ALIGN: Word Association LearnIng for Cross-Cultural Generalization in Large Language Models

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly mediate cross-cultural communication, their behavior still reflects the distributional bias of the languages and viewpoints that are over-represented in their pre-training corpora. Yet, it remains a challenge to model and align culture due to limited cultural knowledge and a lack of exploration into effective learning approaches. We introduce a cost-efficient, cognitively grounded remedy: parameter-efficient fine-tuning on native speakers' free *word–association* norms, which encode implicit cultural schemas. Leveraging English-US and Mandarin associations from the Small-World-of-Words project, we adapt LLAMA-3.1-8B and QWEN-2.5-7B via supervised fine-tuning (SFT) and PPO-based preference optimization. SFT boosts held-out association Precision@5 by 16–20 % in English and 43–165 % in Mandarin, lifts median concreteness by +0.20, and attains human-level valence and arousal. These lexical gains transfer: on World-Values-Survey questions, fine-tuned models shift answer distributions toward the target culture, and on a 50-item high-tension subset, Qwen's Chinese-aligned responses double while Llama's US bias drops by one-third. Our 7–8B models rival or beat vanilla 70B baselines, showing that a few million culture-grounded associations can instill value alignment without costly 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]

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 (). 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.

As large language models (LLMs) become em-

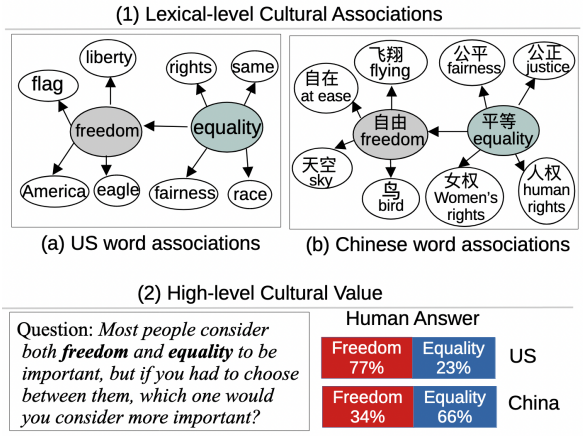


Figure 1: Example of how cultural word associations at the lexical level relate to higher-level cultural value preferences. (1) Word associations show distinct cultural perception around the word of **freedom** and **equality**, with American associations emphasizing individual liberty and patriotic 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.

[1] All code and data will be released upon acceptance.

bedded in global communication, they increasingly engage with users from diverse cultural backgrounds. However, most LLMs are trained primarily on English-language data, leading to an over-representation of Western perspectives and an under-representation of culturally 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 slashes compute demands, yet still needs culture-rich data (Hu et al., 2021; Dettmers et al., 2023).

Recent work has focused on evaluating cultural alignment using survey-based methods (Durmus et al., 2024) and adapting models through prompting or synthetic data generation (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 schemas. When prompted with *red*, U.S. respondents offer *danger*, *stop*, or *anger*, whereas Chinese respondents give *happiness*, *celebration*, or *luck*, illustrating how such spontaneous links reveal culture-specific cognition absent from text corpora. If such lexical links mirror deeper values, aligning them should nudge models toward culture-consistent judgements.

We fine-tune Llama-3.1-8B and Qwen-2.5-7B on English (SWOW.EN) and Mandarin (SWOW.ZH) word associations via SFT and PPO, then test (i) how well it regenerates human associations and (ii) World-Values-Survey alignment. Our findings reveal that (1) vanilla Llama leans toward U.S. associations and values, whereas vanilla Qwen leans toward Chinese; (2) association-tuned models produce markedly more human-like, affective, and concrete associations; and (3) this lexical gain translates into stronger value alignment with the target culture, most notably when the original model lacked that knowledge.

This study makes three key contributions:

1. We conduct the first head-to-head study of cultural fine-tuning, contrasting LoRA-based supervised fine-tuning with preference-optimized PPO on the same English and Chinese-SWOW corpora, and track their impact on valence, arousal and concreteness.

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

3. We commit to releasing – upon acceptance – the complete training pipeline plus the top-performing LoRA adapters so that anyone can plug US- or CN-specific cultural knowledge into their own LLMs.

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 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 tapping 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 multiple cultures (Liu et al., 2025; Adilazuarda et al., 2024). Researchers draw on cross-cultural surveys such as Hofstede's dimensions (Geert et al., 2020), the Pew Global Attitudes Survey and the World Values Survey (WVS) (Haerpfer et al., 2020). Recent benchmarks build on WVS to score LLMs across nations (Durmus et al., 2024; Zhao et al., 2024; Giuliani et al., 2024), capitalizing on its large sample sizes, 200-country coverage, and breadth of topics. We likewise adopt WVS for our value-alignment tests in Section 5.

2.2 Word Associations and Their Value

Word associations and their value Word association tasks elicit the first responses that come to

mind for a cue, exposing the spontaneous, affect-laden 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 edition spans 12 000 cues and 3 M 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., 2024) and other languages provide language-specific resources that ground 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). Whether such lexical-level signals can also steer LLMs toward higher-level value alignment, however, remains open. We tackle this gap by fine-tuning models on cross-cultural association data and testing transfer from word associations to World-Values-Survey 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 from association-level signals to high-level value alignment.

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 multiple training strategies and model families. We then assess each model on two tiers: (i) association generation and (ii) value alignment via survey questions. This section covers data and training, while the evaluation setups are given in Sections 4 and 5.

3.1 Language and Culture Selection

We focus on English (US) and Mandarin (CN) because they provide a clear cultural contrast for transfer experiments. These cultures differ in individualism vs. collectivism, emotional expression norms, and conceptual associations (as shown in our "red" example).

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.

3.2 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., 2024). SWOW.EN (2011–2019) provides 12 k cues and 3.6 M responses from 90 k native speakers in the United States, United Kingdom, Canada, and Australia (≈50 % U.S.). Each cue was answered by 100 participants with three free associations. For our U.S. analyses we retain only respondents whose country *and* native language are "United States," calling this subset SWOW.US. SWOW.ZH (2016–2023) comprises 10 k cues and 2 M responses from 40 k Mainland Chinese speakers. Both SWOW.US and SWOW.ZH are split **by cue** into 80 % train, 10 % validation, and 10 % test.

3.3 Model Selection

We choose widely used English-centric and Chinese-centric models 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 select Llama3.1-8B-Instruct as the English-centric model and Qwen2.5-7B-Instruct as the Chinese-centric model.[3] While the specific proportion of English data in Llama3.1 is unknown, Llama 3 is trained on a dataset comprising approximately 95% English content (Meta AI, 2023), and prior work has shown that it tends to reflect a strong Western cultural bias in its outputs (Aksoy, 2025). In contrast, the Qwen2.5 family (Qwen et al., 2025), developed by the Chinese company Alibaba, exhibits more Chinese-centric behavior across a range of

[2]We treat language-specific word associations as culturally grounded signals, reflecting the conceptual organization shaped by speakers' cultural experiences.

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

evaluation tasks in Chinese understanding and reasoning (Guo et al., 2025; Hong et al., 2025).

3.4 Fine-tuning LLMs on Cultural Associations

To investigate how models acquire culturally grounded knowledge from word associations, we employ two approaches that leverage different signals from human association data. First, the list of associations itself captures how native speakers understand a word. 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 use supervised fine-tuning (SFT) to train models to generate associations that are more aligned with these associations. Second, some associations are more commonly produced than others in human word association data (e.g., *nation* is more frequent than *farm*), which can serve as a signal of human preference. We designed an approach using reinforcement learning with PPO (Schulman et al., 2017) to train models to rank the importance of associated words in a way that aligns with human-produced frequency rankings.

From an imitation-learning perspective, SFT aims for broad coverage of the training data distribution, whereas PPO fine-tuning is more mode-seeking, making it particularly effective for improving LLM reasoning capabilities in tasks demanding precise and accurate answers (Xiao et al., 2025).

However, it remains unclear how these approaches differ in acquiring cultural knowledge. In this study, we compare their effectiveness in learning language-specific word associations and their impact on downstream cultural alignment. Next, we describe the two training approaches and tasks.

Supervised Fine-tuning We implement *the word association prediction task* directly in the supervised fine-tuning (SFT) framework.[4] 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.[5]

PPO Training We formulate PPO training as a ranking task, motivated by the observation that certain associations are more commonly produced than others in human word association data. Given a cue word, the model is tasked with ranking a list of associated words according to their relative prominence based on frequency in the SWOW dataset. Formally, each training example is represented as $x = \langle c, \mathbf{w} \rangle$, where c is the cue word and $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$ is a list of candidate associated words. The ground-truth ranking is denoted as $\mathbf{r} = \langle r_1, r_2, \ldots, r_n \rangle$, where r_i indicates the empirical rank of word w_i based on human association frequency.

We implement a reward function that reflects human preferences by prioritizing associations that appear more frequently in human responses. Specifically, we compute the Spearman rank correlation between the model's predicted rankings and the ground-truth ranks from SWOW to determine the reward. This reward signal guides policy updates via Proximal Policy Optimization (PPO), encouraging the model to produce association rankings that better align with human judgments[6].

4 Association-level Evaluation

We test whether fine-tuning taught the models human-like word associations. For this, we run two complementary evaluations: **Intrinsic**: Generation accuracy for SFT models and ranking accuracy for PPO models and **Extrinsic**: Psychological quality of generated associates, measured on valence, arousal and concreteness.

4.1 Experimental Setup

Data & prompts Each language's SWOW corpus is split **by cue** into 80 / 10 / 10; the 10 % test cues drive all evaluations, using the same prompt templates as training.

Metrics For generation we report Precision@K (overlap with human top-K). For ranking we compute Spearman ρ against human frequency ranks.

Psychological scoring Following Xiang et al. (2025), we attach ratings from large norms and compare the resulting distributions with the human baseline. English norms: Warriner et al. (13 k lemmas, 1–9 V/A) and Brysbaert et al. (40 k lemmas,

[4] We provide more details in Appendix C.

[5] The details of the hyperparameter setting for SFT are provided in Appendix F.

[6] Initially, we conducted preliminary experiments with multiple task formats to determine the most effective design for PPO training. See details in Appendix B and Appendix F

M Type	M Class	SWOW	P@5	P@10	P@40
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

Table 1: Word Association Generation for English

M Type	M Class	SWOW	P@5	P@10	P@40
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 2: Word Association Generation for Chinese

M Type	M Class	SWOW	Spearman ρ
Vanilla	Llama	-	0.241
PPO	Llama	US	0.270
Vanilla	Qwen	-	0.292
PPO	Qwen	US	**0.321**

Table 3: Results of the Ranking Task in English

M Type	M Class	SWOW	Spearman ρ
Vanilla	Llama	-	0.211
PPO	Llama	ZH	0.226
Vanilla	Qwen	-	0.291
PPO	Qwen	ZH	**0.323**

Table 4: Results of the Ranking Task in Chinese

1–5 concreteness). Chinese norms: Xu et al. (11 k V/A) and Xu & Li (9.9 k concreteness). Scales are min–max or inverted to match the English spans.

Results Tables 1–2 show generation scores. All models score higher in English than in Mandarin, and Mandarin- centric Qwen outperforms Llama on Chinese cues. Supervised fine-tuning is decisive: P@5 jumps by 16–20 % in English and 43–165 % in Mandarin. Ranking results (Tables 3, 4) show PPO gives minor gains over Vanilla but remains far below SFT.

4.2 Results on Psychological Attributes

We compute per-cue medians of Valence, Arousal, and Concreteness over each model's top-10 generated associations, pairing them with human medians via Wilcoxon tests. Tables 5 (English) and 6 (Chinese) summarize these results. **Valence**: SFT variants (**Llama-SFT**, **Qwen-SFT**) reach human level in both languages. **Valence**: SFT variants (**Llama-SFT**, **Qwen-SFT**) reach human level in both languages. **Arousal**: in English, Vanilla/PPO Llama match humans; in Mandarin only SFT does.[7]. **Concreteness**: SFT raises concreteness by $+0.20$–0.21 but remains 0.06–0.11 below human medians; non-SFT models stay more abstract. Together with the Precision@K gains, these results show that SFT not only boosts association accuracy but also aligns models with human psycholinguistic profiles.

5 Experiment 2: Cultural Value Alignment Evaluation

We have shown that fine-tuning on language-specific word associations embeds cultural patterns at the lexical level. However, the key question is whether this internalized knowledge supports higher-order reasoning about cultural values and beliefs. In RQ2, 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., 2020), focusing on the United States and China—the dominant cultures in our training data. Models are tested in the original survey languages (English for US, Chinese for China). From the original 290 WVS questions, we removed individual demographic questions (questions 260-290) and retained only questions being asked in both countries, yielding 221 questions for evaluation. We adopted the prompts that the WVS was presented to the participants.

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 re-normalize them to get the distribution of the answer options (Robinson and Wingate, 2023). We measure the alignment using the distance between the model predicted probability distribution and human answer distribution. We use two metrics that are

[7] Cue-level violin and box plots for valence, arousal, and concreteness are in Appendix G

Metric	Human	$\text{Llama}_{\text{van}}$	$\text{Llama}_{\text{ppo}}$	$\text{Llama}_{\text{sft}}$	Qwen_{van}	Qwen_{ppo}	Qwen_{sft}
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

Table 5: Emotion and concreteness scores on SWOW.EN. * Bold indicates no significant difference from human medians ($p \geq 0.05$, Wilcoxon paired test).

Metric	Human	$\text{Llama}_{\text{van}}$	$\text{Llama}_{\text{ppo}}$	$\text{Llama}_{\text{sft}}$	Qwen_{van}	Qwen_{ppo}	Qwen_{sft}
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 6: Emotion and concreteness scores on SWOW.ZH. * Bold indicates no significant difference from human medians ($p \geq 0.05$, Wilcoxon paired test).

used separately in prior work (Durmus et al., 2024; Zhao et al., 2024) to calculate the distance: (a) **Jensen-Shannon distance** and (b) **Earth Mover's distance**, which is location-aware—it can't tell whether a prediction is "almost right" or "very wrong" if both are confident and equally wrong. For a finer comparison, we also measure the alignment by calculating the percentage of questions with distances below different thresholds. We use the language that is aligned with the target culture to prompt the language models (Chinese for both the World Values Survey questionnaire and the models trained on Chinese SWOW).[8]

Approaches We use Vanilla models as our **baseline** to understand to what extent the models are currently aligned with the specified culture. We apply the same prompts as the Vanilla models on our fine-tuned models to understand how the impact of fine-tuning on word associations does on transferring the cultural values encoded. We also included 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

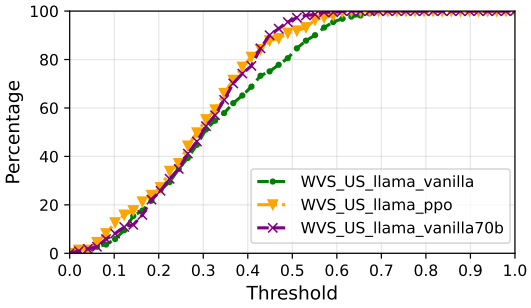
Table 7 presents our experimental results on the World Values Survey. We observe that Vanilla models exhibit different degrees of cultural alignment with the target populations. In the **US setting (English)**, the Llama model shows better alignment with the ground-truth human responses compared to Qwen. In contrast, under the **ZH setting (Chinese)**, the alignment trend reverses: the Qwen model outperforms Llama, achieving a notably lower alignment score. These findings align with prior work that Llama models tend to be western-value centric and less capable in understanding Chinese (Xiang et al., 2025; Aksoy, 2025), and Qwen is more Chinese-centric.

Interestingly, models trained on the **Chinese SWOW** data (i.e., SWOW.ZH) exhibit consistent and significant improvements on the Vanilla models (both Llama and Qwen). Specifically, **SWOW.zh supervised fine-tuning improves Chinese (ZH) performance** for Qwen across both met-

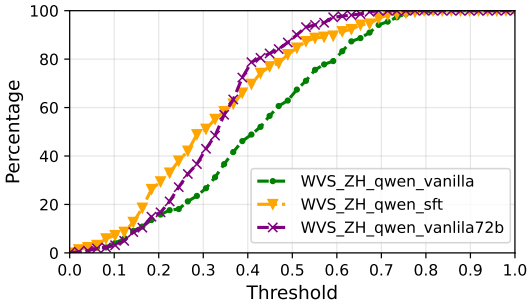
M Type	M Class	Nation	JS	EMD
Vanilla	Llama	US	0.324	0.102
SFT	Llama	US	0.392	0.114
PPO	Llama	US	**0.288***	0.092
Vanilla	Qwen	US	0.388	0.131
SFT	Qwen	US	0.355*	0.118
PPO	Qwen	US	0.353*	0.125
Vanilla	Llama3.1_70b	US	0.294*	0.094
Vanilla	Qwen2.5_72b	US	0.262*	0.109*
Vanilla	Llama	ZH	0.459	0.152
SFT	Llama	ZH	0.421*	0.129*
PPO	Llama	ZH	0.445*	0.143*
Vanilla	Qwen	ZH	0.415	0.139
SFT	Qwen	ZH	**0.325***	**0.100***
PPO	Qwen	ZH	<u>0.412</u>*	0.139
Vanilla	Llama3.1_70b	ZH	0.333*	0.100*
Vanilla	Qwen2.5_72b	ZH	0.328*	0.116*

Table 7: World Values Survey results. Top: SWOW.EN fine-tuned; bottom: SWOW.ZH fine-tuned. * indicates significant improvement over Vanilla (Wilcoxon test). Survey language matches dataset (EN for US, ZH for CN).

[8] We collected the English and Chinese WVS questionnaire from the official website https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

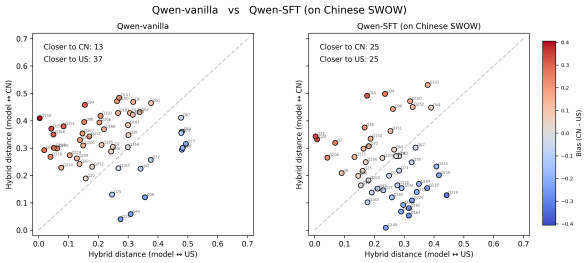


(a) WVS-US under Jensen Shannon

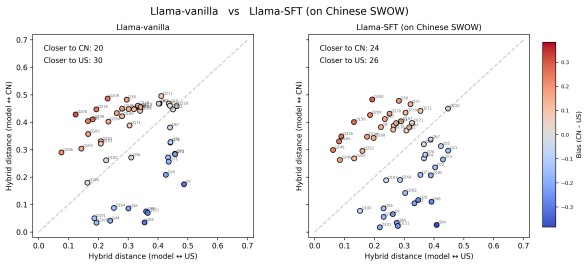


(b) WVS-ZH performance under Jensen Shannon

Figure 2: 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. 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) **Qwen-7B**



(b) **Llama-8B**

Figure 3: Comparison of shifts after SFT for **Qwen-7B** and **Llama-8B** on SWOW.ZH (ZH prompts). Each dot = one WVS question; blue (red) indicates that the question is more towards Chinese (English). Table 8 presents concrete examples that illustrate the shifts.

rics—achieving the best alignment overall. Moreover, after fine-tuning, the alignment of the Llama model towards Chinese values is closer, even better than the Vanilla Qwen model on EMD, suggesting that training on Chinese word association data steers the model towards more of the higher-level Chinese value. This highlights the cross-lingual transferability of semantic associations when the training data has a strong cultural grounding. Meanwhile, the task used to train models also matters, as we can see the improvements from PPO training are more on the US while SFT is more on Chinese.

In English, training on SWOW.EN brings significant improvements (except for SFT Llama). The best-performing model is PPO Llama, which even achieves comparable or better results than the larger 70B models. We also find that the overall degree of improvement on the US set is smaller than that on the ZH set, suggesting that English associations might provide a weaker cultural signal than Chinese associations. This might be because the models are already highly exposed to English during pre-training and less so to Chinese data, or it could be due to the greater cultural diversity in the US, which makes alignment more challenging.

Interestingly, our best-performing trained 7/8B models not only hold their ground against the much larger 70B models, but in some cases even surpass them. For English, the PPO-tuned Llama (8B) outperforms the Vanilla Qwen2.5_72B, while in Chinese, the SFT-tuned Qwen (7B) outperforms the Vanilla Llama3.1_70B. Figure 2 further illustrates how well different models align with human responses, evaluated under varying thresholds of Jensen-Shannon distance. For both US and ZH settings, we include the best-performing fine-tuned model, its vanilla counterpart, and a larger model version. In the US setting, the PPO-tuned model outperforms the vanilla model and even slightly surpasses the 70B model. In the ZH setting, the SFT model largely improved the vanilla model across thresholds. For example, at a JS distance of 0.3, only about 20% of questions are aligned for the vanilla model, compared to approximately 50% for the SFT model. Notably, the SFT model even outperforms the 72B model under stricter conditions (e.g., JS < 0.3). These results are promising, highlighting the powerful potential of culturally grounded fine-tuning as a lightweight yet effective alternative to scaling up.

Id	WVS question (full wording + choice labels)	US	CN	Qwen_{van}	Qwen_{sft}	$\text{Llama}_{\text{van}}$	$\text{Llama}_{\text{sft}}$
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 8: WVS questions where SFT on Chinese SWOW shifts Qwen's distribution toward Chinese responses. Shaded cells highlight the fine-tuned model's probabilities.

5.3 Cross-Cultural Value Alignment Evaluation

Beyond comparing a model's answers to a single culture, we examine how its responses shift from one culture toward the target culture whose word-association data it was fine-tuned on. To do so, we evaluate the model's answers with respect to both the US and China. To capture the shifts, we focus on WVS questions where Chinese and U.S. participants' responses diverge strongly. We ranked the divergence by the average of both Jensen–Shannon divergence and Earth Mover's distance and chose the top 50 most divergent questions.[9] Concentrating on such "high-tension" questions provides maximal sensitivity: even a small cultural shift in the model becomes observable, whereas questions answered similarly by both populations offer little diagnostic signal.

Results Figures 3a (Qwen-7B) and 3b (Llama-8B) present the models' prediction shifts before and after training in Chinese.[10] For each of the 50 questions, we compare the model's response distance to U.S. answers (x-axis) against its distance to Chinese answers (y-axis). For **Qwen-7B**, we find that Chinese-leaning responses increase from *13,/,50* in the Vanilla model to *25,/,50* after SFT, indicating a marked shift toward Chinese cultural preferences. For **Llama-8B**, the Vanilla model's predictions are clustered along the diagonal and skewed toward the U.S., while the SFT-tuned Llama shifts more modestly—still increasing from *20* to *24* Chinese-leaning responses, thereby reducing roughly one-third of its initial U.S. bias. Table 8 presents concrete 'before-and-after' examples with human answer distributions (US, ZH) and model prediction distribution, illustrating how supervised fine-tuning 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 can serve as a source of cultural knowledge. We develop several approaches to train models to learn cultural signals and evaluate their alignment at both the lexical level and in terms of high-level values. We find that fine-tuning mid-sized LLMs on language-specific word-association norms (English and Mandarin SWOW) yields clear improvements in both lexical and value alignment. Fine-tuned models retrieve human associations with higher precision and more closely match human valence, arousal, and concreteness 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 deep, cross-lingual cultural understanding—enhancing reasoning about values without costly retraining.

[9] The details of selecting the tension-set are provided in Appendix H.2

[10] More results in US are provided in Appendix H.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., 2024). 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 narrow selection was chosen for tractability but limits the generalizability of our findings. Given the positive results from our analysis, future work should extend to additional languages and model architectures.

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.

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.

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.

Nevan Giuliani, Cheng Charles Ma, Prakruthi Pradeep, and Daphne Ippolito. 2024. CAVA: A tool for cultural alignment visualization & analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 153–161, Miami, Florida, USA. Association for Computational Linguistics.

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.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. Care: Aligning language models for regional cultural awareness. *arXiv preprint arXiv:2504.05154*.

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.

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.

Mengze Hong, Wailing Ng, Di Jiang, and Chen Jason Zhang. 2025. Qualbench: Benchmarking chinese llms with localized professional qualifications for vertical domain evaluation. *arXiv preprint arXiv:2505.05225*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

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

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.

Meta AI. 2023. Meta LLaMA 3: Advancing Open-Source AI. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-01-10.

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*.

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.

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.

Chaoyi Xiang, Chunhua Liu, Simon De Deyne, and Lea Frermann. 2025. Comparing moral values in western english-speaking societies and llms with word associations. *arXiv preprint arXiv:2505.19674*.

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.

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 C 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:

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

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[11] 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:

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

where

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

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

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

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.

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

B 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 9 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.

C 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,
```

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 9: An example illustrating MCQ task categories.

```
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.

D 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

E Reward function details

```python
% 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)
```

F 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 10.

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 10: Hyperparameters for SFT Training

For PPO training, we used OpenRLHF library. The hyperparameters are provided in Table 11.

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 11: Hyperparameters for PPO Training

G Evaluation on the Emotions and Concreteness

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

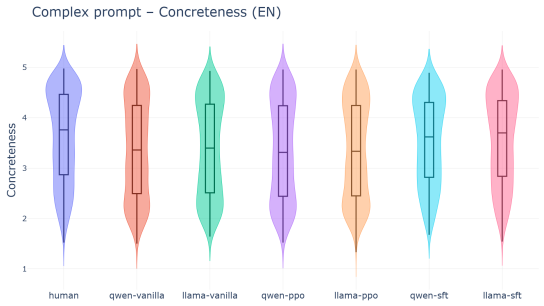
Cue-level medians are compared with a paired Wilcoxon signed-rank test to determine whether the model's lexical profile is *indistinguishable* from that of humans.

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".

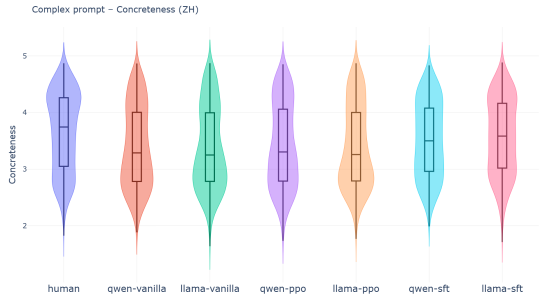
Null hypothesis H_0: $\tilde{x}_{\textbf{model}} = \tilde{x}_{\textbf{human}}$ **(assumes equality).**
Alternative H_1: $\tilde{x}_{\textbf{model}} \neq \tilde{x}_{\textbf{human}}$ **(assumes a non-zero gap).**

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

Cue-level Valence, Arousal and Concreteness (Complex prompt)

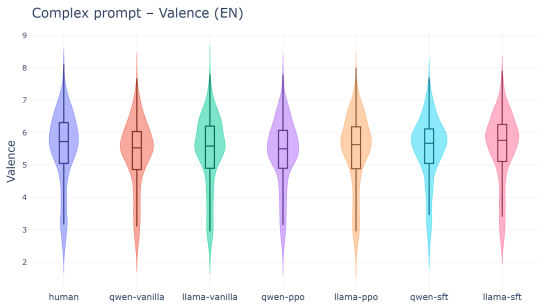


(a) English: association concreteness (1 = abstract, 5 = concrete).

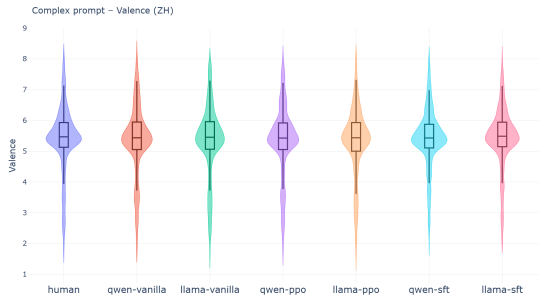


(b) Mandarin: association concreteness on the rescaled 1–5 range.

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.

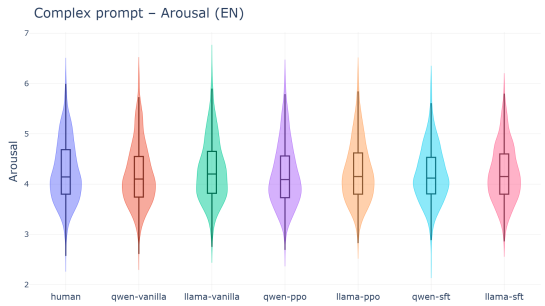
Figure 5: Violin + box plots of per-cue **valence** medians for the *Complex* prompt. Left: English (1 = unpleasant, 9 = pleasant); Right: Mandarin (rescaled to 1–9).

Metric	Human	Vanilla Llama	Llama PPO	Llama SFT	Qwen Vanilla	Qwen PPO	Qwen SFT
Valence	5.514	**5.495***	**5.489***	5.572	**5.543***	**5.544***	**5.614***
Arousal	4.244	4.292	4.281	**4.276***	**4.247***	**4.250***	**4.181***
Concreteness	3.644	3.478	3.460	**3.573***	3.419	3.415	3.762
Emotional %	84.6%	80.0%	78.7%	69.3%	80.6%	80.1%	72.3%
%Conc | %Abs | %Unk	64.3/29.8/5.9	53.9/35.3/10.7	52.5/35.5/12.0	51.6/26.9/21.5	53.3/36.7/10.0	53.5/36.5/10.1	59.1/23.4/17.4

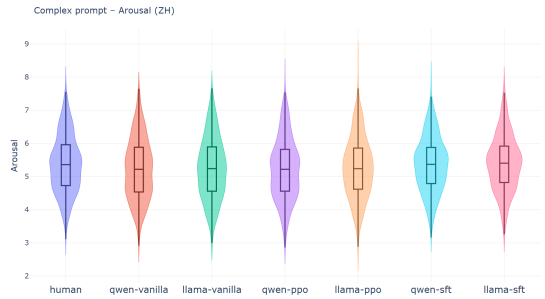
Table 12: Emotional and concreteness metrics for the **Simple** prompt on U.S. SWOW English. * Bold cells indicate no significant difference from human medians ($p \geq 0.05$, Wilcoxon paired test).

Metric	Human	Vanilla Llama	Llama PPO	Llama SFT	Qwen Vanilla	Qwen PPO	Qwen SFT
Valence	0.290	**0.316***	**0.320***	0.348	0.377	0.367	**0.327***
Arousal	2.189	2.164	**2.174***	**2.198***	**2.173***	**2.172***	**2.183***
Concreteness	2.343	2.449	2.479	2.423	2.572	2.541	2.429
Emotional %	53.3%	35.4%	38.9%	37.3%	52.0%	52.7%	40.2%
%Conc | %Abs | %Unk	35.9/15.8/48.3	22.8/11.4/65.8	24.9/12.8/62.4	24.5/11.8/63.7	30.7/19.7/49.6	31.7/19.5/48.7	25.7/13.0/61.3

Table 13: Emotional and concreteness metrics for the **Simple** prompt on Mandarin SWOW. * Bold cells indicate no significant difference from human medians ($p \geq 0.05$, Wilcoxon paired test).



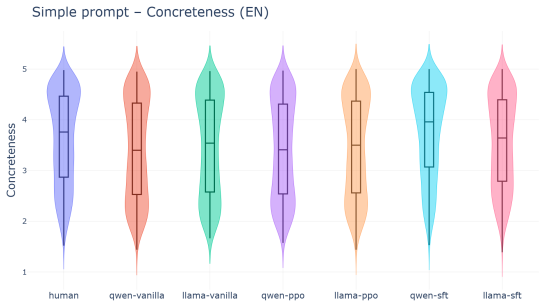
(a) English: association **arousal** (1 = calm, 9 = excited).



(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).

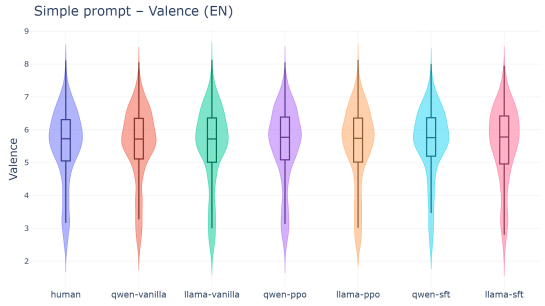
Cue-level Valence, Arousal and Concreteness (Simple prompt)



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

(b) Mandarin: concreteness, rescaled to 1–5.

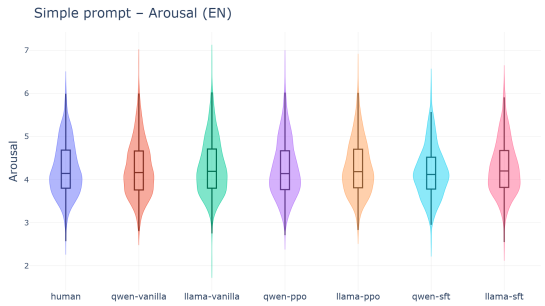
Figure 7: Violin + box plots of per-cue **concreteness** medians for the *Simple* prompt. Left: English (1 = abstract, 5 = concrete); Right: Mandarin (rescaled to 1–5).



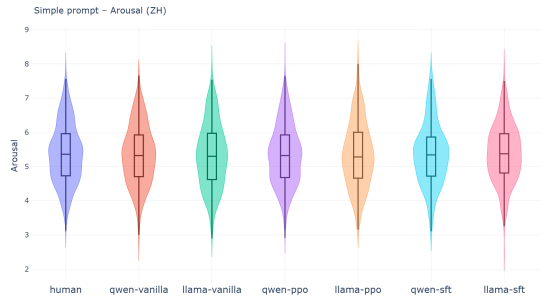
(a) English: **valence** (1 = unpleasant, 9 = pleasant).

(b) Mandarin: valence, rescaled to 1–9.

Figure 8: Violin + box plots of per-cue **valence** medians for the *Simple* prompt. Left: English (1 = unpleasant, 9 = pleasant); Right: Mandarin (rescaled to 1–9).



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

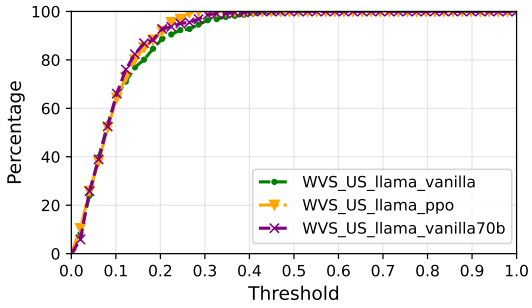


(b) Mandarin: arousal, rescaled to 1–9.

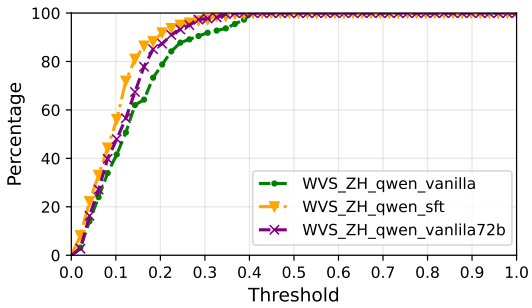
Figure 9: Violin + box plots of per-cue **arousal** medians for the *Simple* prompt. Left: English (1 = calm, 9 = excited); Right: Mandarin (rescaled to 1–9).

H Evaluation results on the world values survey

H.1 Breakdown results on EMD



(a) WVS-us under Jensen Shnnon



(b) WVS-zh performance under Jensen Shannon

Figure 10: 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.

H.2 Tension Set Selection

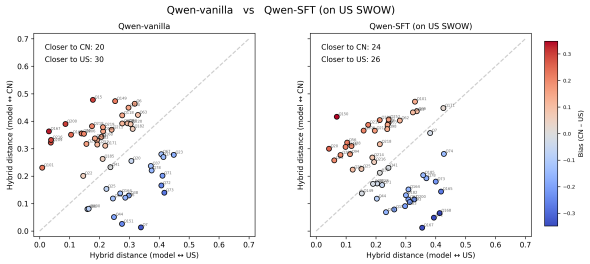
Given the participants' answer distributions for China (q) and the United States (p), we first normalise each to a probability vector i.e. we divide each count by the total number of respondents for that question so the values now represent probabilities (fractions between 0 and 1). Divergence is then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

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

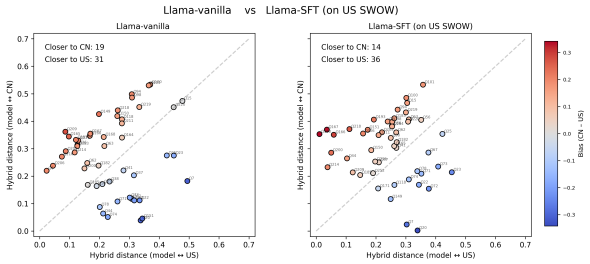
Sorting the WVS questions by this score and retaining the top 50 yields our fixed *tension set*.

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

Beyond Mandarin prompts, we also evaluate cultural shifts with English prompts. Figures 11a and 11b mirror the same layout used for Chinese



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



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

Figure 11: Shifts after SFT on US SWOW (EN prompts). Each dot = one WVS question; colour = bias (CN–US).

prompts: hybrid distances to U.S. answers (x-axis) and Chinese answers (y-axis) are plotted across 50 high-tension WVS questions.

- **Qwen-7B.** The vanilla model already exhibits strong alignment with U.S. responses; fine-tuning on U.S. SWOW slightly reduces this alignment (from 30 to 26 U.S.-aligned points).
- **Llama-8B.** Supervised fine-tuning increases U.S. alignment, shifting the number of U.S.-aligned points from 31 to 36.

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

H.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 12 and 13 (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.

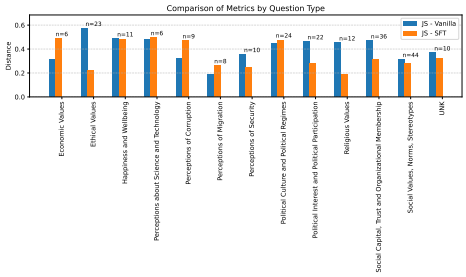


Figure 12: Jensen–Shannon distance by WVS topic (Vanilla vs. SFT Qwen-7B on ZH prompts).

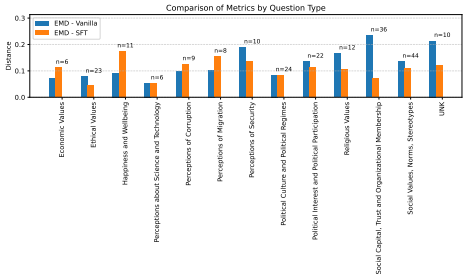


Figure 13: Earth Mover's distance by WVS topic (Vanilla vs. SFT Qwen-7B on ZH prompts).

Table 14 presents concrete examples of distribution shifts from the vanilla Qwen-2.5 model to the SFT Qwen-2.5 model. For example, in the 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.

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