From Word to World: Evaluate and Mitigate Culture Bias in LLMs via Word Association Test

Xunlian Dai[1], Li Zhou[1]*, Benyou Wang[1], Haizhou Li[1,2]
[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Research Institute of Big Data
225040015@link.cuhk.edu.cn; lizhou21@cuhk.edu.cn

Abstract

The human-centered word association test (WAT) serves as a cognitive proxy, revealing sociocultural variations through culturally shared semantic expectations and implicit linguistic patterns shaped by lived experiences. We extend this test into an LLM-adaptive, free-relation task to assess the alignment of large language models (LLMs) with cross-cultural cognition. To address culture preference, we propose *CultureSteer*, an innovative approach that moves beyond superficial cultural prompting by embedding cultural-specific semantic associations directly within the model's internal representation space. Experiments show that current LLMs exhibit significant bias toward Western (notably American) schemas at the word association level. In contrast, our model substantially improves cross-cultural alignment, capturing diverse semantic associations. Further validation on culture-sensitive downstream tasks confirms its efficacy in fostering cognitive alignment across cultures. This work contributes a novel methodological paradigm for enhancing cultural awareness in LLMs, advancing the development of more inclusive language technologies.[1]

1 Introduction

Word associations, often perceived as commonplace neural activities (Pulvermüller, 1999; Anderson et al., 2017; Khanna et al., 2024), are deeply rooted in lived environments and personal experiences. The neurocognitive research (Schneider et al., 2024) find out that transient human associations reflect not only situational factors but also ingrained in sociocultural identities. Multilingual and cross-cultural WAT also unveil how culture modulates perceptual and interactive patterns (Szalay and Deese, 2024; Garimella et al., 2017, 2016).
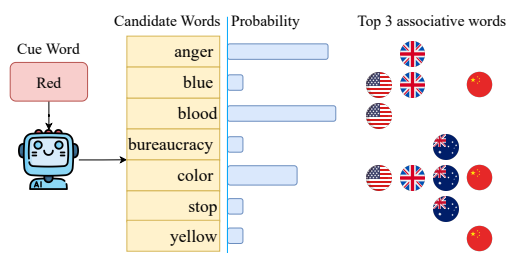

Figure 1: Cross-cultural comparison of word associations for the cue "Red" between LLM predictions and human responses.

For instance, as shown in Figure 1, when prompted with "red," beyond general associative words such as "blue" and "color", which reflects semantic relations, British respondents tend to activate idiomatic expressions like *see red* (**anger**), while Australians are more likely to mention bureaucratic metaphors such as *red tape*(**bureaucracy**). This cross-cultural divergence in word associations reveals rich cognitive-semantic information embedded within linguistic conditioning.

Multilingual LLMs exhibit persistent cultural biases, ranging from stereotypical associations (Abid et al., 2021; Bano et al., 2025) to value misalignments (Masoud et al., 2023; Cao et al., 2023a; Jiang et al., 2024). Prior research has largely concentrated on addressing these biases by aligning multicultural cognition through prompt-based methods (Choenni and Shutova, 2024; Wang et al., 2024; Sato et al., 2024), which depend on explicitly providing cultural context to elicit targeted responses. On the other hand, fine-tuning approaches (Xu et al., 2024; Yao et al., 2025) are knowledge-driven but still require explicit cultural settings during inference.

Building on this perspective, our approach follows the opinion that cultural competence in NLP is often best evaluated through diagnostic tasks that

---

reveal latent cultural patterns (Zhou et al., 2025b). By simulating human word association tasks, we directly capture latent cultural relationships and assess the semantic spaces of cultural preferences within LLMs themselves. Extending beyond existing word association datasets, our work adapts these implicit, word-level cultural patterns to the evaluation of LLMs. Our **contributions** can be summarized as follows:

- We design an LLM-adaptive, free-relation word association task and a quantitative evaluation metric to assess the cross-cultural cognitive abilities of LLMs (§3).

- We introduce *CultureSteer*, an innovative approach that integrates a culture-aware steering mechanism to guide semantic representations toward culturally specific spaces (§ 4).

- Our experiments show that LLMs exhibit cultural bias in word association tasks, while our model is able to effectively mitigate this bias, outperforming prompt-based methods (§ 5).

- Further analysis demonstrates the effectiveness of the culture-aware steering mechanism and the generalizability of the model (§ 6).

2 Related Work

Cultural Bias and Alignment in LLMs Extensive research has investigated cultural biases in LLMs across domains such as value preference (Benkler et al., 2023; Cao et al., 2023a; Jiang et al., 2024; Durmus et al., 2023), knowledge perception (Palta and Rudinger, 2023; Shen et al., 2024), and moral measurement (Jinnai, 2024; Rao et al., 2023, 2024), revealing persistent Western-centric value preferences and regionally divergent commonsense knowledge. To improve cultural perception, current approaches predominantly employ prompt engineering (Wang et al., 2024; AlKhamissi et al., 2024; Zhou et al., 2025a; Masoud et al., 2025), explicitly embedding cultural context in prompts to evoke culturally sensitive responses. Alternatively, fine-tuning methods such as Culture-Bank (Shi et al., 2024), CultureLLM (Li et al., 2024b), CultureSPA (Xu et al., 2024) and Sim-LLMCultureDist (Cao et al., 2025a) have been proposed to enhance cultural alignment in different ways. However, these methods rely on survey data (particularly from word value surveys), data-driven SFT, and still require culture-specific prompts for downstream tasks.

Word Association in LLMs Prior to the LLM era, previous studies have also revealed cultural preferences from the perspectives of word usage and associations (Garimella et al., 2016), and have attempted to use models, such as the Composite Skip-gram Model (C-SGM) (Garimella et al., 2017), to predict the word association process.

On the other hand, recent work adopts word association tasks to evaluate the capabilities of LLMs across various domains. For instance, uncovering gender stereotypes (Abramski et al., 2024b) and assessing color associative perception (Fukushima et al., 2024). These studies primarily focus on word association mappings under specific relationships, emphasizing explicit word generation and partially relying on manual interpretation during evaluation (Abramski et al., 2024a; Vintar et al., 2024).

In contrast, we not only extend word association datasets to four cultures, going beyond the two cultures (Chinese and English) which is focused on in the pre-LLM era, but also propose metrics that focus on the predicted probabilities of association words without manual intervention, enabling a more comprehensive evaluation of LLMs' associative capabilities.

Steering Control of LLMs Steering methods aim to control the internal activation states of LLMs at inference time to influence the generation style of their outputs (Bo et al., 2025). These have been applied to reduce harmful responses (Arditi et al., 2024; Cao et al., 2024) and perform style transfer (Song et al., 2025; Konen et al., 2024). Our approach introduces a steerable layer that modulates associative outputs to reflect culture-specific cognitive patterns in its lexical associations.

3 Word Association Test Task

Word Association Test (Woodworth and Wells, 1911; Gough, 1976) is a psychological assessment that reveals subconscious thoughts, emotional states, or cognitive patterns through an individual's quick associative responses to stimulus words. The cultural background influences the associated words and response patterns, with different cultures often producing vastly different associations to the same stimulus, reflecting the profound impact of culture on language, thinking, and emotions.

3.1 Human-centered task

The WAT for human participants is conceptualized as an open-ended activity in a free-relation format, allowing for flexible and diverse associations that more accurately reflect human cognitive processes.[2] Before conducting the WAT, demographic information, such as age, gender, education, native language, and region, is collected for each participant. Given a cue word w, the participant u is asked to provide the first word that comes to mind and may optionally add a second and third word sequentially, resulting in up to three associated words, represented as $u_w^c = \{a_1^u, a_2^u, a_3^u\}$, where c denotes the participant's cultural background. Within this cultural context, the collective set of associative words for a given cue word w, aggregated across all participants $U_w^c = \{u_w^c\}$, is represented as $A_w^c = \{(a_1, f_1), \cdots, (a_n, f_n)\}$. Here, $f_i \geqslant f_{i+1}$, and f_i represents the frequency with which the associative word a_i appears in U_w^c.

3.2 LLM-adaptive task

Directly employing the human-centered task paradigm to conduct WAT with LLMs proves insufficient for effectively capturing variations across entire cultural groups. Inspired by Fierro et al. (2024) and Zhou et al. (2025a), we adapt this task into a word prediction framework, designed as a retrieval process for candidate associative words. Specifically, we employ a template q, such as "When {cue_word} is mentioned, people often associate it with the following words:", to prompt the LLM and elicit its predictions of associated words. The input is represented as $x = q(w)$, and the output is denoted as $o = M(q(w))$, where the generalized function $M(\cdot)$ represents the inference process of the LLM.

Unlike the human-centered task, which gathers associative words directly from participants, our method computes the probability of each candidate associative word being predicted by the LLM and ranks them based on their relative probabilities.[3] To address potential uninformative prefixes in the LLM's next-token predictions (e.g., phrases like 'I think'), we move beyond relying solely on the likelihood of a candidate associative word appearing as the immediate next token. Instead, we compute its probability within a fixed window of next-token

predictions. By evaluating the candidate word's probability at each position within the prediction window, we select the maximum value as its final probability, $P_w(a_i)$. The mathematical formulation of this process is as follows:

$$P_w(a_i) = \max_{m \in [0, k-t]} \left(\frac{1}{t} \sum_{j=0}^{t-1} p\left(\hat{y}_{m+j} = a_i^j\right) \right) \tag{1}$$

where k is the size of the fixed prediction window, t is the number of sub-tokens in the candidate word a_i, a_i^j is the j-th sub-token, and $\hat{y}_{m+j}$ is the token predicted at position $m + j$. Based on these probabilities, we construct the predicted ranked list of associative words, denoted as $\hat{A}_w = [\hat{a}_0, \hat{a}_1, \dots]$, where $P_w(\hat{a}_i) \geqslant P_w(\hat{a}_{i+1})$.

3.3 Evaluation metric

To evaluate the alignment between human cultural associations A_w^c and the LLM's predicted associations $\hat{A}_w$, we propose a novel evaluation metric, Position-Weighted Recall (PWR@K), which is an extension of Top-K Recall (R@K).

Baseline Metric Given that WAT in LLMs is a retrieval task involving a large number of truly relevant associative words and a ranking process close to full ranking, we adopt R@K as the baseline evaluation metric, emphasizing full coverage. Formally, it is defined as:

$$\begin{aligned} \text{R@K} &= \frac{\sum_{i=1}^{K} \mathbb{I}(\hat{a}_i \in A_w^c)}{N} \\ &= \frac{\sum_{i=1}^{N} \mathbb{I}\left(a_i \in \text{Top-K}\left(\hat{A}_w^c\right)\right)}{N} \end{aligned} \tag{2}$$

where $N = |A_w^c|$ denotes the total number of truly relevant associative words, and $\mathbb{I}(\cdot)$ is the indicator function.

Proposed Metric While R@K effectively measures coverage, it fails to account for ranking order. In our task, higher-ranked words in A_w^c signify stronger associative strength and thus hold greater importance. To address this limitation, we propose PWR@K, which incorporates positional weighting by assigning higher weights to words ranked earlier, specifically using the inverse of their position as the weight. It is formally defined as:

$$\text{PWR@K} = \frac{\sum_{i=1}^{N} \frac{1}{i} \cdot \mathbb{I}(a_i \in \text{Top-K}(\hat{A}_w))}{\sum_{i=1}^{N} \frac{1}{i}} \tag{3}$$

---

[2] https://smallworldofwords.org
[3] This can be viewed as an open-ended multi-choice question-answering task without predefined options.
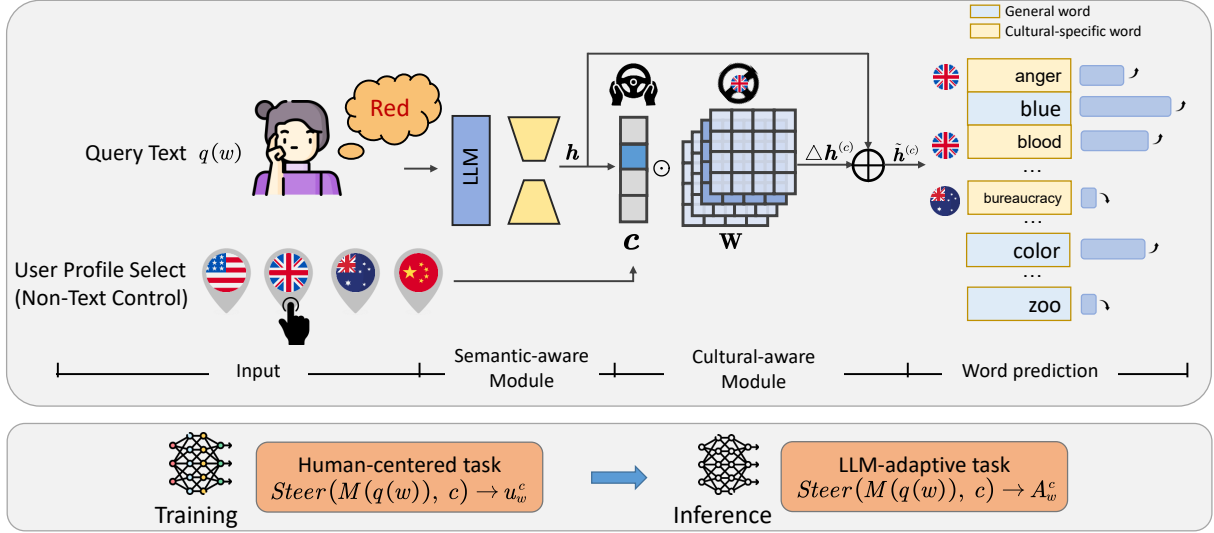
Figure 2: The framework of *CultureSteer* model. Top: model pipeline; Bottom: training and inference process based on WAT.

where the denominator normalizes the positional weights across all N positions.

A detailed comparison with other position-weighted metrics such as DCG@K is provided in Appendix B.

4 *CultureSteer*: Modeling and Controlling Cultural Awareness in LLMs

Inspired by the findings of LM-Steer (Han et al., 2024), which suggest that, with certain assumptions, shifting styles in language models is equivalent to a linear transformation in the word embedding space, we extend this perspective to the cultural domain. Specifically, we hypothesize that different cultures define distinct semantic association spaces shaped by cultural preferences, which can be modeled through culture-specific linear transformations within the embedding space. Based on this, we introduce the *CultureSteer* model, which leverages both human-centered and LLM-adaptive WAT tasks to (1) equip LLMs with human-like word association perception, thereby enhancing their cognitive abilities, and (2) enable the word association capabilities of LLMs to align across different cultures. The overall architecture of the model is illustrated in Figure 2.

Cultural Control Paradigms Previous approaches primarily control cultural preferences explicitly at the input level by defining the cultural context c through textual prompts. The output in these methods can be denoted as:

$$o = M\left(q\left(w, c\right)\right) \qquad (4)$$

In contrast, our model performs post-processing on the LLM's output, mapping the semantic representation space to a specific cultural perception space. This operation is represented as:

$$o = \text{Steer}\left(M\left(q(w)\right), c\right) \qquad (5)$$

Cultural-Aware Steering Mechanism Building on this paradigm, we detail the *CultureSteer* framework, focusing on how semantic representations are steered into culturally-specific spaces. First, we obtain the semantic representation $\boldsymbol{h} = M\left(q\left(w\right)\right)$, which encodes the general meaning without any culture-specific preferences. To model culture-specific semantic associations, we apply a transformation to the base semantic representation $\boldsymbol{h}$. Specifically, the cultural preference adjustment $\triangle\boldsymbol{h}^{(c)}$ is computed as:

$$\triangle\boldsymbol{h}^{(c)} = \boldsymbol{h} \cdot (\boldsymbol{c} \odot \mathbf{W}) = \boldsymbol{h} \cdot \mathbf{W}_c \qquad (6)$$

Here, $\boldsymbol{c} \in \mathbb{R}^{|C|}$ is the cultural control vector, where each element corresponds to a specific culture. In our work, we use a one-hot encoding for different cultural contexts. The operation $\boldsymbol{c} \odot \mathbf{W}$ selects the relevant cultural subspace from $\mathbf{W}$ called $\mathbf{W}_c$, which contains learnable parameters that encode culture-specific semantic directions. The final culture-aware representation is then computed as:

$$\tilde{\boldsymbol{h}}^{(c)} = \boldsymbol{h} + \epsilon \triangle \boldsymbol{h}^{(c)} \qquad (7)$$

where ϵ is a scaling factor that modulates the influence of the culture-specific adjustment.[4] The

---
[4]We set $\epsilon = 1e-3$ as the default steering value (Han et al., 2024).

24513

adjusted representation $\tilde{\boldsymbol{h}}^{(c)}$ is then used to predict the associated words for a given query, considering the specified cultural context.

Training and Inference Process During training, we use the human-centered WAT pattern to train the *CultureSteer* model SFT, where the true human associations are fed back into the model at each time step t. The loss function, which minimizes the discrepancy between the predicted associations $\tilde{a}_t^u$ and the true human responses a_t^u. used is cross-entropy During the inference phase, we employ the LLM-adaptive WAT task to evaluate the LLM's cross-cultural word association ability, using the proposed metric PWR@K.

5 Experiment

5.1 Dataset Creation

Data Source We utilized the officially released datasets from the Small World of Words Project (SWOW)[5]. For our study, we selected data in two languages: English and Mandarin Chinese, referred to as SWOW-EN (De Deyne et al., 2019) and SWOW-ZH (Li et al., 2024a), respectively. The SWOW-EN dataset comprises over 3 million responses collected from more than 90,000 participants across various English-speaking countries or regions, covering more than 12,000 cues. Moreover, the SWOW-ZH dataset contains over 2 million responses corresponding to 10,192 cues, with participants primarily from the Chinese mainland.

Culture Selection To ensure sufficient participant representation across different cultures, we first filter the three most prominent cultural groups from SWOW-EN: the United States (USA), the United Kingdom (UK), and Oceania (OC)[6]. Additionally, SWOW-ZH serves as the representative for China (CN). As a result, we construct datasets representing four distinct cultural groups: USA, UK, OC, and CN.

Data Preprocessing To ensure fairness in evaluating word association perception across different cultures, we utilize the Intercontinental Dictionary Series (IDS) (jung Yu and Wang, 2023) to align the cue words between English and Chinese, ultimately retaining 881 cue words that are common across all

four cultural groups. Given the substantial differences in participant numbers across cultures, which lead to significant variation in the number of associated words (A_w^c), we standardize the A_w^c lists by truncating them to the size of the culture with the fewest associated words. Using the IDS framework, we categorize these cue words into 22 semantic categories based on their conceptual meanings. Examples of the final preprocessed data and statistical details are provided in Appendix C.2, Table 4 and Table 5. Finally, we split the cue words in each semantic category into 70% for training and 30% for testing.

5.2 Experimental Setup

We conduct experiments using Llama3.1-8B (Grattafiori et al., 2024) and Qwen2.5-7B (Team, 2024) to simulate the WAT. First, we evaluate the base versions of these models on word association tasks using a base prompt without any cultural conditioning, aiming to assess their human-like associative cognition and the default cultural representation they exhibit. Subsequently, we build *CultureSteer* based on these two models to enhance their cross-cultural perception capabilities. For the LLM backbone, we adopt LoRA (Hu et al., 2022) and train the models on a single A100 GPU.

Furthermore, to comprehensively evaluate the effectiveness of our approach, we employ both prompting-based strategies and cultural-related LLMs as baseline methods for comparison.**1)** Explicit Culture-Aware Prompting Strategies: We implement two prompting approaches to examine how explicit cultural conditioning affects model performance: First, Culture-Specific Prompt (CSP) (Xu et al., 2025) and Cross-Culture Think (CCT). Detailed prompt templates and settings for all approaches are provided in Appendix C.1. **2)** Cultural-Related LLMs: To establish comprehensive baselines, we compare against four cultural-aware language models. These include CultureSPA, Culture-merge[7], SimLLMCultureDist (Cao et al., 2025b) and CultureLLM (Li et al., 2024b)[8].

5.3 Overall Results

LLMs Exhibit Cultural Biases in the WAT Task
The overall comparison results are presented in

---

[5]Designed as a human-centered Word Association Task (WAT).

[6]OC includes participant responses from Australia and New Zealand, which are combined due to their relatively participant numbers and shared historical background.

[7]https://huggingface.co/surbhi21/llama3-8b-cultural-merged

[8]Originally trained on LLaMA2-70B, we migrate the approach to LLaMA 3.1-8B to ensure fair comparison.

		Llama				Qwen			
	PWR@K	K=3	K=5	K=10	K=20	K=3	K=5	K=10	K=20
USA	Baseline	12.37	19.57	29.25	42.54	15.22	19.05	26.18	35.90
	CultureLLM	15.32	19.89	25.10	27.76	6.01	7.48	10.34	16.15
	SimLLMCultureDist	22.30	30.94	42.20	50.73	16.83	21.89	30.84	42.05
	CultureMerge	20.29	29.24	40.33	49.07	-	-	-	-
	CultureSPA	19.30	27.73	40.93	49.71	-	-	-	-
	CSP	30.30	37.61	48.95	60.95	20.34	28.19	39.05	50.65
	CCT	25.92	32.98	44.05	55.54	22.59	30.51	42.31	54.76
	CultureSteer	**38.14**	**50.30**	**67.58**	**80.13**	**36.34**	**46.06**	**59.78**	**72.05**
UK	Baseline	10.69	16.53	24.96	35.92	12.43	16.16	22.05	29.97
	CultureLLM	12.68	16.33	20.65	23.19	5.68	7.14	9.46	14.06
	SimLLMCultureDist	19.24	26.28	35.00	41.43	14.90	19.27	26.04	34.53
	CultureMerge	17.34	24.95	33.58	41.04	-	-	-	-
	CultureSPA	16.46	23.43	33.53	40.96	-	-	-	-
	CSP	24.31	30.64	39.91	49.27	16.43	22.14	30.58	39.86
	CCT	21.58	27.74	35.70	44.96	18.17	24.27	33.54	42.72
	CultureSteer	**29.19**	**38.84**	**51.45**	**61.88**	**28.23**	**34.88**	**46.10**	**57.53**
OC	Baseline	9.53	15.62	23.25	33.67	11.87	15.07	21.35	28.41
	CultureLLM	13.13	16.82	20.29	22.11	5.15	5.97	8.25	13.45
	SimLLMCultureDist	16.63	22.67	31.10	37.80	13.16	17.45	24.10	32.88
	CultureMerge	15.99	22.64	30.64	37.83	-	-	-	-
	CultureSPA	14.42	21.20	31.35	37.78	-	-	-	-
	CSP	23.17	28.62	36.78	45.88	15.58	21.21	29.02	37.60
	CCT	20.21	25.64	33.37	42.41	17.16	22.67	31.61	40.63
	CultureSteer	**27.56**	**35.44**	**47.65**	**58.77**	**25.05**	**32.22**	**41.55**	**52.22**
CN	Baseline	7.31	11.03	18.43	24.01	8.61	11.65	17.16	24.11
	CultureLLM	0.73	0.86	1.01	1.05	4.20	6.87	10.56	14.61
	SimLLMCultureDist	8.42	13.33	19.09	24.72	11.78	16.32	23.82	34.98
	CultureMerge	10.30	15.11	21.45	25.77	-	-	-	-
	CultureSPA	10.10	14.75	20.57	26.35	-	-	-	-
	CSP	10.52	15.83	23.34	29.81	15.56	21.59	31.77	42.81
	CCT	9.71	14.08	21.02	28.00	12.32	17.71	26.57	36.85
	CultureSteer	**12.77**	**19.04**	**27.26**	**33.53**	**21.78**	**30.13**	**44.88**	**58.76**
Average	Baseline	9.98	15.69	23.97	34.04	12.03	15.48	21.69	29.60
	CultureLLM	10.47	13.48	16.76	18.53	5.26	6.87	9.65	14.57
	SimLLMCultureDist	16.65	23.31	31.85	38.67	14.17	18.73	26.20	36.11
	CultureMerge	15.98	22.99	31.50	38.43	-	-	-	-
	CultureSPA	15.07	21.78	31.60	38.70	-	-	-	-
	CSP	22.08	28.18	37.25	46.48	16.98	23.28	32.61	42.73
	CCT	19.36	25.11	33.54	42.73	17.56	23.79	33.51	43.74
	CultureSteer	**26.92**	**35.91**	**48.49**	**58.58**	**27.85**	**35.82**	**48.08**	**60.14**

Table 1: Overall comparison results: (1) LLMs display cultural biases in the WAT task; (2) *CultureSteer* enhances LLMs' cultural awareness and achieves the best performance.

Table 1, where the " Baseline " denotes the default cultural preferences exhibited by the LLMs. Both models consistently reveal a cultural preference pattern: USA > UK > OC > CN.[9] This suggests that, in terms of word-level cognitive abilities, LLMs exhibit a stronger inclination toward American culture, aligning with findings from prior studies (Cao et al., 2023b; Myung et al., 2024; Zhou et al., 2025a). Overall, Llama demonstrates stronger human-like word-level cognitive abilities compared to Qwen, primarily due to its significantly better performance in English-speaking cultural contexts. In contrast, the performance gap between the two models is relatively small in CN.

Interestingly, when evaluating PWR@3, Qwen outperforms Llama across all cultural contexts, suggesting that Qwen excels at predicting highly associated words. However, as K increases, Llama surpasses Qwen, indicating that Llama possesses a more comprehensive associative cognition across a broader range of related words.

***CultureSteer* Enhance LLM Cultural Awareness**
At the methodological level, *CultureSteer* consistently outperforms both prompt-based methods and culture-related models across all cultural settings while some models performed worse after fine-tuning. This demonstrates that the model does not rely on explicit textual configurations but also showcases the superiority of our implicit cultural steering mechanism over traditional fine-tuning approaches that may suffer from overfitting or cul-

[9]To demonstrate that performance differences are due to cultural factors rather than the language used in prompts, we conducted an ablation study. Details are provided in Appendix D.
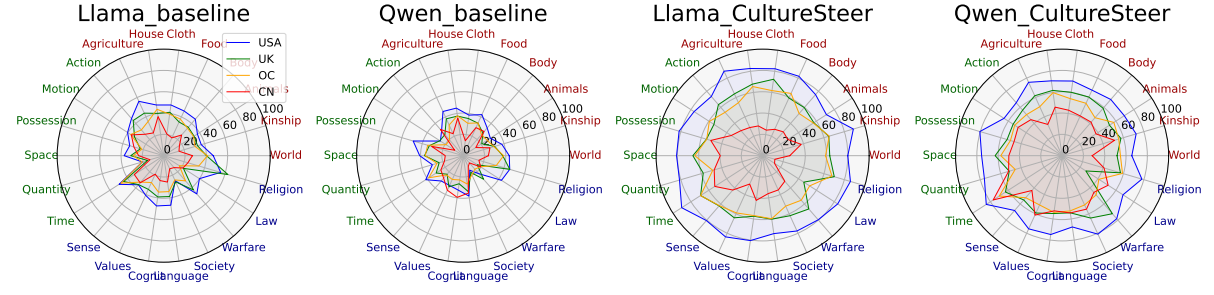
Figure 3: Fine-grained performance comparison across 22 semantic classes in the test set with PWR@20. Red denotes Global Knowledge, green denotes Perceptual Experience, and blue denotes Cultural Ideologies. Other PWR@K (K=3, 5, 10) results are shown in Appendix E.

tural knowledge interference. Interestingly, for the Llama model, the CSP method consistently outperforms the CCT method across all cultural contexts. In contrast, for the Qwen model, the CSP method performs better only in the Chinese cultural context, while the CCT method excels in the other three, particularly in English-speaking settings. This suggests that Llama is more effective with culture-specific prompts tailored to a single context, whereas Qwen demonstrates stronger cross-cultural reasoning, especially for the English-speaking environments.

5.4 Fine-Grained Results

We visualize the fine-grained performance comparison across 22 semantic classes in the test set, as shown in Figure 3. The results reveal notable differences among baseline models in the *Cognition* and *Value* classes, with Qwen demonstrating value orientations more aligned with Chinese culture, while Llama exhibits stronger alignment with English cultural values. Furthermore, based on their semantic meanings, these 22 classes are categorized into three groups: Global Knowledge, Perceptual Experience, and Cultural Ideologies. *CultureSteer* improves cultural awareness in both models with domain-dependent effectiveness. While global knowledge categories (e.g., *World*, *Animals*) remain difficult to align culturally, perceptual experience categories (e.g., *Action*, *Space*) show notable gains, especially within English-speaking cultures. Llama still exhibits gaps between English and Chinese contexts after steering, whereas Qwen achieves more balanced cross-cultural performance, nearing parity in cultural ideologies (*Values*, *Society*). Overall, *CultureSteer* mitigates cultural biases

and enhances associative abilities but cannot fully overcome inherent differences in models' cultural perception.

6 Analysis

6.1 Culture-Specific Semantic Learning Matters

To validate the effectiveness of our proposed cultural-aware steering mechanism, we conduct two controlled experiments: one without cultural preference semantic learning in training and one with cross-cultural semantic space steering in inference. The results are shown in Table 2. We observe a general decline in performance when the cultural semantic preference space is not learned or when a mismatched cultural semantic space is used for steering. This finding confirms the presence of cultural semantic preferences, highlights the differences across cultural semantic spaces, and underscores the importance of implicitly learning these preferences. For more detail, on LLama, no significant performance differences are observed when comparing the absence of cultural preference semantic learning with the use of mismatched cross-cultural semantic space steering. However, on Qwen, mismatched cultural semantic spaces have a negative impact on performance when evaluated at $K \leq 10$. Interestingly, as the evaluation range increases to $K = 20$, a slight improvement in performance is observed, which suggests that broader evaluation scopes may capture additional relevant associations despite the mismatch, partially offsetting the negative effects.

24516

TopK	Model	Culture	w/o $\mathbf{W}$	$\mathbf{W}_{USA}$	$\mathbf{W}_{UK}$	$\mathbf{W}_{OC}$	$\mathbf{W}_{CN}$	PWR@K	Model	Culture	w/o $\mathbf{W}$	$\mathbf{W}_{USA}$	$\mathbf{W}_{UK}$	$\mathbf{W}_{OC}$	$\mathbf{W}_{CN}$
K=3	Llama	USA	36.28	**38.14**	36.21	36.32	36.32	K=5	Llama	USA	47.96	**50.30**	47.99	48.07	47.95
		UK	27.80	27.78	**29.19**	27.96	27.84			UK	37.03	37.03	**38.84**	37.08	37.03
		OC	26.48	26.48	26.36	**27.56**	26.49			OC	33.57	33.55	33.73	**35.44**	33.56
		CN	12.23	12.23	12.23	12.21	**12.77**			CN	18.30	18.30	18.36	18.36	**19.04**
	Qwen	USA	37.14	**38.04**	34.83	34.72	34.74		Qwen	USA	47.24	**47.91**	43.87	43.83	43.89
		UK	26.45	26.72	**27.84**	26.79	26.77			UK	34.01	33.21	**35.69**	33.26	33.27
		OC	**26.82**	25.23	25.20	25.05	25.19			OC	31.35	31.97	31.97	**32.22**	32.01
		CN	21.49	19.54	19.34	19.35	**21.78**			CN	24.81	24.44	24.15	24.14	**30.13**
K=10	Llama	USA	64.43	**67.58**	64.50	64.56	64.40	K=20	Llama	USA	76.11	**80.13**	76.20	76.05	76.08
		UK	48.88	49.02	**51.45**	48.92	48.85			UK	58.87	58.97	**61.88**	58.77	58.87
		OC	45.09	45.10	44.96	**47.65**	45.05			OC	55.89	55.89	56.04	**58.77**	55.89
		CN	26.12	26.12	26.12	26.12	**27.26**			CN	32.09	32.09	32.09	32.09	**33.53**
	Qwen	USA	56.19	**61.88**	56.81	56.83	56.86		Qwen	USA	68.34	**73.94**	68.82	68.92	68.89
		UK	43.61	43.92	**46.45**	44.02	43.87			UK	54.58	54.78	**56.66**	54.74	54.78
		OC	41.19	41.28	41.14	**41.55**	41.11			OC	50.73	50.70	50.73	**52.22**	50.85
		CN	33.53	33.01	32.62	32.73	**44.88**			CN	44.34	44.51	44.59	44.57	**58.76**

Table 2: Results of controlled experiments on cultural-aware steering mechanism: performance declines without cultural preference semantic learning or with mismatched cross-cultural semantic space steering, highlighting the existence and variability of cultural semantic preferences.

		Llama		Qwen	
		Baseline	CultureSteer	Baseline	CultureSteer
WVS	USA	**53.64**	51.11	47.66	**51.43**
	UK	51.69	**53.89**	48.33	**51.57**
	OC	51.43	**53.32**	46.63	**52.35**
	CN	49.11	**53.11**	46.87	**49.01**
BLEnd	USA	46.29	**50.47**	62.73	**68.91**
	UK	43.18	**48.86**	57.89	**62.86**
	CN	41.65	**47.93**	52.84	**56.55**

Table 3: Performance Comparison on Cross-Cultural Benchmarks: WVS and BLEnD. Bold values indicate superior performance under each model. In WVS, we evaluated the Australian Q&A data as the result for OC, while BLEnD lacks data from Australia and New Zealand, so no evaluation was conducted.

6.2 Generalization on Cultural Tasks

To assess whether *CultureSteer*'s cultural awareness and cognitive modeling capabilities benefit downstream culture-related tasks, we perform experiments on two datasets: the World Values Survey (WVS) (Xu et al., 2024), which captures global perspectives on values and beliefs, and BLEnD (Myung et al., 2024), which focuses on everyday knowledge in diverse cultures. The experimental results are shown in Table 3. *CultureSteer* consistently outperforms across all cultural settings on both tasks, except for the USA culture in the WVS task, indicating that our method enhances LLMs' cultural awareness through word-level interventions. The performance gains on both WVS and BLEnD confirm that mitigating cultural bias in word association tasks establishes a foundational cognitive structure for broader cultural competence. By grounding conceptual mappings in cultural contexts through word association mechanisms, *CultureSteer* activates latent cultural perception capabilities in language models, surpassing superficial mitigation of task-specific biases. This validates that word association analysis serves as an effective entry point for enhancing the generalization capacity of multicultural LLMs.

6.3 Case Study

To gain deeper insights into how *CultureSteer* differentially influences the probability distributions of $\hat{A}_w^c$, we perform a targeted case study focusing on the cue word "red" across three English-speaking cultures: USA, UK, and Oceania (OC).[10] We compute the probability differences between *CultureSteer* and the baseline LLM version for A_{red}^c, as illustrated in Figure 4. The results reveal notable probability increases for both culture-specific words (red marks) and general words shared across cultures. This indicates that the model learns not only universal word associations

---

[10]Chinese (CN) is excluded to avoid confounding linguistic differences across language families.
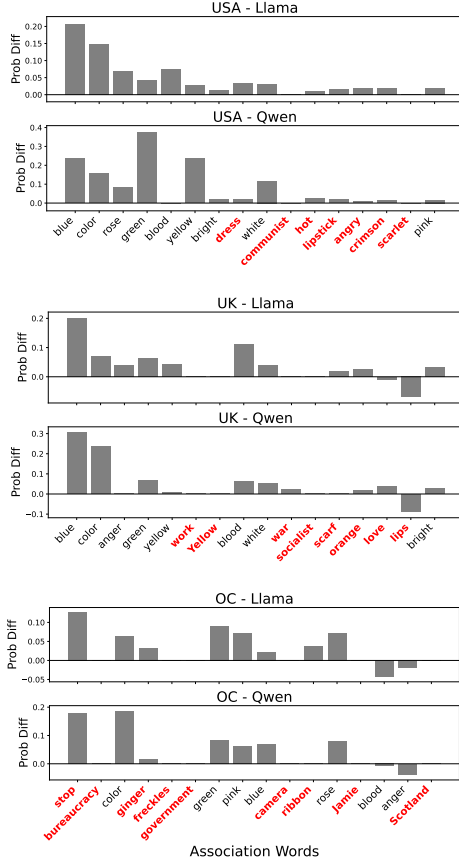
Figure 4: Probability differences between CultureSteer and vanilla models across USA, UK and OC. Red marks denote culture-specific words, while gray marks represent general words common across cultures.

but also culture-specific ones, validating the effectiveness of the *CultureSteer*.

7 Conclusion

Our research aims to utilize the insights from WAT to evaluate and mitigate cultural biases in LLMs. We proposed a novel metrics for assessing the association abilities of LLMs through word prediction task. Additionally, we introduced *CultureSteer* that leverages word association patterns to align LLMs' cultural cognition by guiding semantic representations toward culturally specific spaces and validated effectiveness of this method in downstream tasks. Although we did not explicitly prescribe the specific associative logic underlying word associations, we instead enabled the model to infer these connections through contextual prompts. Remarkably, our training methodology successfully elicited the model's cross-cultural word association capabilities. Future research could concentrate on refining these associative patterns to further aug-

ment models' cultural sensitivity and semantic comprehension.

Limitation

Our research is constrained by its focus on a limited number of languages and cultures when examining word association differences, primarily due to the reliance on word association test datasets sourced from human participants, which poses significant challenges to scalability. Moreover, our control vector mechanism treats cultures as independent entities, without adequately considering potential relationships and interactions between them. Future work could address these limitations by developing scalable methods for collecting word association data, expanding the linguistic and cultural scope, and incorporating inter-cultural connections to provide a more comprehensive and nuanced perspective.

Acknowledgments

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Katherine Abramski, Riccardo Improta, Giulio Rossetti, and Massimo Stella. 2024a. The" llm world of words" english free association norms generated by large language models. *arXiv preprint arXiv:2412.01330*.

Katherine Abramski, Clara Lavorati, Giulio Rossetti, and Massimo Stella. 2024b. Llm-generated word association norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 3–12. IOS Press.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. 2017. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Muneera Bano, Hashini Gunatilake, and Rashina Hoda. 2025. What does a software engineer look like? exploring societal stereotypes in llms. *arXiv preprint arXiv:2501.03569*.

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075*.

Jessica Y Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. 2025. Steerable chatbots: Personalizing llms with preference-based activation steering. *arXiv preprint arXiv:2505.04260*.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025a. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025b. Specializing large language models to simulate survey response distributions for global populations. *arXiv preprint arXiv:2502.07068*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023a. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023b. 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.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Rochelle Choenni and Ekaterina Shutova. 2024. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *arXiv preprint arXiv:2408.16482*.

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.

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

Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. 2024. MuLan: A study of fact mutability in language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 762–771, Mexico City, Mexico. Association for Computational Linguistics.

Makoto Fukushima, Saki KANADA, Shusuke ESHITA, and Hiroshige FUKUHARA. 2024. Evaluating color-word association in llm a comparative study of human and ai. pages 2Q5IS105–2Q5IS105.

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.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Harrison G Gough. 1976. Studying creativity by means of word association tests. *Journal of Applied Psychology*, 61(3):348.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024.

Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430, Bangkok, Thailand. Association for Computational Linguistics.

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.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? *arXiv preprint arXiv:2410.03868*.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? *arXiv preprint arXiv:2406.16316*.

Hsiao jung Yu and Yifan Wang. 2023. Mandarin chinese. In Mary Ritchie Key and Bernard Comrie, editors, *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Arjun R Khanna, William Muñoz, Young Joon Kim, Yoav Kfir, Angelique C Paulk, Mohsen Jamali, Jing Cai, Martina L Mustroph, Irene Caprara, Richard Hardstone, et al. 2024. Single-neuronal elements of speech production in humans. *Nature*, 626(7999):603–610.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language model. *arXiv preprint arXiv:2402.01618*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

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.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

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.

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.

Friedemann Pulvermüller. 1999. Words in the brain's language. *Behavioral and brain sciences*, 22(2):253–279.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A framework for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Kanato Sato, Haruto Kaneko, and Mei Fujimura. 2024. Reducing cultural hallucination in non-english languages via prompt engineering for large language models. *OSF Preprints*, 10.

Julie M Schneider, Mohammad Hossein Behboudi, and Mandy J Maguire. 2024. The necessity of taking culture and context into account when studying the relationship between socioeconomic status and brain development. *Brain Sciences*, 14(4):392.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.

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.

Bingqing Song, Boran Han, Shuai Zhang, Hao Wang, Haoyang Fang, Bonan Min, Yuyang Wang, and Mingyi Hong. 2025. Effectively steer llm to follow preference via building confident directions. *arXiv preprint arXiv:2503.02989*.

Lorand B Szalay and James Deese. 2024. *Subjective meaning and culture: An assessment through word associations*. Taylor & Francis.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

Špela Vintar, Mojca Brglez, and Aleš Žagar. 2024. How human-like are word associations in generative models? an experiment in Slovene. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 42–48, Torino, Italia. ELRA and ICCL.

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

Robert Sessions Woodworth and Frederic Lyman Wells. 1911. Association tests. *The Psychological Monographs*, 13(5):i.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. *arXiv preprint arXiv:2410.12971*.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Jindong Wang, Zhicheng Dou, and Xing Xie. 2025. Caredio: Cultural alignment of llm via representativeness and distinctiveness guided data optimization. *arXiv preprint arXiv:2504.08820*.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025a. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9840–9867, Albuquerque, New Mexico. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025b. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

A Data Preparation

A.1 Few example of IDS chapters

We list the whole 22 chapters of IDS and a few example words, shown in Table 4

Chapter	Examples(English)
Physical world	flame, ice, light
Kinship	they, girl, male
Animals	bee, ant, fox
The body	blind, wound, dream
Food and drink	fruit, mill, dish
Clothing and grooming	headband, coat, tailor
The house	brick, window, yard
Agricultural and vegetation	crop, plow, garden
Action and technology	wash, untie, glass
Motion	harbor, carriage, road
Possession	market, trade, tax
Spatial relations	wide, south, far
Quantity	second, some, more
Time	often, age, summer
Sense perception	see, color, heavy
Emotion and values	joyful, love, hate
Cognition	know, learn, pupil
Speech and Language	ask, refuse, paper
Society and politics	master, friend, custom
Warfare and hunting	fight, soldier, shoot
Law	court, accuse, steal
Religion and belief	bless, heaven, ghost

Table 4: The 22 chapters names with few sample words

A.2 Data Cleaning

We filter SWOW with IDS by the following steps:

- For one-to-many alignments in IDS, we retain only the first word in the list of multiple terms.

- Align the cue words between English and Chinese in IDS cross-culture word paris

- Remove abnormal characters and meaningless responses, such as '#Missing' or '?'.

- Filter out responses containing multiple words in English and those with more than four characters in Chinese.

- Truncate the association words for each cue word to the number of association words corresponding to the culture with fewest association words.

A.3 Data Overview

The data overview(frequency) of our dataset has been shown in Table 5. By truncating the association word, we ensure that the number of tuples for each association sorting A_w corresponding to each cue word is consistent across cultures.

Chap	Cue	Asso (avg.)	Users			
			USA	UK	OC	CN
World	52	15.96	3895	645	513	1995
Kinship	34	14.21	2246	342	266	1282
Animals	65	14.2	3416	706	459	2420
Body	92	13.12	4988	897	679	3462
Food	53	14.25	3046	519	379	2066
Cloth	44	12.95	2296	450	273	1631
House	31	13.81	1784	267	234	1205
Agriculture	36	13.81	2285	409	270	1337
Action	39	14.08	2336	362	353	1436
Motion	47	15.36	2910	465	488	1617
Possession	33	12.97	1841	237	313	1175
Space	59	12.8	3544	406	529	1970
Quantity	29	12.69	1456	166	296	924
Time	42	12.05	2711	289	370	1455
Sense	34	15.38	1975	292	426	1337
Values	42	13.17	2243	346	319	1670
Cognit	38	13.45	2299	277	323	1332
Language	31	14.13	1978	277	271	1073
Society	27	14.19	1353	229	183	957
Warfare	26	14.46	1428	277	158	940
Law	16	15.19	1078	135	135	682
Religion	11	14.36	592	104	72	440
Overall	881	13.94	51700	8097	7309	32406

Table 5: The data overview(frequency) of our dataset.

B Metric Discussing

We provide a detailed comparison with Discounted Cumulative Gain at K (DCG@K), a widely-used position-weighted evaluation metric in information retrieval.

DCG@K Formulation DCG@K applies a logarithmic discount to the relevance scores based on their ranking positions, with the intuition that items ranked lower should contribute less to the overall score. For our Word Association Task (WAT), we adapt DCG@K as follows. Given the ground truth associative words $A_w^c = \{a_1, a_2, \ldots, a_N\}$ ranked by their associative strength, and the LLM's predicted ranking $\hat{A}_w$, DCG@K is defined as:

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{\text{rel}(i)}{\log_2(i+1)} \quad (8)$$

where rel(i) represents the relevance score of the item at position i in the predicted ranking. In our adaptation, we define:

$$\text{rel}(i) = \begin{cases} \frac{N-\text{rank}(\hat{a}_i)+1}{N} & \text{if } \hat{a}_i \in A_w^c \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where rank$(\hat{a}_i)$ denotes the ground truth ranking position of the predicted word $\hat{a}_i$ in A_w^c. This formulation assigns higher relevance scores to words that appear earlier in the ground truth ranking.

Experimental Results with DCG@K We re-evaluate all our main experiments using DCG@K and present the results in Table 6. The experimental setup and datasets remain identical to those reported in the main paper.

The results obtained using NDCG@K are consistent with those derived from our proposed PWR@K metric. Both metrics demonstrate similar trends across different models, cultural contexts, and experimental conditions. This consistency validates that our key findings and conclusions are robust to the choice of position-weighted evaluation metric.

C Experiment Settings

C.1 Template Using

The templates used in this work are shown in Table 7. Notably, the CN-t condition also employs English prompts.

C.2 Hyperparaments

The detailed setting of hyperparaments are shown in Table 8.

D Language Ablation Study

In § 5, we performed an ablation study introducing a CN-t (Chinese-to-English translated) group to investigate the impact of linguistic differences on our results, controlling for language while preserving English context. Chinese-associated words were translated using MUSE bilingual dictionaries (Lample et al., 2017), with Opus-MT-zh-en (Tiedemann et al., 2024) covering uncovered pairs. Results in Table 9 suggest that Chinese contexts more effectively activate Chinese word association capabilities, indicating that the main experimental performance differences stem from cultural rather than linguistic factors.

E Granular Performance

The granular performance is shown in Figure 5. When K=3, 5, 10, the CultureSteer results consistently exhibit comparably weaker performance in Global Knowledge categories, while cultural disparities in Llama remain greater than those in Qwen.

|		Llama				Qwen			
	PWR@K	K=3	K=5	K=10	K=20	K=3	K=5	K=10	K=20
USA	Baseline	0.88	1.17	1.54	1.78	0.82	1.00	1.23	1.51
	CultureLLM	0.78	0.93	1.08	1.12	0.29	0.35	0.44	0.54
	SimLLMCultureDist	0.98	1.27	1.63	1.87	0.82	1.00	1.23	1.51
	CultureMerge	0.92	1.20	1.57	1.79	-	-	-	-
	CultureSPA	0.87	1.15	1.53	1.76	-	-	-	-
	CSP	1.04	1.31	1.65	1.90	1.02	1.22	1.50	1.77
	CCT	0.83	1.06	1.33	1.49	1.04	1.21	1.46	1.72
	CultureSteer	**1.49**	**1.87**	**2.36**	**2.66**	**1.38**	**1.71**	**2.13**	**2.40**
UK	Baseline	0.94	1.18	1.35	0.68	0.70	0.82	1.00	1.20
	CultureLLM	0.70	0.81	0.85	0.62	0.25	0.30	0.37	0.45
	SimLLMCultureDist	1.03	1.28	1.43	0.77	0.69	0.82	1.00	1.20
	CultureMerge	0.97	1.22	1.38	0.74	-	-	-	-
	CultureSPA	0.95	1.20	1.37	0.68	-	-	-	-
	CSP	1.00	1.24	1.41	0.81	0.79	0.95	1.16	1.35
	CCT	0.85	1.04	1.17	0.68	0.81	0.95	1.15	1.33
	CultureSteer	**1.38**	**1.71**	**1.94**	**1.16**	**1.03**	**1.27**	**1.55**	**1.77**
OC	Baseline	0.68	0.91	1.18	1.33	0.70	0.83	1.01	1.20
	CultureLLM	0.62	0.72	0.83	0.86	0.28	0.33	0.40	0.49
	SimLLMCultureDist	0.77	0.98	1.22	1.39	0.70	0.83	1.01	1.20
	CultureMerge	0.74	0.95	1.20	1.37	-	-	-	-
	CultureSPA	0.68	0.90	1.17	1.32	-	-	-	-
	CSP	0.81	1.00	1.22	1.39	0.82	0.98	1.19	1.37
	CCT	0.68	0.85	1.04	1.16	0.79	0.93	1.10	1.28
	CultureSteer	**1.16**	**1.44**	**1.77**	**1.99**	**1.04**	**1.28**	**1.56**	**1.79**
CN	Baseline	0.62	0.80	1.01	1.16	0.77	0.98	1.26	1.58
	CultureLLM	0.06	0.06	0.07	0.07	0.29	0.38	0.52	0.65
	SimLLMCultureDist	0.58	0.76	0.96	1.11	0.77	0.98	1.26	1.58
	CultureMerge	0.59	0.76	0.95	1.08	-	-	-	-
	CultureSPA	0.62	0.79	1.00	1.14	-	-	-	-
	CSP	0.80	1.03	1.30	1.47	1.13	1.45	1.94	2.35
	CCT	0.81	1.01	1.24	1.39	0.94	1.24	1.68	2.06
	CultureSteer	**1.06**	**1.35**	**1.69**	**1.87**	**1.45**	**1.88**	**2.54**	**3.06**
Average	Baseline	0.90	0.91	1.18	1.37	0.86	0.86	1.06	1.29
	CultureLLM	0.71	0.60	0.70	0.73	0.33	0.32	0.40	0.50
	SimLLMCultureDist	0.99	0.97	1.22	1.41	0.85	0.86	1.06	1.29
	CultureMerge	0.95	0.93	1.19	1.37	-	-	-	-
	CultureSPA	0.90	0.91	1.17	1.36	-	-	-	-
	CSP	1.02	1.03	1.29	1.50	1.00	1.07	1.33	1.61
	CCT	0.84	0.89	1.11	1.27	0.98	1.01	1.24	1.50
	CultureSteer	**1.45**	**1.44**	**1.80**	**2.07**	**1.31**	**1.43**	**1.78**	**2.13**

Table 6: Re-evaluated experimental results using DCG@K show consistent performance trends with our proposed PWR@K metric across models and conditions, confirming the validity and robustness of PWR@K for assessing ranking effectiveness.

| Language	Mode	Template
EN,CN-t	Base	When "{cue_word}" is mentioned, people often think of the following words:
	CSP	You are a person with {culture} cultural background. You will be performing a word association task. Please directly answer the association word. When "{cue_word}" is mentioned, people often think of the following words:
	CCT	You are a person with {culture} cultural background. You will be performing a word association task. Please directly answer the association word. Before you respond, think about how {culture} culture is different from {cultures} cultures. When "{cue_word}" is mentioned, people often think of the following words
CN	Base	当提起"{cue_word}",人们往往会想到的词是:
	CSP	你是一个中国文化背景的人。 你将进行词联想任务,请直接说出你联想到的词。 当提起"{cue_word}",人们往往会想到的词是:
	CCT	你是一个中国文化背景的人。 你将进行词联想任务,请直接说出你联想到的词。 在回答之前,请你注意中国文化与美国、英国、大洋洲文化的不同。 当提起"{cue_word}",人们往往会想到的词是:

Table 7: Templates for different tasks and regions, where red denotes additional cultural information, blue indicates cross-cultural thinking instructions, and orange marks task descriptions aimed at preventing LLMs from generating redundant responses.
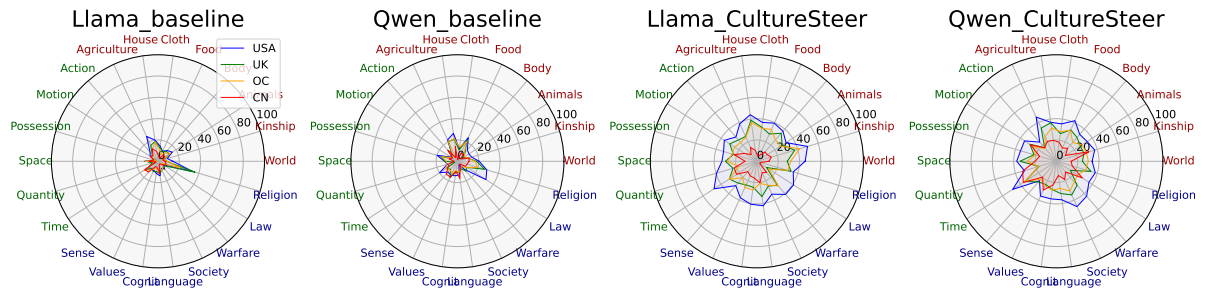
(a) Performance in PWR@3
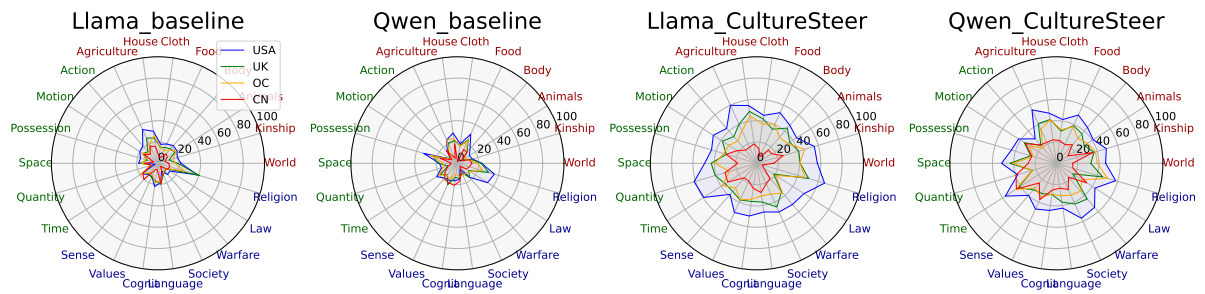
(b) Performance in PWR@5
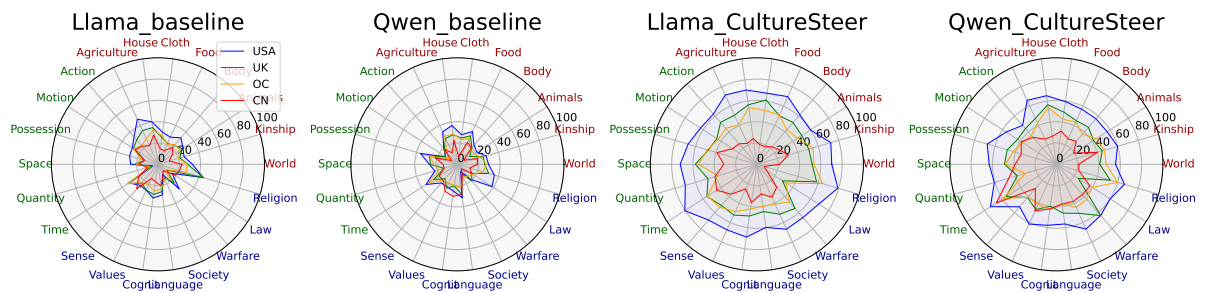
(c) Performance in PWR@10

Figure 5: Graunalr performance in PWR@3, 5 and 10.

Phase	Hyperparameters	Value
Train	learn rate	1e-4
	epsilon	1e-3
	max length	64
	l2 rate	0
	LoRA rank	8
	LoRA alpha	16
	LoRA dropout	0.05
	batch size	8
	epoch	5
Generate	max new tokens	5
	temperature	1
	epsilon	1e-3

Table 8: Relative hyperparameters

Llama(baseline)				
PWR@K	3	5	10	20
CN	**7.31**	**11.03**	**18.43**	**24.01**
CN-t	5.51	8.79	15.62	23.75
Qwen(baseline)				
CN	**8.61**	**11.65**	**17.16**	**24.11**
CN-t	5.55	7.21	10.33	15.26

Table 9: Comparison results after translating Chinese prompts and lexical items into English. Bold values indicate stronger activation of word association capabilities in Chinese linguistic contexts.