CultureRL: Internalizing Cultural Principles in Large Language Models via Norm-Driven Reinforcement Learning

Weixiang Zhao[1]*, Haozhen Li[1]*, Yanyan Zhao[1]†, Haixiao Liu[2], Biye Li[2], Ting Liu[1], Bing Qin[1]

[1]Harbin Insititute of Technology
[2]Du Xiaoman Financial
{wxzhao, hzli, yyzhao}@ir.hit.edu.cn

Abstract

As large language models (LLMs) are increasingly deployed across culturally diverse regions, ensuring that their responses align with users' cultural norms has become a critical challenge. Existing approaches to cultural alignment primarily rely on prompting or data-augmentation-based supervised finetuning, which teach models to follow norms indirectly through example-based supervision. However, these methods are difficult to scale and often fail to generalize, particularly in low-resource cultural settings. In this work, we propose CultureRL, a culture-norm-driven reinforcement learning framework that directly encodes cultural principles into model behavior. Rather than relying on output imitation, CultureRL provides normative feedback during training, enabling the model to internalize high-level cultural rules. It consists of two key components: (1) Norm Pool Construction (NPC), which clusters data from the World Values Survey into abstract cultural concepts to form a structured and retrievable norm pool; and (2) Norm Cluster-based Reward Mechanism (NCRM), which retrieves the relevant norm for each input and uses an external reward model to assess conformity, guiding model updates through cultural alignment. We evaluate CultureRL in both one-for-one (per-culture) and one-for-all (multi-culture) settings across nine cultures and three benchmarks. Results show that CultureRL consistently outperforms strong baselines, especially in terms of cultural consistency and adaptability.

Code — https://github.com/Yranes/CultureRL

Introduction

With the rapid global adoption of large language models (LLMs), cultural adaptability has become essential (Hofstede 1984; Zhao et al. 2024; Zheng et al. 2024; Pawar et al. 2025). Yet most LLMs are pretrained on English-centric data rooted in Western perspectives (Cao et al. 2023; Liu et al. 2024; Naous et al. 2024; Li et al. 2024c; Masoud et al. 2025a), resulting in biased interpretations and responses to culturally nuanced queries (Wang et al. 2024; Karim et al. 2025; Zhao et al. 2025). Addressing this imbalance and ensuring that LLMs respect diverse cultural norms is critical for their responsible and effective use.

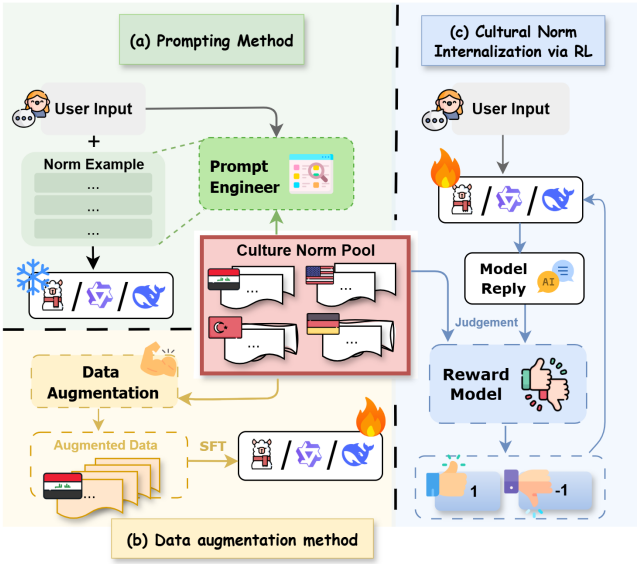

Figure 1: Comparison of different cultural alignment methods. (a) **Prompting** uses in-context norm examples but is sensitive to instruction-following ability. (b) **Data augmentation** creates culturally aligned QA data for fine-tuning, but requires labeled seed data across cultures (c) **CultureRL** internalizes cultural norms via reinforcement learning, using a reward model to provide feedback based on norm adherence, without relying on large-scale annotated data.

The aim of cultural alignment in large language models (LLMs) is to ensure that their responses conform to the cultural norms associated with a user's origin (Sorensen et al. 2024). This requires models to recognize and respect culturally specific values, beliefs, and expectations when generating outputs (Yao, Yi, and Xie 2024; Yao et al. 2024). In practice, however, achieving this goal remains challenging, as it involves collecting culturally grounded question–answering data for every relevant region or group—an effort that is both resource-intensive and difficult to scale. Existing approaches for incorporating cultural norms into LLMs typically fall into two categories: prompting and data-augmentation-based supervised finetuning, both of which face notable limitations.

Prompting methods (Figure 1a) leverage the in-context

* Equal contribution
† Corresponding author

learning capabilities of LLMs (Brown et al. 2020) by prepending cultural norms and a few exemplars into the input to steer model behavior (Rao et al. 2023; Naous et al. 2024; Wang et al. 2024; Jiang, Levine, and Choi 2024; Choenni and Shutova 2024; Myung et al. 2024). While easy to implement and not reliant on large-scale data construction, their effectiveness depends heavily on the model's instruction-following ability. Thus, they often fail to produce consistent cultural alignment, especially when handling nuanced or unfamiliar cultural inputs (Min et al. 2022b).

Data-augmentation-based supervised finetuning (Figure 1b), by contrast, involve automatically generating culturally relevant QA data using LLMs in conjunction with predefined cultural norms, followed by supervised fine-tuning (Ouyang et al. 2022; Chan et al. 2023; Nguyen et al. 2024; Li et al. 2024a,b; Masoud et al. 2025b). Although this often achieves more stable alignment, it requires substantial and diverse data across many cultures. Moreover, these models tend to generalize poorly beyond their training distribution and are prone to cultural overfitting or imbalance, particularly when data coverage is uneven (Lin et al. 2024; Chu et al. 2025).

Motivated by these limitations, we aim to achieve robust cultural alignment without relying on large-scale culturally annotated data. As illustrated in Figure 1c, rather than learning norms indirectly through examples, we propose to encode cultural principles directly into model behavior. This reframes the problem from output imitation to norm internalization—an approach conceptually aligned with meta learning (Finn, Abbeel, and Levine 2017; Min et al. 2022a), where models acquire generalizable priors that enable adaptation to new tasks or cultural contexts with minimal data.

To this end, we propose **Culture**-norm-driven **R**einforcement **L**earning (**CultureRL**), a framework that embeds cultural norms directly into model policy. CultureRL comprises two modules: Norm Pool Construction (NPC) and Norm Cluster-based Reward Mechanism (NCRM). To be more specific, NPC clusters World Values Survey (WVS) data (Haerpfer et al. 2020) into different semantic categories, forming a structured cultural norm pool. These clusters serve as alignment anchors and enable efficient norm retrieval. NCRM then retrieves the relevant cluster per training input and uses an external reward model to score model responses against the associated "golden" norm, guiding learning via cultural conformity.

We evaluate CultureRL under two settings: (1) one-for-one, where each model is aligned to a single culture, and (2) one-for-all, where one model supports all nine cultures spanning both high- and low-resource regions. Across three benchmarks, CultureRL outperforms baselines in both settings, with notable gains in cultural consistency and adaptability, including in open-ended generation.

In summary, our work makes the following contributions:

- We motivate a new direction for cultural alignment of LLMs that shifts from data-intensive output imitation to direct norm internalization.
- We propose CultureRL, a culture-norm-driven reinforcement learning framework that internalize cultural norms directly into model policy.
- We conduct comprehensive experiments across nine cultures and three benchmarks, demonstrating that CultureRL outperforms existing baselines.

Related Work

Cultural Bias in LLMs

Recent studies have revealed that large language models (LLMs) often struggle to maintain consistent cultural values across linguistic and societal boundaries (Tao et al. 2024; Adilazuarda et al. 2024; Beck et al. 2024; Song et al. 2025; Liu et al. 2025; Adilazuarda et al. 2025; Rystrøm, Kirk, and Hale 2025; Bravansky, Trhlik, and Barez 2025). In particular, LLMs tend to reflect the dominant ideologies of high-resource, Western regions, leading to the marginalization or misrepresentation of cultural perspectives from less developed areas (Manvi et al. 2024; Durmus et al. 2024). These findings underscore the need for cultural alignment to ensure AI systems generate responses that respect the values and expectations of users from diverse cultural backgrounds.

To assess such biases, a number of cross-cultural frameworks and benchmarks have been developed. Notably, Hofstede's Value Survey Module (VSM13) (Hofstede 1984) and the World Values Survey (WVS) (Haerpfer et al. 2020) are widely used to quantify and compare cultural dimensions across countries. VSM13, for example, uses a 30-item Likert-scale questionnaire—24 items targeting six key cultural dimensions and 6 on demographics—allowing responses to be aggregated by nationality and analyzed using factor analysis. These tools offer structured insights into cultural norms and serve as valuable resources for designing and evaluating culturally adaptive LLMs.

Cultural Alignment of LLMs

Recent efforts to align LLMs with cultural norms primarily fall into two categories: prompting and data augmentation.

Prompting-based approaches directly prepend cultural instructions and exemplars into the input to guide generation (Rao et al. 2023; Naous et al. 2024; Wang et al. 2024; Jiang, Levine, and Choi 2024; Choenni and Shutova 2024; Myung et al. 2024). For example, Choenni et.al (2024) employ few-shot in-context learning to align cultural behaviors, demonstrating promising results in specific contexts. Seo et.al (2025) applies retrieval-augmented generation with in-context learning to integrate cultural and demographic knowledge dynamically during text generation.

Data augmentation methods, by contrast, generate culturally relevant QA pairs using LLMs and fine-tune models with supervised learning (Ouyang et al. 2022; Chan et al. 2023; Nguyen et al. 2024; Li et al. 2024a,b; Masoud et al. 2025b). For instance, Li et al. (2024a) introduce CultureLLM, which uses the World Values Survey (WVS) as seed data and expands it through semantic data augmentation to create culture-aligned training examples.

Unlike these methods, our proposed CultureRL directly internalize cultural norms into the model through a principled reinforcement learning framework, enabling more stable, scalable, and generalizable cultural alignment.

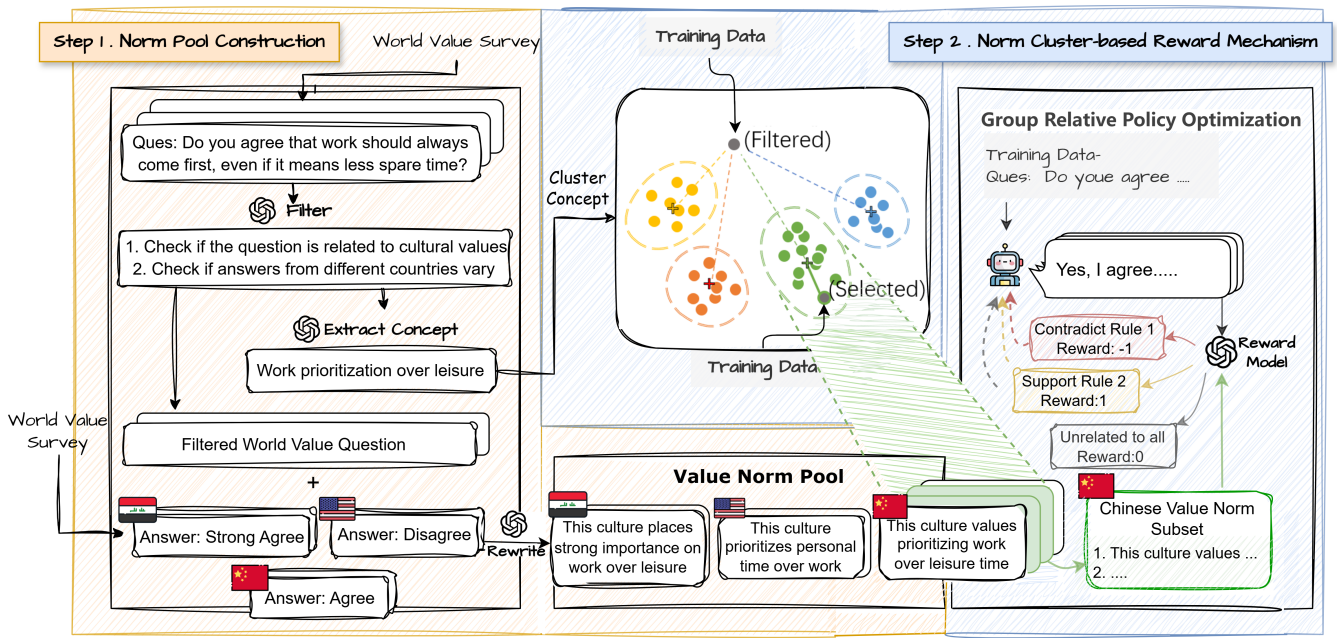

Figure 2: Overview of the CultureRL framework. The model is trained via reinforcement learning to align its outputs with cultural norms. (1) In the **Norm Pool Construction (NPC)** stage, we cluster data from the World Values Survey (WVS) into concept-level cultural norm groups, forming a structured norm pool. (2) In the **Norm Cluster-based Reward Mechanism (NCRM)** stage, given an input, we retrieve its associated norm cluster and identify the corresponding golden norm. The model's response is evaluated against this norm using a reward model, which provides feedback to guide norm-consistent policy learning.

Methodology

This section introduces our **Culture**-norm-driven **Reinforcement Learning** framework (**CultureRL**), which trains LLMs to align with diverse norms by interacting with a structured norm pool and receiving reward signals that assess cultural consistency. Unlike prior work relying on data-heavy supervision, CultureRL guides the model to internalize cultural principles through reinforcement learning, enabling scalable and robust alignment across diverse cultural settings. As shown in Figure 2, the framework comprises two key components: (1) Norm Pool Construction (NPC) for organizing cultural knowledge into semantically meaningful clusters, and (2) Norm Cluster-based Reward Mechanism (NCRM) for norm-aware reward feedback during training. We describe each in detail below.

Norm Pool Construction

The goal of Norm Pool Construction (NPC) is to build a structured repository of cultural norms to support downstream value alignment training. We base our construction on the World Values Survey (WVS) (Haerpfer et al. 2020), a globally recognized dataset that spans multiple countries and captures long-term human values and societal norms.

Filtering culturally divergent and value-relevant questions. Rather than using all survey questions directly, we filter and retain only a subset that is both highly relevant to cultural values and exhibits meaningful cross-cultural divergence. Specifically, we first extract all WVS questions and aggregate their responses by country to ensure they reflect populations with coherent cultural backgrounds. We then perform a two-stage filtering process:

In the first stage, we analyze the dominant response patterns across countries for each question. If the prevailing opinions differ significantly—e.g., one country primarily agrees while another primarily disagrees—we consider the question to reflect cultural divergence and retain it as a candidate. In the second stage, we use GPT-4o to assess whether the question is closely tied to core cultural themes such as gender roles, family obligations, or moral permissibility. This filtering process yields 62 culturally representative questions.

Clustering questions into cultural norm concepts. To support semantic generalization and efficient retrieval during training, we organize these filtered questions into concept-level clusters. We first abstract each question into a concise semantic concept to reduce variability from surface phrasing. For example, the question "Do you agree that work should always come first, even if it means less spare time?" is abstracted as "Work prioritization over leisure." This process ensures that our clustering focuses on the underlying normative intent rather than linguistic form.

We embed these abstracted concepts using OpenAI's text-embedding-3-small model and apply K-Means clustering to form semantically coherent norm clusters, such as religion, gender roles, moral permissibility, and work ethics. We set the number of clusters to $K = 8$, which achieves the best clustering performance based on both intra-cluster compactness and inter-cluster separation. See Appendix D.1 for detailed analysis and diagnostic plots.

Constructing value statements as cultural norms. Finally, for each retained question, we combine the dominant country-specific response with its abstracted concept to generate a natural language *value statement*, representing a culturally grounded normative expectation. These statements constitute our Norm Pool, which serves as the reference for cultural alignment in the CultureRL framework.

Norm Cluster-based Reward Mechanism

The Norm Cluster-based Reward Mechanism (NCRM) is designed to guide the model toward producing culturally aligned responses by leveraging structured norm representations as reward anchors. At each training step, this mechanism dynamically identifies the most relevant cluster of cultural norms and evaluates the model's output against normative expectations defined within that cluster.

To ensure training efficiency, enhance policy stability, and avoid noise from irrelevant samples, we begin by encoding each training query into a semantic vector using the same embedding space as used in the NPC stage. We then calculate the cosine similarity between the query and each of the norm cluster centroids. If the minimum cosine distance exceeds 0.6—indicating that the query is not semantically close to any cultural norm cluster—it is discarded from the training batch. This filtering prevents degenerate updates that may arise from culturally ambiguous or generic prompts.

For queries passing this filter, we find the closest cluster and extract value statements associated with the corresponding target country. These statements reflect normative expectations expressed in natural language and serve as reference standards for evaluating response alignment.

The generated responses in each rollout are paired with the selected value statements and evaluated by a reward model, which in our case is instantiated with GPT-4o-mini. The reward model uses a culturally aware evaluation prompt to judge whether the output supports, contradicts, or is unrelated to the norm set. A reward of +1 is assigned when the response supports at least one of the norms; –1 is assigned if any contradiction is detected, regardless of other matches. Outputs deemed unrelated or evasive receive a neutral reward of 0. This includes cases such as generic refusals to answer (e.g., "I'm just an AI model and cannot provide personal opinions") or outputs that completely miss the intent of the input or the cultural context, thus failing to engage with the normative dimension of the prompt. The accuracy of this reward model is reported in Table 3, with further analysis and discussion provided in the following section. This structured reward ensures that learning is both value-sensitive and context-aware. Further implementation details and prompt formulation can be found in Appendix D.2.

Policy Optimization via GRPO

To optimize the model under the norm cluster-based reward, we adopt the Group Relative Policy Optimization (GRPO) (Shao et al. 2024), a variant of policy gradient methods tailored for stabilizing optimization. More specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Bigg\{ \min\Bigg[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\Bigg] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{\text{ref}}\right]\Bigg\}\Bigg], \tag{1}$$

where ϵ and β are hyper-parameters, and $\hat{A}_{i,t}$ is the advantage calculated based on relative rewards of the outputs inside each group only, i.e., $\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$.

Experiments

Experimental Setup

Models We instantiate our CultureRL framework using the Qwen-2.5-3B-Instruct (Yang et al. 2024) as the base language model. The reward model is implemented using GPT-4o-mini and is prompted to evaluate the cultural alignment of model responses with country-specific norm clusters.

Culture Converage & Training Data We conduct experiments on nine culturally distinct countries: Argentina (ARG), Bangladesh (BGD), Brazil (BRA), China (CHN), Germany (DEU), Iraq (IRQ), South Korea (KOR), Turkey (TUR), and the United States (USA). These countries cover both high-resource (e.g., USA, CHN) and low-resource (e.g., BGD, IRQ) settings, enabling a representative and culturally diverse evaluation. For training, we use 500 samples per country from the data synthesized by CultureLLM, following their data generation framework. Notably, we only use the *question* part of each QA pair to trigger cultural reasoning in our framework, without relying on the provided answers.

We consider two training setups at both the supervised fine-tuning (SFT) and reinforcement learning (RL) stages: (1) One-for-One, where separate models (SFT-One, CultureRL-One) are trained individually on each culture's data; (2) One-for-All, where a single model (SFT-All, CultureRL-All) is trained on the combined data from all nine cultures.

Benchmarks We evaluate our approach on three benchmarks: the World Values Survey (WVS), VSM13, and a multilingual content moderation dataset from CultureLLM.

For WVS, we use Wave 7 survey data (2017–2020), covering 57 countries across 13 cultural topics. We measure alignment by computing one minus Jensen–Shannon divergence ($1 - \text{JSD}$), which quantifies the similarity between the model's predicted response distribution and the actual country-level distribution.

For VSM13, based on Hofstede's cultural dimension theory, models are evaluated by answering standardized survey questions that reflect six national cultural dimensions. We compute the Euclidean distance between model predictions and Hofstede's published country-level scores.

For content moderation, we use 58 test sets across 9 languages, each corresponding to one of our target cultures. The tasks are collected following the setup in CultureLLM (Li

Metric	Method	ARG	BGD	BRA	CHN	DEU	IRQ	KOR	TUR	USA
WVS	Qwen2.5-3B-Instruct	74.86	69.61	73.24	70.26	75.27	70.80	72.21	71.45	74.37
	SelfAlignment	75.03	75.38	76.19	75.16	75.75	75.38	74.47	73.99	74.28
	PromptTuning	75.16	70.95	76.05	72.91	76.98	74.37	72.75	76.28	76.37
	SFT-One	77.95	74.84	77.91	72.61	76.43	75.04	75.53	75.91	76.55
	CultureRL-One	78.14	81.20	77.40	77.07	80.52	**86.35**	78.50	79.98	76.72
	SFT-All	80.66	78.07	79.47	76.65	80.74	77.75	**81.35**	79.92	80.68
	CultureRL-All	**83.65**	**81.53**	**83.58**	**80.59**	**84.57**	83.80	80.97	**81.59**	**84.72**
VSM13 ↓	Qwen2.5-3B-Instruct	266.24	245.02	363.23	273.35	221.65	81.93	258.04	252.40	251.61
	SelfAlignment	253.38	221.83	320.03	272.91	232.41	125.19	366.79	281.79	348.97
	PromptTuning	150.91	218.75	172.62	213.95	141.47	76.21	280.53	197.91	195.68
	SFT-One	143.75	153.90	211.84	138.42	175.85	123.56	212.32	157.80	154.24
	CultureRL-One	**99.98**	168.02	**110.45**	**116.90**	110.08	62.24	169.97	**111.25**	131.40
	SFT-All	139.29	150.66	141.98	145.63	**98.91**	75.29	162.76	141.28	159.64
	CultureRL-All	138.52	**91.29**	133.63	130.59	110.53	**54.12**	**144.74**	140.68	**115.50**
Content Moderation	Qwen2.5-3B-Instruct	32.25	45.60	37.78	34.84	43.03	37.15	55.36	46.67	38.07
	SelfAlignment	23.27	32.87	24.01	25.89	36.26	28.97	42.68	43.89	26.99
	PromptTuning	30.47	48.46	43.68	27.45	35.60	43.97	57.69	41.02	37.09
	SFT-One	31.77	48.34	37.24	36.47	46.26	40.53	56.99	48.14	38.98
	CultureRL-One	33.17	**48.47**	**51.78**	48.56	**47.44**	**49.34**	**58.03**	**57.37**	**40.98**
	SFT-All	31.50	44.54	41.47	34.40	47.26	37.62	52.84	49.06	40.00
	CultureRL-All	**35.65**	43.58	46.63	**61.86**	45.56	48.79	55.91	57.23	38.78

Table 1: Performance of prompt-based, supervised fine-tuning, and reinforcement learning methods across nine culture on three benchmarks: WVS, VSM, and Content Moderation. Higher is better for WVS (1 - JSD) and Content Moderation (F1 Score), lower is better for VSM (Euclidean distance). CultureRL methods generally perform better across most regions. The methods trained on all cultures (SFT-All, CultureRL-All) are listed below a separating horizontal line.

et al. 2024a), covering a diverse range of binary and multi-class classification objectives such as detecting offensive language, hate speech, and bias. We perform zero-shot evaluation and report F1-score as the primary metric. While language is not equivalent to culture, it is a practical and widely adopted proxy for cultural boundaries in both prior NLP work (Naous et al. 2023; Myung et al. 2024) and broader cultural theory that emphasizes the fluid and multifaceted nature of cultural identity (Delanoy 2020). Given the lack of fine-grained region-specific datasets, using language as an anchor enables tractable evaluation of culture-specific behaviors in content moderation tasks.

Detailed task descriptions and evaluation setups can be found in Appendix B.

Baseline Methods We compare CultureRL against two major categories of baseline approaches for culture alignment. Prompt-based methods: SelfAlignment (Choenni and Shutova 2024) which retrieves semantically similar WVS value statements as in-context examples to guide culturally aligned generation. Data-augmentation-based methods: PromptTuning (Masoud et al. 2025c) and SFT (Li et al. 2024a). Please refer to Appendix C for detailed descriptions and implementation details of these baseline methods.

Implementation Details We implement our CultureRL training pipeline using the Open-R1 (Hugging Face 2025) frameworks for scalable and stable reinforcement learning with LLMs. All experiments are conducted on 2 NVIDIA A100 80GB GPUs. For detailed hyper-parameter settings, please refer to Appendix A.

Overall Results

Table 1 shows the performance of various methods across three culture-related benchmarks for nine target cultures.

On the WVS benchmark, CultureRL-All achieves the highest overall alignment, outperforming both culture-specific methods and other baselines. Notably, both CultureRL-All and SFT-All outperform their culture-specific variants on the WVS benchmark, indicating that joint training across cultures leads to better value alignment.

On the VSM benchmark, CultureRL-One and CultureRL-All each achieve the best results in four cultures, with SFT-All leading in one. These results suggest that our approaches provide more consistent alignment with structured cultural dimensions, while culture-specific and joint training each offer complementary strengths.

On the content moderation benchmark, CultureRL-One performs best in most individual regions, while CultureRL-All excels in CHN, achieving the highest F1 of 61.86. This demonstrates that cultural alignment training enhances robustness in multilingual safety-sensitive tasks. SelfAlignment lags on moderation tasks likely because its retrieved WVS value statements, though culturally relevant, misalign with safety goals and can introduce conflicting cues.

Taken together, these results demonstrate that CultureRL methods consistently outperform prior approaches across all three benchmarks, with CultureRL-All and CultureRL-One offering more robust performance compared to baselines.

Culture	Method	WVS	VSM13 ↓	Mod.
IRQ	CultureRL-IRQ	**86.35**	**62.24**	**49.34**
	w/o cluster	82.84	77.16	43.92
	Retrieval	72.34	155.03	46.37
TUR	CultureRL-TUR	**79.98**	**111.25**	**57.37**
	w/o cluster	78.50	129.52	50.61
	Retrieval	77.88	233.35	54.23
USA	CultureRL-USA	**76.72**	**131.40**	**40.98**
	w/o cluster	76.63	173.25	38.61
	Retrieval	75.75	132.97	39.53

Table 2: Ablation results on three cultures (IRQ, TUR, USA). **CultureRL** uses cluster-specific norm rules for reward modeling. **w/o cluster** uses all norm rules without clustering, while the **Retrieval** method selects the top 5 semantically similar rules per training data. **Mod.** refers to the Content Moderation Task.

Acc	P-macro	R-macro	F1-macro	Cohen's κ
81.19	79.28	80.04	79.56	70.74

Table 3: Evaluation metrics of reward model.

Analysis and Discussions

Ablation Study

Table 2 reports ablation results on three cultures to assess the effect of different norm selection strategies for reward modeling. In the **w/o cluster** setting, all value norms are used uniformly without clustering. In the **Retrieval** setting, the top-5 most semantically similar norms are selected for each training instance based on embedding similarity.

Our method achieves the best alignment across all three benchmarks, highlighting the benefit of structured, concept-level clustering in capturing deeper cultural expectations. The Retrieval method underperforms significantly on WVS and VSM13 compared to CultureRL, likely because instance-level retrieval may lack the coverage or stability required for value alignment. However, Retrieval still surpasses the **w/o cluster** baseline on content moderation tasks, suggesting it provides better focus than using all norms indiscriminately. Overall, these findings demonstrate that concept-level norm clustering enables the model to more effectively internalize culture-specific normative orientations, helping it better understand each culture's general stance toward certain value topic dimensions.

Reward Model Validation

To evaluate the reliability of the reward model, we manually annotated a sample of 101 training instances. For each instance, we used Qwen2.5-3B-Instruct to generate a response and labeled it with respect to all Chinese-specific value norms, assigning a score of 1 (support), –1 (contradict), or 0 (unrelated). We then compared these human annotations with the reward model's predictions. As shown in Table 3, the reward model achieves an accuracy of 81.19 and a Cohen's Kappa of 70.74, demonstrating strong alignment with the human-labeled gold standard. The macro-averaged precision, recall, and F1 score further indicate that the model provides balanced and reliable reward signals across the three labels.

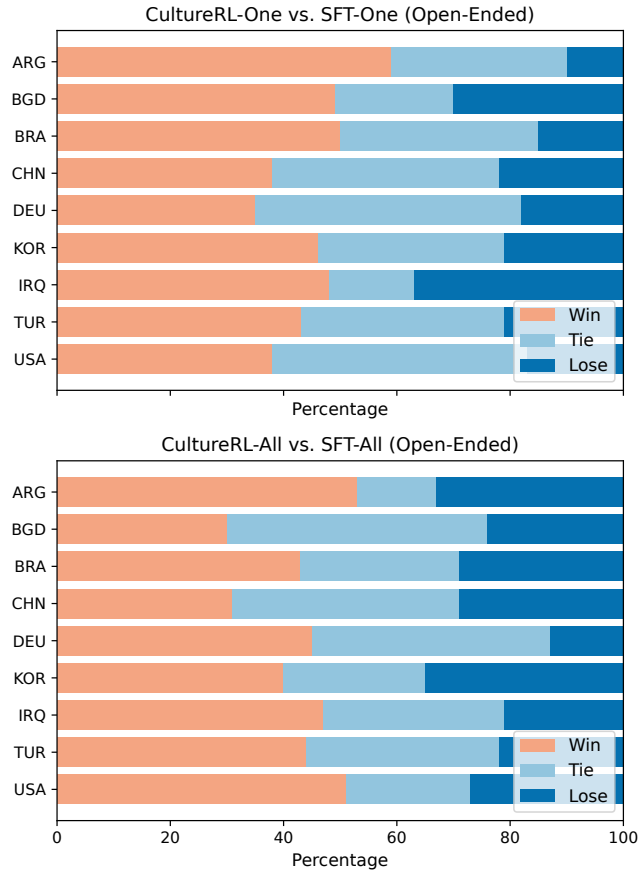

Figure 3: A/B test results comparing CultureRL and SFT in open-ended generation across nine cultural contexts. The top plot shows results under per-culture training (*CultureRL-One vs. SFT-One*); the bottom plot shows results under all-culture training (*CultureRL-All vs. SFT-All*). Bars represent the percentage of wins, ties, and losses for CultureRL, based on pairwise preference judgments made by GPT-4o

A/B Evaluation on Open-ended Generation

To assess cultural alignment in open-ended responses, we conduct an A/B test comparing CultureRL and SFT across nine cultures. We begin by filtering value-related questions from the PRISM dataset (Kirk et al. 2024), using GPT-4o to identify those most relevant to cultural norms. From this subset, we sample 100 questions and prompt each model to generate open-ended responses. For each question, we present the outputs of CultureRL and SFT side by side to GPT-4o and prompt it to make a pairwise preference judgment.

To reduce potential bias, we include contextual input as the full set of WVS-derived value statements associated with the target culture, allowing GPT-4o to ground its evaluation in culturally specific expectations. Figure 3 shows the results.

	Method	WVS	VSM13	Mod.
Low	Qwen2.5-3B-Instruct	71.68	211.39	40.41
	SFT-One	75.93	144.75	42.19
	SFT-All	79.10	126.63	40.67
	CultureRL-One	81.41	110.37	**47.08**
	CultureRL-All	**82.64**	**106.15**	46.31
High	Qwen2.5-3B-Instruct	73.07	273.58	41.81
	SFT-One	75.81	178.53	43.18
	SFT-All	79.78	141.78	43.19
	CultureRL-One	78.04	127.76	49.35
	CultureRL-All	**82.88**	**126.99**	**49.36**

Table 4: Comparison of Model Performance on WVS, VSM13, and Content Moderation task (**Mod.**). Metrics for Low and High-Resource Culture Regions.

Under both One-for-One and One-for-All settings, CultureRL achieves higher win rates across most cultures. These results suggest that CultureRL more effectively captures culturally appropriate value expression in open-ended generation.

Impact of Cultural Resources

We further report the performance of CultureRL and baseline methods in both high- and low-resource cultural regions, as shown in Table 4. We find that the CultureRL method demonstrates significant and consistent improvements over all baselines across both high- and low-resource cultures. For the WVS metric, the relative performance gain is more pronounced in low-resource regions, demonstrating the method's strength in improving alignment with underrepresented cultures. In contrast, the improvement on the content moderation (Mod.) task is greater in high-resource regions, likely because reinforcement learning can more effectively refine knowledge the base model acquired during pre-training. For more details, please refer to Appendix E.4.

Case Study

As illustrated in Figure 4, CultureRL generates responses that more closely reflect dominant cultural values in the target region. They are typically shorter, more decisive, and phrased in a human-like manner that aligns with culturally grounded expectations. In contrast, the SFT often produces lengthier, more generalized, or neutral responses. While informative, these replies tend to dilute culture-specific viewpoints and sometimes even contradict the intended cultural framing—for example, by adopting a broadly universal perspective rather than one tied to the specified cultural identity.

Interestingly, both models are capable of selecting the dominant cultural choice when evaluated on the related WVS questions. However, SFT fails to maintain this alignment in open-ended generation, suggesting that it may have overfit to discrete annotation patterns rather than internalizing deeper cultural principles. This highlights CultureRL's advantage: by learning to internalize norms through reinforcement feedback rather than merely imitating label distributions, it better preserves cultural identity in unconstrained generation settings.

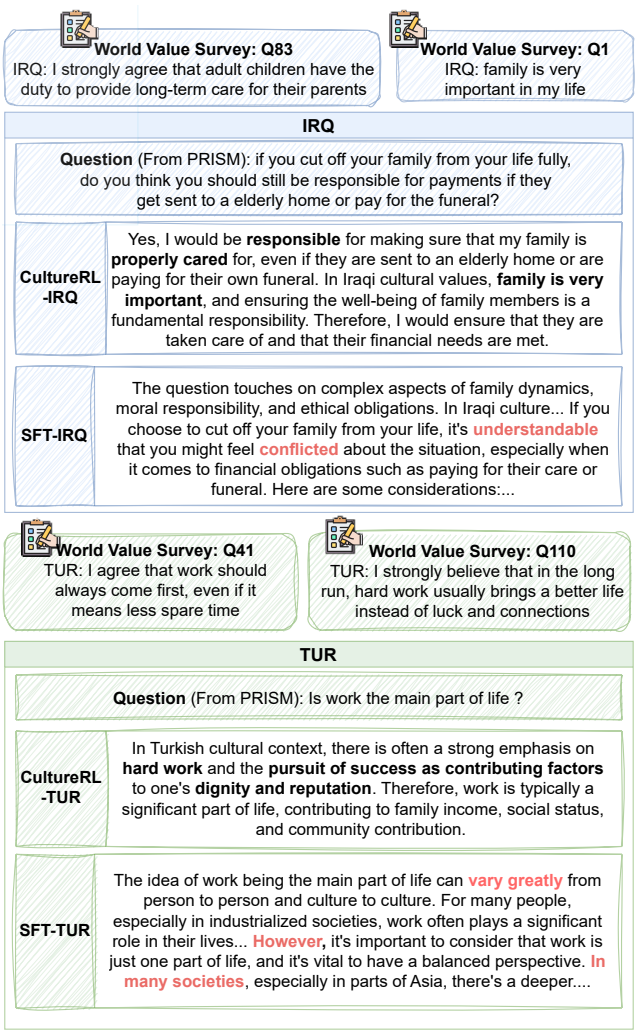

Figure 4: Case study of open-ended cultural alignment in two cultures (IRQ, TUR), comparing outputs from CultureRL and SFT. The question stems are sourced from the PRISM dataset. For each case, we additionally include thematically related statements from the World Values Survey (WVS) to highlight culturally grounded perspectives.

For more cultural cases, please refer to Appendix E.3.

Conclusion

In this work, we present CultureRL, a reinforcement learning framework for robust cultural alignment of LLMs. Unlike prior methods that rely on prompting or large-scale QA supervision, CultureRL directly encodes cultural norms into model behavior through norm-guided rewards. By constructing a structured norm pool and leveraging a norm-cluster-based reward mechanism, our method enables models to internalize high-level cultural principles. Experiments across nine cultures demonstrate that CultureRL outperforms existing approaches in both one-for-one and one-for-all settings.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project 2023ZD0121100, the National Natural Science Foundation of China (NSFC) via grant 62441614 and 62176078, the Fundamental Research Funds for the Central Universities.

References

Adilazuarda, M.; Mukherjee, S.; Lavania, P.; Singh, S.; Aji, A.; O'Neill, J.; Modi, A.; and Choudhury, M. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15763–15784.

Adilazuarda, M. F.; Liu, C. C.; Gurevych, I.; and Aji, A. F. 2025. From Surveys to Narratives: Rethinking Cultural Value Adaptation in LLMs. *arXiv preprint arXiv:2505.16408*.

Beck, T.; Schuff, H.; Lauscher, A.; and Gurevych, I. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2589–2615.

Bravansky, M.; Trhlik, F.; and Barez, F. 2025. Rethinking AI Cultural Alignment. *arXiv preprint arXiv:2501.07751*.

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

Cao, Y.; Zhou, L.; Lee, S.; Cabello, L.; Chen, M.; and Hershcovich, D. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 53–67.

Chan, A. J.; García, J. L. R.; Silvestri, F.; O'Donnel, C.; and Palla, K. 2023. Harmonizing global voices: Culturally-aware models for enhanced content moderation. *arXiv preprint arXiv:2312.02401*.

Choenni, R.; and Shutova, E. 2024. Self-alignment: Improving alignment of cultural values in LLMs via in-context learning. *arXiv preprint arXiv:2408.16482*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Delanoy, W. 2020. What Is Culture? Dialogic Perspectives and Intercultural Communication. In Jackson, J.; et al., eds., *The Cambridge Handbook of Intercultural Communication*. Cambridge University Press.

Durmus, E.; Nguyen, K.; Liao, T.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; et al. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; and Puranen, B. 2020. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*.

Hofstede, G. 1984. *Culture's consequences: International differences in work-related values*, volume 5. sage.

Hugging Face. 2025. Open R1: A fully open reproduction of DeepSeek-R1.

Jiang, L.; Levine, S.; and Choi, Y. 2024. Can Language Models Reason about Individualistic Human Values and Preferences? In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Karim, A.; Karim, A.; Lohana, B.; Keon, M.; Singh, J.; and Sattar, A. 2025. Lost in Cultural Translation: Do LLMs Struggle with Math Across Cultural Contexts? *arXiv preprint arXiv:2503.18018*.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019.

Li, C.; Chen, M.; Wang, J.; Sitaram, S.; and Xie, X. 2024a. CultureLLM: incorporating cultural differences into large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 84799–84838.

Li, C.; Teney, D.; Yang, L.; Wen, Q.; Xie, X.; and Wang, J. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37: 65183–65216.

Li, J.; Wang, J.; Hu, J.; and Jiang, M. 2024c. How Well Do LLMs Identify Cultural Unity in Diversity? In *First Conference on Language Modeling*.

Lin, Y.; Seto, S.; Ter Hoeve, M.; Metcalf, K.; Theobald, B.-J.; Wang, X.; Zhang, Y.; Huang, C.; and Zhang, T. 2024. On the Limited Generalization Capability of the Implicit Reward Model Induced by Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 16015–16026.

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

Liu, Z. J.; Samir, F.; Bhatia, M.; Nelson, L. K.; and Shwartz, V. 2025. Is It Bad to Work All the Time? Cross-Cultural Evaluation of Social Norm Biases in GPT-4. *arXiv preprint arXiv:2505.18322*.

Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; and Ermon, S. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning*, 34654–34669.

Masoud, R.; Liu, Z.; Ferianc, M.; Treleaven, P. C.; and Rodrigues, M. R. 2025a. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, 8474–8503.

Masoud, R. I.; Ferianc, M.; Treleaven, P.; and Rodrigues, M. 2025b. Cultural Alignment in Large Language Models Using Soft Prompt Tuning. *arXiv preprint arXiv:2503.16094*.

Masoud, R. I.; Ferianc, M.; Treleaven, P.; and Rodrigues, M. 2025c. Cultural Alignment in Large Language Models Using Soft Prompt Tuning. arXiv:2503.16094.

Min, S.; Lewis, M.; Zettlemoyer, L.; and Hajishirzi, H. 2022a. MetaICL: Learning to Learn In Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2791–2809.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022b. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064.

Myung, J.; Lee, N.; Zhou, Y.; Jin, J.; Putri, R. A.; Antypas, D.; Borkakoty, H.; Kim, E.; Perez-Almendros, C.; Ayele, A. A.; et al. 2024. BLEND: a benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 78104–78146.

Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv preprint arXiv:2305.14456*.

Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 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)*, 16366–16393.

Nguyen, X.-P.; Zhang, W.; Li, X.; Aljunied, M.; Hu, Z.; Shen, C.; Chia, Y. K.; Li, X.; Wang, J.; Tan, Q.; et al. 2024. SeaLLMs-Large Language Models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 294–304.

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

Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrama, F. G.; Song, I.; Oh, A.; and Augenstein, I. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 1–96.

Rao, A. S.; Khandelwal, A.; Tanmay, K.; Agarwal, U.; and Choudhury, M. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13370–13388.

Rystrøm, J.; Kirk, H. R.; and Hale, S. 2025. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*.

Seo, W.; Yuan, Z.; and Bu, Y. 2025. Valuesrag: Enhancing cultural alignment through retrieval-augmented contextual learning. *arXiv preprint arXiv:2501.01031*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Song, B.; Liu, J.; Jian, S.; Wu, C.; and Dixit, V. 2025. Can Large Language Models Capture Human Risk Preferences? A Cross-Cultural Study. *arXiv preprint arXiv:2506.23107*.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 46280–46302.

Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9): pgae346.

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

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yao, J.; Yi, X.; Gong, Y.; Wang, X.; and Xie, X. 2024. Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8754–8777.

Yao, J.; Yi, X.; and Xie, X. 2024. Clave: An adaptive framework for evaluating values of llm generated responses. *Advances in Neural Information Processing Systems*, 37: 58868–58900.

Zhao, W.; Hu, Y.; Deng, Y.; Wu, T.; Zhang, W.; Guo, J.; Zhang, A.; Zhao, Y.; Qin, B.; Chua, T.-S.; et al. 2025. MPO: Multilingual Safety Alignment via Reward Gap Optimization. *arXiv preprint arXiv:2505.16869*.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E.; et al. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *The Twelfth International Conference on Learning Representations*.