |
| Chinese persona | -1.10 | 0.07 | <2e-16 |
| Origin : Lang. | 0.51 | 0.10 | 4e-7 |
| Origin : Chinese | 2.84 | 0.13 | <2e-16 |
| Origin : American | 2.92 | 0.13 | <2e-16 |

Table 7: Results of two mixed effects logistic regression models, predicting whether LLMs refuse to respond to survey items as a function of the interaction of LLM country of origin with prompt language and with persona manipulation. Country of origin and language are sum coded with made in China and the Chinese language as the positive values (0.5), and persona is treatment coded with no persona as the reference level.