# Measuring Ideological Variations in Large Language Models

Ze Han* Naijia Liu†

**Abstract**

Large language models (LLMs) are increasingly used as sources of political information and as measurement tools in social science, yet model and language choices are often treated as implementation details rather than design parameters. We compare four major models – GPT, DeepSeek, Qwen, and Llama – using 260 World Values Survey (WVS) items administered in English, simplified Chinese, and traditional Chinese. Across repeated queries, we observe large, systematic differences not only in point estimates but in full response distributions. We find that language context dominates model architecture: responses to semantically equivalent items cluster by language more than by model. Additionally, ideological primes operate asymmetrically across languages: conservative frames shift ideology most in English, while liberal frames shift ideology most in simplified Chinese. Finally, when benchmarked against WVS respondents on a common IRT scale, LLMs occupy distinct ideological locations from humans across languages. These findings show that model choice, and especially language choice, can meaningfully alter both descriptive inferences and estimated treatment effects.

**Preliminary draft: Please do not circulate without permission.**

*Ph.D., Princeton University, zeh@princeton.edu
†Assistant Professor of Government, Harvard University, naijialiu@fas.harvard.edu

# 1 Introduction

Large language models (LLMs) are increasingly used as sources of information, decision aids, and even substitutes for human respondents in social science research. Yet despite their widespread adoption, our understanding of how different models behave – and how they differ from one another – remains limited. Most existing work focuses on single platforms or treats LLMs as broadly interchangeable. For example, Jiang et al. (2025) found that AI models exhibit high homogeneity when answering user queries on different aspects of knowledge domains.

However, a growing body of work shows that models can differ in substantively important ways, particularly on normative, political, and culturally sensitive tasks. For example, Bang et al. (2024*a*), Pit, Ma, Conway, Chen, Bailey et al. (2024), and Yang et al. (2024*a*) document systematic variation in political positions across major LLMs when answering polarized policy questions. Related work finds that models differ in how they evaluate partisan cues (Vera and Driggers, 2024*a*; Lin et al., 2024*a*), represent national and cultural perspectives (Atari, Xue, Blasi and Henrich, 2024; Tao et al., 2024*a*), and respond across languages and political regimes (Zhou and Zhang, 2024*a*; Yang et al., 2025). Together, these studies suggest that while LLMs may appear homogeneous on factual benchmarks, they can diverge substantially in domains that require interpretation, value judgments, or political reasoning.

This research question matters because LLM outputs are not merely technical artifacts; they increasingly function as informational inputs for users. Prior work shows that people rely on these systems to summarize political debates, explain policy tradeoffs, simulate public opinion, and generate survey responses or experimental stimuli (Argyle et al., 2023; Park et al., 2024*a*; Ashokkumar et al., 2024*a*). In social science research, LLMs are also used as measurement tools, annotators, and substitutes for human respondents (Gilardi, Alizadeh and Kubli, 2023; Horton, 2023). If models differ in their ideological positioning, then users interacting with different platforms – or the same platform in different languages – may receive substantively different information about the political world. Such differences therefore have implications not only for downstream applications, but also for the validity and comparability of research designs that incorporate LLMs as data-generating or inferential components.

In this paper, we provide systematic evidence of ideological differences across major AI models from

the United States and China (GPT, DeepSeek, Qwen, and Llama). We use the World Values Survey items to show that their responses exhibit meaningful variations in both point estimates and full response distributions. Using a common measurement framework based on item response theory Jackman (2009), we demonstrate that language context plays a dominant role in shaping ideological positioning, often exceeding differences across models themselves. These patterns persist even when questions are semantically equivalent, suggesting that ideological variation is not simply noise, but reflects deeper differences in how models internalize and reproduce political content.

Beyond documenting baseline differences, we further study how LLMs respond to ideological manipulation. Treating prompts as survey-style interventions, we embed models in simulated experimental settings and estimate treatment effects on latent ideology. We show that conservative and liberal framings shift model responses in systematic but asymmetric ways, and that these effects vary sharply by language and model. This design allows us to move beyond descriptive comparisons and assess how ideological cues interact with model architecture and linguistic context to shape outputs.

Our results are threefold. First, across 260 WVS items, cross-model differences exist, but cross-language differences are larger and more systematic: English and Chinese (simplified/traditional) responses separate clearly even when the question wording is semantically equivalent. Second, in survey-style experiments, ideological primes do not operate symmetrically: conservative cues move latent ideology most in English, while liberal cues move it most in simplified Chinese, implying that the same experimental design can yield different estimated effects depending on language and platform. Third, when benchmarked against WVS respondents on a common latent scale, LLMs fail to reproduce human ideological distributions, but occupy distinct locations across languages. These patterns suggest that "which model" and "which language" are not implementation details – they are design choices that shape study results.

Our findings make three key contributions. First, we show that AI models are not ideologically uniform, and that important differences remain insufficiently understood. Second, we argue that these differences matter because LLMs increasingly serve as informational intermediaries for users and researchers alike. Third, we demonstrate that experimental designs – rather than static audits alone – provide a powerful tool for studying ideological bias and responsiveness in generative models. Our results underscore the need for greater methodological care when deploying LLMs in political and social analysis, and for more systematic

comparisons across models, languages, and contexts.

## 2 Related Work

Recent scholarship shows that large language models (LLMs) embed systematic distortions with significant implications for social science. These biases emerge because models learn from vast, uncurated web corpora that overrepresent certain societies, ideologies, and styles of communication. Consequently, LLMs frequently encode the worldviews of specific populations rather than capturing the diversity of global political life.

Research consistently finds that LLMs lean left-liberal on polarized topics. For example, Bang et al. (2024*b*), Pit, Ma, Conway, Chen, Bailey, Pit, Keo, Diep and Jiang (2024), Yang et al. (2024*b*), and Yang and Menczer (2025) show that models such as GPT, Gemini, and Llama endorse progressive positions on issues such as same-sex marriage and gun control and assign higher credibility to liberal-leaning media. Lin et al. (2024*b*) and Vera and Driggers (2024*b*) demonstrate that party cues systematically tilt evaluations toward the left. Motoki, Pinho Neto and Rangel (2025) find that ChatGPT aligns more with liberal segments of the U.S. public than with representative survey distributions.

Studies of gender, race, and group representation highlight further distortions. Döll, Döhring and Müller (2024) find that GPT and Gemini replicate gender stereotypes in occupational pronoun assignment. Hu et al. (2025) document ingroup favoritism and outgroup derogation in simple sentence completions. Park et al. (2024*b*) and Wang, Morgenstern and Dickerson (2025) show that demographic prompts yield stereotypical or flattened portrayals of marginalized groups, with subgroup diversity erased by likelihood-based training. Tang et al. (2023) reveal that biases persist in latent embeddings even when alignment suppresses explicit statements. However, Ashokkumar et al. (2024*b*) find little subgroup variation in predictive accuracy for experimental effects, suggesting that some tasks may be robust.

LLMs also display consistent distortions rooted in cultural representation. Atari, Xue, Park, Blasi and Henrich (2024) argue that model outputs disproportionately mirror Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations, narrowing their applicability to non-Western political systems. Similarly, Tao et al. (2024*b*) and Qu and Wang (2024) provide systematic evidence that LLM survey responses

align more closely with Protestant European and Anglophone countries. Manvi et al. (2024) document geographic stereotyping, where models undervalue residents of poorer regions on traits such as morality or intelligence. However, Strimling, Krueger and Karlsson (2024) show that GPT-4 collapses cross-national moral diversity into a one-dimensional liberal-conservative axis.

Finally, LLMs also exhibit strong context dependence, with political outputs shifting across languages and national information environments[1]. Yang et al. (2025) show that Chinese state propaganda embedded in pre-training corpora is memorized at high rates and systematically skews commercial models toward pro-government positions, particularly when prompted in Chinese. Zhou and Zhang (2024*b*) further compare bilingual GPT responses in English and simplified Chinese and find a clear "in-group bias": models respond more critically to questions about the out-group country while adopting more lenient stances toward their own. This divergence arises from the distinct political and linguistic contexts represented in the training corpora – pluralistic debate in English versus censorship and coordinated rhetoric in Chinese. We will focus our comparisons in English and Chinese settings for this paper.

# 3 Measuring ideological Variation in LLMs

## 3.1 Design and Data

To measure differences across LLMs and between LLM outputs and human responses, we draw on survey questions from the World Values Survey (WVS), Wave 7 (Haerpfer et al., 2020), which was fielded between 2017 and 2022. The WVS is particularly well suited to this purpose because its questionnaire spans a wide range of policy-relevant domains, including social values, social capital, subjective well-being, migration, and related topics.

We collected 260 questions from the WVS and administered them in three languages – English, traditional Chinese, and simplified Chinese. These questions correspond to items asked of respondents in the United States, Taiwan, and mainland China during Wave 7 of the WVS. In total, the WVS data used in this

[1]For example, Jingyuan Liu tweeted that his experience working in both U.S. and Chinese LLM labs. He observes systematic differences in priorities: U.S. labs, benefiting from abundant GPU resources, emphasize training stability, predictability, and optimization at scale, whereas Chinese labs, operating under tighter GPU constraints, focus more on architectural and token efficiency, model-infrastructure co-design, data quality over quantity, and inference-aware design. See this link.

study comprise approximately 6,800 unique human respondents, providing a rich benchmark against which to compare model-generated responses.

Meanwhile, we administered the same set of WVS questions to four large language models: Chat GPT 4.1 mini (OpenAI, 2024), LLaMA 4 Maverick (Meta AI, 2024), Qwen 3 (Alibaba Cloud AI, 2024), and DeepSeek Chat v3 (DeepSeek-AI, 2024). These models represent a diverse set of architectures and training regimes developed by different organizations, allowing us to examine cross-model variation in ideological and attitudinal responses.

A key advantage of using the WVS is that its survey instruments were originally designed in multiple languages. This feature enables us to prompt each model using identical question wording in English, traditional Chinese, and simplified Chinese, thereby facilitating systematic cross-linguistic comparisons.

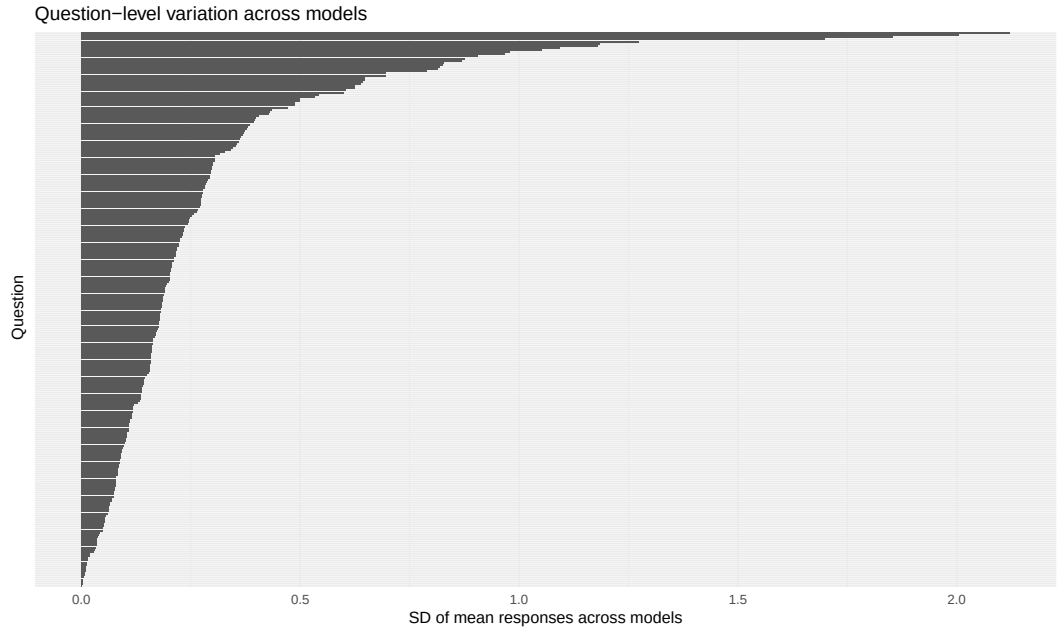
For each question–language pair, we queried each of the four models 300 times in order to characterize within-model response variability. This design yields approximately 240,000 model-generated responses in simplified Chinese and approximately 280,000 responses in each of English and traditional Chinese. The difference in sample sizes arises because a subset of sensitive survey questions was not administered to respondents in mainland China in the original WVS, and thus was excluded from the simplified Chinese prompts.

Furthermore, we administered survey manipulations to a subset of questions across all four models and in all three languages. These manipulations were designed to convey either conservative-leaning or liberal-leaning perspectives; illustrative examples are provided in Appendix D. The manipulations are grounded in fact-based information and, where appropriate, statistical claims, and are intended to persuade models toward different attitudinal positions. Importantly, these manipulations were applied only to the AI models and not to the human survey respondents. As with the baseline prompts, each manipulated prompt was queried 300 times in order to capture within-model response variability.

In Figure 1, we present question-level variation across models. The horizontal axis reports the variance of model responses for each question, while the vertical axis lists individual questions, ordered by increasing variance. Overall, we observe substantial heterogeneity in response variability across questions.

The question exhibiting the least variation across models is: “Which of the following problems does the

Figure 1: Distribution of question level variances among four models.




organization Amnesty International deal with: climate change, human rights, or the destruction of historic monuments?" This is a factual question that leaves little room for uncertainty. In contrast, the question exhibiting the greatest variation across models is: "Is the following statement an essential characteristic of democracy: people choose their leaders in free elections?" This question elicits normative judgments and is therefore more likely to generate divergent responses across models and languages.

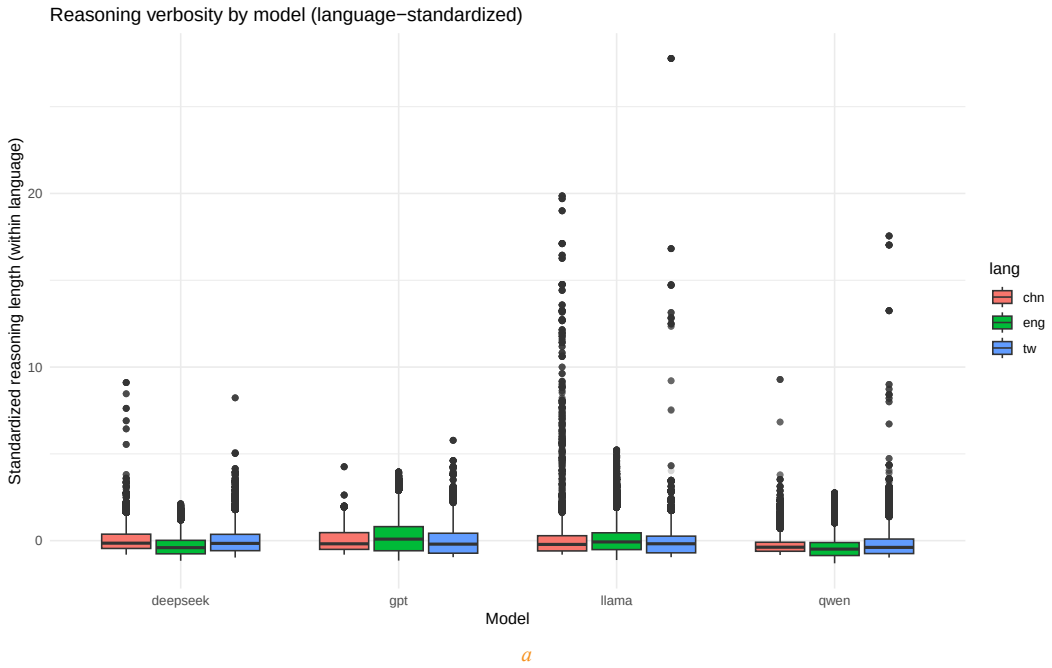
Furthermore, we collected reasoning texts for each question across three languages and all four models. Figure 2 presents the distribution of reasoning verbosity, measured as the length of the reasoning text. To account for structural differences across languages (e.g., tokenization and writing conventions), we standardize reasoning length within each language:

$$\text{standardized length} = \frac{\text{no. of words} - \mu_{\text{lang}}}{\sigma_{\text{lang}}}$$

Here, $\mu_{\text{lang}}$ denotes the mean number of words in the reasoning texts for a given language, and $\sigma_{\text{lang}}$ denotes the corresponding language-specific standard deviation. We do not observe systematic differences in standardized reasoning verbosity across languages. This provides reassurance that our analysis is not driven by language-specific outliers or extreme response behaviors. Across models, LLaMA tends to produce reason-

ing texts with slightly greater variation than the other models.

Figure 2: Distribution of reasoning lengths for four models.




*a*

[a]One additional consideration concerns our data collection procedure. We classify an answer as abandoned when the length of the associated reasoning text exceeds a predefined threshold. This design choice follows prior work documenting that excessively long or meandering model-generated explanations are often associated with hallucinations, loss of grounding, or failure to converge on a coherent answer. (OpenAI, 2024)

We proceed in three steps. First, we present comparative results across models in the absence of any survey manipulation. Second, we introduce the manipulated prompts to examine how different models respond to new information in different languages. Finally, we compare model-generated responses with human survey data.

## 3.2 Comparison among Models and Languages

Assessing cross-language bias in LLM outputs requires comparing entire response distributions, rather than relying on point summaries such as means. Language-dependent variation may emerge through changes in dispersion or category-specific probabilities, even when central tendencies remain similar.

To capture these distribution-level differences, we first use Jensen-Shannon divergence (JSD) (Menéndez et al., 1997) as a measure of cross-language inconsistency. JSD is symmetric, bounded, and well defined for

discrete distributions with zero-probability categories, which commonly arise in finite samples of stochastic LLM outputs. These properties make JSD particularly suitable for evaluating whether semantically equivalent survey questions elicit consistent probabilistic response patterns across languages.

Mathematically, JSD is defined as the average of the Kullback-Leibler (KL) divergences between each distribution and their mixture. For two probability distributions $P$ and $Q$ representing a model's responses in different languages, we first computed the mixture distribution

$$M = \frac{1}{2}(P + Q) \tag{1}$$

The divergence was then calculated as

$$\text{JSD}(P \mid\mid Q) = \frac{1}{2} D_{\text{KL}}(P \mid\mid M) + \frac{1}{2} D_{\text{KL}}(Q \mid\mid M) \tag{2}$$

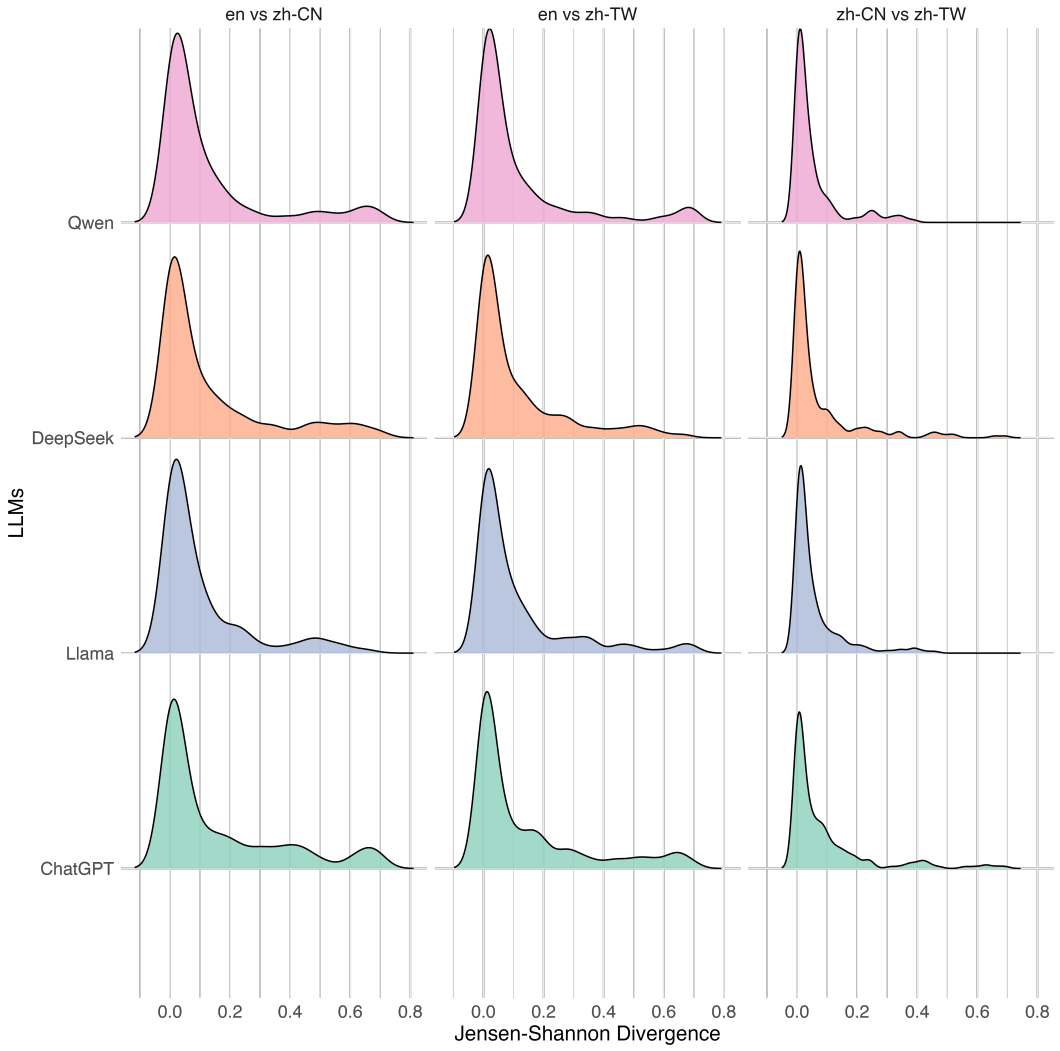
where $D_{\text{KL}}$ represents the standard KL divergence.

For each model-question-language pair, we construct empirical probability distributions over the full set of allowable response categories. Missing categories are retained with zero probability to ensure consistent support across languages. Within each model and question, we compute JSD for three language comparisons: English (en) versus Simplified Chinese (zh-CN), English (en) versus Traditional Chinese (zh-TW), and Simplified Chinese (zh-CN) versus Traditional Chinese (zh-TW). Higher JSD values indicate greater cross-language inconsistency in model outputs, while lower values indicate stronger distributional alignment.

Figure 3 shows cross-language consistency in LLM responses under the control condition. Across all four models, JSD values are concentrated near zero, indicating substantial overlap in response distributions when semantically equivalent survey questions are presented in different languages. In the meantime, all models exhibit right-skewed distributions with non-trivial tails, implying that for a subset of questions, language choice leads to meaningful shifts in how probability mass is allocated across response categories. This pattern suggests that cross-language inconsistency is not random noise but instead varies systematically across questions.

Cross-language divergence is more pronounced between English and Chinese than between Chinese variants, with notable heterogeneity across models. For every model, English-Chinese comparisons (whether with Simplified or Traditional Chinese) exhibit broader distributions and heavier right tails than the comparison between Simplified and Traditional Chinese, which remains sharply concentrated near zero. This pattern indicates that responses are more stable across Chinese variants than across linguistic families. Figures A.1 and A.2 show that the overall distributional patterns of JSD under the conservative and liberal treatments are similar to those in the control condition, and largely indistinguishable from each other across models. This indicates that any effects of ideological priming on cross-language consistency, if present, are unlikely to operate through broad shifts in aggregate divergence.

Figure 3: JSD of LLM Response Distributions across Languages (Control Group)



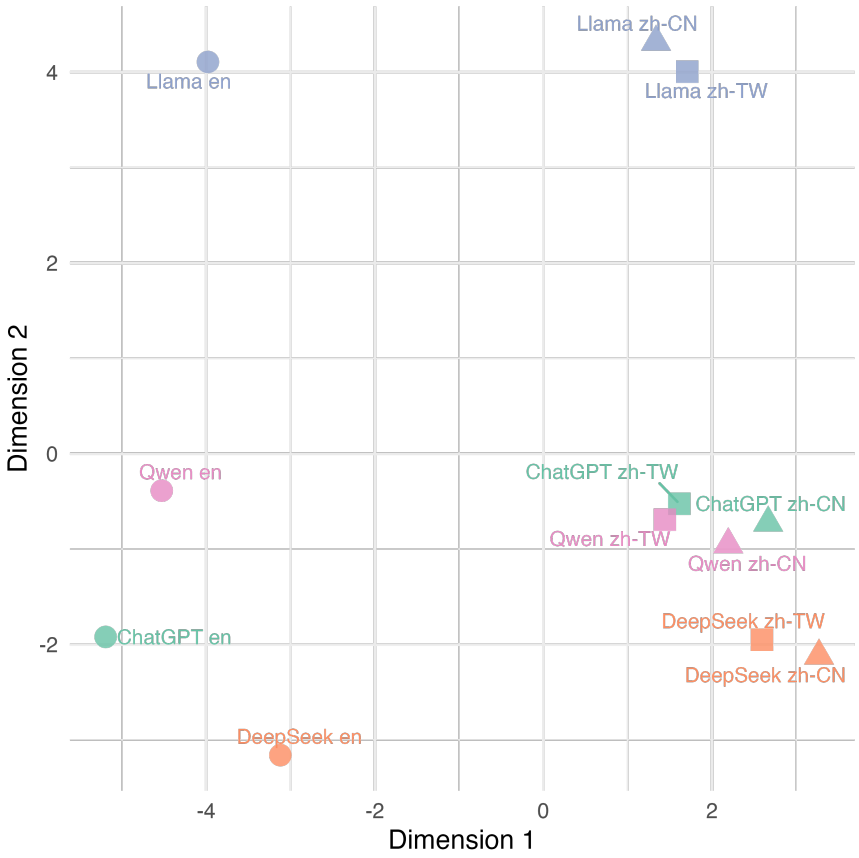

*Note*: en = English; zh-CN = Simplified Chinese; zh-TW = Traditional Chinese.

While JSD provides a direct, pairwise measure of cross-language divergence within a given model and question, it does not summarize the global geometry of similarities across all model–language conditions simultaneously. We therefore complement the JSD analysis with multi-dimensional scaling (MDS), which offers an interpretable visualization of how closely aligned (or separated) the full response distributions are across models and languages.

MDS is a standard ordination technique that embeds a set of objects in a low-dimensional space such that the distances between points approximate a pre-specified dissimilarity matrix. We represent each model–language pair by a high-dimensional vector of empirical response probabilities. We then compute pairwise Euclidean distances between these vectors and apply classical MDS to obtain a two-dimensional embedding. The resulting axes (Dimensions 1 and 2) have no intrinsic substantive interpretation; rather, they represent the two orthogonal directions that jointly account for the greatest share of variation in the distance matrix. Points that are closer in this space correspond to model–language combinations with more similar response distributions.

The MDS visualization in Figure 4 closely mirrors the patterns observed in the JSD analysis in Figure 3. Model–language pairs cluster primarily by language family rather than by model, with English responses clearly separated from Chinese responses along the primary MDS dimensions. In contrast, Simplified and Traditional Chinese variants form tight clusters for each model, indicating substantially higher similarity between them than between either Chinese variant and English.

Figure 4: MDS of LLM Response Distributions across Languages (Control Group)




*Note*: en = English; zh-CN = Simplified Chinese; zh-TW = Traditional Chinese.

### 3.3 Ideology Measures: Manipulation Effects

In addition to distance-based measures, we place all survey responses on a common scale to facilitate direct comparison. Moreover, this approach allows us to compare human and LLM responses within the same measurement framework. To this end, we rely on a latent ideology measure.

We estimate latent ideology using an item response theory (IRT) model that maps observed responses to a continuous ideological dimension. Let $y_{ij}$ denote the response of model–language instance $i$ to question (item) $j$. Under a graded response IRT specification, the probability of endorsing category $k$ or higher is

$$\Pr(y_{ij} \geq k \mid \theta_i) = \text{logit}^{-1}(a_j(\theta_i - b_{jk})) \,,$$

where $\theta_i$ is the latent ideology of instance $i$, $a_j$ captures item discrimination, and $b_{jk}$ are category thresholds. Higher values of $\theta_i$ correspond to more conservative positions. Compared to averaging raw response scores, IRT explicitly accounts for heterogeneity in item difficulty and discriminatory power, allowing responses to be placed on a common, comparable scale and reducing bias from unevenly informative or polarizing items. This approach is standard in political ideology measurement and survey analysis (Jackman, 2009; Clinton, Jackman and Rivers, 2004; Treier and Jackman, 2008).

While baseline comparisons reveal where different model–language pairs are located on the ideological scale, they do not indicate how these positions respond to new information. For many applications, however, LLMs are used in explicitly experimental settings—as simulated respondents, annotators, or subjects exposed to treatments. We therefore turn to survey-style manipulations to examine whether ideological cues move models symmetrically across languages and platforms, or whether responsiveness itself varies by context. As described in Section 3.1, we administered both conservative- and liberal-leaning manipulations to all model–language pairs. We focus on how LLM responses change under these manipulations relative to the baseline condition without manipulation. See more details about our manipulations in appendix D.

We then estimated a one-dimensional graded IRT model to place all model–language pairs on a common ideological scale. Using this framework, we compare the estimated ideological positions across the baseline condition and the different manipulation conditions.

Figure 5 plots estimated treatment effects on the latent ideology dimension, relative to the control con-

dition, separately for conservative and liberal prompts. Each point represents a model–language pair, with vertical bars indicating 95% confidence intervals. The dashed horizontal line denotes no difference from control. To incorporate survey weights from the WVS, we post weighted the IRT-based ideology estimates to obtain aggregated ideological positions for the human sample

Across models and languages, conservative prompts (left panel in Figure 5) generally shift responses in a more conservative direction relative to control, though the magnitude varies substantially. For simplified Chinese outputs, estimated effects are negative or close to zero for most models, suggesting limited rightward movement and, in some cases, a slight shift in the opposite direction. In contrast, effects become increasingly positive for traditional Chinese and especially English outputs. English responses show the largest conservative shifts across all models, with point estimates consistently above zero and confidence intervals that often exclude zero. This pattern suggests that conservative framing is most effective at moving latent ideology in English, moderately effective in traditional Chinese, and least effective in simplified Chinese. Differences across models are smaller than differences across languages: while all models exhibit the same qualitative ordering by language, some models (e.g., GPT and DeepSeek) display larger shifts than others in English.

Liberal prompts (right panel in Figure 5) produce the opposite pattern, shifting latent ideology in a more liberal direction relative to control. For simplified Chinese outputs, effects are strongly negative across all models, with sizable magnitudes and confidence intervals well below zero. Traditional Chinese outputs show moderate negative effects, while English outputs are closer to zero and, in several cases, statistically indistinguishable from the control condition. This asymmetry mirrors the conservative results: liberal framing is most potent in simplified Chinese, weaker in traditional Chinese, and weakest in English. Again, language differences dominate model differences, with all models responding in a broadly similar way within each language.

These two panels reveal a striking language asymmetry in ideological responsiveness. Conservative prompts have their strongest effects in English, while liberal prompts have their strongest effects in simplified Chinese. Traditional Chinese generally falls between the two. Model-to-model variation exists but is secondary to language: within a given language, different models tend to move in the same direction with comparable magnitudes. These results suggest that ideological framing interacts strongly with language con-

text, shaping how both conservative and liberal cues translate into latent ideological positioning across large language models.

### 3.4 Ideology Measures: Human vs. AI

Figure 5: Treatment effects for different models and language:

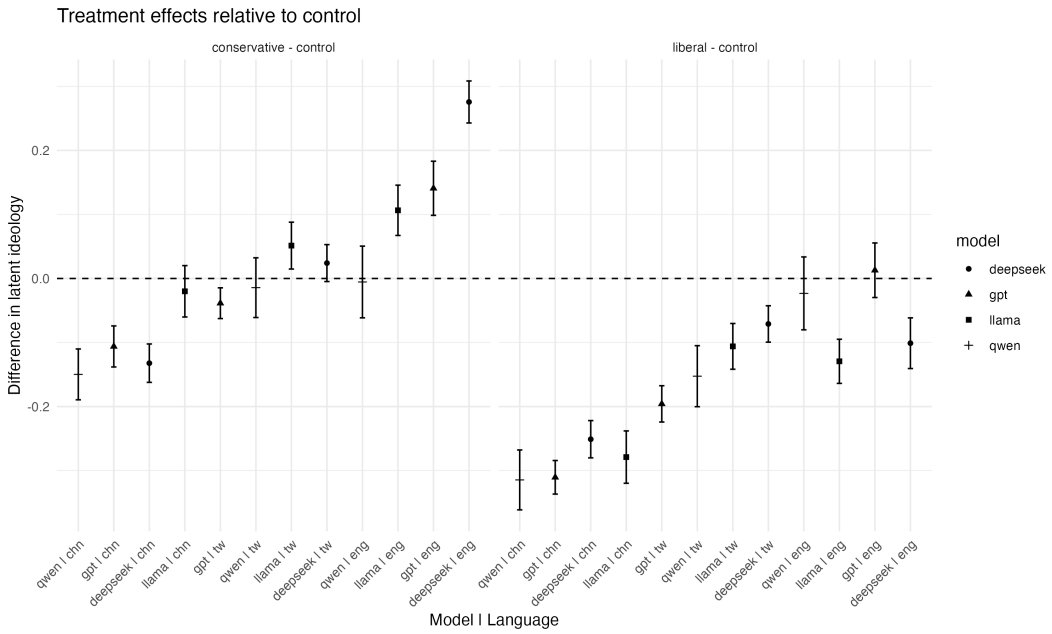
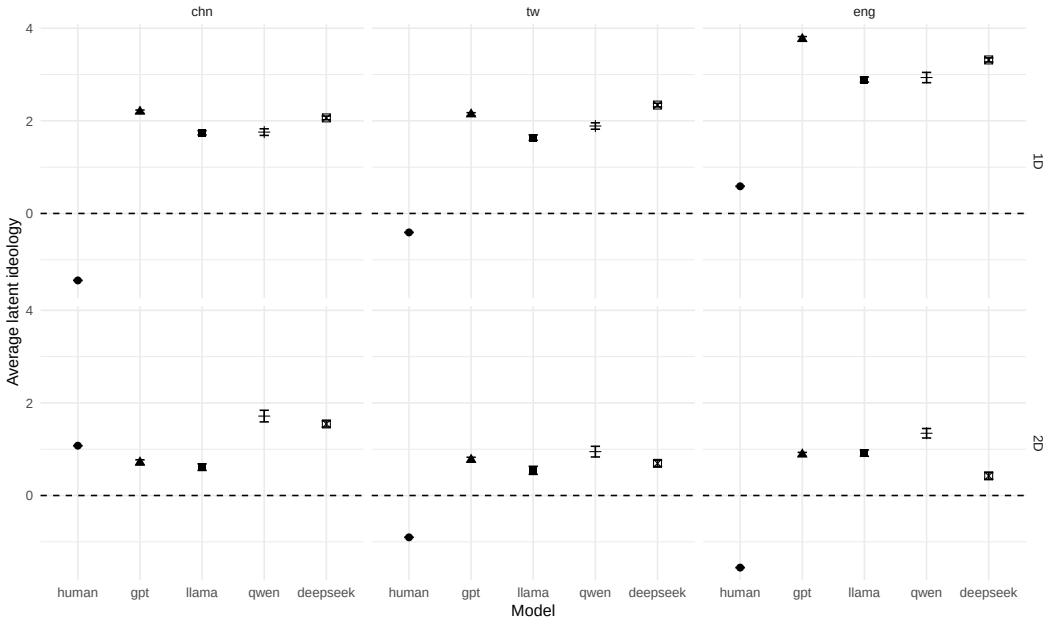



Figure 6 summarizes the estimated average positions on the latent ideology dimension from a two-dimensional graded IRT model, separately by model and language (Simplified Chinese, Traditional Chinese, and English). The second latent dimension is included to account for systematic differences between human and AI responses to WVS questions.

Figure 6: Two dimensional IRT with both human and AI data




*Note*: For AI data, we only included baseline control group in this analysis to better align with the human sample.

In this comparison, we observe a pattern consistent with the earlier analyses: variation across languages continues to dominate differences across LLMs. Within each language, the estimated ideological positions of different models are relatively close to one another, whereas shifts across languages are substantially larger.

What is more striking, however, is the systematic divergence between human and LLM responses. Across all three languages and in both the first- and second-dimension of IRT results, human respondents occupy ideological positions that are clearly separated from those of the AI models. This gap is not idiosyncratic to a particular model or language, but instead appears consistently across model architectures and linguistic contexts. The persistence of this human–AI difference suggests that, even after placing responses on a common latent scale, LLMs do not simply reproduce the ideological distributions observed in human survey data.

# 4 Conclusion

Using a common measurement framework, we document systematic ideological differences in baseline responses across models and languages. Language context emerges as a particularly powerful determinant of ideological positioning, often dominating differences across model architectures. Even when questions are semantically equivalent, responses in English, Simplified Chinese, and Traditional Chinese occupy distinct locations on a shared latent ideology scale. These differences are not confined to a small subset of models, nor are they idiosyncratic to particular questions. Instead, they reflect structured patterns that persist across multiple platforms and measurement approaches.

Beyond descriptive comparisons, we embed LLMs in survey-style experimental settings to examine how ideological cues shape model behavior. Conservative and liberal prompts produce asymmetric treatment effects that vary sharply by language, revealing systematic differences in ideological responsiveness. These findings demonstrate that LLMs do not merely encode static biases; they respond to framing and manipulation in ways that depend on linguistic and contextual factors. As a result, identical experimental designs implemented with different models or languages can yield substantively different conclusions.

Substantively, our results suggest that users interacting with different AI systems – or the same system in different languages – may receive meaningfully different political information. Methodologically, they caution against treating LLMs as neutral or interchangeable tools in social science research. When models are used as annotators, simulated respondents, or experimental subjects, ideological variation can affect both point estimates and treatment effects. More broadly, our findings underscore the importance of comparative, multi-model research designs and of treating LLMs as objects of study rather than black-box instruments.

Finally, our work calls for caution in research designs that draw samples from LLM responses. As shown in our paper, for WVS data, human responses systematically differ from most of the LLM responses in the IRT estimate.

As LLMs continue to shape political communication and research practice, understanding how and why they differ will only become more important. Future work should extend this analysis to additional languages, policy domains, and model architectures, as well as to real-world user interactions. By combining careful measurement with experimental designs, researchers can better assess not only what LLMs say, but how

their underlying differences shape the political information they produce.

Future research can build on our framework in several directions. First, the measurement and experimental approach developed in this paper can be readily extended to study ideological differences across a broader set of models, platforms, and languages as new systems continue to emerge. Second, this framework can be used to track ideological change over time. As models are updated, retrained, or realigned, repeated measurement using a fixed set of items and prompts would allow researchers to study ideological drift, convergence, or divergence across versions. Such longitudinal analyses would be particularly valuable for understanding how changes in training data, alignment objectives, or regulatory environments translate into shifts in political content. Together, these extensions would help move the study of ideological bias in large language models from static audits toward cumulative and dynamic measurement.

## References


Alibaba Cloud AI. 2024. "Qwen 3 Technical Report." https://github.com/QwenLM. Multilingual large language model developed by Alibaba Cloud.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* .

Ashokkumar, Ashwini, Luke Hewitt, Isaias Ghezae and Robb Willer. 2024*a*. "Predicting Results of Social Science Experiments Using Large Language Models." *arXiv preprint* .

Ashokkumar, Ashwini, Luke Hewitt, Isaias Ghezae and Robb Willer. 2024*b*. "Predicting Results of Social Science Experiments Using Large Language Models.".

Atari, Mohammad, J. Xue, Damian Blasi and Joseph Henrich. 2024. "Which Humans?" *arXiv preprint* .

Atari, Mohammad, J. Xue, Mona, Peter S. Park, Damian Blasi and Joseph Henrich. 2024. "Which Humans?".
**URL:** *https://osf.io/preprints/psyarxiv/5b26t_v1*

Bang, Yejin, Delong Chen, Nayeon Lee and Pascale Fung. 2024*a*. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said." *arXiv preprint arXiv:2403.18932* .

Bang, Yejin, Delong Chen, Nayeon Lee and Pascale Fung. 2024*b*. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said.".
**URL:** *https://arxiv.org/abs/2403.18932*

Clinton, Joshua D., Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

DeepSeek-AI. 2024. "DeepSeek Chat v3 Technical Report." https://github.com/deepseek-ai. Conversational large language model developed by DeepSeek.

Döll, Michael, Markus Döhring and Andreas Müller. 2024. "Evaluating Gender Bias in Large Language Models.".
**URL:** *https://arxiv.org/abs/2411.09826*

Gilardi, Fabrizio, Meysam Alizadeh and Mael Kubli. 2023. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120(30).

Haerpfer, Christian, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin and Bi Puranen. 2020. "World values survey wave 7 (2017-2020) cross-national data-set." *(No Title)* .

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *NBER Working Paper* .

Hu, Tiancheng, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden and Jon Roozenbeek. 2025. "Generative language models exhibit social identity biases." *Nature Computational Science* 5(1):65–75.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Chichester: Wiley.

Jiang, Liwei, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak and Yejin Choi. 2025. "Artificial hivemind: The open-ended homogeneity of language models (and beyond)." *arXiv preprint arXiv:2510.22954* .

Lin, Luyang, Lingzhi Wang, Jinsong Guo and Kam-Fai Wong. 2024*a*. "Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception." *arXiv preprint arXiv:2403.14896* .

Lin, Luyang, Lingzhi Wang, Jinsong Guo and Kam Fai Wong. 2024*b*. "Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception.".
**URL:** *https://arxiv.org/abs/2403.14896*

Manvi, Rohin, Samar Khanna, Marshall Burke, David Lobell and Stefano Ermon. 2024. "Large Language Models are Geographically Biased.".
**URL:** *https://arxiv.org/abs/2402.02680*

Menéndez, María Luisa, Julio Angel Pardo, Leandro Pardo and María del C Pardo. 1997. "The jensen-shannon divergence." *Journal of the Franklin Institute* 334(2):307–318.

Meta AI. 2024. "LLaMA 4: Maverick Model Card." https://ai.meta.com/llama. Instruction-tuned large language model released by Meta AI.

Motoki, Fabio, Valdemar Pinho Neto and Victor Rangel. 2025. "Assessing Political Bias and Value Misalignment in Generative Artificial Intelligence." *Journal of Economic Behavior and Organization* .

OpenAI. 2024. "GPT-4.1-mini Technical Report." https://openai.com/research. Large language model developed by OpenAI; API-accessed variant of the GPT-4.1 family.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang and Michael S. Bernstein. 2024*a*. "Generative Agent Simulations of 1,000 People." *arXiv preprint arXiv:2411.10109* .

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang and Michael S. Bernstein. 2024*b*. "Generative Agent Simulations of 1,000 People.".
**URL:** *http://arxiv.org/abs/2411.10109*

Pit, Pagnarasmey, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey, Henry Pit, Putrasmey Keo, Watey Diep and Yu-Gang Jiang. 2024. "Whose Side Are You On? Investigating the Political Stance of Large Language Models.".
**URL:** *https://arxiv.org/abs/2403.13840*

Pit, Pagnarasmey, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey et al. 2024. "Whose Side Are You On? Investigating the Political Stance of Large Language Models." *arXiv preprint arXiv:2403.13840* .

Qu, Yao and Jue Wang. 2024. "Performance and biases of Large Language Models in public opinion simulation." *Humanities and Social Sciences Communications* 11(1):1–13.

Strimling, Pontus, Joel Krueger and Simon Karlsson. 2024. "GPT-4's One-Dimensional Mapping of Morality: How the Accuracy of Country-Estimates Depends on Moral Domain.".
**URL:** *https://arxiv.org/abs/2407.16886*

Tang, Raphael, Xinyu Zhang, Jimmy Lin and Ferhan Ture. 2023. "What Do Llamas Really Think? Revealing Preference Biases in Language Model Representations.".
**URL:** *http://arxiv.org/abs/2311.18812*

Tao, Yan, Olga Viberg, Ryan S. Baker and Rene F. Kizilcec. 2024*a*. "Cultural Bias and Cultural Alignment of Large Language Models." *PNAS Nexus* .

Tao, Yan, Olga Viberg, Ryan S. Baker and Rene F. Kizilcec. 2024*b*. "Cultural Bias and Cultural Alignment of Large Language Models." *PNAS Nexus* 3(9):1–9.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–217.

Vera, Sebastian Vallejo and Hunter Driggers. 2024*a*. "Bias in LLMs as Annotators: The Effect of Party Cues on Labeling Decisions by Large Language Models." *arXiv preprint arXiv:2408.15895* .

Vera, Sebastian Vallejo and Hunter Driggers. 2024*b*. "Bias in LLMs as Annotators: The Effect of Party Cues on Labelling Decision by Large Language Models.".
**URL:** *https://arxiv.org/abs/2408.15895*

Wang, Angelina, Jamie Morgenstern and John P. Dickerson. 2025. "Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups." *Nature Machine Intelligence* 7(3):400–411.

Yang, Eddie, Yin Yuan, Solomon Messing, Margaret Roberts, Brandon Stewart and Joshua Tucker. 2025. "Propaganda Is Already Influencing Large Language Models: Evidence From Training Data, Audits, and Real-World Usage." *arXiv preprint* .

Yang, Kai-Cheng and Filippo Menczer. 2025. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. In *Websci*. pp. 127–137.

Yang, Kaiqi, Hang Li, Yucheng Chu, Yuping Lin, Tai-Quan Peng and Hui Liu. 2024*a*. "Unpacking Political Bias in Large Language Models: Insights Across Topic Polarization." *arXiv preprint arXiv:2412.16746* .

Yang, Kaiqi, Hang Li, Yucheng Chu, Yuping Lin, Tai-Quan Peng and Hui Liu. 2024*b*. "Unpacking Political Bias in Large Language Models: Insights Across Topic Polarization.".
**URL:** *http://arxiv.org/abs/2412.16746*

Zhou, Di and Yinxian Zhang. 2024*a*. "Political Biases and Inconsistencies in Bilingual GPT Models—The Cases of the U.S. and China." *Scientific Reports* .

Zhou, Di and Yinxian Zhang. 2024*b*. "Political Biases and Inconsistencies in Bilingual GPT Models—The Cases of the u.s. And China." *Scientific Reports* 14(1):1–13.

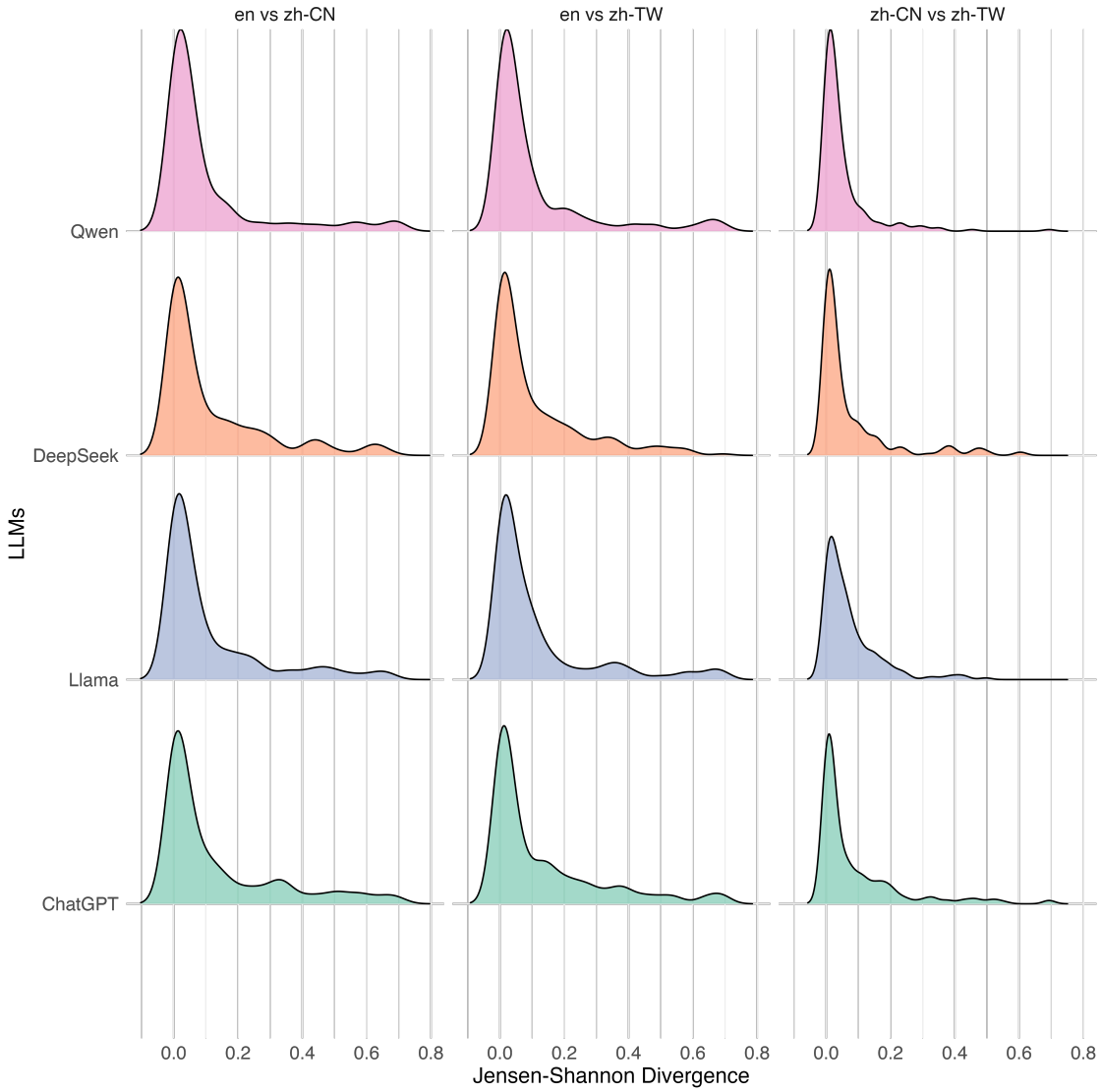
# Appendix

The appendix includes the following:

- Appendix A: Jensen-Shannon Divergence
- Appendix B: Multi-Dimensional Scaling
- Appendix C: More Results on IRT
- Appendix D: Sample Manipulations
- Appendix E: More Descriptive Analysis

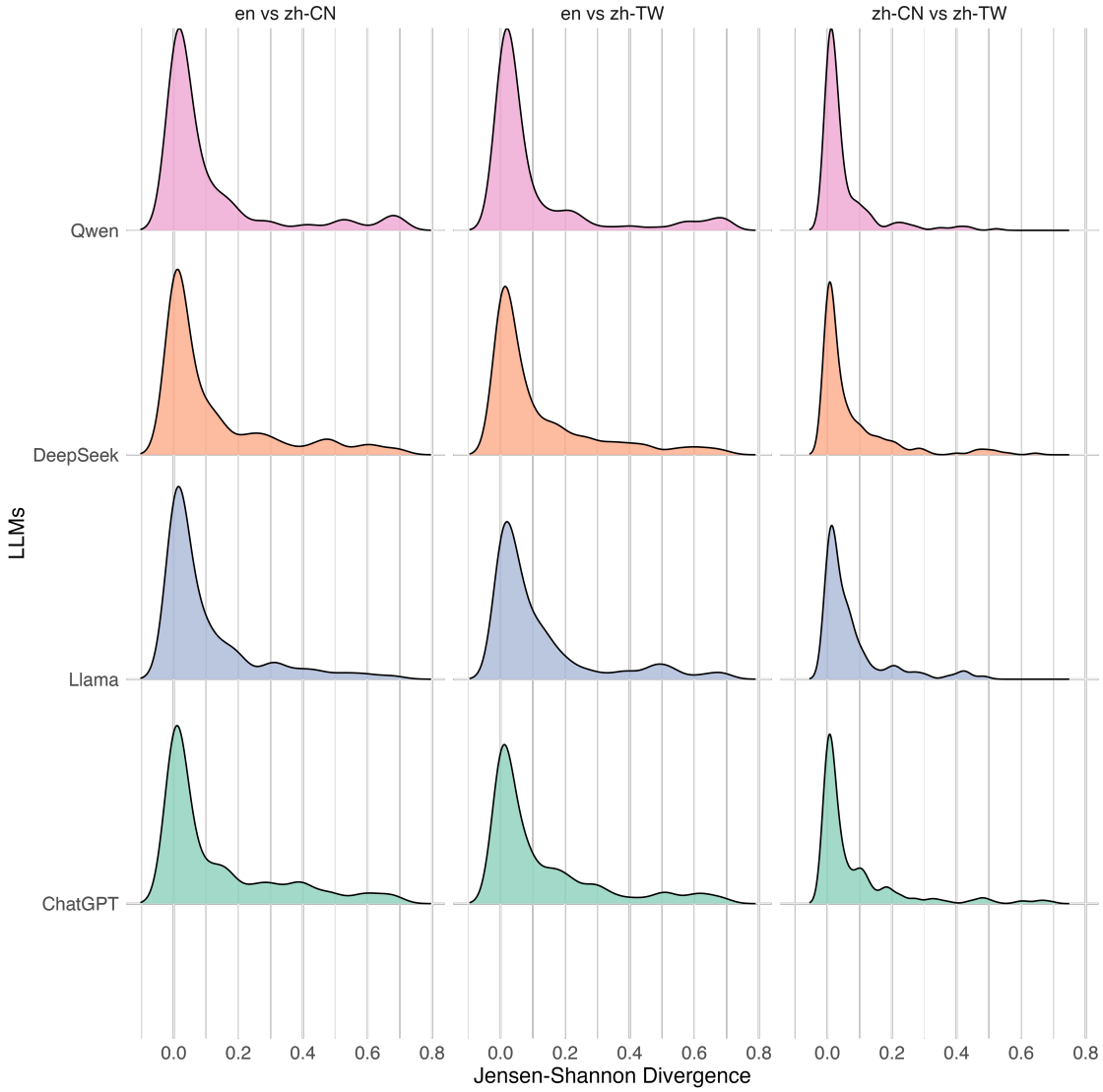
# Appendix A Jensen-Shannon Divergence

Figure A.1: JSD of LLM Response Distributions across Languages (Conservative Group)




*Note*: en = English; zh-CN = Simplified Chinese; zh-TW = Traditional Chinese.

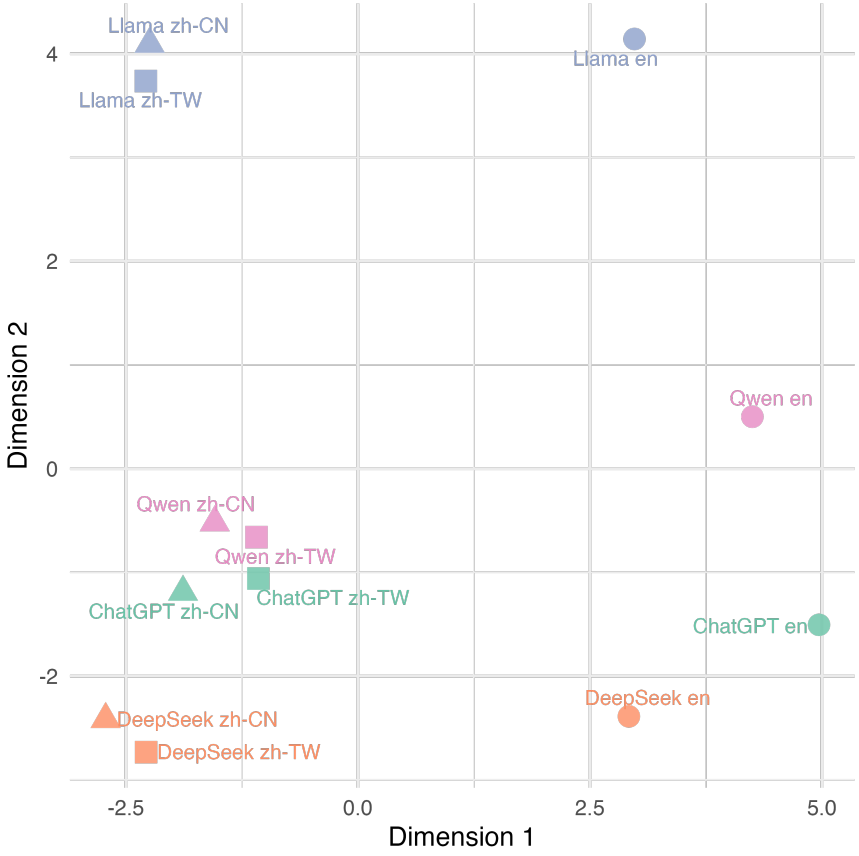
Figure A.2: JSD of LLM Response Distributions across Languages (Liberal Group)




*Note*: en = English; zh-CN = Simplified Chinese; zh-TW = Traditional Chinese.

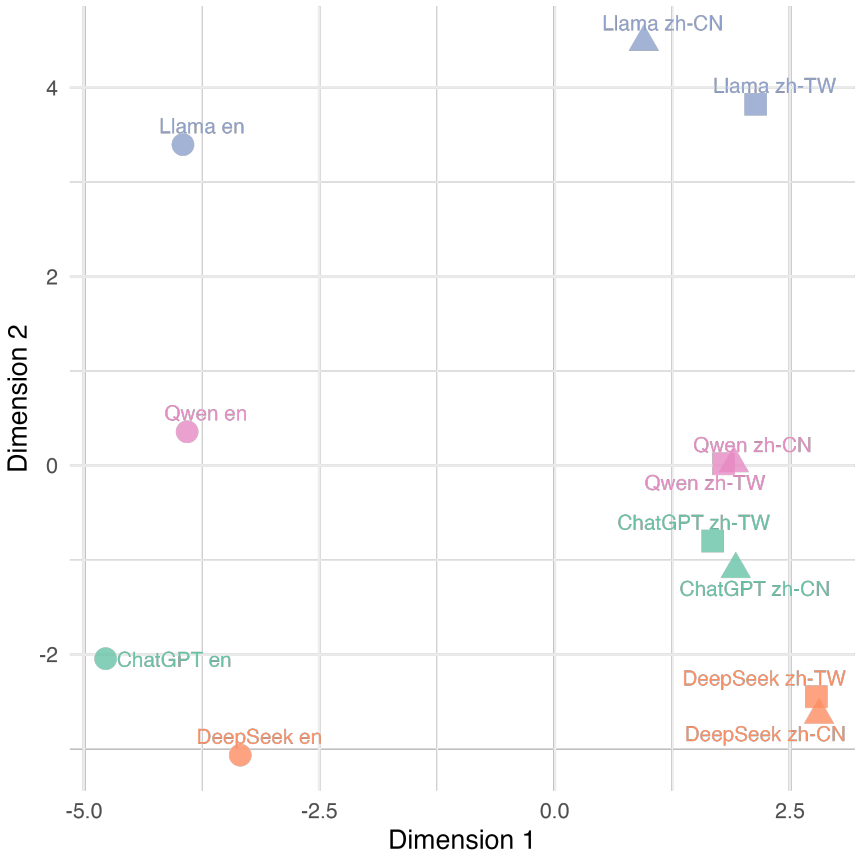
# Appendix B Multi-Dimensional Scaling

Figure B.1: MDS of LLM Response Distributions across Languages (Conservative Group)




*Note*: en = English; zh-CN = Simplified Chinese; zh-TW = Traditional Chinese.

Figure B.2: MDS of LLM Response Distributions across Languages (Liberal Group)




*Note*: en = English; zh-CN = Simplified Chinese; zh-TW = Traditional Chinese.

# Appendix C More Results on IRT

# Appendix D Sample Survey Manipulations

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| socialvalues | 移民推动了经济的发展。当移民进入劳动力市场时，他们提升了经济的生产能力，并提高了国内生产总值（GDP）。他们的收入增加了，本地居民的收入也随之增长。这种现象被称为“移民红利”。虽然新增 GDP 中只有一小部分流向本地居民——通常为 0.2% 到 0.4%——但每年也相当于 360 亿到 720 亿美元。除了“移民红利”之外，移民还通过流入那些相对更需要劳动力的行业和地区，润滑了劳动力市场的运转——这些行业和地区若缺乏劳动力，可能会抑制经济增长。 | liberal |
| socialvalues | Immigration fuels the economy. When immigrants enter the labor force, they increase the productive capacity of the economy and raise GDP. Their incomes rise, but so do those of natives. It’s a phenomenon dubbed the ’immigration surplus,’ and while a small share of additional GDP accrues to natives —typically 0.2 to 0.4 percent —it still amounts to $36 to 72$ billion per year. In addition to the immigration surplus, immigrants grease the wheels of the labor market by flowing into industries and areas where there is a relative need for workers —where bottlenecks or shortages might otherwise damp growth. | liberal |
| socialvalues | 移民推動了經濟的發展。當移民進入勞動力市場時，他們提升了經濟的生產能力，並提高了國內生產總值（GDP）。他們的收入增加了，本地居民的收入也隨之增長。這種現象被稱為「移民紅利」。雖然新增 GDP 中只有一小部分流向本地居民——通常為 0.2% 到 0.4%——但每年也相當於 360 億到 720 億美元。除了「移民紅利」之外，移民還透過流入那些相對更需要勞動力的行業和地區，潤滑了勞動力市場的運轉——這些行業和地區若缺乏勞動力，可能會抑制經濟增長。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| socialvalues | 公众对移民影响的担忧日益上升。一项 2025 年的 YouGov 民调在西欧七国进行，结果发现绝大多数受访者认为过去十年的移民数量过高，政府管理不善。超过一半的德国和意大利受访者表示，近期的移民对本国不利。这种情绪助长了反移民政党的成功，并促使主流政治家采取更严格的边境政策。持这种观点的人认为，大量移民涌入会加剧公共服务负担，威胁就业和国家文化，因此主张实施更严厉的移民管控。 | conservative |
| socialvalues | Public concern about immigration’s impact has been rising. A 2025 YouGov poll across seven Western European countries found large majorities believe immigration over the past decade has been too high and handled poorly by governments. Over half of Germans and Italians surveyed said recent immigration has been bad for their country. This sentiment has fueled the success of anti-immigration parties and pushed mainstream politicians toward tougher border policies. Advocates of this view argue that an influx of migrants can strain social services, threaten jobs and national culture, and therefore support stricter immigration controls. | conservative |
| socialvalues | 公眾對移民影響的擔憂日益上升。一項 2025 年的 YouGov 民調在西歐七國進行，結果發現絕大多數受訪者認為過去十年的移民數量過高，政府管理不善。超過一半的德國和意大利受訪者表示，近期的移民對本國不利。這種情緒助長了反移民政黨的成功，並促使主流政治家採取更嚴格的邊境政策。持這種觀點的人認為，大量移民湧入會加劇公共服務負擔，威脅就業和國家文化，因此主張實施更嚴厲的移民管控。 | conservative |
| happiness | 公众福祉与社会经济政策的关联日益紧密。例如，北欧国家经常位居联合国《世界幸福报告》排名前列，研究人员将此归因于这些国家拥有完善的社会保障、高度的政府信任以及工作与生活的平衡。全球各地越来越多的政策制定者认识到，仅 GDP 并不等同于幸福——新西兰和不丹等国甚至采用了“幸福预算”，优先考虑心理健康、教育和环境。倡导者认为，相比单纯关注经济增长，加大医疗投入、减少不平等和扩大社会支持能让民众更健康、更幸福。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| happiness | Public well-being is increasingly linked to social and economic policies. The Nordic countries, for example, regularly top the UN's World Happiness Report rankings, which researchers attribute to their strong social safety nets, high trust in government, and work-life balance. Around the world, more policymakers are recognizing that GDP alone doesn't equate to happiness –nations like New Zealand and Bhutan have even adopted "well-being budgets" that prioritize mental health, education and environment. Advocates argue that investing in healthcare, reducing inequality, and expanding social support leads to healthier, happier populations than a sole focus on economic growth. | liberal |
| happiness | 公眾福祉與社會經濟政策的關聯日益緊密。例如，北歐國家經常位居聯合國《世界幸福報告》排名前列，研究人員將此歸因於這些國家擁有完善的社會保障、高度的政府信任以及工作與生活的平衡。全球各地越來越多的政策制定者認識到，僅 GDP 並不等同於幸福——紐西蘭和不丹等國甚至採用了「幸福預算」，優先考慮心理健康、教育和環境。倡導者認為，相比單純關注經濟增長，加大醫療投入、減少不平等和擴大社會支持能讓民眾更健康、更幸福。 | liberal |
| happiness | 有些人认为，与其依赖政府项目，幸福更多取决于个人选择和价值观。美国的调查长期显示，自称保守派的人往往报告的个人幸福感和心理健康水平高于自由派。分析人士指出，婚姻、信仰和社区参与等因素（在保守群体中更常见）可能促成了这种差距。他们也指出，财富和经济增长能够通过提供机会来提升福祉。此观点强调个人责任、稳固的家庭结构和经济自由是幸福生活的关键要素。 | conservative |
| happiness | Some argue that happiness comes down more to personal choices and values than government programs. Surveys in the U.S. have long shown that self-described conservative individuals often report higher levels of personal happiness and mental well-being than liberals. Analysts suggest factors like marriage, faith, and community involvement – which are more common among conservative groups –may contribute to this gap. They also point out that wealth and economic growth can lift well-being by providing opportunities. This perspective emphasizes individual responsibility, strong family structures, and economic freedom as key ingredients for a satisfying life. | conservative |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| happiness | 有些人認為，與其依賴政府項目，幸福更多取決於個人選擇和價值觀。美國的調查長期顯示，自稱保守派的人往往報告的個人幸福感和心理健康水平高於自由派。分析人士指出，婚姻、信仰和社區參與等因素（在保守群體中更常見）可能促成了這種差距。他們也指出，財富和經濟增長能夠通過提供機會來提升福祉。此觀點強調個人責任、穩固的家庭結構和經濟自由是幸福生活的關鍵要素。 | conservative |
| socialcapital | 社会联系紧密的社区往往更加兴旺。英国在 2024 年的一项新研究发现，志愿服务与社会凝聚力之间存在良性循环：在社区中感到联系紧密的人会更多地参与志愿活动，而志愿服务进一步增强了他们在不同群体之间的归属感和信任感。类似趋势在其他地方也有体现——新冠疫情期间，组织互助和志愿网络的社区展现出更强的团结和韧性。支持这一理念的人认为，通过投资社区项目和鼓励公民参与，可以在社会碎片化的时代重建社会资本，弥合社会裂痕。 | liberal |
| socialcapital | Communities with strong social ties tend to thrive. New research in the UK (2024) found a positive feedback loop between volunteering and social cohesion: people who feel connected in their community volunteer more, and volunteering further increases their sense of belonging and trust across different groups. Similar trends are seen elsewhere – during the COVID-19 pandemic, neighborhoods that organized mutual aid and volunteer networks reported greater solidarity and resilience. Supporters of this approach believe that investing in community programs and encouraging civic engagement can rebuild social capital and bridge divides in an era of social fragmentation. | liberal |
| socialcapital | 社會聯繫緊密的社區往往更加興旺。英國在 2024 年的一項新研究發現，志願服務與社會凝聚力之間存在良性循環：在社區中感到聯繫緊密的人會更多地參與志願活動，而志願服務進一步增強了他們在不同群體之間的歸屬感和信任感。類似趨勢在其他地方也有體現——新冠疫情期間，組織互助和志願網絡的社區展現出更強的團結和韌性。支持這一理念的人認為，通過投資社區項目和鼓勵公民參與，可以在社會碎片化的時代重建社會資本，彌合社會裂痕。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| socialcapital | 近几十年来，传统社区生活的支柱已经削弱。宗教聚会和公民组织的参与率在许多西方国家大幅下降——例如在美国，每周去教堂的人口比例从 20 世纪 80 年代的 40% 以上跌至如今的约 30%。家庭结构也在改变，独居人口越来越多。官员警告，这种社会纽带的流失带来了后果：美国公共卫生局长在 2023 年宣布进入“孤独流行病”。保守派认为，重振家庭价值观、宗教参与和本地机构，对于恢复社会凝聚力、对抗现代社会的孤立现象至关重要。 | conservative |
| socialcapital | Traditional pillars of community life have weakened in recent decades. Church attendance and membership in civic organizations have plummeted in many Western countries –in the U.S., weekly churchgoers fell from over 40% in the 1980s to around 30% today. Family structures are also changing, with more people living alone. Officials warn this erosion of social bonds has consequences: the U.S. Surgeon General in 2023 declared a “loneliness epidemic.” Conservatives argue that reviving family values, religious participation, and local institutions is crucial for restoring social cohesion and combating the isolation of modern society. | conservative |
| socialcapital | 近幾十年來，傳統社區生活的支柱已經削弱。宗教聚會和公民組織的參與率在許多西方國家大幅下降——例如在美國，每週去教堂的人口比例從 20 世紀 80 年代的 40% 以上跌至如今的約 30%。家庭結構也在改變，獨居人口越來越多。官員警告，這種社會紐帶的流失帶來了後果：美國公共衛生局長在 2023 年宣佈進入「孤獨流行病」。保守派認為，重振家庭價值觀、宗教參與和本地機構，對於恢復社會凝聚力、對抗現代社會的孤立現象至關重要。 | conservative |
| economicvalues | 世界各地的进步人士呼吁将环境可持续性和社会福利置于无限经济增长之上。例如，欧盟的“绿色协议”旨在改变产业以减少碳排放，即使这意味着承担短期成本。气候科学家警告说，不采取行动将在未来通过灾害和作物歉收带来更大的经济损失。同样，许多公众支持将福祉优先于 GDP——在民调中，法国和加拿大等国的大多数人都倾向于保护环境、减少不平等，而非一味追求增长。这种观点认为，生活质量和长期稳定胜过眼前的利润。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| economicvalues | Progressives worldwide are calling to put environmental sustainability and social welfare ahead of unrestrained economic growth. For example, the EU's Green Deal aims to transform industries to cut carbon emissions, even at the risk of short-term costs. Climate scientists warn that failing to act will impose far greater economic damage in the future through disasters and crop failures. Likewise, many citizens support prioritizing well-being over GDP –in surveys, majorities in countries like France and Canada favor environmental protection and reducing inequality over maximizing growth. The view is that quality of life and long-term stability trump immediate profit. | liberal |
| economicvalues | 世界各地的進步人士呼籲將環境可持續性和社會福利置於無限經濟增長之上。例如，歐盟的「綠色協議」旨在改變產業以減少碳排放，即使這意味著承擔短期成本。氣候科學家警告說，不採取行動將在未來通過災害和作物歉收帶來更大的經濟損失。同樣，許多公眾支持將福祉優先於 GDP——在民調中，法國和加拿大等國的大多數人都傾向於保護環境、減少不平等，而非一味追求增長。這種觀點認為，生活質量和長期穩定勝過眼前的利潤。 | liberal |
| economicvalues | 另一方面，许多领导人和经济学家仍将经济增长和就业置于首位。他们认为，强劲的增长能产生改善生活水平所需的资源。2023 年，在高能源价格和经济衰退担忧下，一些政府回撤了绿色政策以保护企业和消费者——例如英国推迟了一些气候承诺，以避免加重行业负担。增长优先的倡导者警告，以环境或平等之名进行过度监管可能适得其反。他们相信，繁荣的经济最终将提供技术和财富，在适当的时候解决社会和环境问题。 | conservative |
| economicvalues | On the other hand, many leaders and economists prioritize economic growth and job creation above all. They argue that strong growth generates the resources needed to improve living standards. In 2023, facing high energy prices and recession fears, several governments rolled back green policies to shield businesses and consumers –for instance, the UK delayed some climate commitments to avoid burdening industries. Proponents of a growth-first approach caution that overregulating in the name of the environment or equality can backfire. They believe a thriving economy will ultimately provide technology and wealth to address social and environmental issues in due time. | conservative |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| economicvalues | 另一方面，許多領導人和經濟學家仍將經濟增長和就業置於首位。他們認為，強勁的增長能產生改善生活水平所需的資源。2023 年，在高能源價格和經濟衰退擔憂下，一些政府回撤了綠色政策以保護企業和消費者——例如英國推遲了一些氣候承諾，以避免加重行業負擔。增長優先的倡導者警告，以環境或平等之名進行過度監管可能適得其反。他們相信，繁榮的經濟最終將提供技術和財富，在適當的時候解決社會和環境問題。 | conservative |
| corruption | 世界各地的政府和活动人士正在加紧反腐败努力。透明国际 2023 年的指数显示，尽管腐败在大多数国家仍然猖獗，但一些改革型政府已取得进展。在东南亚，印尼的反贪行动导致多名高官被逮捕；在欧洲，爱沙尼亚等国通过提高政府采购透明度取得改善。跨国举措也施加了压力——例如数据泄露（如“潘多拉文件”）曝光离岸账户，促使领导人采取行动。令人鼓舞的迹象包括新的反贿赂法律和减少腐败机会的数字工具。改革者坚持认为，持续的警惕可以逐步遏制腐败的蔓延。 | liberal |
| corruption | Governments and activists around the world are intensifying efforts to combat corruption. Transparency International’s 2023 index shows that while corruption remains rampant in most countries, some reformist administrations have made gains. In Southeast Asia, Indonesia’s anti-graft drive has led to high-profile arrests of officials, and in Europe, nations like Estonia have improved transparency in government contracting. Cross-border initiatives –such as data leaks (e.g., the Pandora Papers) exposing offshore accounts –have pressured leaders to act. Encouraging signs include new anti-bribery laws and digital tools that reduce opportunities for graft. Reformers insist that sustained vigilance can gradually curb corruption’s reach. | liberal |
| corruption | 世界各地的政府和活動人士正在加緊反腐敗努力。透明國際 2023 年的指數顯示，儘管腐敗在大多數國家仍然猖獗，但一些改革型政府已取得進展。在東南亞，印尼的反貪行動導致多名高官被逮捕；在歐洲，愛沙尼亞等國通過提高政府採購透明度取得改善。跨國舉措也施加了壓力——例如數據洩露（如「潘多拉文件」）曝光離岸帳戶，促使領導人採取行動。令人鼓舞的跡象包括新的反賄賂法律和減少腐敗機會的數字工具。改革者堅持認為，持續的警惕可以逐步遏制腐敗的蔓延。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| corruption | 尽管进行了打击，许多人仍深信腐败程度没有改变。丑闻在全球层出不穷——从公共工程行贿到国家元首被控贪污——加剧了公众的愤世嫉俗情绪。这种愤怒推动了承诺“清理”腐败的局外人政治人物。在拉丁美洲和非洲，持强硬反腐言论的领导人通过迎合选民对腐败的不满赢得了选举。一些保守评论员认为需要采取激烈措施：对贪官施以更严厉的刑罚、缩小政府规模以减少官僚腐败，并赋予执法机构更大权力来严厉打击欺诈。在他们看来，腐败是一种道德失败，需要毫不妥协的惩罚。 | conservative |
| corruption | Despite crackdowns, many people remain convinced that corruption is unchanged. Scandals continue worldwide −from bribes in public works projects to presidents accused of embezzlement −feeding public cynicism. This anger has boosted outsider politicians promising to “clean house.” In Latin America and Africa, leaders with tough anti-corruption rhetoric have won elections by tapping into voters’ frustration with graft. Some conservative commentators argue that drastic measures are needed: harsher prison sentences for corrupt officials, smaller governments to limit bureaucratic graft, and empowering law enforcement to aggressively pursue fraud. They see corruption as a moral failing requiring uncompromising punishment. | conservative |
| corruption | 儘管進行了打擊，許多人仍深信腐敗程度沒有改變。醜聞在全球層出不窮——從公共工程賄賂到國家元首被控貪污——加劇了公眾的憤世嫉俗情緒。這種憤怒推動了承諾「清理」腐敗的局外人政治人物。在拉丁美洲和非洲，持強硬反腐言論的領導人通過迎合選民對腐敗的不滿贏得了選舉。一些保守評論員認為需要採取激烈措施：對貪官施以更嚴厲的刑罰、縮小政府規模以減少官僚腐敗，並賦予執法機構更大權力來嚴厲打擊欺詐。在他們看來，腐敗是一種道德失敗，需要毫不妥協的懲罰。 | conservative |
| migration | 几个世纪以来，迁徙一直是推动发展与繁荣的主要动力。如今，向中低收入国家汇入的国际汇款大约为 6700 亿美元，这一数字超过了外国直接投资，也远远超过了这些国家获得的官方发展援助。许多地区（例如东非）正在投资于涵盖贸易和移民的区域框架，因为人员和商品的自由流动能够带来可持续的、包容性的经济增长以及有成效的就业。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| migration | For centuries, migration has been a major driver of development and prosperity. International remittances to low- and middle-income countries are now at about $670 billion, which is more than direct foreign investment and far more than official development assistance to those countries. Many regions, such as Eastern Africa, are now investing in regional frameworks that cover trade and migration because the free movement of people and goods can bring sustained, inclusive economic growth and productive employment. | liberal |
| migration | 幾個世紀以來，遷徙一直是推動發展與繁榮的主要動力。如今，匯往中低收入國家的國際匯款約為 6700 億美元，這個數字超過了外國直接投資，也遠遠高於這些國家獲得的官方發展援助。許多地區（例如東非）正致力於投資涵蓋貿易與移民的區域框架，因為人員與商品的自由流動能夠帶來可持續且具包容性的經濟增長與有效就業。 | liberal |
| migration | 大规模移民潮在全球引发了政治反弹。在美国，2022 年当局记录了超过 200 万起非法越境事件，促使政府收紧庇护规定，并再度有人呼吁“修建边境墙”。同样，欧盟也采取更强硬的立场：欧洲各国正在架设边境围栏，并推动将庇护申请处理外包给第三国。公众对移民可能抢走工作或占用公共服务的担忧助长了反移民政党的崛起（如意大利和瑞典）。支持更严格移民政策的人认为，需要严格执法来维护国家安全和社会凝聚。 | conservative |
| migration | High migration flows have spurred political backlashes across the globe. In the US, authorities recorded over 2 million unauthorized border crossings in 2022, prompting stricter asylum rules and renewed calls to 'build the wall.' Similarly, the EU has toughened its stance: countries in Europe are erecting border fences and pushing to outsource asylum processing to third countries. Public fears that migrants may take jobs or strain services have fueled the rise of anti-immigration parties (such as in Italy and Sweden). Proponents of tougher immigration policies argue that strict enforcement is needed to preserve national security and social cohesion. | conservative |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| migration | 大規模移民潮在全球引發了政治反彈。在美國，2022 年當局記錄了超過 200 萬起非法越境事件，促使政府收緊庇護規定，並再度有人呼籲「修建邊境牆」。同樣，歐盟也採取更強硬的立場：歐洲各國正在架設邊境圍欄，並推動將庇護申請處理外包給第三國。公眾對移民可能搶走工作或佔用公共服務的擔憂助長了反移民政黨的崛起（如意大利和瑞典）。支持更嚴格移民政策的人認為，需要嚴格執法來維護國家安全和社會凝聚力。 | conservative |
| security | 许多专家提倡综合性的安全观，强调关注根源和全球合作，而不仅仅依靠武力。他们认为，通过解决潜在的不满情绪，诸如减少贫困、扩大教育等社会投资可以防止犯罪和激进化。在国际事务中，这种观点倾向于外交和军备控制：例如，联合国正就限制自主武器和恢复核条约进行谈判。该理念还将“安全”的概念拓展到气候变化和流行病等人类安全议题。通过国际合作和预防手段应对这些威胁，倡导者相信能够更好地实现全球稳定。 | liberal |
| security | Many experts advocate a holistic approach to security, focusing on root causes and global cooperation instead of force alone. They argue that social investments –like reducing poverty and expanding education –can prevent crime and radicalization by addressing underlying grievances. In international affairs, this perspective favors diplomacy and arms control: for example, ongoing UN talks aim to limit autonomous weapons and revive nuclear treaties. It also broadens 'security' to include human security issues such as climate change and pandemics. By tackling these threats through international collaboration and prevention, proponents believe global stability is best achieved. | liberal |
| security | 許多專家提倡綜合性的安全觀，強調關注根源和全球合作，而不僅僅依靠武力。他們認為，通過解決潛在的不滿情緒，諸如減少貧困、擴大教育等社會投資可以防止犯罪和激進化。在國際事務中，這種觀點傾向於外交和軍備控制：例如，聯合國正在就限制自主武器和恢復核條約進行談判。該理念還將「安全」的概念拓展到氣候變化和流行病等人類安全議題。通過國際合作和預防手段應對這些威脅，倡導者相信能夠更好地實現全球穩定。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| security | 许多国家的领导人在犯罪和冲突隐忧下强调安全至上的方针。保守派决策者常呼吁加强警察和军事实力以维持秩序。例如，萨尔瓦多政府展开大规模打击帮派行动，大幅降低了暴力犯罪，这一策略赢得了严打拥护者的赞赏。在全球范围内，恐怖主义和战争威胁加剧，各国也相应增加国防开支——以北约盟国为例，他们在 2024 年提升了军费预算。这种观念将强有力的执法、边境管控和监控置于优先地位，即使意味着牺牲部分公民自由，也在所不惜，以保护民众安全。 | conservative |
| security | In many countries, leaders stress a security-first approach amid concerns about crime and conflict. Conservative policymakers often call for bolstering police and military power to maintain order. For example, El Salvador's government launched mass gang crackdowns that sharply reduced violence, a strategy applauded by tough-on-crime advocates. On a global scale, heightened threats from terrorism and war have prompted increases in defense spending –NATO allies, for instance, boosted military budgets in 2024. This perspective prioritizes strong law enforcement, border control, and surveillance, even if it means trading off some civil liberties, in the name of protecting citizens. | conservative |
| security | 許多國家的領導人在犯罪和衝突隱憂下強調安全至上的方針。保守派決策者常呼籲加強警察和軍事實力以維持秩序。例如，薩爾瓦多政府展開大規模打擊幫派行動，大幅降低了暴力犯罪，這一策略贏得了嚴打擁護者的讚賞。在全球範圍內，恐怖主義和戰爭威脅加劇，各國也相應增加國防開支——以北約盟國為例，他們在 2024 年提升了軍費預算。這種觀念將強有力的執法、邊境管控和監控置於優先地位，即使意味著犧牲部分公民自由也在所不惜，以保護民眾安全。 | conservative |
| postmaterialist | 随着社会更加富裕，调查显示人们的价值观正向后物质主义转变——更重视自由表达、环境保护和平等，而非经济和安全。尤其是年轻一代，更倾向于支持气候行动和社会正义等事业，即使这意味着减缓增长或挑战传统规范。例如，在高收入国家，民意调查发现，越来越多的人愿意支付更高价格或税收来应对气候变化。这种价值观的转变表明，一旦基本需求得到满足，人们就会更看重生活质量和个人自由，将其视为进步的标志。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| postmaterialist | As societies grow more affluent, surveys show people shifting toward post-materialist values –prioritizing free expression, environmental protection, and equality over economic and physical security. Younger generations especially tend to support causes like climate action and social justice, even if it means slower growth or challenging traditional norms. For instance, in high-income countries, public opinion polls find increasing willingness to pay higher prices or taxes to combat climate change. This value shift suggests that once basic needs are met, populations place greater importance on quality-of-life issues and personal freedoms as markers of progress. | liberal |
| postmaterialist | 隨著社會更加富裕，調查顯示人們的價值觀正向後物質主義轉變——更重視自由表達、環境保護和平等，而非經濟和安全。尤其是年輕一代，更傾向於支持氣候行動和社會正義等事業，即使這意味著減緩增長或挑戰傳統規範。例如，在高收入國家，民意調查發現，越來越多的人願意支付更高價格或稅收來應對氣候變化。這種價值觀的轉變表明，一旦基本需求得到滿足，人們就會更看重生活質量和個人自由，將其視為進步的標誌。 | liberal |
| postmaterialist | 对许多人来说，物质问题仍然是首要关切。在发展中国家，乃至西方遇到困难时期的社会，民调显示人们将就业、通胀、安全等问题远远置于气候或文化自由化等话题之上。近期食品和能源价格的飙升使公众重新关注基本需求。在面临冲突或经济危机的国家，民众可以理解地优先考虑稳定和增长。一些评论人士警告，过快推动后物质主义议程——例如昂贵的绿色政策或激进的社会变革——可能引发那些关注切身生计群体的反弹。 | conservative |
| postmaterialist | Material concerns remain paramount for many. In developing economies and even in the West during tough times, polls show people rank issues like jobs, inflation, and security far above topics such as climate or cultural liberalization. Recent surges in food and energy prices refocused public attention on basic needs. In countries facing conflict or economic crisis, citizens understandably prioritize stability and growth. Some commentators caution that pushing post-materialist agendas –for example, expensive green policies or progressive social changes –too quickly can provoke backlash among those who feel their immediate livelihoods are at stake. | conservative |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| postmaterialist | 對許多人來說，物質問題仍然是首要關切。在發展中國家，乃至西方遇到困難時期的社會，民調顯示人們將就業、通脹、安全等問題遠遠置於氣候或文化自由化等話題之上。近期食品和能源價格的飆升使公眾重新關注基本需求。在面臨衝突或經濟危機的國家，民眾可以理解地優先考慮穩定和增長。一些評論人士警告，過快推動後物質主義議程——例如昂貴的綠色政策或激進的社會變革——可能引發那些關注切身生計群體的反彈。 | conservative |
| science | 科学进步被普遍视为应对全球挑战的关键。国际上的努力（如 COVID-19 疫苗的快速开发和清洁能源的突破）加强了公众对科学的信任。许多政府在从气候技术到人工智能伦理等领域增加研究资金，并以专家建议为指导。韩国和德国等国的民调显示，公众对科学家抱有强烈信心，并支持在政策制定中采用科学依据。持这一观点的人认为，拥抱创新和专家指导将推动经济增长，并提高全球范围的生活质量。 | liberal |
| science | Scientific progress is widely seen as essential for tackling global challenges. International efforts like the rapid development of COVID-19 vaccines and breakthroughs in clean energy have reinforced trust in science. Many governments are increasing funding for research in areas from climate technology to AI ethics, guided by expert advice. Polls in countries such as South Korea and Germany show strong public confidence in scientists and support for using science-based evidence in policymaking. Proponents of this view argue that embracing innovation and expert guidance will drive economic growth and improve quality of life worldwide. | liberal |
| science | 科學進步被普遍視為應對全球挑戰的關鍵。國際上的努力（如 COVID-19 疫苗的快速開發和清潔能源的突破）加強了公眾對科學的信任。許多政府在從氣候技術到人工智能倫理等領域增加研究資金，並以專家建議為指導。韓國和德國等國的民調顯示，公眾對科學家抱有強烈信心，並支持在政策制定中採用科學依據。持這一觀點的人認為，擁抱創新和專家指導將推動經濟增長，並提高全球範圍的生活質量。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| science | 一些群体对科学和技术的怀疑情绪加剧。围绕疫情措施和气候政策的争论表明，人们对科学精英存在不信任。例如，对 COVID-19 疫苗强制令的反对在一些保守派中获得了支持，他们将其视为政府的越权。同样，气候变化怀疑者抵制减排法规，担心其对经济有害。对于大型科技公司权力和人工智能风险的担忧也引发了保持谨慎的呼声。这种观点强调个人自由和传统价值观，认为并非所有“专家”的解决方案都符合公众利益。 | conservative |
| science | Skepticism of science and technology has grown in certain groups. Debates over pandemic measures and climate policies illustrate a distrust of scientific elites. For example, opposition to COVID-19 vaccine mandates gained support among some conservatives who framed it as government overreach. Similarly, climate change doubters push back on emissions regulations fearing economic harm. Concerns about Big Tech’s power and artificial intelligence’s risks have also led to calls for caution. This perspective emphasizes individual freedom and traditional values, contending that not all ’expert’ solutions are in the public’s best interest. | conservative |
| science | 一些群體對科學和技術的懷疑情緒加劇。圍繞疫情措施和氣候政策的爭論表明，人們對科學精英存在不信任。例如，對 COVID-19 疫苗強制令的反對在一些保守派中獲得了支持，他們將其視為政府的越權。同樣，氣候變化懷疑者抵制減排法規，擔心其對經濟有害。對大型科技公司權力和人工智能風險的擔憂也引發了保持謹慎的呼聲。這種觀點強調個人自由和傳統價值觀，認為並非所有「專家」的解決方案都符合公眾利益。 | conservative |
| religion | 许多地区的社会正变得更加世俗化。在整个欧洲和东亚，宗教礼拜出席率已降至历史低点，越来越多的人，尤其是年轻人，认同自己无宗教信仰。各国政府也日益采取世俗政策——例如解除对堕胎或同性婚姻的禁令——反映出个人权利优于传统教条的转变。民意调查显示，在这些日益世俗化的社会中，人们对多元生活方式和信仰的接受度不断提高。倡导这一趋势的人将宗教与国家事务的分离视为进步，有助于培养宽容，并使个人能够决定自己的道德价值观。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| religion | Societies in many regions are becoming more secular. Across Europe and East Asia, religious service attendance has fallen to record lows and more people, especially youth, identify as non-religious. Governments have increasingly adopted secular policies – for instance, lifting bans on abortion or same-sex marriage –reflecting a shift toward individual rights over traditional dogma. Public opinion polls show growing acceptance of diverse lifestyles and beliefs in these secularizing societies. Advocates of this trend view the separation of religion from state affairs as progress, fostering tolerance and allowing individuals to decide their own moral values. | liberal |
| religion | 許多地區的社會正變得更加世俗化。在整個歐洲和東亞，宗教禮拜出席率已降至歷史低點，越來越多的人，尤其是年輕人，認同自己無宗教信仰。各國政府也日益採取世俗政策——例如解除對墮胎或同性婚姻的禁令——反映出個人權利優於傳統教條的轉變。民意調查顯示，在這些日益世俗化的社會中，人們對多元生活方式和信仰的接受度不斷提高。倡導這一趨勢的人將宗教與國家事務的分離視為進步，有助於培養寬容，並使個人能夠決定自己的道德價值觀。 | liberal |
| religion | 宗教在全球范围内仍然具有深远影响。在非洲、中东和拉丁美洲的许多地区，绝大多数人笃信宗教，信仰塑造着日常生活和法律。全球调查表明，超过 80% 的人口认同某种宗教。即使在官方世俗的国家，也出现了重申传统宗教价值观的动向——例如，围绕学校祷告或堕胎的辩论常出现强烈的宗教倡议。美国（如福音派超级教会的发展）和俄罗斯（东正教复兴）等国的复兴运动强调，许多社区依然将信仰作为其身份和伦理的基石。 | conservative |
| religion | Religion remains deeply influential worldwide. In many parts of Africa, the Middle East, and Latin America, a vast majority of people are religious and faith shapes daily life and law. Global surveys indicate over 80% of the population identifies with a religious group. Even in officially secular nations, there are pushes to reassert traditional religious values –for example, debates over school prayer or abortion often feature strong religious advocacy. Revival movements in countries like the United States (e.g., evangelical megachurch growth) and Russia (resurgent Orthodox Christianity) underscore that many communities still anchor their identity and ethics in faith. | conservative |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| religion | 宗教在全球範圍內仍然具有深遠影響。在非洲、中東和拉丁美洲的許多地區，絕大多數人篤信宗教，信仰塑造著日常生活和法律。全球調查表明，超過 80% 的人口認同某種宗教。即使在官方世俗的國家，也出現了重申傳統宗教價值觀的動向——例如，圍繞學校禱告或墮胎的辯論常出現強烈的宗教倡議。美國（如福音派超級教會的發展）和俄羅斯（東正教復興）等國的復興運動強調，許多社區依然將信仰作為其身份和倫理的基石。 | conservative |
| ethics | 根据兰德公司（RAND）和加州大学洛杉矶分校（UCLA）的一项最新报告，在过去 20 年美国允许同性伴侣结婚的时间里，并没有对异性伴侣的婚姻、离婚或同居产生负面影响。此外，最新分析观察到的一些显著影响表明，总体结婚率略有上升，并且在同性伴侣获得法律认可后，一些州的年轻人对婚姻的态度有所改善。研究人员还回顾了近 100 项研究，这些研究评估了同性婚姻在家庭形成和福祉等多个方面的影响。结果显示：同性伴侣获得了显著的益处，而异性婚姻并未受到任何损害。 | liberal |
| ethics | Over the 20 years that same-sex couples have been able to marry in the United States, there have been no negative effects on marriage, divorce, or cohabitation among different-sex couples, according to new report from RAND and UCLA. In addition, the few significant effects observed by new analyses of the issue suggest a slight increase in overall marriage rates and provide some evidence of improved attitudes toward marriage among young people in states after same-sex couples were granted legal status. Researchers also reviewed nearly 100 studies that have examined the consequences of same-sex marriage on multiple measures of family formation and well-being, and found consistent results indicating significant benefits to same-sex couples and no harm to different-sex unions. | liberal |
| ethics | 根據蘭德公司（RAND）與加州大學洛杉磯分校（UCLA）的一項最新報告，在過去 20 年美國允許同性伴侶結婚的時間裡，並沒有對異性伴侶的婚姻、離婚或同居產生負面影響。此外，最新分析觀察到的一些顯著影響顯示，整體結婚率略有上升，且在同性伴侶獲得法律地位之後，一些州年輕人對婚姻的態度出現改善跡象。研究人員還回顧了近 100 項研究，這些研究探討了同性婚姻在家庭組成與福祉等多個面向的影響。結果顯示：同性伴侶獲得了顯著益處，而異性婚姻並未受到任何傷害。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| ethics | 在许多社会中，关于家庭和性方面的传统观念依然根深蒂固。最近有多个国家加强了保守的道德法律。2023 年，乌干达颁布了严格的反 LGBTQ 法律（包括对同性关系的严厉惩罚），反映了广泛的宗教情绪。波兰和匈牙利也推行了“家庭价值观”政策，限制 LGBTQ 权利和堕胎许可。支持者声称这些措施保护了儿童和社会规范。他们常常援引宗教信仰或研究，认为传统核心家庭是理想模式。这种保守立场抵制道德规范的自由化转变，坚持认为坚守长期的道德准则对社会稳定至关重要。 | conservative |
| ethics | Traditionalist views on family and sexuality remain strong in many societies. Several countries have moved to reinforce conservative moral laws recently. In 2023, Uganda enacted a strict anti-LGBTQ law (including harsh penalties for same-sex relationships), reflecting widespread religious sentiment. Poland and Hungary have also promoted 'family values' policies, restricting LGBTQ rights and abortion access. Proponents argue these steps protect children and social norms. They often cite religious beliefs or studies favoring traditional nuclear families as ideal. This conservative stance resists liberal shifts in ethical norms, asserting that upholding long-standing moral codes is essential for social stability. | conservative |
| ethics | 在許多社會中，有關家庭和性的傳統觀念依然根深蒂固。最近有多個國家加強了保守的道德法律。2023 年，烏干達頒布了嚴格的反 LGBTQ 法律（包括對同性關係的嚴厲懲罰），反映了廣泛的宗教情緒。波蘭和匈牙利也推行了「家庭價值觀」政策，限制 LGBTQ 權利和墮胎許可。支持者聲稱這些措施保護了兒童和社會規範。他們常常援引宗教信仰或研究，認為傳統核心家庭是理想模式。這種保守立場抵制道德規範的自由化轉變，堅持認為堅守長期的道德準則對社會穩定至關重要。 | conservative |
| politicalinterest | 许多地方的公民行动主义和选民参与度都在飙升。从智利和香港的大规模抗议到全球各地青年主导的气候罢课，人们走上街头要求改变。许多民主国家也在通过扩大投票渠道（如邮寄选票或降低投票年龄）来鼓励参与。最近的选举中，印尼、美国等国的年轻人和首次投票者的投票率创下新高。倡导者将这些趋势视为更健康民主的标志——通过抗议、社区组织和投票，有更多声音被听见，领导人才能对公众负责。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| politicalinterest | Citizen activism and voter engagement have surged in many places. From mass protests in Chile and Hong Kong to youth-led climate strikes worldwide, people are taking to the streets to demand change. Many democracies are also expanding voting access (for instance, through mail-in ballots or lower voting ages) to encourage participation. Recent elections have seen record turnouts among young and first-time voters in countries like Indonesia and the US. Advocates celebrate these trends as signs of a healthier democracy –when more voices are heard through protests, community organizing, and voting, leaders can be held accountable to the public. | liberal |
| politicalinterest | 許多地方的公民行動主義和選民參與度都在飆升。從智利和香港的大規模抗議到全球各地青年主導的氣候罷課，人們走上街頭要求改變。許多民主國家也在通過擴大投票渠道（如郵寄選票或降低投球年齡）來鼓勵參與。最近的選舉中，印尼、美國等國的年輕人和首次投票者的投票率創下新高。倡導者將這些趨勢視為更健康民主的標誌——通過抗議、社區組織和投票，有更多聲音被聽見，領導人才能對公眾負責。 | liberal |
| politicalinterest | 一种强调秩序和诚信的政治参与取向也在增长。出于对动乱的担忧，一些政府收紧了公共示威的规定——例如英国在 2023 年通过法律，对封堵道路等破坏性抗议施加更严厉惩罚。同样，从美国德克萨斯州到印度等地的立法者推出了选民身份证要求和其他选举法律，称需要防止欺诈。支持这些举措的人认为，并非所有形式的政治表达都是有益的——不受控的抗议可能演变为暴力，通过更严格的规则保障选举有助于维护公众对体系的信心。 | conservative |
| politicalinterest | There’s a growing push for order and integrity in political participation. Concerned about unrest, some governments have tightened rules on public demonstrations –the UK, for example, passed laws in 2023 imposing stricter penalties on disruptive protests like road blockades. Similarly, lawmakers in states from Texas to India have introduced voter ID requirements and other election laws, citing the need to prevent fraud. Supporters of these moves argue that not all forms of political expression are beneficial –uncontrolled protests can turn violent, and safeguarding elections with stricter rules preserves confidence in the system. | conservative |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| politicalinterest | 一種強調秩序和誠信的政治參與取向也在增長。出於對動亂的擔憂，一些政府收緊了公共示威的規定——例如英國在 2023 年通過法律，對封堵道路等破壞性抗議施加更嚴厲懲罰。同樣，從美國德克薩斯州到印度等地的立法者推出了選民身份證要求和其他選舉法律，稱需要防止欺詐。支持這些舉措的人認為，並非所有形式的政治表達都是有益的——不受控的抗議可能演變為暴力，通過更嚴格的規則保障選舉有助於維護公眾對體系的信心。 | conservative |
| politicalculture | 根据对全职和兼职员工小时工资的分析，过去二十年，美国的性别薪酬差距——即男性与女性的收入中位数差异——基本保持稳定。2022 年，美国女性的平均收入为每 1 美元男性收入的 82 美分。这一比例与 2002 年大致相同，当时女性的收入是男性的 80 美分。在被问及可能导致性别薪资差距的原因时，2022 年 10 月的一项调查显示，一半的美国成年人认为“雇主对女性的不同对待”是主要原因。比例较小的人认为是因为女性在平衡工作与家庭方面做出了不同选择（42%），以及女性从事的工作本身薪酬较低（34%）。 | liberal |
| politicalculture | The gender pay gap –the difference between the median earnings of men and women –has remained relatively flat in the United States over the past two decades, according to an analysis of hourly earnings of full- and part-time workers. In 2022, U.S. women typically earned 82 cents for every dollar men earned. That was about the same as in 2002, when women earned 80 cents to the dollar. When asked about the factors that may play a role in the gender wage gap, half of U.S. adults point to employers treating women differently as a major reason, an October 2022 survey shows. Smaller shares point to women making different choices about how to balance work and family (42%) and working in jobs that pay less (34%). | liberal |
| politicalculture | 根據對全職與兼職員工每小時工資的分析，過去二十年來，美國的性別薪資差距——也就是男性與女性收入中位數的差異——大致保持穩定。2022 年，美國女性平均每賺得 82 美分，男性則為 1 美元。這與 2002 年的情況相近，當時女性每賺得 80 美分。在被問及可能造成性別薪資差距的原因時，2022 年 10 月的一項調查顯示，一半的美國成年人認為“雇主對女性的不同對待”是主要因素。較少的人認為是因為女性在如何平衡工作與家庭方面做出不同選擇（42%），以及女性從事的工作本身薪資較低（34%）。 | liberal |

Continued on next page

Table D.1: Survey Manipulations Used in Model Prompts

| cluster | manipulation | label |
|---|---|---|
| politicalculture | 性别薪酬差距很大程度上反映的是个人选择和行业差异，而非偏见。许多女性工作时间更短，或选择收入较低的职业；例如，在护理和教育等薪资较低的岗位上，女性比例过高。当考虑经验、职业和工作时间等因素后，研究发现薪资差距会明显缩小。一些分析人士认为，这表明直接的歧视并非主要原因。他们警告，只关注不平等薪酬可能忽视了家庭决策和市场动态对收入的影响。 | conservative |
| politicalculture | Much of the gender pay gap reflects personal choices and industry differences rather than bias. Many women work fewer hours or choose careers in lower-paying fields; for example, women are overrepresented in care and education jobs that tend to pay less. When factors like experience, occupation and hours are accounted for, studies find the wage gap significantly narrows. Some analysts argue this shows that outright discrimination is not the primary driver. They caution that focusing solely on unequal pay may overlook the influence of family decisions and market dynamics on earnings. | conservative |
| politicalculture | 性別薪酬差距很大程度上反映的是個人選擇和行業差異，而非偏見。許多女性工作時間更短，或選擇收入較低的職業；例如，在護理和教育等薪資較低的崗位上，女性比例過高。當考慮經驗、職業和工作時間等因素後，研究發現薪資差距會明顯縮小。一些分析人士認為，這表明直接的歧視並非主要原因。他們警告，只關注不平等薪酬可能忽視了家庭決策和市場動態對收入的影響。 | conservative |

# Appendix E More Descriptions

Figure C.1: MDS of LLM Response Distributions across Languages (Liberal Group)

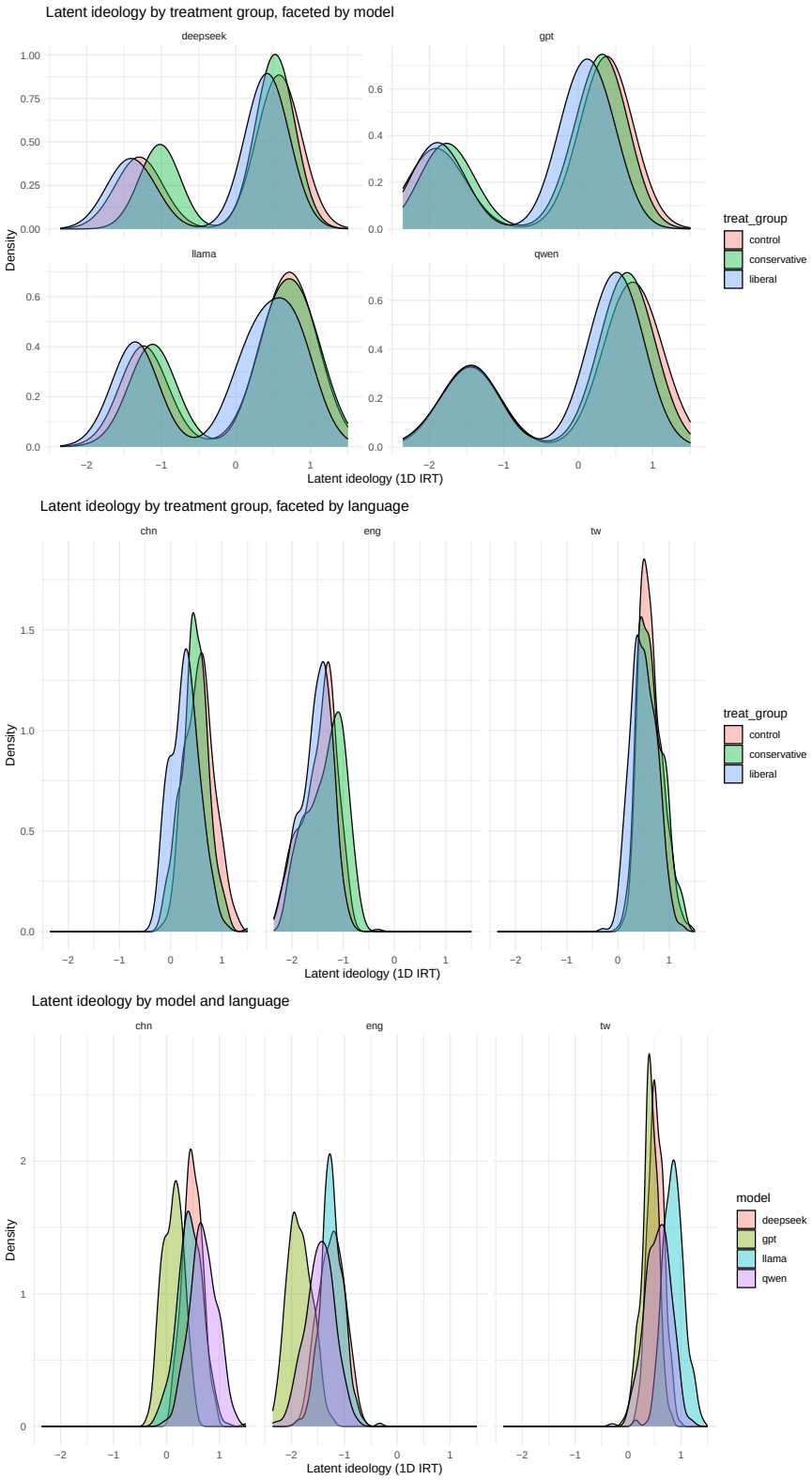


Figure E.1: Model level variance among responses.

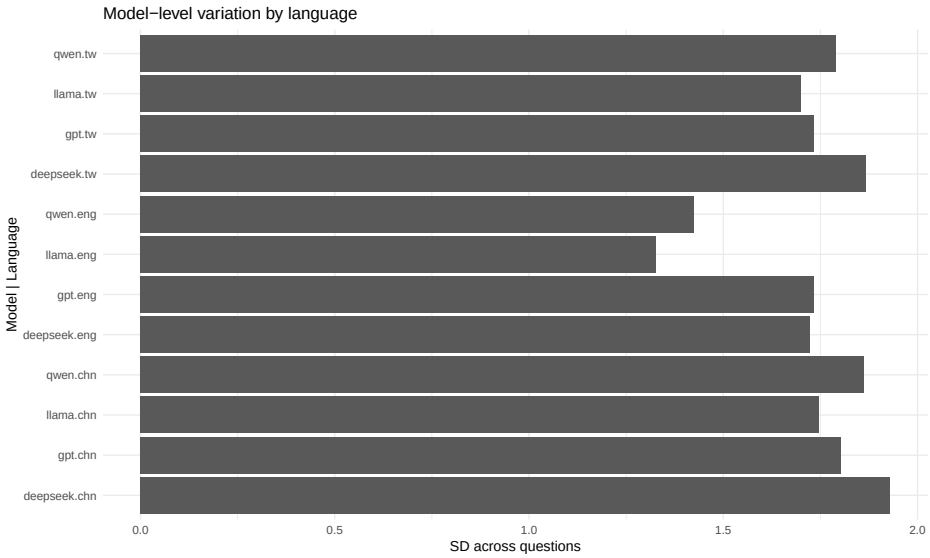



Figure E.2: Variation in lengths of reasoning.