ALIGN: Word Association Learning for Cultural Alignment in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit cultural bias from over-represented viewpoints in training data, yet cultural alignment remains a challenge due to limited cultural knowledge and a lack of exploration into effective learning approaches. We introduce a cost-efficient and cognitively grounded method: fine-tuning LLMs on native speakers' word-association norms, leveraging cognitive psychology findings that such associations capture cultural knowledge. Using word association datasets from native speakers in the US (English) and China (Mandarin), we train Llama-3.1-8B and Qwen-2.5-7B via supervised fine-tuning and preference optimization. We evaluate models' cultural alignment through a two-tier evaluation framework that spans lexical associations and cultural value alignment using the World Values Survey. Results show significant improvements in lexical alignment (16–20% English, 43–165% Mandarin on Precision@5) and high-level cultural value shifts. On a subset of 50 questions where US and Chinese respondents diverge most, fine-tuned Qwen nearly doubles its response alignment with Chinese values (13 → 25). Remarkably, our trained 7–8B models match or exceed vanilla 70B baselines, demonstrating that a few million of culture-grounded associations achieve value alignment without expensive retraining. Our work highlights both the promise and the need for future research grounded in human cognition in improving cultural alignment in AI models.

1 Introduction

Every culture creates its own unique lens for understanding the world (Boroditsky, 2011). While we all share the same basic human brain, the way we use it—how we think, feel, and make sense of reality—is fundamentally shaped by our cultural environment (Park and Huang, 2010). Through years of immersive experience, culturally specific ways of thinking become internalized (Nisbett and Masuda, 2003). These deep mental frameworks automatically guide how we interpret concepts, perceive situations, and make decisions. At the same time, this long-term internalization makes cultural knowledge difficult to capture systematically. Much of this knowledge operates as common sense within a culture—deeply embedded and rarely articulated (Acharya et al., 2021). While some cultural information exists online, e.g., holidays and traditions, this represents only the visible surface (Hall, 1976). The deeper layers of cultural cognition, including unspoken assumptions, subtle social cues, and the implicit ways people naturally connect concepts, remain hidden within the minds of cultural insiders.

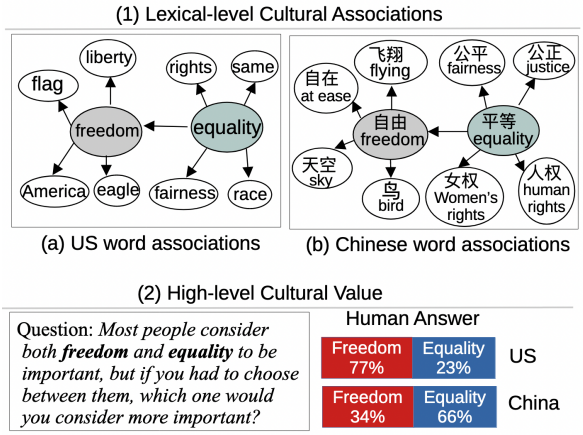


Figure 1: Example of how cultural word associations at the lexical level relate to higher-level cultural values. (1) Word associations show distinct cultural perception around the word of **freedom** and **equality**, with American associations emphasizing individual liberty and patriotic symbols, versus Chinese associations focusing on collective harmony and institutional frameworks. (2) These lexical differences correspond to opposing value preferences in responses to the survey question.

As large language models (LLMs) become embedded in global communication, they increasingly engage with users from diverse cultures. However, most LLMs are trained primarily on English-

language data, leading to an over-representation of Western perspectives and an under-representation of cultural-specific concepts (Cao et al., 2023; Naous et al., 2024). This bias not only limits their effectiveness in culturally grounded applications (Nguyen et al., 2024), but also risks ethical issues and inappropriate responses (e.g., suggesting drinking wine after Maghrib prayer (Naous et al., 2024)). Ensuring LLMs are culturally aware is crucial for fostering diversity and effective communication in today's AI ecosystem (Hershcovich et al., 2022). Full retraining, however, is prohibitive: frontier models consume hundreds of petaFLOPs-days and tens of millions of dollars (Hoffmann et al., 2022), exacerbating carbon costs and the global "AI compute divide" (Faiz et al., 2024). Parameter-efficient fine-tuning (LoRA, QLoRA) touches <1% of weights and substantially reduces compute demands, yet still needs culture-rich data (Hu et al., 2022; Dettmers et al., 2023). Moreover, prior work (Li et al., 2024b) shows that training one universal model for all cultures is challenging and often less effective than culturally tailored models, as cultural knowledge can conflict or intertwine. We adopt the same view that culture-specific models are central to cultural alignment.

Recent work has focused on evaluating cultural alignment using surveys (Durmus et al., 2024) and adapting models through prompting or synthetic data (Cao et al., 2024; Shi et al., 2024), but without lived-experience corpora, true cultural grounding remains elusive (Liu et al., 2025).

In response, we turn to native speakers' free word associations—a classic psycholinguistic lens on implicit cultural representations. When prompted with *red*, US respondents offer *danger*, *stop*, or *anger*, whereas Chinese respondents give *happiness*, *celebration*, or *luck*, illustrating how such spontaneous links reveal culture-specific representations. If such lexical links mirror deeper cultural values, aligning them should steer models toward cultural judgments. Figure 1 provides an example of such transfer.

We use two training approaches to fine-tune Llama-3.1-8B and Qwen-2.5-7B on two word association datasets, English (SWOW.US) and Mandarin (SWOW.ZH). Then we test (i) how well it regenerates human associations and (ii) World Values Survey alignment. Our findings reveal that (1) vanilla Llama initially leans more toward US associations and values than Qwen, whereas vanilla Qwen leans more toward Chinese associations and values than Llama; (2) association-tuned models produce markedly more human-like affective associations; and (3) this lexical gain translates into higher value alignment with the target culture, most notably when the original model lacked that knowledge. This work makes three key contributions:

1. We present the first head-to-head study of cultural fine-tuning, contrasting LoRA-based supervised fine-tuning with preference-optimized models on the English and Chinese SWOW associations, demonstrating their potential as valuable cultural resources.

2. We show how lexical-level association training shifts models toward target-culture value judgments using a two-tier evaluation.

3. We will release[1] the training pipeline and the top-performing models, to support plugging US- or CN-specific adapters into other LLMs and extending it to new cultures.

2 Related Work

2.1 Cultural Alignment in LLMs

Cultural Bias in LLMs LLMs inherit the skew of their training corpora; the English-heavy web thus pushes models toward Western-centric values (Naous et al., 2024; Adilazuarda et al., 2024). In the absence of broad, authentic datasets, researchers mine cultural proxy sources such as Wikipedia (Nguyen et al., 2023) and online communities (Shi et al., 2024), or ask LLMs to fabricate synthetic cultural data (Bhatia and Shwartz, 2023; West et al., 2022). Yet, as Liu et al. (2025) notes, lived-experience corpora remain scarce. We fill this gap by exploring large-scale native word-association norms as a direct, culturally grounded resource.

Cultural Alignment Evaluation Alignment is typically judged by comparing model outputs with human responses from specific cultures (Liu et al., 2025; Adilazuarda et al., 2024). These evaluations broadly fall into two categories: assessing cultural knowledge (e.g., food, customs) and evaluating high-level cultural values. Several recent benchmarks have been proposed for various cultural knowledge. However, these datasets are often either (a) domain-specific, e.g., FORK (Palta and Rudinger, 2023) only focuses on cutlery and food; (b) being verified/annotated by only a few (typically 2–5) native speakers, e.g., FORK was verified

[1] https://github.com/acl-anon-2025/cultural-lexis-anon

by two annotators, BLEnD (Myung et al., 2024) and CulturalBench (Chiu et al., 2025) were annotated by five annotators; or (c) the questions being asked are predominantly English-centric (e.g., CulturalBench and FORK only include English). On the value evaluation direction, researchers draw on cross-national surveys such as Hofstede's dimensions (Geert et al., 2020) and the World Values Survey (WVS) (Haerpfer et al., 2020). These surveys are usually conducted within a large scale of native speakers within one country, and the resulting response often reflects the population level distribution. Recent benchmarks build on WVS to evaluate LLMs across nations (e.g., GlobalOpinionQA (Durmus et al., 2024), WorldValueBench (Zhao et al., 2024) both provide English questions) capitalizing on its large sample sizes and 200-country coverage. We likewise adopt WVS for our value-alignment tests in Section 5 and extend it beyond English with Chinese, matching the native language of Chinese participants.

2.2 Word Associations and Their Value

In a word association task, participants provide the first (three) responses that come to mind for a cue, exposing the spontaneous links that structure semantic memory. Large normative datasets now exist: the University of South Florida norms (Nelson et al., 2004) and the crowd-sourced Small-World-of-Words (SWOW) corpus, whose English version spans 12K cues and 3M responses (De Deyne et al., 2019). Compared with distributional embeddings, human associations convey richer affective and multimodal information (De Deyne et al., 2021). Parallel SWOW collections in Dutch (De Deyne et al., 2013), Spanish (Cabana et al., 2024), Chinese (Li et al., 2024a) and other languages provide language-specific resources that reflect culture directly in speakers' lived experience.

Word Association and Culture Association norms already illuminate cultural contrasts: *food* evokes cuisine-specific terms across groups (Guerrero et al., 2010; Son et al., 2014), and *health* links to *wealth* in India but to *sick* in the United States (Garimella et al., 2017). Large SWOW corpora further identify culture-defining keywords in Spanish, Dutch, English and Chinese (Lim et al., 2024) and recover language-specific moral values (Ramezani and Xu, 2024). However, whether such lexical-level signals can also steer LLMs toward higher-level value alignment remains open. We tackle this gap by fine-tuning models on cultural associations and testing their transfer to value judgments. While drafting this paper, we noticed a concurrent work (Dai et al., 2025) that also uses word associations to steer language models via linear transformations. Unlike their primary focus on culturally aware association generation, our work explores different learning approaches to scale and transfer.

3 Framework Overview

We aim to investigate the extent to which models trained on association-level cultural knowledge can transfer to higher-level value alignment. To this end, we train language models on language-specific human word associations[2] using two training strategies and two model families. We then assess each model on two tiers: (i) association generation and (ii) value alignment via survey questions. This section introduces data and training, while the evaluations are presented in Sections 4 and 5.

Language and Culture Selection We focus on English for US and Mandarin for China (CN) because they provide a clear cultural contrast. These cultures differ in individualism vs. collectivism, emotional expression norms, and conceptual associations (as illustrated in Figure 1). Additionally, both languages have large-scale, high-quality native speaker word association datasets available, making this a practically significant test case for cultural transfer learning. While our study focuses on two cultures, the methodology can also be applied to others. Here, we focus on the mechanisms of training and evaluation framework.

Word-Association Datasets We train on the largest *Small-World-of-Words* corpora: English SWOW (SWOW.EN; De Deyne et al., 2019) and Mandarin SWOW (SWOW.ZH; Li et al., 2024a). SWOW.EN (2011–2019) provides 12K cues and 3.6M responses from 90K native speakers in the United States (≈50%), United Kingdom, Canada, and Australia. Each cue was answered by 100 participants with three free associations. For our US analyses we retain only respondents whose country *and* native language are United States, hereafter SWOW.US. SWOW.ZH (2016–2023) comprises 10K cues and 2M responses from 40K Mainland Chinese speakers. Both SWOW.US and

[2] As culture and language are closely intertwined, we approximate cultures by their primary spoken language (Delanoy, 2020). We treat language-specific word associations as culturally grounded signals, reflecting the conceptual organization shaped by speakers' cultural experiences.

SWOW.ZH are randomly split **by cue** into 80 % train, 10 % validation, and 10 % test (used in Section 4 as the test set).

Model Selection We choose two widely used model families as the subjects of our study to examine how language-specific word associations influence a model's cultural behavior given its initial representations. Specifically, we use Llama3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Qwen et al., 2025) as our baseline models and then fine-tune them on SWOW datasets.[3][4]

3.1 Training LLMs on Cultural Associations

To investigate how models acquire culturally grounded knowledge from word associations, we leverage two signals from human association data: (a) what associations people produce and (b) their relative production frequencies. We employ one learning approach for each signal, described below.

Supervised Fine-tuning (SFT) The first approach leverages the association lists themselves of a cue word, which capture how native speakers understand the cue. For example, for the cue word *country*, English associations include *nation*, *state*, *America*, and *farm*. For its Chinese equivalent 国家, associations include 中国 (China), 人民 (people), 国旗 (flag), and 富强 (wealthy and powerful). We implement *the word association generation task* in the SFT framework, training models to generate associations that are more aligned with human associations.[5] Given a training example $x = \langle c, \mathbf{w} \rangle$, where c is a cue word and $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$ is a list of associated words, the model is trained to generate $\mathbf{w}$ conditioned on the cue word c. The objective of SFT is to maximize the likelihood of the training data.[6]

Proximal Policy Optimization (PPO) Training The second approach leverages the observation that some associations are produced more frequently than others in human word association data (e.g., *nation* is more frequent than *farm*). We use reinforcement learning with PPO (Schulman et al., 2017) to train models to rank associated words according to their frequency, framing the task as a ranking problem. Given a cue word c and its randomized associated words $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$, the model predicts a list ranking $\mathbf{r}' = \langle r'_1, r'_2, \ldots, r'_n \rangle$ to match the ground-truth ranking $\mathbf{r} = \langle r_1, r_2, \ldots, r_n \rangle$, where r_i indicates word w_i's empirical rank based on human association frequency in SWOW. To reflect these human preferences, we use the Spearman rank correlation between $\mathbf{r}'$ and $\mathbf{r}$ to determine the reward. This reward signal guides policy updates via PPO, encouraging the model to produce association rankings that better align with human preferences.[7]

4 Association-level Evaluation

We test whether fine-tuning taught the models human-like word associations. To this end, we conduct two complementary evaluations: (a) **intrinsic** evaluation that measures the model performance on the task they are trained on, i.e., word association generation task for SFT models and ranking task for PPO models; and (b) **extrinsic** evaluation that focuses on the psychological attributes of generated word associations and measures to what extent they align with native speakers' associations regarding the emotional intensity (valence/arousal) and concreteness of meaning.[8][9] For the **evaluation set** in this section, we use **10% of held-out testing cues and their associations**, i.e., 10% of English data from SWOW.US and 10% of Chinese data from SWOW.ZH (see Section 3 for details).

4.1 Intrinsic Evaluation

For each cue in the test set, we use the same prompt as the training stage to elicit the model associations. We use Precision@K (overlap with human top-K) on the word association generation task and Spearman ρ against human frequency ranks on

[3]Due to computational resource constraints, we limited our study to training models of 7/8B parameters.

[4]While our evaluation includes US and Chinese cultural datasets, we do not assume Llama or Qwen to be clean proxies for any national culture due to the multilingual and multicultural composition of their training data. Instead, we quantify each model's initial alignment (Section 4–5) and investigate the relative shifts after fine-tuning on SWOW datasets.

[5]We provide more training details in Appendix B.

[6]See Appendix E for SFT hyperparameter setting details.

[7]Initially, we conducted preliminary experiments with multiple task formats to determine the most effective design for PPO training. See details in Appendix A.3 and E.

[8]This psychological attributes evaluation is inspired by a recent study (Xiang et al., 2025) where they found that the word associations generated by the vanilla model (Llama3.1-8B-Instruct) tend to be less emotional and concrete than humans in English word associations, revealing a gap. We extend the analysis to US and Chinese fine-tuned models.

[9]Prior study in cross-cultural study also shows that different cultural emotional connotations can be reflected in word associations (Tham et al., 2020), e.g., "green" → "envy" in US (from phrase "green with envy" means jealous) and "green" → "hat" in China (from the saying "wearing a green hat", symbolizing unfaithfulness).

Test	M Type	M Class	Train$_{swow}$	P@5	P@10	P@40
SWOW.US	Vanilla	Llama	-	0.754	0.609	0.295
	SFT	Llama	US	**0.875**	**0.773**	**0.437**
	Vanilla	Qwen	-	0.633	0.502	0.238
	SFT	Qwen	US	0.761	0.651	0.327
SWOW.ZH	Vanilla	Llama	-	0.260	0.181	0.057
	SFT	Llama	ZH	**0.689**	0.556	0.277
	Vanilla	Qwen	-	0.481	0.364	0.159
	SFT	Qwen	ZH	**0.689**	**0.559**	**0.279**

Table 1: Word Association Generation Results.

Test	M Type	M Class	Train$_{swow}$	Spearman ρ
SWOW.US	Vanilla	Llama	-	0.241
	PPO	Llama	US	0.270
	Vanilla	Qwen	-	0.292
	PPO	Qwen	US	**0.321**
SWOW.ZH	Vanilla	Llama	-	0.211
	PPO	Llama	ZH	0.226
	Vanilla	Qwen	-	0.291
	PPO	Qwen	ZH	**0.323**

Table 2: Word Association Ranking Results.

Test	Metric	Human	Llama$_{van}$	Llama$_{ppo}$	Llama$_{sft}$	Qwen$_{van}$	Qwen$_{ppo}$	Qwen$_{sft}$
SWOW.US	Valence	5.514	5.398	5.403	**5.543***	5.337	5.352	**5.484***
	Arousal	4.244	**4.272***	**4.238***	4.214	4.192	4.183	4.192
	Concreteness	3.644	3.378	3.355	3.582	3.368	3.349	3.535
	Emotional %	84.6%	78.2%	77.5%	75.5%	73.5%	73.5%	74.9%
	%Conc | %Abs | %Unk	64.3/29.8/5.9	52.8/37.9/9.3	51.1/38.7/10.2	56.8/29.0/14.2	50.5/37.0/12.5	50.4/37.2/12.4	56.7/29.6/13.7
SWOW.ZH	Valence	5.386	5.341	5.311	**5.427***	5.352	5.332	**5.411***
	Arousal	5.378	5.258	5.270	**5.408***	5.233	5.220	**5.370***
	Concreteness	3.657	3.370	3.394	3.576	3.391	3.412	3.516
	Emotional %	53.3%	31.8%	33.8%	41.9%	42.3%	41.6%	47.9%
	%Conc | %Abs | %Unk	35.9/15.8/48.3	17.9/12.7/69.4	19.3/13.2/67.5	27.6/13.2/59.2	24.1/16.6/59.3	24.2/15.8/60.0	30.4/15.9/53.8

Table 3: Emotion and concreteness scores on SWOW.US (top) and SWOW.ZH (bottom). *Bold indicates no significant difference from human medians ($p \geq 0.05$).

the ranking task by following prior work on word association evaluations (Yao et al., 2022).

Table 1 shows the results on word association generation. Overall, all models achieve higher performance in English than in Chinese, reflecting their stronger English capability. On the Chinese test set, vanilla Qwen outperforms Llama, showing that Qwen has stronger Chinese capability. Models trained on SWOW.US and SWOW.ZH achieved substantial gains, with SFT models improving P@5 by 16–20% in English and 43–165% in Chinese. Table 2 shows the results of the ranking task on PPO models, which exhibit similar trends of improvement but to a lesser degree.

4.2 Extrinsic Evaluation

We examine three psychological attributes of associations: valence (pleasantness), arousal (emotional intensity), and concreteness (tangibility). Our **approach** is as follows: (1) for a cue c, we obtain its top-10 model-generated associations; (2) for each association, we look up its emotion and concreteness scores from existing norms (EN: Warriner et al. (2013) for emotion, Brysbaert et al. (2014) for concreteness; ZH: Xu et al. (2022) for emotion, Xu and Li (2020) for concreteness); (3) we compute the median emotion/concreteness score of model-generated and human associations for c; and (4) we compare these medians across all cues.[10]

Table 3 presents the experimental results.[11] Overall, associations generated by SFT models exhibit a similar degree of valence (pleasantness) and arousal (emotional intensity) as human associations (e.g., when prompted with *Halloween*, both humans and fine-tuned models evoke pleasant associations such as *party* and *holiday*). Yet, a persistent gap remains in concreteness, with all model associations being more abstract than human associations (lower concreteness scores) in both languages. While SFT training increases concreteness by +0.20–0.21 from vanilla models, they remain below human medians. For example, for the cue word *emotions*, human associations include both abstract words such as *feelings* (1.68 concreteness score) and *sadness* (1.82), as well as more concrete ones like *tears* (4.56). In contrast, model associations are dominated by abstract concepts, such as *feelings* (1.68), *empathy* (1.63) and *love* (2.07).[12] These analyses highlight the advances achieved and the remaining challenges in aligning cultural conceptual representations with native speakers.

5 Cultural Value Alignment Evaluation

Fine-tuning on language-specific word associations embeds lexical cultural patterns, but *does this*

[10]We use the Wilcoxon signed-rank tests (Wilcoxon, 1992) to examine if the two sets of median scores are significantly different or not. See detailed methodology in Appendix F.

[11]Violin plots in Appendix F.2 are provided to show a finer-grained view of the distributions of the three attributes.

[12]More concrete examples on Valence, Arousal and Concreteness are provided in Table 9 in Appendix F.3.

knowledge support higher-order reasoning about cultural values and beliefs? Next, we evaluate this transfer using the World Values Survey (WVS). Successful transfer of association-driven cues to value-based scenarios would demonstrate deeper cultural understanding; failure would imply the need for explicit training on higher-level cultural reasoning tasks. We first measure how well models align with target-culture responses, then analyze prediction shifts on a curated "tension-set" of questions to probe fine-grained cultural differences.

5.1 Experimental Setup

Dataset We evaluate cultural value alignment using the WVS (Haerpfer et al., 2022), which contains 290 questions systematically designed to cover twelve cultural topics and have surveyed native speakers of each country (wave7: 2,596 in US and 3,036 in China). The WVS provides two critical advantages that are not available in other datasets (Palta and Rudinger, 2023; Myung et al., 2024; Cabana et al., 2024): (1) *population-level* value distributions enabling reliable ground-truth for cultural value estimation, and (2) *parallel questions in native languages* that reveal cultural differences in responses to the same set of questions, allowing us measure how the models align with different cultural values. We focus on the two cultures in our training data: the United States and China. From the 290 original questions, we removed demographic items (Q260–290) and retained only those asked in both countries, yielding 221 questions for evaluation. During evaluation, we use the language that is aligned with the target culture to prompt the language models (Chinese for both the WVS questionnaire and the models trained on SWOW.ZH; English for US).[13]

Evaluation We use vllm with constrained sampling to generate answers. For a given question, we constrain the output tokens to be the symbols of the options (e.g., 1,2,3,4) and constrain the output token number to be 1. Then we take the token logprob across the specified options and renormalize them to get the distribution of the answer options (Robinson and Wingate, 2023). We measure the alignment using the distance between human answer distribution $P = (P_1, P_2, \ldots, P_n)$ and the model predicted probability distribution $Q = (Q_1, Q_2, \ldots, Q_n)$.[14] We use two distance metrics that are used separately in prior work (Durmus et al., 2024; Zhao et al., 2024): (a) **Jensen-Shannon distance (JSD)**, which measures the distributions differences using category-wise probability divergence without considering any relationship between categories;[15] and (b) **Earth Mover's distance (EMD)**, which measures the accumulated cost of moving probability mass along the ordered scale.[16] Both $0 \leq \mathrm{JSD}(P, Q) \leq 1$ and $0 \leq \mathrm{EMD}(P, Q) \leq 1$, closer to 0 means better alignment. JSD is more suitable for categorical distributions, whereas EMD is more appropriate for ordinal distributions as it considers the accumulated distance between options. As WVS answers include both types, we use both metrics. We evaluate performance on the test set at two levels: (a) aggregate overall scores computed over the entire set, and (b) a finer breakdown measuring the percentage of questions with distances below different thresholds of JSD/EMD.

Test	M Type	M Class	$\text{Train}_{\text{swow}}$	JSD↓	EMD↓
WVS.US	Vanilla	Llama	-	0.324	0.102
	SFT	Llama	US	0.392	0.114
	PPO	Llama	US	**0.288***	0.092
	Vanilla	Qwen	-	0.388	0.131
	SFT	Qwen	US	0.355*	0.118
	PPO	Qwen	US	0.353*	0.125
	Vanilla	Llama3.1_70b	-	0.294*	0.094
	Vanilla	Qwen2.5_72b	-	0.262*	0.109*
WVS.CN	Vanilla	Llama	-	0.459	0.152
	SFT	Llama	ZH	0.421*	0.129*
	PPO	Llama	ZH	0.334*	0.112*
	Vanilla	Qwen	-	0.415	0.139
	SFT	Qwen	ZH	**0.325***	**0.100***
	PPO	Qwen	ZH	0.374*	0.123*
	Vanilla	Llama3.1_70b	-	0.333*	0.100*
	Vanilla	Qwen2.5_72b	-	0.328*	0.116*

Table 4: World Values Survey results on US and CN. ↓ indicates that lower is better (higher alignment). * indicates the improvement over Vanilla is significant ($p < 0.05$).Prompting language matches the survey language (EN for US, ZH for CN).

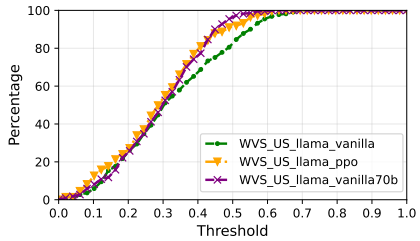
Approaches We use Vanilla models as our **baseline** to understand their *initial alignment* towards a specific culture. We apply the same prompts

[13] We collected the English and Chinese WVS questionnaire from the official website. We also adopted the prompts that the WVS was presented to the participants.

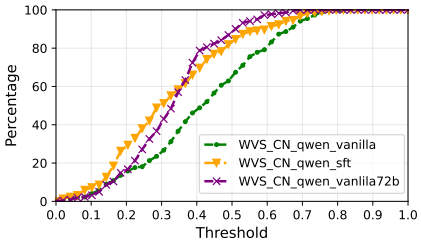
[14] P and Q are two discrete probability distributions on the same n-point ordered support ($\sum_{i=1}^{n} P_i = \sum_{i=1}^{n} Q_i = 1$).

[15] $\mathrm{JSD}(P, Q) = \sqrt{\frac{D(P \| M) + D(Q \| M)}{2}}$, where $M = \frac{1}{2}(P + Q)$. The $D(. \| .)$ is the KL-Divergence of two distributions, e.g., $D_{\mathrm{KL}}(P \| M) = \sum_{i=1}^{n} P_i \log \frac{P_i}{M_i}$.

[16] $EMD(P, Q) = \frac{1}{n} \sum_{i=1}^{n} |\delta_i|$, where $\delta_0 = 0$ and $\delta_{i+1} = \delta_i + P_i - Q_i, \quad i = 0, \ldots, n-1$.

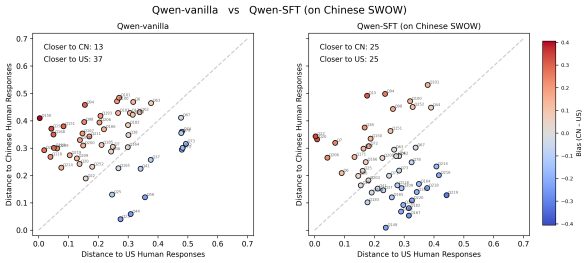


(a) WVS.US performance under JSD.

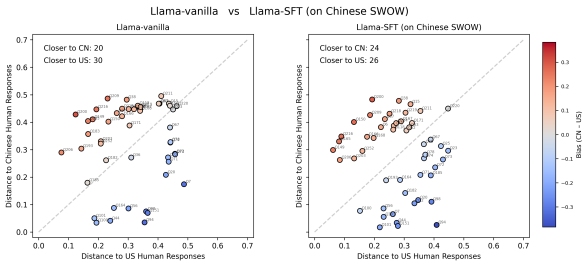


(b) WVS.CN performance under JSD.

Figure 2: Breakdown comparison of model alignment with cultural values across United States (top) and China (bottom) based on the World Values Survey. Results are shown for the Vanilla and trained (SFT and PPO) versions of Qwen2.5 and Llama 3.1. The x-axis is the threshold for what counts as a "good" match, and the y-axis shows the percentage of questions where the model's answer was within that threshold.



(a) Qwen2.5-7B



(b) Llama3.1-8B

Figure 3: Comparison of shifts after SFT for Qwen and Llama on SWOW.ZH (ZH prompts). Each dot = one WVS question; blue (red) indicates that the question answer is more towards Chinese (English). Table 5 presents concrete examples that illustrate the shifts.

as the Vanilla models to our fine-tuned models to measure the extent of cultural value transfer. We also include two 70B-scale models for zero-shot prompting, which allows us to contextualize our results more broadly and estimate the potential upper bound that word associations can provide.

5.2 Overall Results

Table 4 presents our results on WVS. Vanilla models exhibit different *initial* degrees of cultural alignment with the target populations. In the **US setting (English)**, the Llama model shows better alignment with the ground-truth human responses compared to Qwen. While in the **CN setting (Chinese)**, the alignment trend reverses: the Qwen model outperforms Llama. These findings align with our results in Section 4 and prior work that Llama models tend to be less eastern-value centric and less capable in understanding Chinese (Xiang et al., 2025; Aksoy, 2025), and Qwen has stronger Chinese capability (Qwen et al., 2025).

Models trained on the **SWOW.ZH** exhibit substantial improvements over their Vanilla models. Notably, **Qwen SFT model achieves the best performance on WVS.CN across the board.** Moreover, fine-tuned Llama aligns more closely with Chinese values, surpassing vanilla Qwen. This shows that SWOW.ZH provides strong cultural grounding and that training on it effectively steers the model toward high-level Chinese values.

Training on SWOW.US also brings significant improvements on WVS.US (except for SFT Llama). **The best-performing model is PPO Llama, which even achieves comparable or better results than the 70B models.** We also observe smaller overall gains on the US set than on the CN set, suggesting that SWOW.US might provide a weaker cultural signal than SWOW.ZH. We hypothesize this may stem from models being highly exposed to English during pre-training compared to Chinese, or from the greater cultural diversity within the US increasing alignment difficulty.

Surprisingly, our best-performing trained 7/8B models even surpass some of the 70B models. For WVS.US, the 8B PPO-tuned Llama outperforms the Vanilla Qwen2.5 72B, while in WVS.CN, the 7B SFT-tuned Qwen outperforms the Vanilla Llama3.1 70B. Figure 2 further illustrates how well different models align with human responses, evaluated under varying thresholds of JSD.[17] For both US and ZH settings, we include the best-performing fine-tuned model, its vanilla counterpart, and a larger model version. On WVS.US, the PPO-tuned model outperforms the vanilla model

[17] A similar trend on EMD is shown in Appendix G.1.

Id	WVS question (full wording + choice labels)	US	CN	$Qwen_{van}$	$Qwen_{sft}$	$Llama_{van}$	$Llama_{sft}$
Q149	Most people consider both freedom and equality important, but if you had to choose between them, which would you consider more important? *{1: Freedom; 2: Equality}*	[77%,23%]	[34%,66%]	[83%,17%]	[33%,67%]	[93%,7%]	[83%,17%]
Q168	In which of the following do you believe, if you believe in any? – Heaven *{1: Yes; 2: No}*	[65%,35%]	[12%,88%]	[71%,29%]	[18%,82%]	[97%,3%]	[85%,15%]
Q165	In which of the following do you believe, if you believe in any? – God *{1: Yes; 2: No}*	[79%,21%]	[17%,83%]	[41%,59%]	[29%,71%]	[94%,6%]	[87%,13%]
Q118	How often do ordinary people in your neighborhood have to pay a bribe, give a gift, or do a favor to local officials/service-providers to get needed services? *{1: Never; 2: Rarely; 3: Frequently; 4: Always}*	[28%,55%,15%,2%]	[4%,34%,36%,26%]	[33%,55%,10%,2%]	[5%,19%,67%,10%]	[93%,4%,2%,1%]	[77%,9%,8%,6%]
Q166	In which of the following do you believe, if you believe in any? – Life after death *{1: Yes; 2: No}*	[69%,31%]	[12%,88%]	[90%,10%]	[36%,64%]	[95%,5%]	[87%,13%]

Table 5: WVS questions where SFT on Chinese SWOW shifts Qwen's distribution toward Chinese responses. Shaded cells highlight the fine-tuned model's probabilities.

and even slightly surpasses the 70B model. On WVS.CN, the SFT model largely improves over the vanilla model across thresholds. For example, at a JSD threshold of 0.3, the SFT model achieves ≈50% of the questions aligned, outperforming both the vanilla model (20%) and the 72B model (40%). These promising results highlight the potential of culturally grounded fine-tuning as a lightweight yet effective alternative to scaling up.

5.3 Cross-Cultural Shifts

Beyond assessing a model's answers to a single culture, we track how the responses shift across cultures before and after fine-tuning on word-association data. Each model has its initial cultural leanings (e.g., Llama vanilla models align more closely with US values than Qwen, while Qwen aligns more with CN), so fine-tuning reveals both a model's adaptation to a target culture and its shift from its initial bias. We evaluate a model's answers with respect to both US and China. To capture the shifts, we focus on WVS questions where Chinese and US participants' responses diverge strongly. We ranked divergence by the average of JSD and EMD, selecting the top 50 divergent questions.[18] These "high-tension" questions provide greater sensitivity for detecting cultural shifts, allowing small changes in the model's alignment to become observable, whereas questions answered similarly by both populations offer little diagnostic value.

Results Figures 3a (Qwen2.5-7B) and 3b (Llama3.1-8B) present the models' prediction shifts before and after training in on SWOW.ZH.[19] For each of the 50 questions, we compare the model's response distance to US answers (x-axis) against its distance to Chinese answers (y-axis). For **Qwen2.5-7B**, we find that Chinese-leaning responses increase from *13* in the Vanilla model to *25* after SFT, indicating a marked shift toward Chinese cultural preferences. For **Llama3.1-8B**, the Vanilla model's predictions are clustered along the diagonal and skewed toward the US, while the SFT-tuned Llama shifts more modestly, increasing from *20* to *24* Chinese-leaning responses, thereby reducing roughly one-third of its initial US bias. Table 5 presents concrete 'before-and-after' examples with human answer distributions (US, CN) and model prediction distributions, illustrating how SFT consistently shifts Qwen (and, to a lesser extent, Llama) away from the US majority proportions and toward the Chinese ones.

6 Conclusion

This study investigates how native speakers' word associations serve as cultural knowledge resources. We fine-tuned 8 language models (across two languages, two LLMs, and two training approaches) to learn cultural signals and evaluate their cultural alignment. We find that fine-tuned mid-sized LLMs on language-specific word-association norms (English and Mandarin SWOW) yield clear improvements in both lexical and value alignment. Fine-tuned models retrieve human associations with higher precision and more closely match human valence and arousal ratings, while their World Values Survey responses shift toward target-culture distributions. These findings demonstrate that grounding LLMs in a few million associative cues can instill authentic cultural understanding and enhance value reasoning without costly retraining.

[18] See details on selecting the tension-set in Appendix G.2.

[19] More results in US are provided in Appendix G.3.

7 Limitations

Focusing on country-level alignment. Our evaluation aggregates cultural values at the national level (United States vs. China) and does not employ persona- or demographic-based prompting. While this choice simplifies the analysis, it may mask important regional, social, or demographic variations within each country.

Temporal gap between data and model training. We rely on WVS Wave 7 surveys conducted during 2017–2022 (Haerpfer et al., 2020), English SWOW associations collected in 2011–2018 (De Deyne et al., 2019), and Mandarin SWOW data from 2016–2023 (Li et al., 2024a). In contrast, Llama 3.1 (8B) and Qwen 2.5 (7B) were trained on web data up to late 2023/early 2024. This temporal mismatch means our human cultural benchmarks may not fully reflect the information learned by the models, and shifts in cultural values or associations after the data collection periods are not captured.

Limited scope of languages and models. We focus on two high-resource languages (English and Mandarin) and two open-source models (Llama 3.1 and Qwen 2.5). This selection was chosen for tractability, but the findings might not generalize to other open-sourced model families or commercial models. Furthermore, the generalizability to low-resource languages and cultures requires further exploration. We consider cultural alignment research of using word associations as a two-step program: (1) establish whether word associations can serve as transferable cultural knowledge, and (2) explore how this transfers to low-resource languages and determine minimal data requirements for effective transfer. Our study focuses on step (1), which provides the foundation for step (2). Given the positive results from our analysis and the open-sourced code, future work should extend to additional languages and model architectures.

The Double-edged sword of cultural alignment. In this study, we focus on the mechanism of learning cultural knowledge from the word associations dataset, and we observed closer cultural value alignment on the evaluated dataset. At the same time, we acknowledge that alignment might also bring risks, such as reinforcing stereotypes and amplifying existing biases. However, we argue that context-aware alignment mitigates harm more effectively than generic, one-size-fits-all models in contexts that require culturally specific adaptation. Universal models often default to Anglophone-majority norms; when user context differs, this mismatch is a common source of insensitive or inappropriate outputs. Our intent is to reflect cultural norms accurately within context, not to promote or endorse them. Accordingly, we recommend per-culture LoRA adapters used only in their intended context (e.g., ZH adapter for Chinese settings), with baseline fallback otherwise. While not part of this study, we suggest a deployment pattern that further mitigates harms: wrapping the per-culture adapters in lightweight agent safeguards (e.g., context routing/opt-in, a safety critic for stereotyping/-generalizations, uncertainty-based abstention, and baseline fallback).

Impact of fine-tuning. Our study evaluates how fine-tuning with culture-specific data directs the model toward a target culture. However, we do not evaluate the impact of fine-tuned models on non-targeted cultures. This is by design: our scope is not to train a universal model for all cultures, but to develop culture-specific models, as prior work shows that one-size-fits-all LLMs are inferior to cultural tailored models (Li et al., 2024b). In line with our stance on cultural alignment, as described in the Introduction, we regard culture-specific models as essential. Therefore, we train culture-specific models (e.g., via adapters, or specific checkpoints from PPO models) on a shared base model and evaluate them within the target culture. For languages without an available adapter (e.g., Dutch, German), we recommend using the baseline model rather than applying adapters tuned for other cultures (e.g., ZH-tuned or US-tuned adapters). Our released pipeline can also facilitate future work on expanding to more cultures.

Learning efficiency across cultures. While our empirical results show that cultural association training improves alignment, they do not fully explain why certain learning approaches perform better under specific cultural contexts. For instance, SFT and PPO exhibit different learning efficiencies across cultures: SFT achieves optimal alignment with Chinese values, whereas PPO performs best on US values. These findings point to promising directions for future work to explore how training methods and cultural contexts interact, for example through analyses of model internal states.

References

Anurag Acharya, Kartik Talamadupula, and Mark Finlayson. 2021. Toward an atlas of cultural commonsense for machine reasoning.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Meltem Aksoy. 2025. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, page 100172.

Mehar Bhatia and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Lera Boroditsky. 2011. How language shapes thought. *Scientific American*, 304(2):62–65.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Álvaro Cabana, Camila Zugarramurdi, Juan C Valle-Lisboa, and Simon De Deyne. 2024. The" small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, 56(2):968–985.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 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)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *The Second Conference on Parsimony and Learning (Recent Spotlight Track)*.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias via word association test. *arXiv preprint arXiv:2505.18562*.

Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior research methods*, 45(2):480–498.

Simon De Deyne, Danielle J Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Werner Delanoy. 2020. *What Is Culture?* Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei

Jiang. 2024. LLMCarbon: Modeling the end-to-end carbon footprint of large language models. In *The Twelfth International Conference on Learning Representations*.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark. Association for Computational Linguistics.

Hofstede Geert, Hofstede Hofstede, Gert Jan, and Michael Minkov. 2020. *Cultures and organizations: Software for the mind*. McGraw-Hill.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Luis Guerrero, Anna Claret, Wim Verbeke, Geraldine Enderli, Sylwia Zakowska-Biemans, Filiep Vanhonacker, Sylvie Issanchou, Marta Sajdakowska, Britt Signe Granli, Luisa Scalvedi, et al. 2010. Perception of traditional food products in six european regions using free word association. *Food quality and preference*, 21(2):225–233.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Björn Puranen, et al. 2020. World values survey: Round seven–country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 7:2021.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, et al. 2022. World values survey: Round seven–country-pooled datafile version 5.0.

Edward T Hall. 1976. *Beyond culture*. Anchor.

Alexander Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. In *AI for Math Workshop @ ICML 2024*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Bing Li, Ziyi Ding, Simon De Deyne, and Qing Cai. 2024a. A large-scale database of mandarin chinese word associations from the small world of words project. *Behavior Research Methods*, 57(1):34.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024b. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Zheng Wei Lim, Harry Stuart, Simon De Deyne, Terry Regier, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. A computational approach to identifying cultural keywords across languages. *Cognitive Science*, 48(1):e13402.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

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

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1774–1784, New York, NY, USA. Association for Computing Machinery.

Richard E. Nisbett and Takahiko Masuda. 2003. Culture and point of view. *Proceedings of the National Academy of Sciences*, 100(19):11163–11170.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Denise Park and Chih-Mao Huang. 2010. Culture wires the brain. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 5:391–400.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aida Ramezani and Yang Xu. 2024. Moral association graph: A cognitive model for moral inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Jung-Soo Son, Vinh Bao Do, Kwang-Ok Kim, Mi Sook Cho, Thongchai Suwonsichon, and Dominique Valentin. 2014. Understanding the effect of culture on food representations using word associations: The case of "rice" and "good rice". *Food quality and Preference*, 31:38–48.

Diana Su Yun Tham, Paul T Sowden, Alexandra Grandison, Anna Franklin, Anna Kai Win Lee, Michelle Ng, Juhyun Park, Weiguo Pang, and Jingwen Zhao. 2020. A systematic investigation of conceptual color associations. *Journal of Experimental Psychology: General*, 149(7):1311.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. breakthroughs in statistics. *Springer Series in Statistics. Kotz S, Johnson NL (ed): New York*, 1992:196–202.

Chaoyi Xiang, Chunhua Liu, Simon De Deyne, and Lea Frermann. 2025. Comparing moral values in Western English-speaking societies and LLMs with word associations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3521–3536.

Teng Xiao, Yige Yuan, Mingxiao Li, Zhengyu Chen, and Vasant G Honavar. 2025. On a connection between imitation learning and RLHF. In *The Thirteenth International Conference on Learning Representations*.

Xu Xu and Jiayin Li. 2020. Concreteness/abstractness ratings for two-character chinese words in meld-sch. *PloS one*, 15(6):e0232133.

Xu Xu, Jiayin Li, and Huilin Chen. 2022. Valence and arousal ratings for 11,310 simplified chinese words. *Behavior research methods*, 54(1):26–41.

Peiran Yao, Tobias Renwick, and Denilson Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

A Fine-tuning LLMs on Cultural Associations

Fine-tuning directly on word association reshapes the model's behavior by adjusting its weight parameters. This approach has two key benefits:

Independence from external KB: Fine-tuning eliminates the need for an external retrieval system during inference. RAG relies on real-time access to a knowledge base, which may not always be *available* and can significantly slow down inference due to retrieval latency. In contrast, a fine-tuned model carries its learned associations internally, making it faster and more self-contained.

Generalization beyond the dataset: Fine-tuning enables the model to generalize to unseen examples by learning patterns and semantic relationships during training. For example, since "gorilla" and "monkey" are close in the word embedding space due to their shared features, a model fine-tuned on "monkey" or other nearby words—whether as cue words or associations—can implicitly infer associations for "gorilla", even if it's absent from the dataset.

In the following sections, we discuss the types of fine-tuning techniques and the associated task designs we employ for LLMs to learn word associations.

A.1 Supervised Fine-tuning

To provide context, we consider autoregressive LMs such as the GPT (Brown et al., 2020) and Llama (Grattafiori et al., 2024) series, which generate tokens in a left-to-right, autoregressive manner. Let $\mathbf{x}< i$ be the first $i-1$ tokens of a sequence $\mathbf{x}$, and let x_i be the i-th token. The probability that the LLM predicts token x_i at position i can be written as $LM\theta(\hat{x}_i = x_i \mid \mathbf{x}< i)$, where $LM\theta(\cdot)$ is the model's probability distribution over the vocabulary, and θ represents the model parameters.

We implement *a word association prediction task* directly in the supervised fine-tuning (SFT) framework. Given a training example $x = \langle c, \mathbf{w}\rangle$, where c is a cue word and $\mathbf{w} = \langle w_1, w_2, \ldots, w_n\rangle$ is a list of associated words, the model is trained to generate the associated words $\mathbf{w}$ conditioned on the cue word c. The objective of SFT is to maximize the likelihood of the training data, which is formalized as:

$$J(\theta) = \max_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathcal{X}} \left[\sum_{i=1}^{|\mathbf{x}|} \log LM_\theta(x_i \mid \mathbf{x}_{<i}) \right] \tag{1}$$

where $\mathcal{X}$ denotes the training dataset, and $|\mathbf{x}|$ is the length of the token sequence.

While this formulation captures the core learning objective, in practice we reformat each training instance into a more natural, instruction-style prompt that aligns with how LLMs are typically used. For example, we add constraints to the prompt to further guide the model's generation process, such as "do not generate words conditioned on the presence of other words, but focus solely on the cue word." See Appendix B for details.

A.2 PPO training

To further align LLMs with culturally-informed word associations, we explore reinforcement learning from human feedback (RLHF), using Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). RLHF has proven to be a powerful technique for fine-tuning LLMs by aligning them with preferences defined by a reward model, which is either trained on human feedback or based on predefined rules (Ouyang et al., 2022; DeepSeek-AI et al., 2025). Recent studies indicate that RLHF surpasses supervised fine-tuning (SFT) in enhancing LLMs' reasoning capabilities, as RLHF encourages exploration beyond explicit solutions found in training data, whereas SFT focuses on broad imitation of human-provided examples (Havrilla et al., 2024; Chu et al., 2025). From an imitation-learning viewpoint, RLHF exhibits mode-seeking behavior, prioritizing precise modes of response distributions, which makes it particularly effective for reasoning tasks demanding accuracy (Xiao et al., 2025). For further details on the differences between these fine-tuning approaches, we refer readers to Xiao et al. (2025).

We use a rule-based reward function designed to reflect the fulfillment of designed tasks. Before we turn into the task design, we first introduce the three components of RLHF framework:

1. a language model (policy) LM_θ generating candidate outputs,

2. a reward model $r(q, a)$ evaluating those outputs, where q is the question and a is the generated answer, and

3. a reinforcement learning algorithm (e.g., PPO) that updates the model to maximize the received reward.

Formally, RLHF fine-tunes the language model LM_θ by optimizing the following objective:

$$\max_\theta \mathbb{E}_{a\sim LM_\theta(a|q)}\left[r(q,a)\right] - \beta D_{\text{KL}}\left[LM_\theta(a \mid q)||LM_{\text{ref}}(a \mid q)\right] \quad (2)$$

where LM_{ref} is a frozen reference model (typically the initial SFT model), and β is a scaling factor controlling the KL penalty that discourages large divergences from the reference model so as to maintain the model stability.

Ranking-based format[20] Ultimately, we settled on a ranking task, where the model was asked to rank a list of association words of a cue word based on its frequency in the SWOW dataset. This design offers a middle ground: (1) It is more structured and constrained than free-form generation, improving training stability and (2) It is more challenging than MCQ, providing useful reward gradients for learning.

The reward function evaluates the alignment between the model's ranked list and ground truth rankings using Spearman's rank correlation coefficient.

The objective of PPO is formalized as:

$$L_{\text{PPO}}(\theta) = \mathbb{E}_{(c,\mathbf{w})\sim\pi_\theta}\left[\min\big(r(\mathbf{w})A, \text{clip}(r(\mathbf{w}), 1-\epsilon, 1+\epsilon)A\big) - \beta \log q(\mathbf{w})\right] \quad (3)$$

where

$$r(\mathbf{w}) = \frac{\text{LM}_\theta(\mathbf{w} \mid c)}{\text{LM}_{\theta_{-1}}(\mathbf{w} \mid c)}, \quad (4)$$

$$q(\mathbf{w}) = \frac{\text{LM}_\theta(\mathbf{w} \mid c)}{\text{LM}_{\text{ref}}(\mathbf{w} \mid c)}, \quad (5)$$

$$A = R_{\text{spearman}}(x) - V_{\text{critic}}(x) \quad (6)$$

While our main results focus on evaluating cultural alignment in downstream tasks, we also assess the LLMs' performance on the training tasks themselves—namely, supervised fine-tuning (SFT) for word association prediction and PPO training for ranking tasks. These results provide hints into whether models have successfully learned word association patterns during fine-tuning.

[20] Initially, we conducted preliminary experiments with multiple task formats to determine the most effective design for PPO training. See details in Appendix A.3.

Category	Example Words (Cue: *apple*)
High-frequency	fruit, red, pear, tree
Low-frequency	stem, sauce, farm, healthy
Indirect association	internet, mouse, machine (from word *computer*)
Random	house, planet, justice, notebook

Table 6: An example illustrating MCQ task categories.

A.3 Preliminary Experiments on Task Formats for PPO Training

One of the important preliminary experiments is to identify suitable task formats for PPO training, ensuring the complexity was balanced — neither trivially solvable nor excessively challenging. Tasks that are too easy yield minimal gradients for learning, whereas excessively difficult tasks also prevent LLMs from exploring the correct answer.

We considered three task formats: Multiple Choice Questions (MCQ), Free-form Association Word Prediction, and Ranking-based Association Prediction. Below we discuss each format in detail along with our experimental findings.

Experiment 1: MCQ Format. We initially designed an MCQ-style task to evaluate candidate answers consisting of different categories of word associations. Specifically, the model was presented with a cue word and required to choose the option (a set of associated words) most closely related to it. Each MCQ contained four categories of candidate answers:

- Category 1: High-frequency direct associations
- Category 2: Low-frequency direct associations
- Category 3: Indirect associations (frequent associations of the cue's frequent associations)
- Category 4: Random unrelated words

Table 6 provides an illustrative example of this MCQ format.

We hypothesized that Category 2 (low-frequency direct associations) and Category 3 (indirect associations) would serve as hard negative distractors, enhancing task difficulty. However, our experiments

revealed that Vanilla LLMs were able to solve these MCQs easily, achieving accuracy consistently near 100%. Thus, we concluded that the MCQ format was too simplistic to generate meaningful reward gradients for PPO training.

Experiment 2: Free-form Word Prediction. Our next experiment involved training PPO directly on the original word-association prediction task used for supervised fine-tuning (SFT). Here, the model freely generated association words conditioned solely on the cue word without explicit constraints.

This task proved to be overly challenging. The space of potential actions and states was extremely large, causing PPO training to suffer from poor convergence. The model rarely explored words sufficiently close to the ground-truth associations, leading to sparse reward signals, which hindered effective training.

Final Selection: Ranking-based Format. Ultimately, we selected a ranking-based format (as described in the main text), where the model ranks a provided list of association words for each cue word, ordered by their frequency in the SWOW dataset. This task strikes a suitable balance between structured guidance (to avoid sparse reward signals) and sufficient complexity (to prevent trivial performance), enabling effective gradient signals to guide PPO optimization.

B Prompts for Supervised Fine-tuning

We reformat each training instance into a more natural, instruction-style prompt that aligns with how LLMs are typically used. Below is a sample prompt for the cue word "mosquito" and its associated words:

Supervised Fine-tuning Example for English SWOW word association prediction

```
[CONTEXT]
You are a sophisticated language model
designed to explore word associations
comprehensively.
Given a cue word, your task is to
generate a comprehensive list of words
associated with the cue word.  Aim
to cover as many relevant contexts,
uses, and meanings as possible without
repeating similar concepts.  List a
target of [LOWER BOUND SIZE] to [UPPER
BOUND SIZE] words that together provide
a broad and insightful representation
of all significant associations.  Focus
on revealing both common and unique
aspects related to the cue word
to ensure a balanced and thorough
exploration of potential associations.
Words should be distinct from each
other.  Your response shall only be
the list of associated words.  Do
not generate words conditioned on the
presence of other words but rather
focus on the cue word itself.
[CUE WORD]
mosquito
[ASSOCIATED WORDS]
bite, bug, itch, buzz, malaria, insect,
blood, net, fly, annoying, pest, summer,
ouch, itchy, buzzing, repellent, small,
swat, irritating, gnat, netting,
camping, midge, proboscis, river,
pain, lump, sting, flight, disease,
spray, slap, swamp, fever, allergy,
annoyance, worthless, nest, crunchy,
smack, huge in canada, dead, amazonian,
insect bite, awake, tropical, water,
female, anopheles, coast, valentine,
doug, tent, jungle, whine, bumblebee,
bored, nozzle, blood sucker, noisy,
nasty, skin, vampire, torment, hawk,
ear, itchy welt, pinch, needle, dengue,
africa, bloodsucker, annoying bug,
mosquito net, australia, horrible, kill,
ugly, genetics
```

Supervised Fine-tuning Example for Mandarin SWOW word association prediction

```
[CONTEXT]
您是一款专为全面探索词语关联而设计的高级
语言模型。给定一个提示词，你的任务是生成
一个与该提示词相关联的全面词汇列表。目标
是尽可能涵盖所有相关的语境、用法和含义，
避免重复相似的概念。列出目标数量为[LOWER
BOUND SIZE] 到[UPPER BOUND SIZE]个词，这
些词共同提供对所有重要关联的广泛而深刻的
表示。专注于揭示与提示词相关的常见和独特
的方面，以确保对潜在关联进行平衡而彻底的
探索。词语应彼此不同。你的回答只能是相关
联的词语列表。不要生成受其他词语存在影响
的词语，而是专注于提示词本身。
[CUE WORD]
狱警
[ASSOCIATED WORDS]
监狱，警察，警棍，囚犯，制服，罪犯，犯
人，凶，看守，坐牢，严厉，警犬，暴力，很
凶，手铐，监管，刑警，局长，公安，强悍，
抹布，铁窗泪，打架，叮当作响，囚服，斯雷
因，管理，刑罚，敬业，可怕，辛苦，工作，
黑暗，霸王，钥匙，牢饭，SM，冷漠，凶恶，
逃狱，逃跑，强壮，酷刑，狱都市变，坏人，
凶悍，男人，刑法，条纹服，黑猫警长，铁
牢，卓别林，狱卒，反派，美剧，狱中杂记，
法律，僻静，虐待，劳改，悔恨，棍棒，牢
房，殴打，性虐待，女警，典狱长，警装，严
格，帅哥，肉文，铁棍，警服，电网，高墙，
严肃，警司，很辛苦，害怕，抓人，阳光，美
国，斯坦福大学，越狱
```

To prevent overfitting and pattern memorization during training, we randomly set the lower and upper bounds for the number of associated words required in each training instance. The associated words are not shuffled; instead, they are ordered by frequency from the SWOW dataset, with the most frequent words listed first. This ordering introduces an inductive bias, encouraging the model to think of the most common associations first.

C Prompts for PPO training

The task for PPO training is to rank a list of association words of a cue word based on its frequency in the SWOW dataset. The prompt for PPO training is similar to that of SFT, but with a different instruction.

PPO training Example for English SWOW ranking task

```
[CONTEXT]
You are a sophisticated language model
designed to explore word associations
comprehensively.
Given the cue word, rank the following
associated words from the most strongly
related (rank 1) to the least strongly
related (rank 10).
Important Notes:  1.  Rank ONLY
the provided associated words from
strongest (1) to weakest (10) in
relation to the cue word.  2.  Do
NOT introduce any new words that aren't
in the provided list.
Think step by step, comparing each
associated word to the others to
determine their relative strength of
association with the cue word.
**Your final answer should at the end
of the response and be in the following
format:**
Final Ranking:  Rank 1:  [Associated
Word] Rank 2:  [Associated Word] ...
Rank 10:  [Associated Word]
[CUE WORD]
dislike
[TARGET ANSWER]
Rank 1:  detest
Rank 2:  orange
Rank 3:  flavor
Rank 4:  displeasure
Rank 5:  be well
Rank 6:  kid refusing to eat
Rank 7:  ugh
Rank 8:  boss
Rank 9:  peeve
Rank 10:  gas
```

D PPO Reward function details

```
% def compute_reward(queries, prompts, labels):
    """
    Computes reward scores for PPO training based on Spearman's rank correlation
    between predicted and ground-truth word association rankings.

    Args:
        queries: List of model responses (each includes both prompt and response).
        prompts: List of prompt texts.
        labels: List of ground-truth ranked word lists.

    Returns:
        A tensor of Spearman correlation scores, one per example.
    """
    rewards = []
    for query, prompt, label in zip(queries, prompts, labels):
        # Extract the response by removing the prompt part
        response = query[len(prompt) - 1:]

        # Parse predicted rankings (e.g., "1: cat, 2: dog, ...")
        predicted_words = parse_ranked_words(response)

        # Normalize and filter ground truth
        ground_truth = [w.lower() for w in eval(label)]
        predicted_filtered = [w for w in predicted_words if w.lower() in ground_truth]

        # Convert to rank indices
        pred_ranks, gt_ranks = map_to_rank_indices(predicted_filtered, ground_truth)

        # Compute Spearman correlation
        score = spearmanr(pred_ranks, gt_ranks).correlation
        rewards.append(score if not pd.isnull(score) else -1.0)

    return torch.tensor(rewards, dtype=torch.float32)
```

E Experiment Settings

The experiments were conducted using two compute nodes equipped with 4 NVIDIA A100 GPUs per node. For SFT, we used Llama Factory library. The hyperparameters are provided in Table 7.

Hyperparameters	Value
Fine-tuning method	LoRA
LoRA Rank	64
LoRA Alpha	256
Learning rate	1.0e-5
Scheduler	Cosine (warmup ratio=0.1)
Batch size per GPU	18
Gradient accumulation	2
Number of epochs	1.5
Precision	bf16
Max sequence length	2048

Table 7: Hyperparameters for SFT Training

For PPO training, we used OpenRLHF library. The hyperparameters are provided in Table 8.

Hyperparameters	Value
Actor learning rate	5e-7
Critic learning rate	9e-6
Initial KL coefficient	0.1
Micro train batch size	8
Train batch size	32
Micro rollout batch size	16
Rollout batch size	64
Max training samples	1,000,000
Max epochs	1
Prompt max length	1024
Generation max length	1024
Zero optimization stage	3
Precision	bf16
Gradient checkpointing	Enabled
Optimizer offload	Adam offload
Attention implementation	Flash attention
VLLM tensor parallel size	2

Table 8: Hyperparameters for PPO Training

F Evaluation on the Emotions and Concreteness

F.1 Psychological Norms

For English, we evaluate the emotions in associations using the Valence, Arousal, Dominance (VAD) dataset (Warriner et al., 2013) with 13,915 English lemmas. A score close to 1 suggests that the concept tends to evoke a relaxed, bored, or sleepy emotional state, indicating a low arousal response, whereas a score near 8 signifies that the concept tends to be associated with feelings of excitement, happiness, or high arousal. Concreteness score is obtained from a lexicon with 40K English word lemmas (Brysbaert et al., 2014). Highly concrete concepts (a score within the range of 4 to 5) are defined as those that can be directly experienced through the senses, such as objects, actions, or sensations that are easily experienced.

For Chinese, we use a lexicon with 11K simplified Chinese words for the Valence and Arousal (Xu et al., 2022). For valence ratings, each word is rated on a seven-point scale: "–3" = extremely negative, "0" = neutral, and "+3" = extremely positive. For arousal ratings, each word is rated on a five-point scale: "0" = very low arousal and "4" = very high arousal. For concreteness in Chinese, we use a lexicon of 9877 Two Character Chinese words (Xu and Li, 2020). Each word is mapped into a 1 to 5 score, where "1" = "very concrete" and "5" = "very abstract".

Pre-processing

- **Token cleaning**: d-case, strip punctuation; English tokens are WordNet-lemmatised using NLTK (Bird, 2006), while Mandarin tokens remain in surface form after Chinese punctuation removal.
- **Lexicon look-up**: tokens are matched against the English VAD norms (Warriner et al., 2013) and concreteness norms (Brysbaert et al., 2014), or the corresponding Mandarin lexicons (Xu et al., 2022; Xu and Li, 2020). Tokens absent from a lexicon are ignored for that metric.

Hypothesis testing

Cue-level medians are compared with a paired Wilcoxon signed-rank test to determine whether the model's lexical profile is *indistinguishable* from that of humans.

We test whether a model's typical score is *statistically indistinguishable* from the human baseline, so the null states "no difference" while the alternative states "some difference".

Null hypothesis H_0: $\tilde{x}_{\text{model}} = \tilde{x}_{\text{human}}$ **(assumes equality).**
Alternative H_1: $\tilde{x}_{\text{model}} \neq \tilde{x}_{\text{human}}$ **(assumes a non-zero gap).**

Cells with $p \geq 0.05$ (i.e. we *fail to reject* H_0) are highlighted in **bold**.

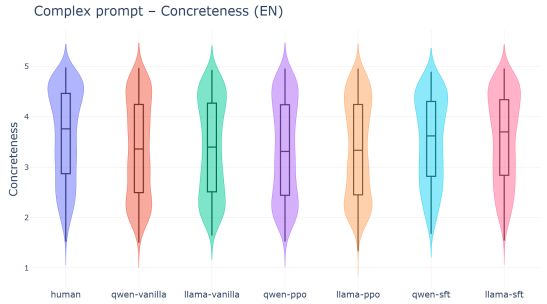
F.2 Cue-level Valence, Arousal and Concreteness

Section 4 (Table 3) presents the median scores for the psychological attribute evaluation. Since the median tends to compress information, we further visualize the distributions of these scores across all cues to better illustrate variations in each attribute. Figures 4 (Concreteness), 6 (Arousal), and 5 (Valence) show the distributions of the psychological attributes.

In these violin plots, we can clearly see that models fine-tuned on association datasets tend to exhibit distribution shapes more similar to those of humans (shown on the far left). For example, in English, the concreteness scores (Figure 4) of both SFT models display a noticeable bulge in the upper range—resembling the human distribution—whereas the vanilla models show a more evenly dispersed pattern.

F.3 Examples of psychological attributes

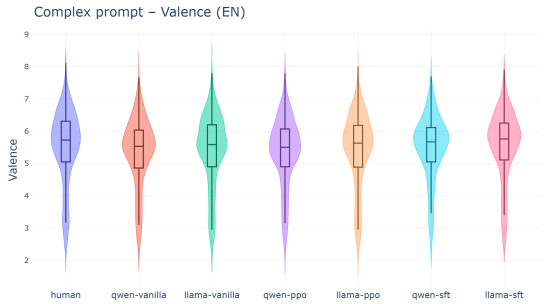
Table 9 presents examples of the various sources of associations for two cues: Halloween and emotions.



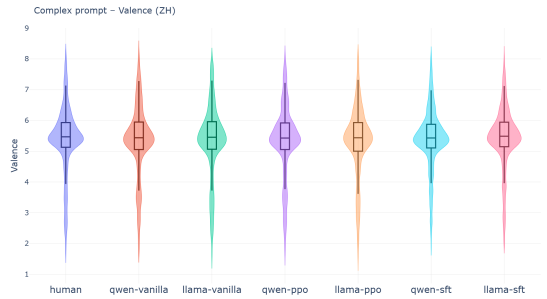
(a) English: association **concreteness** (1 = abstract, 5 = concrete).

(b) Mandarin: association **concreteness** on the rescaled 1–5 range.

Figure 4: Violin + box plots of per-cue **concreteness** medians for the *Complex* prompt. Left: English (1 = abstract, 5 = concrete); Right: Mandarin (rescaled to 1–5).



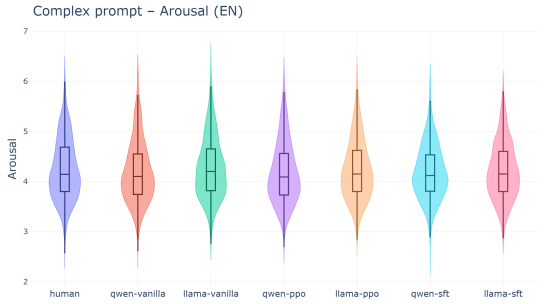
(a) English: association **valence** (1 = unpleasant, 9 = pleasant).



(b) Mandarin: association **valence**, rescaled to the English 1–9 range.

Figure 5: Violin + box plots of per-cue **valence** medians for the *Complex* prompt. Left: English (1 = unpleasant, 9 = pleasant); Right: Mandarin (rescaled to 1–9).

Below, we summarize key patterns observed in these examples.



(a) English: association **arousal** (1 = calm, 9 = excited).

(b) Mandarin: association **arousal**, rescaled to the English 1–9 range.

Figure 6: Violin + box plots of per-cue **arousal** medians for the *Complex* prompt. Left: English (1 = calm, 9 = excited); Right: Mandarin (rescaled to 1–9).

Analysis on Valence. Fine-tuned models align their associations more closely with human representations, as reflected in the valence scores of their associations. For instance, when prompted with *Halloween*, US participants tend to produce highly pleasant associations (median valence = 7), such as *candy* (7.27), *holiday* (7.18), and *party* (7.18). In contrast, Vanilla Llama model often evokes less pleasant associations, including *monster* (2.55), *skeleton* (4.37), and *spider* (3.35). Models fine-tuned on SWOW.US narrow this gap: the Llama SFT model, for example, generates high-valence associations like *holiday* (7.18) and *kid* (7.23), more closely mirroring human affective patterns. A similar pattern is observed for the cue *emotions*. A substantial valence gap exists between human associations (median = 2.83) and those of the vanilla Llama model (6.90). The vanilla model tends to produce overly pleasant (high Valence scores) associations, whereas human associations reflect a more complex emotional landscape—mixing both positive and negative feelings such as *sad* (2.1), *happy* (8.47), *anger* (2.5), and *cry* (3.22). The best-performing fine-tuned model, Llama-PPO, better captures this complexity, generating associations such as *happiness* (8.48), *sadness* and *anxiety* (2.38).

Cue	Source	Attribute	Associations (Score)	Median
Halloween	Human	Valence	candy (7.27), pumpkin (7.0), costume (6.05), costumes (6.05), holiday (7.18), October (-), ghosts (4.23), orange (6.81), pumpkins (7.0), party (7.18)	7.00
	Llama-8B (Vanilla)	Valence	pumpkin (7.0), costume (6.05), trick-or-treat (5.87), candy (7.27), ghost (4.23), monster (2.55), skeleton (4.37), bat (4.81), spider (3.35), witch (3.14)	4.59
	Llama-8B (SFT)	Valence	costume (6.05), october (-), night (6.68), leyton (-), pumpkins (7.0), masks (4.81), holiday (7.18), trick-or-treat (5.87), scare (3.55), kid (7.23)	6.37
	Llama-8B (PPO)	Valence	costume (6.05), pumpkin (7.0), candy (7.27), trick-or-treat (5.87), ghost (4.23), spider (3.35), witch (3.14), candy corn (6.61), bats (4.81), black cat (6.18)	5.96
	Human	Arousal	candy (5.03), pumpkin (3.43), costume (4.78), costumes (4.78), holiday (4.93), October (-), ghosts (5.7), orange (4.04), pumpkins (3.43), party (6.08)	4.78
	Llama-8B (Vanilla)	Arousal	pumpkin (3.43), costume (4.78), trick-or-treat (5.29), candy (5.03), ghost (5.7), monster (5.55), skeleton (4.55), bat (4.57), spider (6.91), witch (5.3)	5.16
	Llama-8B (SFT)	Arousal	costume (4.78), october (-), night (3.57), leyton (-), pumpkins (3.43), masks (3.26), holiday (4.93), trick-or-treat (5.29), scare (7.1), kid (4.71)	4.76
	Llama-8B (PPO)	Arousal	costume (4.78), pumpkin (3.43), candy (5.03), trick-or-treat (5.29), ghost (5.7), spider (6.91), witch (5.3), candy corn (4.23), bats (4.57), black cat (4.04)	4.91
	Human	Concreteness	candy (4.83), pumpkin (4.9), costume (4.57), costumes (4.57), holiday (2.86), October (2.81), ghosts (3.19), orange (4.66), pumpkins (4.9), party (3.89)	4.57
	Llama-8B (Vanilla)	Concreteness	pumpkin (4.9), costume (4.57), trick-or-treat (3.36), candy (4.83), ghost (3.19), monster (3.72), skeleton (4.97), bat (5.0), spider (4.97), witch (4.17)	4.70
	Llama-8B (SFT)	Concreteness	costume (4.57), october (2.81), night (4.52), leyton (-), pumpkins (4.9), masks (4.96), holiday (2.86), trick-or-treat (3.36), scare (2.96), kid (4.56)	4.52
	Llama-8B (PPO)	Concreteness	costume (4.57), pumpkin (4.9), candy (4.83), trick-or-treat (3.36), ghost (3.19), spider (4.97), witch (4.17), candy corn (4.9), bats (5.0), black cat (4.31)	4.70
emotions	Human	Valence	feelings (6.5), sad (2.1), happy (8.47), love (8.0), sadness (2.4), anger (2.5), angry (2.53), tears (3.14), cry (3.22), mad (2.47)	2.83
	Llama-8B (Vanilla)	Valence	feelings (6.5), sentiments (6.2), moods (5.29), sincerity (7.9), empathy (7.29), compassion (7.9), love (8.0), anger (2.5), fear (2.93), joy (8.21)	6.90
	Llama-8B (SFT)	Valence	feelings (6.5), happy (8.47), sad (2.1), love (8.0), mood (5.29), feelings and thoughts (6.63), heart (6.95), thoughts (6.76), feelings* (6.5), face (6.36)	6.56
	Llama-8B (PPO)	Valence	happiness (8.48), sadness (2.4), anxiety (2.38), love (8.0), fear (2.93), anger (2.5), empathy (7.29), jealousy (2.58), guilt (2.29), shame (2.62)	2.60
	Human	Arousal	feelings (3.86), sad (3.49), happy (6.05), love (5.36), sadness (2.81), anger (5.93), angry (6.2), tears (4.8), cry (5.45), mad (5.59)	5.41
	Llama-8B (Vanilla)	Arousal	feelings (3.86), sentiments (3.54), moods (4.5), sincerity (4.42), empathy (3.62), compassion (4.5), love (5.36), anger (5.93), fear (6.14), joy (5.55)	4.50
	Llama-8B (SFT)	Arousal	feelings (3.86), happy (6.05), sad (3.49), love (5.36), mood (4.5), feelings and thoughts (4.01), heart (5.07), thoughts (4.16), feelings* (3.86), face (4.59)	4.33
	Llama-8B (PPO)	Arousal	happiness (6.5), sadness (2.81), anxiety (4.78), love (5.36), fear (6.14), anger (5.93), empathy (3.62), jealousy (5.45), guilt (4.48), shame (5.4)	5.38
	Human	Concreteness	feelings (1.68), sad (3.07), happy (2.56), love (2.07), sadness (1.82), anger (2.41), angry (2.53), tears (4.56), cry (4.0), mad (2.76)	2.54
	Llama-8B (Vanilla)	Concreteness	feelings (1.68), sentiments (2.1), moods (1.75), sincerity (1.97), empathy (1.63), compassion (1.89), love (2.07), anger (2.41), fear (2.57), joy (2.37)	2.02
	Llama-8B (SFT)	Concreteness	feelings (1.68), happy (2.56), sad (3.07), love (2.07), mood (1.75), feelings and thoughts (1.68), heart (4.52), thoughts (1.97), feelings* (1.68), face (4.87)	2.02
	Llama-8B (PPO)	Concreteness	happiness (2.6), sadness (1.82), anxiety (2.21), love (2.07), fear (2.57), anger (2.41), empathy (1.63), jealousy (1.8), guilt (1.93), shame (2.24)	2.14

Table 9: Examples of cue–association psychological attributes, including Valence, Arousal, and Concreteness, for the cues *Halloween* and *emotions*. The "Median" column is the median score of each cue, computed from the attribute values of its associated words in each row. Numbers in parentheses indicate the corresponding attribute scores. Valence and Arousal values range from 1–9 (higher values indicate more pleasantness and stronger emotional intensity), while Concreteness values range from 1 (more abstract) to 5 (higher concrete). (-) means the word is not found in the corresponding norms.

Analysis on Arousal. Unlike valence, which captures pleasantness, arousal reflects the intensity or activation level of associations. For *Halloween*, human responses indicate moderate excitement (median = 4.78), balancing calm elements like *pumpkin* (3.43) with livelier cues such as *party* (6.08). The vanilla models amplify this excitement, favoring highly stimulating words like *monster* (5.55) and *spider* (6.91), resulting in slightly higher overall arousal (median = 5.16). Fine-tuned variants temper this tendency—Llama-PPO (median = 4.91), for instance, retrieves a steadier mix of associations

spanning both neutral and intense states (*candy* (5.03), *ghost* (5.7), *black cat* (4.04)).

The cue *emotions* shows an example that the Vanilla model tends to diminish the emotional intensity. Human associations cover a broad emotional range, from *sad* (3.49) to *anger* (5.93), yielding a balanced median (5.41). Vanilla Llama compresses this variation, producing a flatter, less expressive pattern (median = 4.50). In contrast, the Llama-PPO model restores much of this dynamic spread (median = 5.38), surfacing high-arousal concepts such as *anger* (5.93), *guilt* (4.48), and *shame* (5.40), which better approximate human affective diversity.

Analysis on Concreteness. As observed previously, models still struggle to align their concreteness scores with human judgments. For *Halloween*, the Llama SFT model achieves a similar median concreteness score to humans, whereas other models produce overly concrete associations. The pattern reverses for the cue *emotions*, where models tend to generate words that are excessively abstract.

G Evaluation Results on World Values Survey

G.1 Breakdown results on EMD

G.2 Tension Set Selection

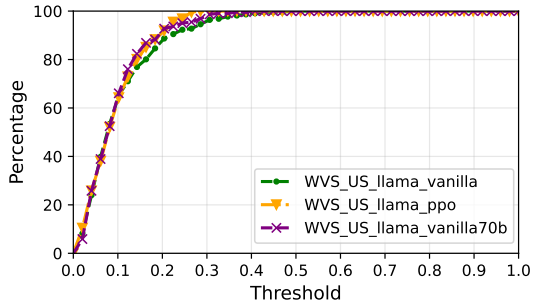
Given the participants' answer distributions for China (q) and the United States (p), we first normalise each to a probability vector i.e. we divide each count by the total number of respondents for that question so the values now represent probabilities (fractions between 0 and 1). Divergence is then measured with a hybrid score that averages an entropy-sensitive component (Jensen–Shannon divergence, JSD) and an ordinal component (normalised Earth-Mover distance, EMD):

$$\text{combo}(p, q) \;=\; \tfrac{1}{2}\,JSD(p, q) \;+\; \tfrac{1}{2}\,EMD(p, q).$$

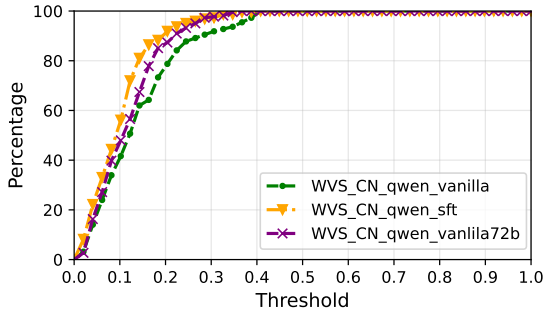
Sorting the WVS questions by this score and retaining the top 50 yields our fixed *tension set*.

G.3 Cross-Cultural Value Alignment Evaluation (EN Prompts)

Beyond Mandarin prompts, we also evaluate cultural shifts with English prompts. Figures 8a and 8b mirror the same layout used for Chinese prompts: hybrid distances to US answers (x-axis) and Chinese answers (y-axis) are plotted across 50 high-tension WVS questions.

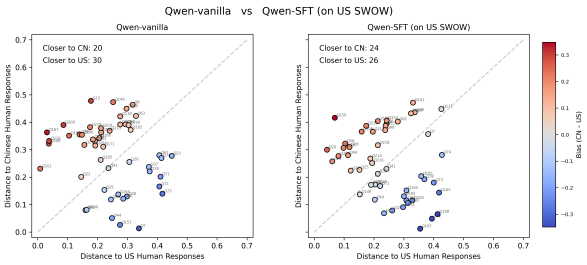


(a) WVS-us under Jensen Shnnon

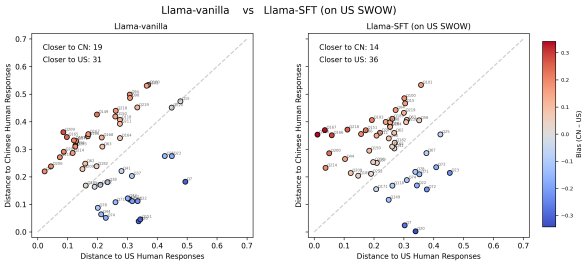


(b) WVS-zh performance under Jensen Shannon

Figure 7: Breakdown comparison of model alignment with cultural values across China and United States based on the World Values Survey. Results are shown for the Vanilla and trained (SFT and PPO) versions of Qwen2.5 and Llama 3.1.



(a) **Qwen-7B**: SFT on US SWOW does not shift the cloud substantially.



(b) **Llama-8B**: minimal movement after SFT on US SWOW.

Figure 8: Shifts after SFT on US SWOW (EN prompts). Each dot = one WVS question; colour = bias (CN–US).

- **Qwen-7B.** The vanilla model already exhibits strong alignment with US responses; fine-tuning on US SWOW slightly reduces this alignment (from 30 to 26 US-aligned points).
- **Llama-8B.** Supervised fine-tuning increases US

Question (ZH)	Prompt (EN)	Survey	Q_{van}	Q_{sft}	JS	JS–SFT	EMD	EMD–SFT	Type
您是否认为有天堂?	In which of the following do you believe, if you believe in any? – Heaven *(1: Yes; 2: No)*	[12%,88%]	[71%,29%]	[18%,82%]	0.437	0.061	0.173	0.062	Religious
您是否相信死后有来生?	In which of the following do you believe, if you believe in any? – Life after death *(1: Yes; 2: No)*	[12%,88%]	[90%,10%]	[36%,64%]	0.596	0.208	0.020	0.246	Religious
您是否信仰佛祖/上帝/真主/神明?	In which of the following do you believe, if you believe in any? – God *(1: Yes; 2: No)*	[17%,83%]	[41%,59%]	[29%,71%]	0.182	0.100	0.232	0.119	Religious
您是否认为有地狱?	In which of the following do you believe, if you believe in any? – Hell *(1: Yes; 2: No)*	[11%,89%]	[47%,53%]	[16%,84%]	0.288	0.049	0.359	0.047	Religious

Table 10: Comparison of survey distributions and model outputs (vanilla vs. SFT) for five religious-belief WVS items. Highlighted cells show metrics after SFT.

alignment, shifting the number of US-aligned points from 31 to 36.

These results suggest that for English prompts, vanilla models—particularly Qwen—may already exhibit strong US alignment, reducing the effect of SFT on US SWOW.

G.4 WVS Answer Shifts Across Topics

To examine fine-grained cultural effects, we group WVS questions into twelve topical domains and compare alignment before and after SFT on Chinese SWOW. Figures 9 and 10 (below) visualize Jensen–Shannon and Earth Mover's distances by topic. Fine-tuning improves alignment in five domains—ethical values, political engagement, religious beliefs, social capital, and safety perceptions—while it slightly reduces alignment for economic values and corruption perceptions. This drop may reflect a mismatch between model training distributions and the nuanced economic attitudes Chinese respondents hold.

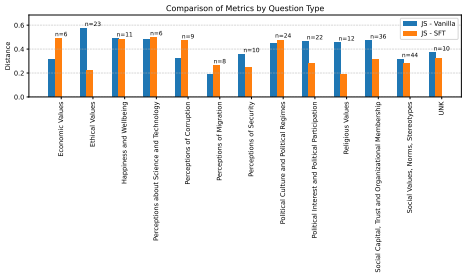


Figure 9: Jensen–Shannon distance by WVS topic (Vanilla vs. SFT Qwen-7B on ZH prompts).

Table 10 presents concrete examples of distribution shifts from the vanilla Qwen-2.5 model to the SFT Qwen-2.5 model. For example, in the

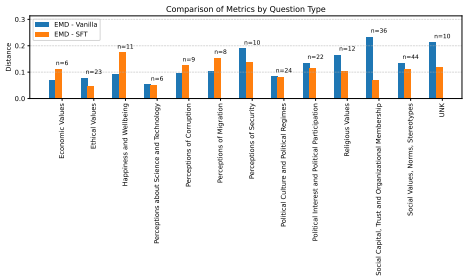


Figure 10: Earth Mover's distance by WVS topic (Vanilla vs. SFT Qwen-7B on ZH prompts).

domain of religious values, the vanilla model's predictions are either overly dispersed or peak at culturally incongruent options, whereas fine-tuning realigns the predicted distributions with human responses. When asked "Do you believe in Heaven?", the vanilla model strongly predicts "Yes" (0.70), while the fine-tuned model shifts to "No" (0.84), closely matching the actual distribution from Chinese participants (0.89 "No"). Notably, although the SFT model rejects Western religious imagery like "Heaven," it also captures Chinese-specific spiritual concepts such as "Life after death." In the SWOW–ZH associations for 死亡 (death), responses like 轮回 (reincarnation) and 新生 (new life) reflect how Chinese speakers conceptualize death, illustrating how association-based fine-tuning contributes to value prediction.