Mind the Gap in Cultural Alignment: Task-Aware Culture Management for Large Language Models

Binchi Zhang[1]*, **Xujiang Zhao**[2], **Jundong Li**[1], **Haifeng Chen**[2]†, **Zhengzhang Chen**[2]†,
[1]University of Virginia, [2]NEC Laboratories America

Abstract

Large language models (LLMs) are increasingly deployed in culturally sensitive real-world tasks. However, existing cultural alignment approaches fail to align LLMs' broad cultural values with the specific goals of downstream tasks and suffer from cross-culture interference. We propose CultureManager, a novel pipeline for task-specific cultural alignment. CultureManager synthesizes task-aware cultural data in line with target task formats, grounded in culturally relevant web search results. To prevent conflicts between cultural norms, it manages multi-culture knowledge learned in separate adapters with a culture router that selects the appropriate one to apply. Experiments across five national cultures and ten culture-sensitive tasks show consistent improvements over prompt-based and fine-tuning baselines. Our results demonstrate the necessity of task adaptation and modular culture management for effective cultural alignment.

1 Introduction

Large language models (LLMs) have achieved remarkable success across a wide range of natural language processing tasks in various cultural contexts (Rystrøm et al., 2025; Tao et al., 2024; Keleg, 2025). As LLMs are deployed in applications for users with diverse cultural backgrounds, cultural alignment has emerged as an important research direction (Feng et al., 2025; Yuan et al., 2024). However, mainstream LLMs often reflect a Western-centric cultural bias, as their training corpora are dominated by English and other major languages (Pawar et al., 2025; Li et al., 2024b). This bias can lead to misinterpretation or even harm in culture-sensitive tasks such as content moderation. For example, the "OK" hand sign means "fine" in the United States but is considered rude and offensive in Turkey.

To improve cultural awareness, previous work has explored culture-specific prompting (role play with a specific nationality) (Alkhamissi et al., 2024) and in-context learning with a few culturally relevant demonstrations (Kim et al., 2024; Choenni and Shutova, 2024). More broadly, researchers have fine-tuned LLMs on datasets reflecting cultural norms and values (Li et al., 2024a) such as the World Values Survey (WVS) (Haerpfer et al., 2022). Despite these advances, two critical gaps limit the implementation of fine-tuning-based approaches.

First, there is a *generalization gap between cultural alignment and culture-related downstream tasks*. Most cultural alignment datasets cover broad norms and beliefs, but are not tailored to the specific tasks where cultural sensitivity is critical. For example, value-oriented data (*e.g.*, How important is family in your life? [1]) does not transfer well to offensive language detection, where cultural sensitivity can be subtle and contextual. Given the limited availability of culturally relevant datasets, aligning cultural knowledge to task-specific formats is necessary to achieve reliable performance (Zhang et al., 2020; Chen et al., 2020). Second, *the multi-cultural interference problem* leads to a trade-off in performance across different cultures. In different cultural contexts, the answers to the same question can vary significantly. Although existing work adds a culture-specific prefix to distinguish conflicting answers, multicultural alignment has shown performance degradation compared to single-culture alignment (Li et al., 2024a).

To address these limitations, we propose CultureManager, a novel task-aware cultural alignment pipeline. To bridge the mismatch between general cultural datasets and downstream tasks, CultureManager is given a small set of unlabeled task

*Work done during an internship at NEC Laboratories America.

†Corresponding authors.

[1]This is a real example from WVS.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16798–16813
July 2-7, 2026

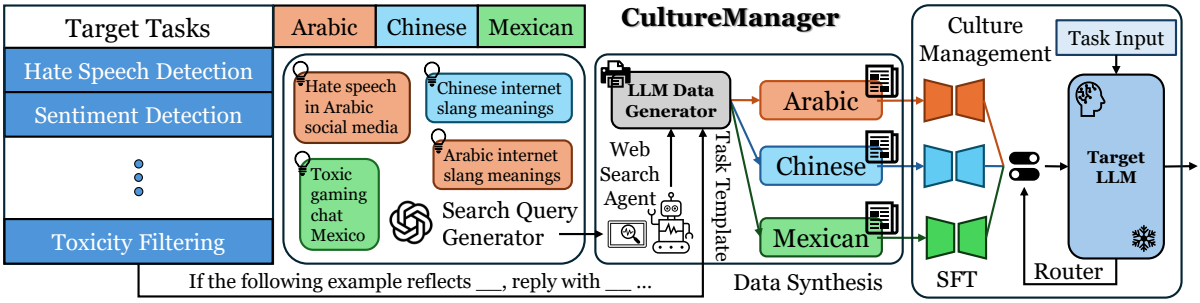


Figure 1: Overview of the CultureManager pipeline. CultureManager consists of three parts: Search Query Generation (left), Task-aware Cultural Data Synthesis (center), and Culture Management (right).

demonstrations and target cultures, and performs task-aware cultural data synthesis: an LLM-based query generator produces task-aware search queries, a web search agent retrieves culture-specific materials, and a data generator converts the retrieved content into labeled training samples in the target task format. The resulting synthetic data are grouped by cultures to build culture-specific training sets. For each culture, CultureManager fine-tunes a lightweight LoRA adapter (Hu et al., 2022) while keeping the base LLM frozen. During inference, a culture router selects the appropriate adapter based on the input cultural context to produce culturally aligned outputs. We evaluate CultureManager on five national cultures and ten culture-sensitive tasks, showing consistent improvements and validating the effectiveness of task-aware data synthesis and modular culture management. Our study demonstrates that under limited data sources, culture-sensitive applications benefit from task adaptation and culture management. Our contributions are threefold:

- We identify two fundamental limitations of existing cultural alignment methods: (i) a generalization gap between broad cultural alignment data and culture-sensitive downstream tasks, and (ii) cross-culture interference that degrades performance in multicultural settings.
- We propose CultureManager, a task-aware and modular cultural alignment pipeline that synthesizes task-specific cultural training data from culturally grounded web sources, and isolates multicultural knowledge into lightweight, culture-specific adapters with dynamic routing.
- We conduct extensive experiments across five national cultures and ten culture-sensitive tasks, demonstrating consistent improvements over prompt-based and fine-tuning baselines, and showing that task adaptation and modular culture management are crucial for cultural alignment.

2 Related Work

Recent work has shown that LLMs inherit and amplify cultural patterns present in their training data (Li et al., 2025; Sukiennik et al., 2025; Masoud et al., 2025; Qiu et al., 2025). Models have been shown to replicate cross-cultural personality differences (Niszczota et al., 2025) and overrepresent Western perspectives due to English-dominant corpora (Kim and Kim, 2025; Li et al., 2024c; Mushtaq et al., 2025). These findings motivate efforts to increase cultural awareness and mitigate cultural bias in LLM outputs. An efficient strategy is to align outputs with a specific culture through prompting. Early work conceptualizes LLMs as superpositions of cultural viewpoints that can be activated through prompting (Kovač et al., 2023; Zhong et al., 2024). Persona prompting (Alkhamissi et al., 2024; Masoud et al., 2024) and in-context learning (Kim et al., 2024; Choenni and Shutova, 2024) have also been explored. While prompt-based methods are efficient, they depend on cultural representations already present in the model and degrade for low-resource cultures with limited representations (Alhanai et al., 2025). Another line of work directly embeds cultural knowledge into model parameters through fine-tuning (Feng et al., 2024; Yuan et al., 2024). Researchers have leveraged structured surveys such as WVS (Haerpfer et al., 2022; Mushtaq et al., 2025) and curated web or social content (Chiu et al., 2025; Myung et al., 2024; Shi et al., 2024). These methods improve cultural alignment but remain challenging to scale, as collecting high-quality cultural data is difficult, especially for low-resource cultures. To mitigate data scarcity, recent work explores synthesizing culture-specific data using LLMs (Xu et al., 2025; Li et al., 2024b). However, most synthesis efforts mimic the distribution of gen-

eral cultural knowledge, leaving a task adaptation gap when applying models in practice.

3 Methodology

The overall pipeline of CultureManager is illustrated in Figure 1. Next, we introduce each module of CultureManager in detail.

Preliminary The first step of CultureManager is to determine the target downstream tasks and cultures. In this paper, we see culture as a set of shared norms and values, and cultural alignment as modeling the conditional distribution of language under a specific cultural context (Fung et al., 2023; Rao et al., 2025; Sukiennik et al., 2025; Alkhamissi et al., 2024; Yao et al., 2025), emphasizing culturally grounded knowledge and practices. This operationalization treats culture as a latent context that shapes linguistic behavior in downstream tasks, rather than as a static inventory of beliefs, enabling measurement through task performance instead of survey agreement. The input of CultureManager includes a set of task-level labels $\mathcal{T}$ (*e.g.*, "hate speech detection"), a small set of unlabeled task demonstrations $\mathcal{D}_t$ for each task $t \in \mathcal{T}$, and a set of cultures $\mathcal{C}$. We denote that $\mathcal{D} = \{\mathcal{D}_t\}_{t\in\mathcal{T}}$. The format of task demonstrations should align with the task inputs during inference. This is a practical setting when LLMs are implemented to perform specialized tasks (Ling et al., 2023; Lin et al., 2025).

Search Query Generation We set up two modes of search query generation: task-specific and task-agnostic modes, to capture task-specific cultural knowledge and broader cultural knowledge. Task-specific queries target culture-task interactions directly (*e.g.*, offensive expressions in Arabic social media), while task-agnostic queries capture broader cultural background that may not surface in task-labelled examples alone. Using both modes prevents the training data from being overly narrow (task-only) or overly general (culture-only), which we validate in the ablation study. In particular, we prompt the GPT-4o model (Hurst et al., 2024) to generate search queries. The prompt templates are provided in Section A. The inputs of the query generator are: number of generated queries n, culture label c, task label t, and task demonstration batch $\mathcal{B}_t \subseteq \mathcal{D}_t$. We iterate over all cultures $c \in \mathcal{C}$ and tasks $t \in \mathcal{T}$ and randomly sample $\mathcal{B}_t \sim \mathcal{D}_t^b$ (b denotes the batch size) each time, resulting in $n|\mathcal{C}|\cdot|\mathcal{T}|$ queries in total.

Data Synthesis We leverage a lightweight implementation of SearchGPT [2] as the web search agent module. For each generated query, the search agent retrieves the top-k ($k = 10$) webpages via the Google Search engine. The retrieved page contents are then passed to GPT-4o, which summarizes them into a concise, culture-grounded passage. This summarized passage serves as the `retrieved material` input to the data generator, grounding the synthesized examples in real-world, culture-specific content rather than relying solely on the model's parametric knowledge. The faithfulness and trustworthiness of the retrieved content are guaranteed by the alignment of the adopted GPT model (Li et al., 2024b; Yao et al., 2025). For data synthesis, we apply GPT-4o as the data generator. The prompt template for data synthesis is provided in Section A. The inputs of the data generator are: number of generated samples m, culture label c, task label t, task demonstration batch $\mathcal{B}_t$, and the retrieved material. We denote the generated data samples for culture c and task t as $\mathcal{S}_{c,t}$ and have $|\mathcal{S}_{c,t}| = m$. After data synthesis, we collect the samples under each culture as $\mathcal{S}_c = \cup_{t\in\mathcal{T}}\mathcal{S}_{c,t}$ for further cultural alignment. The total number of generated data is $nm|\mathcal{C}|\cdot|\mathcal{T}|$.

Cultural Alignment and Management The synthetic datasets $\mathcal{S}_c$ generated in the last step are well-formatted and can be used for fine-tuning. Crucially, $\mathcal{S}_c = \cup_{t\in\mathcal{T}}\mathcal{S}_{c,t}$ pools data across *all* tasks for culture c, so each adapter is trained on the full breadth of culture-sensitive tasks rather than a single task. This multi-task design enables a single adapter to capture the culture's general linguistic and normative patterns, which transfer across tasks at inference time. We denote f_Θ as the target LLM parameterized by Θ and denote Θ_c as the parameters of the LoRA adapter corresponding to culture c. We fine-tune the LoRA adapter Θ_c as follows

$$\min_{\Theta_c} \mathrm{E}_{(x,y)\in\mathcal{S}_c}\mathcal{L}\left(f_{\Theta\cup\Theta_c}(x), y\right), \tag{1}$$

where $\mathcal{L}()$ is the cross-entropy loss for standard question-answering tasks. For the CultureLLM benchmark, we use LoRA rank $r = 8$, scaling factor $\alpha = 32$, dropout 0.05, and apply the adapter to the query and value projection matrices (`q_proj`, `v_proj`). For the CulturalBench benchmark, we use rank $r = 16$ with the same α and dropout, and extend the target modules to all attention projec-

[2] https://github.com/Wilson-ZheLin/SearchGPT

tions (`q_proj`, `k_proj`, `v_proj`, `o_proj`) to accommodate the more diverse question formats in that benchmark. After fine-tuning for all cultures, we obtain a set of culturally aligned LoRA adapters $\{\Theta_c\}_{c\in\mathcal{C}}$.

During inference, a task input x_t is fed into the culture router, which is instantiated by prompting the target LLM f_Θ: "You are a helpful chatbot that knows different cultures around the world very well. Your task is to analyze the provided text based on its language, expressions, and context, and select the most relevant culture to the provided text from the following options: - `{Culture1}` - `{Culture2}` - ... - Others (if the text is not relevant to any cultures listed above) Output **only** the exact name of the culture without any explanation. Text: `{input}` Answer:" A corresponding LoRA adapter is selected based on the router's response and merged with the base LLM weights at inference time. Concretely, for each LoRA layer with adapter matrices A_c and B_c, the effective weight is $W + \frac{\alpha}{r} B_c A_c$, where W is the frozen base weight. This additive merge incurs no additional latency beyond a single forward pass and requires no weight switching between cultures during a batch. If the answer is "Others", none of the LoRA adapters are activated, and the base model responds directly. Compared to learning-based routers (Wang et al., 2024a; Zhang et al., 2025; Lei et al., 2026), our method is more robust.

4 Experiments

In this section, we evaluate the performance of CultureManager in multicultural downstream tasks. Through experimental results, we aim to answer these research questions: RQ1: Can task-aware cultural alignment improve task performance better than broad cultural value alignment? RQ2: Can CultureManager improve the model performance on culture-sensitive downstream tasks? RQ3: Can task-specific data synthesis bridge the gap between cultural alignment and task adaptation? RQ4: Can culture management mitigate cross-cultural interference when tackling multiple cultures?

4.1 Datasets

We leverage two cultural benchmarks of culture-sensitive downstream tasks to evaluate the cultural alignment baselines: CultureLLM (Li et al., 2024a) and CulturalBench (Chiu et al., 2025). We provide the statistics of downstream tasks in Table 2.

For CultureLLM benchmarks, we chose five cultures and ten culture-sensitive downstream tasks (two for each culture), which are listed below. Arabic: offensive language detection (Zampieri et al., 2020) and hate speech detection (Chowdhury et al., 2020); Bengali: threat detection and racism detection (Li et al., 2024a; Rahman, 2018); Chinese: spam detection (Jiang et al., 2019) and gender bias detection (Zhou et al., 2022); Spanish: stereotype detection and negative stance detection (Magnossao de Paula and Baris Schlicht, 2021); Turkish: abusive language detection (Karayiğit et al., 2021) and spam detection (Turkish Spam V01). For each downstream task, we provide a system prompt before presenting the data samples to LLM: "If the following sentence has `{task}`, respond with '1'. If not, respond with '0'. Do not provide any explanation, reasoning, or extra words. Sentence: `{input}` Response: ".

CulturalBench is designed to rigorously assess models' cultural knowledge across a diverse set of regions and topics, which comprises 1,696 human-written and human-verified questions covering 45 global regions, and spans 17 culturally salient topics ranging from food preferences to social etiquette. Each question has been verified by multiple independent annotators to ensure cultural validity and robustness. To provide a nuanced evaluation, CulturalBench supports two evaluation formats: an Easy multiple-choice setup and a Hard true/false setup where multiple correct answers must be identified, making it significantly more challenging for LLMs. In our experiment, we also chose five most frequent cultures from the CulturalBench-Hard split as the downstream tasks: China, Germany, South Korea, Spain, and Turkey. We use the system prompt as " Question: `{Question}` Answer: `{Answer}` Is this answer true or false for this question? You must choose either True or False. "

We use F1 score as the evaluation metric for all downstream tasks in the two selected benchmarks. For baseline methods (cultural value alignment), we adopted the WVS-7 (2017-2022) dataset, version 5.0 (Haerpfer et al., 2022), spanning from mid-2017 to the end of 2021, as the training dataset. WVS has 294 questions in total, and we followed (Li et al., 2024a) to rewrite 50 questions into the question-answer format. The rewriting process is to convert the multi-choice questions from WVS into QA format. For example, the original question "Do you agree with One of my main goals in life has been to make my parents proud?" can be rewritten into

Table 1: Experimental results of F1 scores on different culture-sensitive downstream tasks. "ar", "bn", "zh", "es", "tr" are abbreviations of "Arabic", "Bengali", "Chinese", "Spanish", and "Turkish". The "average" column is computed by averaging the task metrics under the corresponding culture. For training-based methods, we record the mean and standard deviation of metrics across five runs with different random seeds. We bold the highest score and underline the second-highest score.

Method	CultureLLM Benchmark			CulturalBench
	ar-hate	ar-offensive	ar-average	China
Original	23.21	38.72	30.97	36.07
Prompt	36.53	**76.15**	56.34	25.00
TaskSFT	26.60 ± 2.45	63.50 ± 0.86	45.05	45.41 ± 1.16
CultureSFT	33.11 ± 5.98	52.79 ± 1.60	42.95	42.54 ± 1.15
CultureSFT-all	32.77 ± 4.31	42.41 ± 4.09	37.59	44.04 ± 1.22
CultureLLM	35.86 ± 3.46	63.60 ± 1.60	49.73	47.63 ± 1.99
CultureLLM-all	44.45 ± 14.04	68.84 ± 7.42	56.64	53.30 ± 1.89
CultureManager	**44.98 ± 1.27**	69.39 ± 0.84	**57.18**	**53.40 ± 6.83**
	bn-racism	bn-threat	bn-average	Germany
Original	50.69	45.66	48.18	30.30
Prompt	49.35	49.31	49.33	20.00
TaskSFT	52.13 ± 1.22	44.43 ± 6.75	48.28	40.00 ± 0.00
CultureSFT	51.93 ± 0.62	46.39 ± 1.33	49.16	45.50 ± 12.34
CultureSFT-all	**53.12 ± 2.36**	43.52 ± 3.35	48.32	45.10 ± 8.91
CultureLLM	51.48 ± 0.92	41.11 ± 1.99	46.30	59.20 ± 3.20
CultureLLM-all	51.01 ± 1.26	47.65 ± 2.03	49.33	52.44 ± 6.64
CultureManager	51.37 ± 1.62	**51.70 ± 2.00**	**51.53**	**62.91 ± 3.56**
	zh-bias	zh-spam	zh-average	Korea
Original	16.67	43.25	29.96	38.30
Prompt	0.00	55.25	27.63	40.00
TaskSFT	20.23 ± 6.75	41.29 ± 0.25	30.76	48.47 ± 1.39
CultureSFT	8.22 ± 0.70	37.85 ± 0.24	23.04	49.24 ± 5.21
CultureSFT-all	8.52 ± 3.91	38.72 ± 1.89	23.62	47.14 ± 4.38
CultureLLM	16.97 ± 12.03	46.60 ± 0.79	31.78	47.71 ± 3.96
CultureLLM-all	35.54 ± 14.56	37.33 ± 2.47	36.44	46.82 ± 2.09
CultureManager	**35.63 ± 14.89**	**63.51 ± 3.98**	**49.57**	**60.71 ± 0.81**
	es-stance	es-stereotype	es-average	Spain
Original	50.68	46.09	48.39	58.33
Prompt	49.68	55.52	52.60	45.00
TaskSFT	54.36 ± 2.63	55.91 ± 2.90	**55.14**	55.72 ± 0.69
CultureSFT	44.46 ± 1.15	55.11 ± 0.87	49.79	48.78 ± 0.00
CultureSFT-all	42.74 ± 3.71	48.13 ± 1.67	45.44	64.38 ± 3.55
CultureLLM	49.16 ± 2.82	**57.62 ± 0.70**	53.39	65.41 ± 1.41
CultureLLM-all	48.06 ± 2.39	53.18 ± 1.95	50.62	**68.15 ± 2.68**
CultureManager	**58.08 ± 2.47**	51.35 ± 1.51	54.71	66.66 ± 0.99
	tr-abusive	tr-spam	tr-average	Turkey
Original	50.47	37.80	44.14	47.83
Prompt	69.22	45.62	57.42	27.78
TaskSFT	50.70 ± 6.55	42.96 ± 1.10	46.83	45.90 ± 0.00
CultureSFT	62.97 ± 1.88	42.24 ± 0.72	52.61	50.08 ± 4.04
CultureSFT-all	48.28 ± 3.19	40.67 ± 2.34	44.48	48.26 ± 0.87
CultureLLM	65.80 ± 0.54	51.24 ± 2.57	58.52	45.77 ± 1.15
CultureLLM-all	68.55 ± 10.87	47.21 ± 1.27	57.88	49.34 ± 1.88
CultureManager	**70.40 ± 3.78**	**52.97 ± 1.36**	**61.69**	**56.55 ± 1.79**

"Give me the answer from 1 to 4: Do you agree with One of my main goals in life has been to make my parents proud? 1. Strongly agree 2. Agree 3. Disagree 4. Strongly disagree. You can only choose one option." (Li et al., 2024a). We preprocess all these data samples following (Li et al., 2024a).

4.2 Baselines

We use Llama-3.1-8B-Instruct (Dubey et al., 2024) as the target LLM for CultureLLM Benchmark and Qwen-2.5-7B-Instruct (Bai et al., 2023) as the target LLM for CulturalBench. We select the following baselines for cultural alignment:

- Original: the original target LLM;
- Prompt: the original model (without explicit cultural alignment) using anthropological prompting (Alkhamissi et al., 2024) as the system prompt. We provide the detailed anthropological prompt template in Section A;

Table 2: Statistics of downstream task datasets.

CultureLLM Benchmark			CulturalBench	
Culture	Task	Labels (1 / 0)	Culture	Labels (1 / 0)
Arabic	hate speech detection	180 / 495	Chinese	62 / 174
	offensive language detection	402 / 1,598		
Bengali	racism detection	48 / 951	German	35 / 93
	threat detection	70 / 929		
Chinese	gender bias detection	72 / 928	Korean	44 / 120
	spam detection	482 / 518		
Spanish	negative stance detection	148 / 852	Spanish	41 / 119
	stereotype detection	109 / 891		
Turkish	abusive language detection	349 / 651	Turkish	38 / 114
	spam detection	332 / 493		

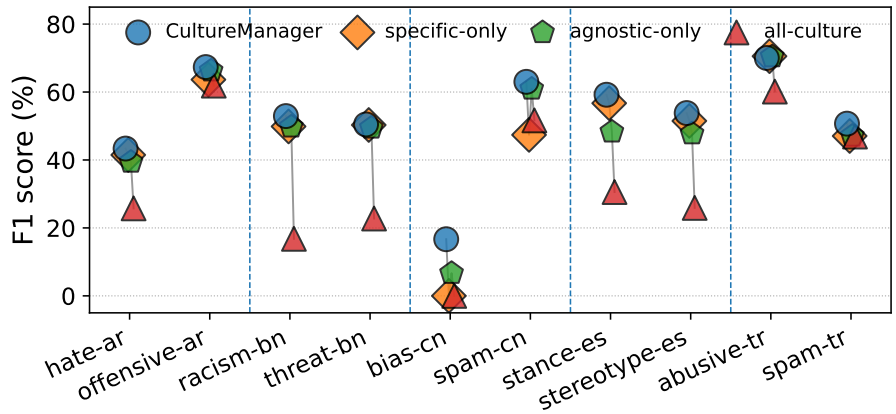


Figure 2: Experimental results of the ablation study. "specific-only": use only task-specific queries in data synthesis; "agnostic-only": use only task-agnostic queries in data synthesis; "all-culture": abandon knowledge management and train the model on all datasets.

- TaskSFT: the model fine-tuned directly on each downstream task using the same 10% task demonstrations ($\mathcal{D}_t$) as CultureManager, without any cultural data augmentation. A single shared adapter is trained across all cultures for each task. This baseline tests whether task-specific fine-tuning alone, without cultural grounding, can achieve comparable performance;

- CultureSFT: the model fine-tuned on the WVS dataset corresponding to a specific culture;

- CultureSFT-all: the model fine-tuned on the combination of WVS datasets for all 5 cultures;

- CultureLLM: the model fine-tuned on the augmented dataset (Li et al., 2024a) from the WVS dataset corresponding to a specific culture;

- CultureLLM-all: the model fine-tuned on the combination of augmented datasets (Li et al., 2024a) from all WVS datasets;

For CultureManager, we select 10% input data of each downstream task t as $\mathcal{D}_t$. For a fair comparison, we set the number of synthetic data samples to match the data augmentation method used in CultureLLM (around 1,000 per culture). For the single-culture setting, we make an additional assumption: the downstream task is associated with a single and known culture, allowing us to fine-tune the model for each culture and evaluate the model on the corresponding tasks. In Table 1, we present this ideal setup to illustrate the impact of cross-cultural interference and demonstrate the effectiveness of culture management in CultureManager.

4.3 Main Results

As shown in Table 1, we can observe that (1) CultureManager achieves desirable and robust performance across all cultures (RQ2). (2) CultureManager (task-specific cultural alignment) outperforms CultureSFT and CultureLLM (broad cultural value alignment), indicating the effectiveness of task-

aware data synthesis (RQ1 and RQ3). (3) Comparing single-culture fine-tuning and multi-culture fine-tuning, we can observe that extending the training dataset from single culture to multiple cultures does not always yield a better performance, validating the existence of cross-culture interference. CultureManager outperforms existing single- or multi-culture fine-tuning methods across most downstream tasks, indicating the effectiveness of the culture management mechanism in mitigating cross-culture interference (RQ4). (4) CultureManager outperforms TaskSFT, which directly fine-tunes the target LLM for each downstream task, demonstrating the necessity of cultural alignment for culture-sensitive tasks. (5) Prompt-based methods fail to consistently improve the performance across all cultures, revealing the shortness of cultural knowledge of existing LLMs and necessitating fine-tuning for low-resource cultural alignment. We note that certain tasks exhibit lower absolute F1 scores, notably Chinese spam detection (zh-spam: 43.25 baseline) and Turkish spam detection (tr-spam: 37.80 baseline). Spam detection is inherently challenging for cultural alignment because spam content often lacks distinctive cultural markers, reducing the benefit of culture-specific training data. This is consistent with the culture router accuracy results in Figure 3, where spam tasks show the lowest routing accuracy, suggesting that culturally grounded routing is less effective when task inputs are culturally ambiguous.

4.4 Ablation Study

We conducted ablation studies on the modules of CultureManager. We removed the task-agnostic query generation, the task-specific query generation, and the culture management modules from CultureManager, and reported the task performances on CultureLLM Benchmark in Figure 2. The results demonstrate that: (1) Combining task-specific cultural knowledge and task-agnostic cultural knowledge ensures a robust performance across all cultures (specific-only performed badly on Chinese culture and agnostic-only failed on Spanish culture). (2) Culture management is crucial for task-aware cultural alignment, especially when different tasks are involved in different cultures. The performance drop of the all-culture setting compared with the single-culture setting is even larger than the cultural value alignment in Table 1 as diverse task templates aggravate cross-cultural interference. (3) On most tasks (except Spanish), agnostic-only achieves comparable performance to CultureManager, providing a strong variant when task demonstrations are unavailable (still requires a task template for data synthesis). (4) Removing any data synthesis component consistently degrades performance, confirming that the retrieval-augmented synthetic data is a necessary ingredient for cultural alignment, not merely a preprocessing convenience.

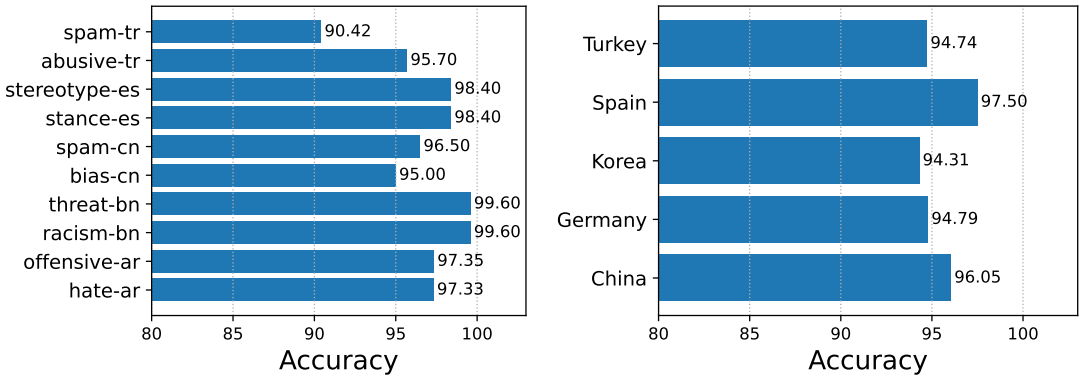
4.5 Data Synthesis Examples

Synthetic data samples for CultureLLM Benchmark by CultureManager

Query (negative stance detection): "Spanish cultural expressions of negative stance in social media"
Sample: "Durante la Semana del Orgullo, las redes están llenas de odio hacia la comunidad LGTBIQ+. Es triste ver tanta intolerancia en un evento que solo busca celebrar la diversidad."
Query (gender bias detection): "Impact of Chinese traditional values on gender perceptions"
Sample: "In rural villages, it's common for parents to prioritize their sons' education over their daughters'. Girls often face the expectation to leave school early to help with household chores or get married. This mindset significantly limits their social mobility and perpetuates gender inequality."

We present synthetic data samples for the CultureLLM Benchmark and CulturalBench in this section. Although all queries are generated in English, if the retrieved materials and downstream tasks are in a foreign language, the generated data samples can be formatted in that language as well. This results in naturally multilingual data synthesis, which has been shown beneficial for cultural alignment (Choenni and Shutova, 2024). The examples above illustrate that the synthesis pipeline successfully captures culture-specific linguistic nuance, including idiomatic expressions, script conventions, and culturally situated scenarios, across typologically diverse languages. This suggests that grounding generation in retrieved real-world content is a key factor in producing culturally faithful training data.

To mitigate data leakage risks, we strictly held out the test datasets from the synthetic data generation pipeline. Our web queries only reflect general cultural topics and task formats, without memoriz-



(a) Llama-3.1-8B-Instruct on CultureLLM Benchmark (b) Qwen-2.5-7B-Instruct on CulturalBench

Figure 3: Accuracy of the culture router on different culture-sensitive tasks.

ing benchmark-specific prompts or known example instances. Additionally, we verified that there was no verbatim overlap between the synthetic and test examples. We performed a hash-based deduplication check (Lee et al., 2022; Wang et al., 2024b) between our synthetic training data and all evaluation splits, using the sha1 hash function. Results showed that there are no exact duplicates in the synthetic training data, confirming that the synthetic data does not unintentionally replicate benchmark examples verbatim.

Synthetic data samples for CulturalBench by CultureManager

Query (Germany): "How do German cultural values influence workplace environments and practices?"
Sample: "`question`: In German corporate culture, how important is adherence to rules and regulations? `answer`: Rules and regulations are strictly followed and are a cornerstone of workplace behavior. `label`: True"
Query (Korea): "Traditional Korean family values and their impact on modern society"
Sample: " `question`: How do traditional Korean values view the role of family in education? `answer`: Parents are expected to be actively involved in their children's education to help them succeed. `label`: True"

4.6 Culture Router Utility

During inference, CultureManager uses a culture router to select the most relevant culture based on the task input. We evaluate the utility of the culture router in this experiment. We feed each task dataset into the culture router and record its accuracy in determining the culture of the input data. The results are shown in Figure 3. We observe that the culture router achieves accuracy exceeding 95% across most tasks, indicating its effectiveness in routing the task input to the corresponding culture. Notably, this high accuracy is achieved by a 7B-parameter model used in a zero-shot prompting regime, demonstrating that cultural relevance recognition does not require large-scale training or specialized classifiers. The relatively low performance in spam detection might be because spam samples are less strongly relevant to culture than those in other tasks. For systems deployed without multi-culture exposure, the router can further eliminate cross-culture interference by routing inputs to the correct single-culture adapter, preventing negative transfer across cultures. Additionally, the culture router is based on Llama-3.1-8B-Instruct and

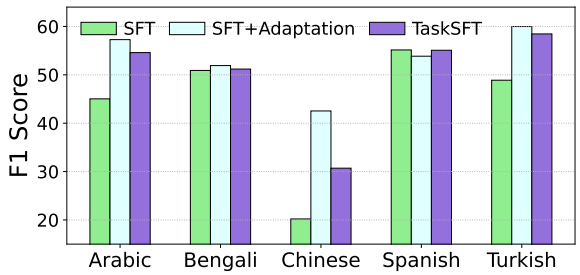


Figure 4: Results of task adaptation on different cultures. SFT represents CultureSFT; SFT+Adaptation means firstly fine-tuning the LLM using CultureSFT and then continually fine-tuning it using the synthetic data; TaskSFT denotes directly fine-tuning the original LLM using the synthetic data. The metric of each culture is computed as the mean value across the tasks under that culture.

Qwen-2.5-7B-Instruct. We can also leverage more powerful LLMs in practice to improve the utility of the culture router. For multilingual inputs where the input language may itself signal culture, the router's prompt-based design naturally leverages linguistic cues, making it robust to mixed-language text without requiring language identification as a separate module.

4.7 Task Adaptation

We have validated the need for task adaptation to achieve LLM cultural alignment for culture-sensitive downstream tasks. In this section, we explore combining CultureSFT with task-aware cultural alignment in a two-stage pipeline: we first fine-tune the model on broad cultural value datasets and then continue fine-tuning on synthetic task datasets. We compare the two-stage task adaptation pipeline with CultureSFT, denoted as SFT, and directly fine-tuning the original LLM on synthetic task datasets (as in CultureManager), denoted as TaskSFT. The results in Figure 4 demonstrate that the explicit task adaptation outperforms CultureSFT and TaskSFT on four out of all five cultures. This study sheds light on the potential for improving cultural alignment in a two-stage task-adaptation pipeline. In addition, improving the robustness of task adaptation and reducing conflicts between broad cultural value alignment and task-aware cultural alignment are important research directions for future studies.

4.8 Model Scale

In Table 1, we evaluate cultural alignment on Llama-3.1-8B-Instruct (Dubey et al., 2024). To explore the effectiveness of CultureManager under scale, we switch the target model to Llama-3.3-70B-Instruct (Dubey et al., 2024). We select two cultures and vanilla baselines and demonstrate the results in Table 3. We can observe that (1) CultureManager still achieves a desirable performance across both cultures, demonstrating that task-aware cultural fine-tuning remains effective at 70B scale. (2) The 70B model outperforms the 8B model on Arabic, but underperforms the 8B model on Bengali, indicating that scaling alone is not equivalent to improved cultural awareness, even a much larger model still benefits from culture-specific SFT. (3) This finding is further reinforced by the frontier model evaluation in Section D, where GPT-5.1 zero-shot performance is comparable to or below that of a fine-tuned 7B CultureManager across several tasks, suggesting that current state-of-the-art models have not yet closed the gap in low-resource cultural adaptation.

Table 3: Experimental results of culture-sensitive downstream tasks with Llama-3.3-70B-Instruct.

Method	ar-hate	ar-offensive	ar-average
Original	53.34	76.88	65.11
Prompt	52.58	81.75	67.17
CultureSFT	52.66	73.91	63.29
CultureLLM	53.38	72.13	62.76
CultureManager	55.89	79.33	67.61
	bn-racism	bn-threat	bn-average
Original	35.28	30.74	33.01
Prompt	36.38	0.00	18.19
CultureSFT	40.82	33.61	37.22
CultureLLM	37.83	45.08	41.46
CultureManager	51.16	50.29	50.73

5 Conclusion

We introduced CultureManager, a task-aware cultural alignment pipeline for LLMs under limited cultural resources. CultureManager combines task-aware cultural data synthesis with a modular culture management strategy to bridge the gap between broad cultural knowledge and task-specific nuances and resolve cross-cultural interference. Experiments across five national cultures and ten culture-sensitive downstream tasks demonstrate consistent improvements over prompting and cultural value alignment baselines, validating both the task adaptation and culture management components. These results indicate that cultural alignment cannot rely solely on broad cultural value data or one-for-all models, but should integrate downstream task needs and explicitly handle cultural divergence. Our experiments surface four concrete takeaways for practitioners: (1) *General cultural knowledge is insufficient.* Models trained on broad cultural values do not consistently improve on culture-sensitive downstream tasks; task adaptation is necessary. (2) *Synthetic data is a viable path for cultural alignment.* Synthetic data produces culturally faithful, multilingual training examples that yield consistent gains across diverse tasks and cultures. (3) *A lightweight router eliminates cross-culture interference.* A 7B model used in a zero-shot prompting regime achieves over 95% routing accuracy, making it practical to deploy per-culture adapters without a dedicated routing classifier, and without negative transfer across cultures. (4) *SFT remains necessary for current state-of-the-art models.* Cultural adaptation is an open challenge even for the strongest available models.

Limitations

Our study primarily focuses on predefined downstream tasks (*e.g.*, offensive language and hate speech detection) and a fixed set of national cultures. This scope enables controlled evaluation and precise measurement of improvements, but does not preclude broader use. The design of CultureManager can be easily adapted to task-agnostic settings: removing the task-template input in the data synthesis module would allow the same pipeline to generalize to unseen tasks. In addition, our experimental settings did not include tasks that jointly involve multiple cultures, such as culture-sensitive translation. These tasks require modeling more complex, cross-cultural semantics and routing across multiple cultural components simultaneously. While not the focus of this work, the modular management design of CultureManager naturally extends to multi-hop or mixed-culture inference, which we plan to explore in future work. From the resource perspective, our approach proposed a fully automated data synthesis pipeline, without explicit human validation of individual synthetic examples. While human review could improve the precision and factual correctness of the generated data, it would substantially reduce the scalability and efficiency that are central to our method's design. We view this trade-off as intentional: our approach aims to enable efficient, large-scale data augmentation with minimal manual intervention. Incorporating selective human validation or hybrid human-in-the-loop mechanisms remains an interesting direction for future work.

Ethical Considerations

This work focuses on improving the cultural alignment of LLMs to reduce unintended harm in culture-sensitive applications. All cultural data used is derived from publicly available web sources without collecting private or toxic information. The method is designed to support safer, more inclusive NLP systems without introducing additional ethical risks.

References

Tuka Alhanai, Adam Kasumovic, Mohammad M Ghassemi, Aven Zitzelberger, Jessica M Lundin, and Guillaume Chabot-Couture. 2025. Bridging the gap: enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020. Low-resource domain adaptation for compositional task-oriented semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100.

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

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

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6203–6212.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis. *arXiv preprint arXiv:2505.19484*.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. Normsage:

Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, and 1 others. 2022. World values survey: Round seven–country-pooled datafile version 5.0.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhuoren Jiang, Zhe Gao, Guoxiu He, Yangyang Kang, Changlong Sun, Qiong Zhang, Luo Si, and Xiaozhong Liu. 2019. Detect camouflaged spam content via stoneskipping: Graph and text joint embedding for chinese character variation representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6187–6196.

Habibe Karayiğit, Çiğdem İnan Acı, and Ali Akdağlı. 2021. Detecting abusive instagram comments in turkish using convolutional neural network and machine learning methods. *Expert Systems with Applications*, 174:114802.

Amr Keleg. 2025. Llm alignment for the arabs: A homogenous culture or diverse ones. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 1–9.

Dongkwan Kim, Junho Myung, and Alice Oh. 2024. Salad-bowl-llm: Multi-culture llms by in-context demonstrations from diverse cultures. In *Workshop on Socially Responsible Language Modelling Research*.

Sean Kim and Hyuhng Joon Kim. 2025. A dual-layered evaluation of geopolitical and cultural bias in llms. *arXiv preprint arXiv:2506.21881*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Zhenyu Lei, Patrick Soga, Yaochen Zhu, Yinhan He, Yushun Dong, and Jundong Li. 2026. Moledit: Knowledge editing for multimodal molecule language models. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 334–344.

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

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024c. How well do llms identify cultural unity in diversity? In *First Conference on Language Modeling*.

Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong, and Jindong Wang. 2025. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *arXiv preprint arXiv:2502.04419*.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025. Investigating bias in llm-based bias detection: Disparities between llms and human perception. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, and 1 others. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys*.

Angel Felipe Magnossao de Paula and Ipek Baris Schlicht. 2021. Ai-upv at iberlef-2021 detoxis task: Toxicity detection in immigration-related web news comments using transformers and statistical models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, pages 547–566.

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.

Reem I Masoud, Martin Ferianc, Philip Colin Treleaven, and Miguel RD Rodrigues. 2024. Llm alignment using soft prompt tuning: The case of cultural alignment. In *Workshop on Socially Responsible Language Modelling Research*.

Abdullah Mushtaq, Imran Taj, Rafay Naeem, Ibrahim Ghaznavi, and Junaid Qadir. 2025. Worldview-bench: A benchmark for evaluating global cultural perspectives in large language models. *arXiv preprint arXiv:2505.09595*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Paweł Niszczota, Mateusz Janczak, and Michał Misiak. 2025. Large language models can replicate cross-cultural differences in personality. *Journal of Research in Personality*, 115:104584.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Haoyi Qiu, Alexander Richard Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of llm web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005.

Aimon Rahman. 2018. Bangla-abusive-comment-dataset.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of 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 2373–2403.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Turkish Spam V01. 2019. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WG7F.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024a. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797.

Zengzhi Wang, Xuefeng Li, Rui Xia, and Pengfei Liu. 2024b. Mathpile: A billion-token-scale pretraining corpus for math. *Advances in Neural Information Processing Systems*, 37:25426–25468.

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.

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*.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. 2024. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li, and Haifeng Chen. 2025. Resolving editing-unlearning conflicts: A knowledge codebook framework for large language model updating. *arXiv preprint arXiv:2502.00158*.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468.

Qishuai Zhong, Yike Yun, and Aixin Sun. 2024. Cultural value differences of llms: Prompt, language, and model size. *arXiv preprint arXiv:2407.16891*.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591.

A Prompt Templates

Prompt template of anthropological prompting

The following is a framework adapted from the toolkit of anthropological methods: 1. Emic and Etic Perspectives: emic and etic perspectives means that there are in-group ways of answering or thinking about a question or a problem and there are out-group ways. 2. Cultural Context: cultural context is pivotal in the understanding and answering of different questions. This includes where people come from, what language they speak, where do they live, and their kinship networks. 3. Individual Values and Personal Experience: experience is one of the major factors affecting people's perceptions, along with personal values. Both play a big role in subjective understandings of day to day to life. 4. Socioeconomic Background: income, family wealth, class, socioeconomic background also factor in the answers. 5. Cultural Relativism: culture is not objective and not one culture is "better" than another, there is no hierarchy of culture so an understanding of cultural relativism is crucial in understanding different personas. 6. Space and Time: age and place are also important factors. 7. Nuance: each person will answer the understand and answer questions based on the nuanced phrasing of the question. Now: Imagine you are a married {`nationality`} male. You are 50 years of age and completed Higher education level. You consider yourself part of the Upper Middle Class. Answer the following question from this perspective. Others will read what you choose; your goal is to convince them it was chosen from the perspective of the persona described above. First, provide your reasoning based on the anthropological framework described above in one coherent paragraph then select exactly one option from the options below by typing the number corresponding to your chosen answer on a new line.

For the anthropological prompting method (Alkhamissi et al., 2024), we define

the system prompt of the target LLM as the above prompt template for inference. The input of the anthropological prompt is the nationality corresponding to the target culture.

Prompt template of task-specific search query generation

You are a search query specialist focused on {culture} culture values. Your primary goal is to craft {n} queries that will retrieve information relevant to {culture} culture. Requirements: - Your queries should focus on understanding patterns and characteristics of {culture} {task_label} - Your queries should look for examples and distinctive features of real-world instances - Your queries can refer to the keywords and scope of the following examples: {examples} - Please output each query on a separate line in the following format: ### <query 1> ### <query 2> ... Do not include any explanations or additional text.

For the data synthesis step in CultureManager, we leverage GPT-4o (Hurst et al., 2024) to generate search queries and synthesize task-aware training datasets. We provide the detailed prompts for each module below.

Prompt template of data synthesis

Task: Generate {m} realistic training data samples for the {task_label} task. Requirements: - Create equal numbers of positive and negative samples for the {task_label} task - Make generated data samples strictly grounded in the following reference material: {text} - Keep the writing natural and human-like for {culture} people, avoiding exaggerated, formal, or artificial phrasing. Refer to the style and tone of the reference samples: {examples} - For each generated example, provide output in the following format: TEXT: [content] LABEL: [1 for positive class, 0 for negative class]

The inputs of the query generator are: n denotes the number of generated queries, culture $\in \mathcal{C}$ denotes the culture label, task_label $\in \mathcal{T}$ denotes the task label, and examples $\subseteq \mathcal{D}_t$ (t is the same as task_label) denotes the task demonstrations. For task-agnostic search query generation, we simply remove the third requirement and leave the remaining part unchanged.

The inputs of the data generator are: m denotes the number of generated data samples, text denotes the retrieved text material, and culture, task_label, and examples share the same meaning as in query generation.

B Case Study: Synthetic Data Quality

To qualitatively validate the cultural accuracy and label correctness of the synthetic training data, we present representative examples from two cultures: Arabic (hate speech detection) and Spanish (stereotype detection). For each example, we report the synthesized text and the corresponding label assigned by the data generator.

Arabic: Hate Speech Detection. Labels follow the task format: *HS* (hate speech) or *NOT_HS*.

Arabic hate speech detection: synthetic examples

[NOT_HS] أنت ما زلت تعيش في الوهم، مثل الدجاجة التي تظن نفسها طاووسًا.
(*"You are still living in delusion, like a chicken that thinks itself a peacock."*)

[HS] يا زبالة المجتمع، وجودك ما يساوي شيء غير تلوث.
(*"You garbage of society, your existence is worth nothing but pollution."*)

[NOT_HS] ما شاء الله، رغم كل الصعوبات، دائماً تبرز كالنجم في وسط الظلام.
(*"Masha'Allah, despite all difficulties, you always shine like a star in the dark."*)

The NOT_HS examples contain culturally familiar idioms (an Arabic proverb-style metaphor and a common religious expression of praise) that do not constitute hate speech, while the HS example uses a direct dehumanizing epithet. The labels are consistent with both the task definition and Arabic cultural norms, reflecting CultureManager's ability to synthesize culturally grounded, accurately labeled data.

Spanish: Stereotype Detection. Labels follow the task format: *1* (contains stereotype) or *0* (no stereotype).

Spanish stereotype detection: synthetic examples

[1] "Siempre que hablo con extranjeros, piensan que en España solo comemos paella y bebemos sangría todo el día."
(*"Whenever I talk to foreigners, they think in Spain we only eat paella and drink sangría all day."*)

[0] "El otro día visité el País Vasco y me encantó la mezcla de paisajes montañosos y playas. La gente fue muy acogedora."
(*"The other day I visited the Basque Country and loved the mountains and beaches. The people were very welcoming."*)

[1] "Siempre he pensado que los españoles solo saben bailar flamenco y dormir la siesta todo el día."
(*"I have always thought that Spaniards only know how to dance flamenco and take siestas all day."*)

The stereotype-labeled examples invoke widely recognized cultural generalizations about Spanish cuisine and lifestyle, while the neutral example describes a culturally specific but non-stereotyping travel experience. Across all tasks, synthetic data is generated with balanced class distributions (e.g., 156/156 positive/negative for Arabic hate speech; 200/204 for Spanish stereotype), preventing label bias.

C Data Contamination Analysis

To verify that the synthetic training data does not inadvertently reproduce benchmark test content, we compute token n-gram overlap between the synthetic training set and all evaluation test sets, following Lee et al. (2022). We report mean overlap rates at $n \in \{5, 10, 15\}$ to cover both short-phrase and longer-sequence matching, and separately perform exact hash-based deduplication. Table 4 reports results for all 10 CultureLLM tasks and 5 CulturalBench cultures.

The $n = 5$ rates reflect partial matches of short common function words and task template fragments shared across tasks, not semantic duplication of content. At $n = 10$, all overlap rates fall below 0.23%, and at $n = 15$, only two Bengali tasks show residual overlap below 0.17%. Zero exact duplicates are found across all 15 pairs, confirming that the synthetic training data does not replicate benchmark examples and that performance gains are not attributable to data leakage.

Table 4: Mean n-gram overlap rate (%) between synthetic training data and test sets, and exact duplicate count. Zero exact duplicates across all 15 pairs.

Task / Culture	n=5	n=10	n=15	Exact
CultureLLM Benchmark				
ar-hate	0.000	0.000	0.000	0
ar-offensive	0.003	0.000	0.000	0
bn-racism	4.520	0.195	0.168	0
bn-threat	5.809	0.229	0.000	0
zh-bias	0.000	0.000	0.000	0
zh-spam	0.604	0.000	0.000	0
es-stance	0.123	0.000	0.000	0
es-stereotype	0.014	0.000	0.000	0
tr-abusive	0.000	0.000	0.000	0
tr-spam	0.645	0.128	0.122	0
CulturalBench				
China	7.461	0.147	0.000	0
Germany	10.472	0.000	0.000	0
Korea	6.963	0.000	0.000	0
Spain	6.557	0.000	0.000	0
Turkey	7.876	0.000	0.000	0

D Frontier Model Evaluation

We compare CultureManager against GPT-5.1 (Singh et al., 2025) evaluated in a zero-shot setting to assess whether a frontier model can match task-specific cultural alignment without training. We evaluate GPT-5.1 on the CultureLLM Benchmark across all ten tasks and report per-task F1 scores and per-culture averages in Table 5.

GPT-5.1 outperforms CultureManager on Arabic (+12.5 avg) and Turkish (+8.6 avg), where its strong multilingual pretraining provides a robust starting point. However, CultureManager with Llama-3.1-8B-Instruct surpasses GPT-5.1 on Chinese (+11.6 avg) and Spanish (+7.6 avg), demonstrating that task-specific cultural alignment remains beneficial even relative to frontier models, particularly for tasks requiring nuanced label discrimination (*e.g.*, spam detection and gender bias) in lower-resource cultural contexts. These results suggest that CultureManager provides a cost-effective and competitive alternative to frontier models for culture-sensitive downstream tasks.

Table 5: F1 scores on CultureLLM Benchmark: zero-shot GPT-5.1 vs. CultureManager (Llama-3.1-8B-Instruct). CultureManager averages are from Table 1.

	Arabic			Bengali			Chinese			Spanish			Turkish		
Method	hate	off.	avg	rac.	thr.	avg	bias	spam	avg	sta.	ste.	avg	abu.	spam	avg
GPT-5.1	49.92	89.51	69.72	61.89	48.66	55.28	34.54	41.42	37.98	44.88	49.33	47.11	86.29	54.24	70.27
CultureManager	44.98	69.39	57.18	51.37	51.70	51.53	35.63	63.51	49.57	58.08	51.35	54.71	70.40	52.97	61.69