CultureScope: From Domain-Level Heuristics to a Stratified and Multi-Theoretical Taxonomy of Culture

Anonymous ACL submission

Abstract

As large language models (LLMs) are increasingly deployed in diverse cultural environments, evaluating their cultural understanding capability has become essential for ensuring trustworthy and culturally aligned applications. However, existing work often lacks systematic synthesis of cultural theories, leaving theoretical guidance largely at the level of domain-level heuristics and resulting in incomplete cultural modeling. Building on established cultural theories and in close collaboration with domain experts, we propose a cultural evaluation framework with the most comprehensive coverage of cultural knowledge dimensions to date. At its core is a stratified hierarchical taxonomy comprising 140 dimensions, which provides systematic guidance for the automated extraction of cultural knowledge and the construction of evaluation datasets across cultural contexts. Experimental results on constructed datasets demonstrate that existing LLMs lack comprehensive cultural competence. Possessing cultural knowledge does not necessarily translate into the ability to apply it appropriately, indicating that cultural competence requires dedicated training. All code and data files are available at `https://github.com/AnonymousUserForSubmission/Culture`

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across many natural language processing tasks, and are widely adopted in applications ranging from virtual assistants to educational tools (Bharathi Mohan et al., 2024; Raiaan et al., 2024; Raza et al., 2025). Despite these achievements, a persistent challenge remains: Current LLMs still exhibit significant shortcomings in cultural understanding, often leading to cultural misalignment when applied in **practical real-world contexts** (Bu et al., 2025; Masoud et al., 2025; Ramezani and Xu, 2025). For example, if an LLM social assistant fails to distinguish between politeness conventions and genuine intentions in Chinese culture, it may generate inappropriate suggestions that harm users' social interactions. Therefore, how to comprehensively and accurately evaluate LLMs' cultural understanding abilities in specific cultural contexts has become an urgent issue that must be addressed for their broader adoption.

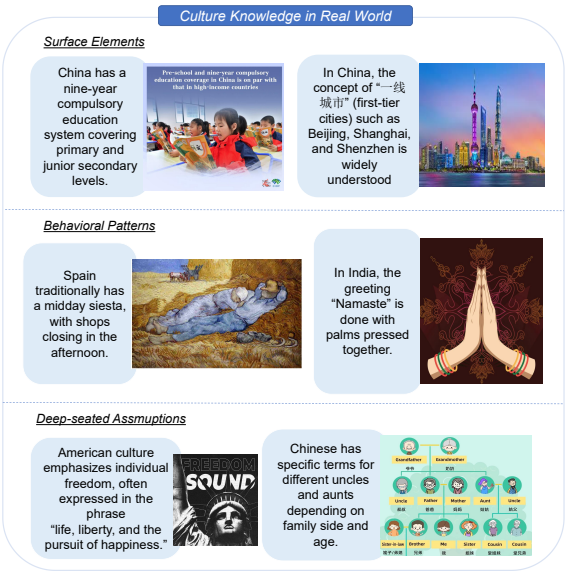


Figure 1: In the real world, the composition of an individual culture is inherently stratified, comprising observable surface elements (e.g., artifacts, concepts), intermediate behavioral patterns (e.g., etiquette, habits), and deep-seated assumptions (e.g., values, society). Existing evaluation datasets often fail to comprehensively and accurately cover these types of knowledge.

Cultural understanding refers to a model's ability to **comprehend** culturally grounded concepts, **identify** cultural stereotypes or biases, and appropriately **apply** cultural knowledge in real-world or context-sensitive scenarios (Earley, 2002). As is shown in Figure 1, culture is an inherently stratified concept (Tylor, 1871; Hall, 1976; Sasu, 2016), comprising observable surface elements (e.g., artifacts, concepts), intermediate behavioral patterns (e.g., etiquette, habits), and deep-seated assump-

Benchmark	Design of Taxonomy				Knowledge Coverage			
	Theory	Domain-level	Dimension-level	Stratification	Norms	Habits	Values	Coverage(%)
AraDiCE-Culture (2025)	✗	✗	✗	✗	✓	✓	✗	13.57
CulturalBench (2024)	✗	✗	✗	✗	✗	✗	✓	16.43
KorNAT (2024)	✗	✗	✓	✓	✓	✗	✓	7.86
BLEnD (2024)	✗	✓	✗	✗	✗	✓	✗	8.57
WorldValuesBench (2024)	✓	✓	✗	✗	✗	✗	✓	22.14
CDEval (2024)	✓	✓	✗	✗	✗	✗	✓	15.71
CultureSynth (2025)	✓	✓	✓	✗	✓	✓	✓	40.00
Ours(CultureScope)	✓	✓	✓	✓	✓	✓	✓	100.00

Table 1: Prior work often relies on intuition or domain-level heuristics and treats cultural knowledge as flat attributes, lacking operational dimensions and explicit modeling of culture's stratified organization. As a result, their taxonomic coverage remains limited; we report each benchmark's relative taxonomic coverage with respect to ours.

tions (e.g., values, beliefs). Existing benchmarks, however, fail to fully cover these elements.

There are two primary reasons for the limitation. **First**, existing works rely on isolated theories that offer only domain-level heuristics, lacking the theoretical synthesis needed to provide operational dimensions. As shown in Table 1, many works (Myung et al., 2024; Lee et al., 2024; Chiu et al., 2024; Mousi et al., 2025) rely on researchers' intuition to categorize cultural facts, while those adopting established theories (Wang et al., 2024; Zhao et al., 2024) treat them as abstract descriptors rather than operational cultural dimensions. For example, Hofstede's cultural dimension(Hofstede, 1984) describe culture only at the level of values (e.g., individualism vs. collectivism), which is too coarse-grained to yield operational evaluation dimensions, leaving many aspects of everyday cultural knowledge uncovered. **Second**, most benchmarks fail to model the stratified structure of culture. Treating cultural knowledge as independent, flat attributes (Nayak et al., 2024; Singh et al., 2025; Zhang et al., 2025) ignores its stratified nature, where distinct levels range from observable surface norms to underlying value orientations.

Meanwhile, many existing evaluation benchmarks are difficult to scale and adapt to other cultural and linguistic contexts, as they **heavily** rely on manual data construction (Palta and Rudinger, 2023; Myung et al., 2024; Mousi et al., 2025). The high cost and labor intensity of manual labeling limit their generalizability.

To address the limitations, we propose CultureScope, a novel evaluation **framework** that operationalizes multiple cultural theories into a stratified knowledge taxonomy for LLMs' cultural understanding. **Our goal is to** rigorously evaluate LLMs' cultural understanding of **specific** countries or regions , which necessitates a systematic clarification of the stratified knowledge that constitutes "culture". We transcend domain-level heuristics by synthesizing and extending classical cultural theories into a stratified and operational classification taxonomy. Unlike flat taxonomies, our framework explicitly models the stratified structure from deep value orientations to observable norms.

CultureScope is labor-efficient, scalable, and currently the most comprehensive in terms of **taxonomic coverage**. Crucially, we define this comprehensiveness by the structural depth and breadth of evaluable cultural knowledge within a given culture, **rather than** the sheer number of regions. Through expert synthesis of multiple theoretical frameworks, our taxonomy represents the most exhaustive mapping of cultural strata to date. We conduct an empirical study to quantify this advantage. As is reported in Table 1, when mapping existing cultural benchmarks to ours, their maximum taxonomic coverage reaches only **40%**, indicating that a substantial portion of operational cultural phenomena evaluated in CultureScope are absent from prior work. Our contributions are as follows:

- We propose a novel stratified taxonomy that transcends domain-level heuristics by integrating multiple cultural theories into actionable, stratified dimensions. This framework achieves the most extensive taxonomic coverage to date, which guides the collection of cultural knowledge and evaluation datasets.
- We extract a substantial volume of cultural knowledge embedded in specific cultural contexts from various data sources and construct the corresponding evaluation datasets. We conduct quality control through source filtering and expert annotation.
- We conduct extensive experiments on commonly used LLMs to evaluate their abilities to understand, master, and apply cultural knowledge. We observe that developing cultural

capability requires specialized training, particularly for the practical applicational tasks.

2 Related Work

Current studies on the evaluation of LLMs' cultural understanding mainly include: (1) Factual knowledge mastery (Onohara et al., 2025; Singh et al., 2025; Nayak et al., 2024), and (2) Cultural values and moral alignment (Hadar-Shoval et al., 2024; Karinshak et al., 2024; Ji et al., 2024; Aksoy, 2025; Jiang et al., 2025). However, current research on the evaluation of LLMs' cultural understanding capabilities still faces notable limitations. As is shown in Table 1, while some existing studies (Wang et al., 2024; Yadav et al., 2025; Zhou et al., 2025) draw inspirations from relevant cultural theories, they often lack theoretical synthesis, resulting in domain-level heuristics rather than operational dimensions. This fragmented approach fails to capture the intricate, stratified structure of cultural knowledge, leading to incomplete taxonomic coverage of cultural knowledge. On the other hand, many existing methods **heavily** rely on **manual data curation**, and some datasets even involve hundreds of annotators (Palta and Rudinger, 2023; Myung et al., 2024; Mousi et al., 2025), which incurs high costs and is difficult to scale.

3 CultureScope Framework

3.1 Taxonomy Design

To address the limited coverage of cultural knowledge in existing evaluation frameworks, we **synthesized** multiple theoretical cultural frameworks to guide the construction of our taxonomy. Specifically, the Cultural Iceberg Model(Hall, 1976) provides the overall stratified perspective that distinguishes observable cultural phenomena from deeper value systems. Hall's High-Context and Low-Context Culture(Hall, 1976), Cultural Patterns(Kroeber and Kluckhohn, 1952), and Value Orientation Theory(Kluckhohn and Strodtbeck, 1961) help organize recurring behavioural and communicative practices at the intermediate level, while Hofstede's Cultural Dimensions(Hofstede, 1984) and Social Memory Theory(Halbwachs, 2020) inform the representation of deeper social relations, values, and belief structures. As shown in Figure 2, this taxonomy consists of 3 layers, each organized into four hierarchical levels, comprising a total of 140 dimensions. It explicitly captures the stratified organization of culture, spanning surface-level Institutional Norms, intermediate Behavioral Patterns, and deep-seated Core Values. Detailed content, theoretical foundation and advantages of this taxonomy are provided in Appendix C.

3.2 Cultural Knowledge Extraction

To ensure both reliability and comprehensiveness, our culture knowledge is derived from two types of sources. 1) *reliability*: manually selected professional cultural websites. 2) *comprehensiveness*: data dynamically retrieved through Google and Bing search engines. The detailed sources of cultural knowledge are provided in Appendix F. Some studies have found that even datasets widely regarded as reliable, such as Wikipedia, still exhibit notable cultural biases (Naous et al., 2024). By combining static, high-quality data sources with dynamic retrieval, we cross-validate information to reduce biases while effectively handling rapidly changing aspects of cultural knowledge, such as expressions of fashion.[1] We convert the leaf nodes of the stratified hierarchical dimensional taxonomy into search queries(e.g., Celebration of Festivals in Spanish culture). For each query, we select the top five web pages, retain their textual content, and apply GPT-4o to extract relevant information from the retrieved texts. Examples of extracted cultural knowledge are provided in the Appendix H.

3.3 Evaluation Dataset Generation

We adapted cognitive science (Wilberschied Ph D, 2015; Avgousti, 2018; Feng et al., 2025) theories and define four types of tasks for our questions, each assessing whether the model: (1)Factual: **knows** cultural facts, (2)Conceptual: **understands** the underlying meaning of cultural phenomena, (3)Misleading: can **identify** cultural biases in **practical** scenarios, (4)Multi-hop: **apply** its knowledge to reason and act appropriately in complex settings. Our evaluation focuses not only the **understanding and mastery** of cultural knowledge but also its **practical application**, providing a comprehensive assessment of the model's cultural understanding. Question generation is based on a RAG (Lewis et al., 2020) approach. We employ Qwen3-Max (Yang et al., 2025) as the generator. For each type of question, we apply specific expert-designed prompts to clearly define the task requirements. All relative prompts are in Appendix I.

[1] Regarding the dynamics of culture, we provide further elaboration in the Appendix E.

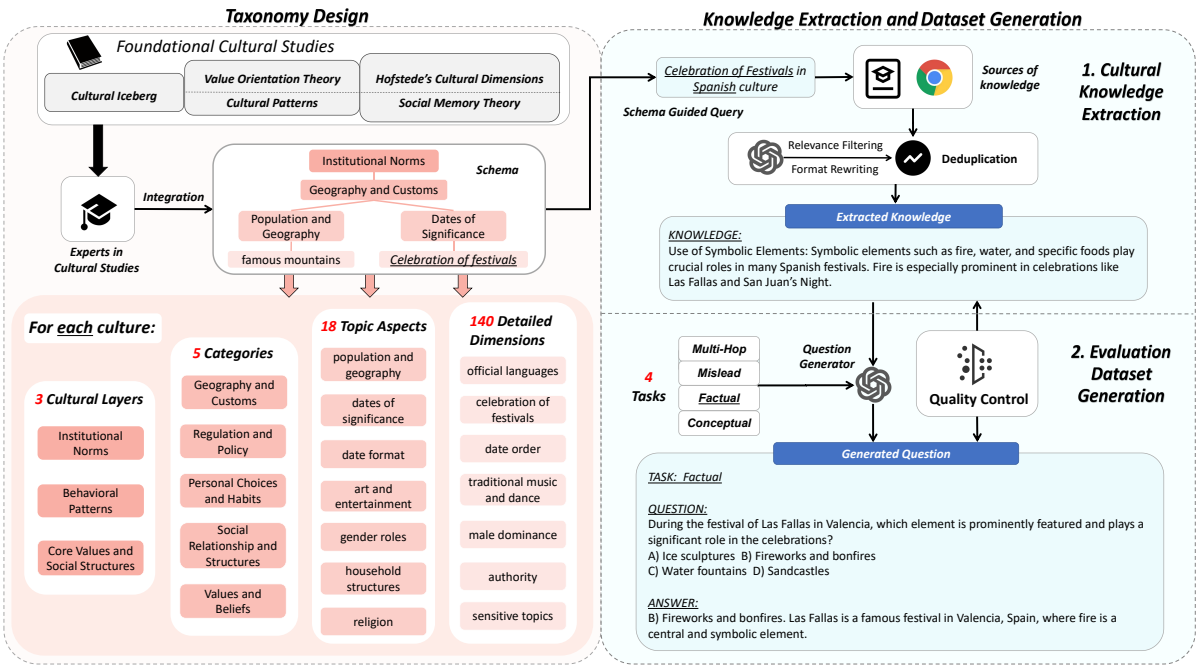


Figure 2: The overall framework of CultureScope. We engage cultural studies experts to develop a taxonomy that models cultural knowledge at a fine-grained level in specific cultural contexts, addressing aspects largely overlooked by previous work, and providing a framework for evaluating LLMs' cultural understanding. This framework allows for comprehensive, fine-grained analysis and interpretation of LLMs' performance across diverse detailed cultural dimensions. By adjusting the name of the target culture in knowledge extraction, this framework can be readily adapted to assess cultural understanding for any given target culture.

Culture	Metric	Knowledge	Questions
Spanish	Annotation	94.29%	97.41%
	Consistency	90.71%	97.68%
Chinese	Annotation	97.84%	97.50%
	Consistency	97.14%	96.79%
French	Annotation	92.50%	98.57%
	Consistency	93.57%	98.93%
Korean	Annotation	93.22%	98.30%
	Consistency	91.00%	98.39%

Table 2: The quality assessment results of the cultural knowledge and the evaluation dataset via manual sampling inspection. Consistency is measured using the simple agreement rate. The validation of the cultural knowledge focuses on identifying factual errors, and the that of the evaluation set emphasizes the consistency among the question, answer, and reference knowledge.

3.4 Quality Control

We conduct quality control from two perspectives: source selection and manual annotation. When using queries such as "eating habits in Spanish culture", the top five results returned by Google and Bing are consistently high-quality webpages that are highly relevant to the target dimension of the query, including widely recognized platforms such as Wikipedia, official releases from relevant government agencies, and so on. Detailed statistics on the information sources are provided in the Appendix F. For each culture, we invited two annotators with the corresponding cultural background to perform sample validation, identifying severe factual errors in the cultural knowledge as well as logical or knowledge inconsistencies in the questions. For the evaluation dataset, we sampled 1,120 instances per culture. The annotation results are presented in Table 2. To further assess the reliability of automatic evaluation, we additionally sampled 1,000 instances per dataset for manual verification of the judgments produced by the LLM Judge, evaluating their quality and alignment with human expectations. Full details, including annotator demographics, instructions, detailed annotation results, and the annotated LLM Judge outputs, are provided in Appendix G.

4 Evaluation Benchmarks

4.1 Evaluation Metric

We evaluate model performance using two metrics. First, **Accuracy** measures the proportion of culturally correct answers. For objective questions (e.g., multiple-choice, true/false), predictions $\hat{y}$ are correct if they match the ground truth y. For sub-

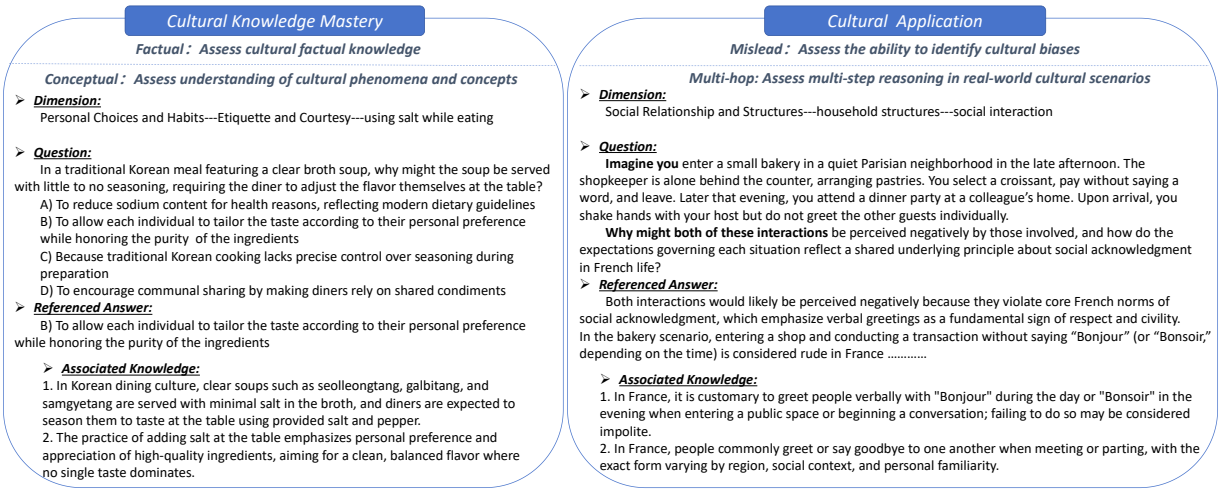
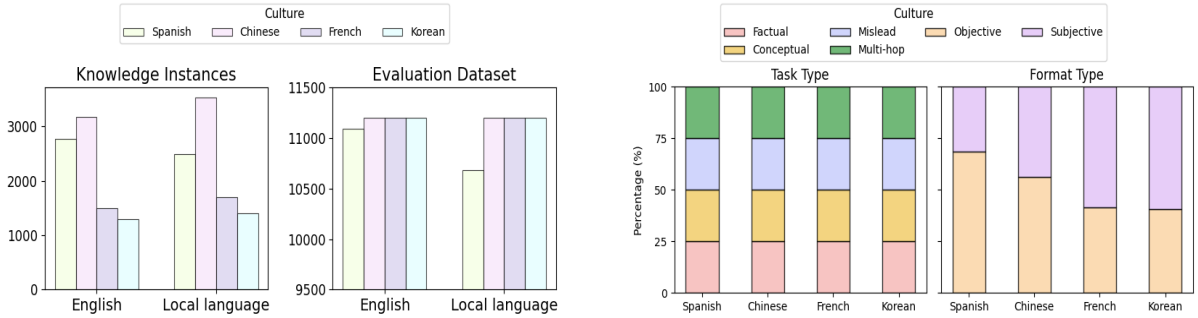


Figure 3: Examples of the generated evaluation dataset. We designed four types of tasks to evaluate LLMs' mastery of cultural knowledge and their ability to apply such knowledge. Mastery of cultural knowledge includes understanding cultural **facts** and grasping cultural **concepts**, while cultural application ability involves identifying cultural **biases** in real-world scenarios and performing multi-turn cultural reasoning.



(a) Data Volume of cultural knowledge and questions.

(b) Data distribution of evaluation dataset.

Figure 4: Statistics of constructed benchmarks. The generated questions are balanced in terms of tasks, with the four types (Factual, Conceptual, Mislead, Multi-hop) represented in equal proportions. In terms of format, the ratio of subjective to objective questions is also reasonably proportioned.

jective questions (e.g., short answers, essays), we use the LLM-as-a-Judge approach (Desmond et al., 2025), where a Judge Model assesses whether the model-generated answer a_{test} aligns with the key content of the reference a_{ref}. Many real-world cultural perspectives have no single correct answer and only indicate tendencies (Geertz, 1973; Hofstede, 1984). Therefore, we consider a response "culturally correct" if it reasonably covers the main ideas. This definition is explicitly incorporated into the prompts for the LLM judge.

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\begin{cases} \hat{y}_i = y_i & \text{Obj} \\ \text{Judge}(q_i, a_i^{\text{ref}}, a_i^{\text{test}}) & \text{Sub} \end{cases}\right) \quad (1)$$

The second metric is the **Bradley–Terry Ability Score (BT-Score)**, a statistical estimate of a model's relative capability obtained by fitting a Bradley–Terry model to pairwise win–loss outcomes against baseline models. For each pairwise outcomes between a target model M and a baseline model B, an instance records one of three possible outcomes: M wins, B wins, or a tie. Let w_{ij} denote the number of times model i is preferred over model j. Following the Bradley-Terry model (Bradley and Terry, 1952), each model i is associated with a latent ability score $\theta_i \in \mathcal{R}$, and the probability that i is preferred over j is given by

$$\Pr(i \succ j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)}. \quad (2)$$

Given the pairwise outcomes between M and the baseline models B_1, B_2, we compute the log-likelihood

$$\mathcal{L}(\theta) = \sum_{(i,j)} w_{ij} \cdot \log \Pr(i \succ j), \quad (3)$$

and obtain the ability estimates θ by maximizing this likelihood. The resulting estimate θ_M is

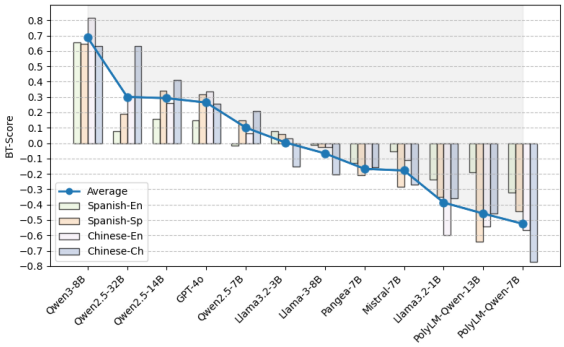


Figure 5: BT-Scores of the tested models, estimated from pairwise comparisons against two baselines(Mistral-7B and Pangea-7B). The BT model fits global ability params from the comparison outcomes; higher scores indicate stronger relative performance within this ability space. The abilities estimated by the BT model align closely with the Acc rankings.

reported as the **BT-Score** of model M.

$$\text{BT-Score}(M) = \theta_M = \arg\max_{\theta_M} \mathcal{L}(\theta). \quad (4)$$

where a higher BT-Score corresponds to stronger relative performance, as inferred from global ability parameters fitted using the baseline comparisons.

4.2 Statistics

Data Volume and Distribution The constructed benchmark consists of 8 knowledge datasets and 8 evaluation datasets grounded in Spanish, Chinese, French and Korean cultures, with two languages for each culture. There are **17.8k** cultural knowledge instances and **88.9k** questions in total. Figure 4 presents the volumes of cultural knowledge and evaluation datasets, as well as the distribution of the evaluation datasets. The distribution of cultural knowledge is provided in the Appendix F.

Data Instances All data have been released as open-source. We provide full examples of all data instances in Appendix H. The complete dimensional taxonomy is presented in Table 13. It consists of 5 Categories, 18 Topic Aspects, and 140 Detailed Dimensions, progressing from coarse to fine-grained levels. For each culture, our dataset is divided into two parts. The first part consists of cultural knowledge extracted and rewritten from various sources; examples are provided in Table 11. The extracted cultural knowledge instances are organized according to the evaluation dimensions. Each instance offers a characteristic description of the corresponding cultural dimension. The second part is a series of evaluation questions designed to assess models' cultural understanding from basic to advanced levels. Each instance in the evaluation set consists of a question, its corresponding answer, and the associated cultural knowledge. Examples of the evaluation data are shown in Figure 3.

5 Experiment Results

We evaluate different models across all dimensions, languages, and question types, and analyzed potential influencing factors. All detailed setups and parameters are shown in Appendix A. The overall performances of all LLMs are listed in Table 3 and Figure 5. We report the statistical reliability of our results in Appendix B.1 using BT confidence intervals. We have the following observations.

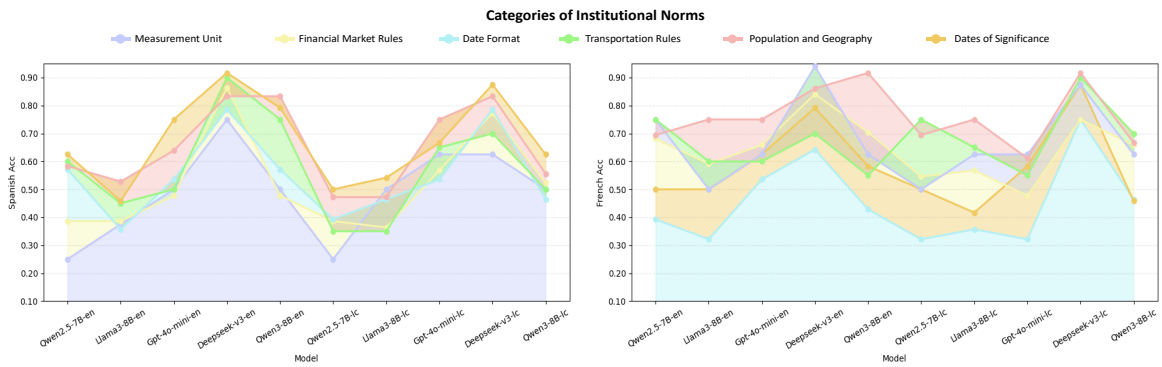
Observation 1: Stronger cultural understanding in English than local languages. Almost all models exhibit stronger cultural capabilities in English than in local languages, with an average performance gap of **6.66%**. We examined model performance on the same questions across different languages and found that this gap is associated with language proficiency. Qualitative analysis of model outputs are presented in the Appendix D.1. When asked to analyze the cultural meaning of bodily behaviors in Korean culture, Qwen2.5-7B provides a mechanism-focused analysis in its English response, mapping specific physical actions to their underlying cultural logic. In contrast, the Korean response is characterized by high redundancy and conceptual vagueness.

Observation 2: Weaker in cultural application. As is shown in Figure 7, most models perform worse on Mislead and Multi-hop, which assess cultural application, compared to other tasks. The average gaps on 4 cultures are **12.9%, 7.8%, 25.3% and 20.7%**. This suggests that current models are relatively weaker at applying cultural knowledge than at merely recalling it, indicating that application not only requires knowledge of the relevant facts but also correct reasoning based on that knowledge within a given context.

Observation 3: Larger models typically perform better on cultural understanding, as they encode a broader range of knowledge. Models from the same series exhibit a generally positive correlation between performance and model size; for example, Llama3-8B outperforms Llama3-1B by an average margin of **26.7%** on tested cultures. The similar phenomenon can also be observed in the Qwen family.

Model	Lan	Spanish Culture				Chinese Culture				French Culture				Korean Culture			
		Acc	Ins	Beh	Val	Acc	Ins	Beh	Val	Acc	Ins	Beh	Val	Acc	Ins	Beh	Val
General-purpose models																	
Qwen2.5-7B	en	0.489	0.525	0.471	0.483	0.580	0.525	0.583	0.645	0.696	0.593	0.721	0.774	0.669	0.581	0.688	0.741
	lc	0.425	0.400	0.434	0.435	0.411	0.408	0.376	0.491	0.616	0.525	0.634	0.693	0.408	0.406	0.387	0.459
Qwen2.5-14B	en	0.629	0.656	0.6327	0.588	0.670	0.647	0.655	0.733	0.778	0.650	0.826	0.838	0.742	0.662	0.750	0.830
	lc	0.514	0.487	0.536	0.500	0.514	0.484	0.505	0.572	0.733	0.625	0.775	0.782	0.507	0.487	0.518	0.508
Qwen2.5-32B	en	**0.632**	0.618	0.637	0.637	**0.694**	0.681	0.681	0.741	**0.791**	0.675	0.833	0.846	**0.775**	0.737	0.760	0.854
	lc	**0.608**	0.606	0.612	0.604	**0.541**	0.537	0.536	0.556	**0.751**	0.668	0.764	0.830	0.535	0.475	0.572	0.532
Llama3.2-1B	en	0.191	0.212	0.188	0.169	0.184	0.176	0.173	0.217	0.376	0.343	0.369	0.435	0.416	0.350	0.445	0.435
	lc	0.137	0.143	0.137	0.129	0.067	0.075	0.087	0.016	0.330	0.318	0.337	0.330	0.233	0.189	0.235	0.282
Llama3.2-3B	en	0.341	0.293	0.369	0.338	0.368	0.358	0.376	0.362	0.553	0.512	0.547	0.621	0.583	0.481	0.615	0.645
	lc	0.287	0.287	0.297	0.266	0.180	0.181	0.159	0.225	0.448	0.400	0.456	0.491	0.257	0.251	0.250	0.282
Llama3-8B	en	0.486	0.418	0.538	0.459	0.474	0.468	0.434	0.572	0.650	0.543	0.695	0.685	0.625	0.531	0.652	0.685
	lc	0.400	0.418	0.405	0.365	0.382	0.437	0.347	0.387	0.598	0.531	0.626	0.621	0.458	0.412	0.467	0.500
GPT-4o-mini	en	0.582	0.568	0.594	0.572	0.567	0.606	0.554	0.548	0.410	0.387	0.434	0.387	0.341	0.325	0.337	0.371
	lc	0.579	0.625	0.576	0.528	0.510	0.537	0.474	0.556	0.396	0.318	0.420	0.443	**0.560**	0.506	0.587	0.572
Deep-reasoning models																	
Deepseek-r1	en	**0.857**	0.909	0.825	0.881	**0.874**	0.880	0.855	0.911	0.535	0.543	0.543	0.508	0.539	0.575	0.500	0.580
	lc	**0.846**	0.818	0.851	0.869	**0.894**	0.886	0.891	0.911	0.516	0.456	0.536	0.548	0.701	0.673	0.692	0.758
Deepseek-v3	en	0.830	0.850	0.803	0.862	0.831	0.830	0.818	0.862	0.503	0.506	0.492	0.524	0.460	0.506	0.434	0.459
	lc	0.801	0.781	0.818	0.788	0.873	0.843	0.880	0.895	0.532	0.537	0.521	0.548	**0.715**	0.717	0.706	0.733
Qwen3-8B	en	0.648	0.656	0.637	0.661	0.706	0.710	0.681	0.758	**0.733**	0.662	0.739	0.814	**0.737**	0.687	0.731	0.814
	lc	0.580	0.531	0.601	0.596	0.658	0.662	0.655	0.661	**0.630**	0.575	0.619	0.725	0.523	0.487	0.532	0.548
Qwen3-8B-f	en	0.731	0.662	0.757	0.764	0.741	0.718	0.750	0.750	0.408	0.381	0.402	0.459	0.396	0.406	0.369	0.443
	lc	0.679	0.625	0.698	0.709	0.796	0.798	0.782	0.822	0.442	0.387	0.445	0.508	0.610	0.581	0.587	0.701
Models Continually-trained on Multilingual Corpus																	
PolyLM-7B	en	0.062	0.081	0.058	0.048	0.060	0.068	0.047	0.080	0.201	0.200	0.224	0.153	0.209	0.162	0.221	0.243
	lc	0.030	0.037	0.029	0.024	0.064	0.063	0.072	0.049	0.117	0.125	0.115	0.112	0.071	0.075	0.072	0.064
PolyLM-13B	en	0.325	0.387	0.297	0.306	0.298	0.343	0.391	0.032	0.301	0.281	0.304	0.322	0.244	0.250	0.231	0.266
	lc	0.173	0.206	0.166	0.145	0.075	0.069	0.072	0.088	0.092	0.106	0.094	0.072	0.225	0.175	0.246	0.241

Table 3: Main evaluation results on 4 cultures and 5 languages. We evaluate the question-answer accuracy of several types of LLMs on different languages and cultural layers mentioned above. Lang refers to the language of the questions. For each culture, we constructed parallel datasets in English(en) and the corresponding local language(lc) for evaluation. "Ins" stands for Institional Norms; "Beh" stands for Behavioral Patterns; "Val" stands for Core Values and Social Structures. Qwen3-8B-f refers to Qwen3-8B without deep reasoning enabled.



(a) Institutional Norms of Spanish culture

(b) Institutional Norms of French culture

Figure 6: Accuracy on the Institutional Norms within the Spanish cultural context and French cultural context. Many categories of cultural knowledge are areas where models typically exhibit limited capability—for example, Measurement Unit in Spanish culture and Date Format in French culture.

Observation 4: Knowledge injection does not promise better cultural application. We construct supervised fine-tuning (SFT) datasets based on extracted cultural knowledge and conduct training on Qwen2.5-7B, with the results presented in the Table 4. The datasets are disjoint from the test set, and settings are provided in Appendix A. The performance of the trained model shows notable improvements(1.08%-5.18%), supporting that cultural understanding is correlated with the acquisition of cultural knowledge, and that enhancing a model's knowledge of culture facilitates the advancement of its cultural understanding. However, training does not yield consistent gains on tasks that evaluate cultural application(Mislead & Multi-hop). This indicates that enhancing cultural application capabilities requires dedicated training rather than mere knowledge injection.

Observation 5: Deep reasoning does not guarantee stronger cultural understanding. Deep rea-

Culture	LLM	Acc	CATs.	Alpaca	MMLU
Spanish	Base	82.14%	76.8%	81.8%	85.4%
	SFT	**87.32%**	76.4%	82.0%	85.2%
Chinese	Base	73.21%	84.9%	81.8%	85.4%
	SFT	**74.29%**	84.8%	81.9%	85.6%

Table 4: SFT-based knowledge injection enhances the model's cultural understanding while preserving its general-purpose abilities. However, for Cultural Application Tasks (CATs)—including Mislead and Multi-hop questions—the injected knowledge does not yield consistent improvements.

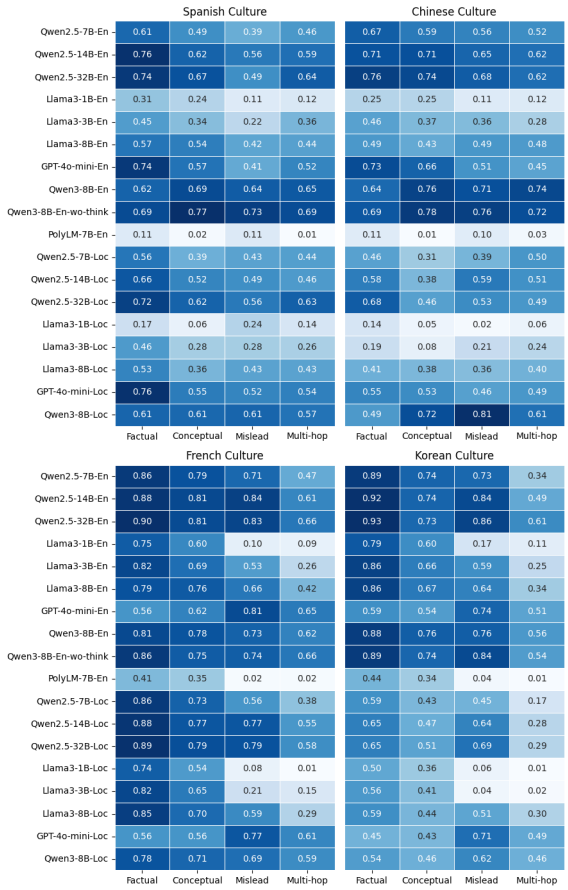


Figure 7: Models' performances across different types of tasks. Many models demonstrate weaknesses in cultural application tasks, like bias identification in misleading environment and multi-hop cultural reasoning.

soning improves performance in only **56%** of the experimental cases, indicating that its benefits are far from universal (see Appendix D.2 for qualitative analysis). A closer comparison between Qwen3-8B and Qwen3-max further reveals a reversal in reasoning effectiveness. The smaller 8B model correctly identifies cultural implications in its standard mode (e.g., giving someone a vacuum cleaner may imply criticism of the host's cleanliness). However, when deep reasoning is enabled, it produces incorrect explanations. By contrast, the larger Qwen3-max consistently generates coherent reasoning under the same setting. These results suggest that deep reasoning requires not only cultural knowledge but also the ability to reliably externalize it through multi-step reasoning. Smaller models may hold correct cultural associations, yet forcing explicit reasoning often triggers over-rationalization or error propagation, while larger models maintain more stable reasoning trajectories.

Observation 6: Language≠Culture. Although PolyLM-Qwen models (Wei et al., 2023) are initialized from Qwen and further optimized with extensive multilingual data, they paradoxically exhibit significant performance drops(**43.5%** for 7B and **41.9%** for 14B) across all cultural tasks despite outperforming their Qwen base on numerous multilingual benchmarks[2]. This inversion indicates that multilingual proficiency does not equate to multicultural understanding. Increasing the proportion of language-specific data may sharpen linguistic fluency but does not inherently catalyze the acquisition of deep-seated cultural logic.

Observation 7: Weakness in Certain Cultural Categories. Many categories of cultural knowledge are areas where models typically exhibit limited capability. As shown in Figure 6, categories involving culture-specific conventions tend to exhibit notably lower performance. In Spanish cultural contexts, models struggle with the *Measurement Unit* category(6.46% on average), while in French contexts, performance on *Date Format* is significantly lower than on other categories(9.91% on average). These patterns suggest that models have difficulty handling implicit cultural conventions that vary across regions, even when the underlying concepts are relatively simple. The complete results of the Categories are in the Appendix B.

6 Conclusion

By proposing a theoretically grounded and extensible taxonomy, we address the critical challenge of accurately and comprehensively model culture knowledge and evaluate the cultural understanding capability of LLMs under a certain cultural context. Our experimental results reveal that current LLMs remain insufficient in cultural understanding, highlighting the need for targeted training, which offers valuable insights for the future enhancement and deployment of culturally aligned LLMs.

[2]As is shown in https://modelscope.cn/models/iic/nlp_polylm_qwen_7b_text_generation

Limitations

Coverage vs. Overlap Culture is not strictly divisible, and some dimensions may naturally overlap. However, this overlap is not our primary concern. Our main objective is to achieve maximal coverage of cultural knowledge for operational evaluation in certain cultural context, and we believe that our taxonomy is well-suited to fulfill this goal.

Use of Acc and LLM judge for "Culture" Cultural questions often do not have a single definitive answer. Taking this into account, we only apply direct matching for factual, objective questions. For subjective questions, the open-ended nature of the items and the characteristics of culture mean that a single question may have multiple reasonable answers. Therefore, we define Accuracy in this context as the absence of direct conflict with the reference answer, allowing some degree of uncertainty in the model's response. This consideration is explicitly incorporated into the prompts given to the LLM judge. Manual annotation and the introduction of pairwise metric BT-Score further confirm that this strategy is reasonable.

Sociodemographics While sociodemographic factors such as age, gender, and education can influence cultural understanding, our study does not aim to evaluate models' performance in fine-grained, realistic and personalized cultural role-playing. Meanwhile, these factors could introduce higher risks of overgeneralization, stereotyping, or reinforcing social prejudices (Zhu et al., 2025). Consequently, these factors are outside the scope of the current work. Instead, our design focus on assessing models' general cultural mastery and reasoning across broader, population-level contexts rather than subgroup-specific behavior.

Use of Manual Annotation While our framework does not completely eliminate human annotation, this choice is intentional. Although manual annotation is costly, it remains a reliable source of high-quality cultural information. Compared to prior approaches that rely almost entirely on human annotation, our framework is suitable for **scenarios where** minimizing manual labeling is needed, as it combines automated knowledge extraction and question generation with selective manual verification to maintain accuracy while lowering costs.

Cultural Knowledge vs. Stereotypes Stereotypes in datasets are a common challenge in related research. Although strategies like human annotation help mitigate them, such biases are still an inherent aspect of real-world data. Our evaluation framework addresses this by combining dedicated multi-source knowledge extraction with targeted manual annotation to reduce the impact of stereotypes. In this process, we design prompts so that the knowledge extracted by LLMs emphasizes descriptive **tendencies** rather than definitive judgments.

Ethical Considerations

Ethical data curation This work involves the construction and use of a large-scale cultural understanding evaluation dataset, which contains content ranging from common cultural norms to sensitive aspects such as values and beliefs. The data were collected exclusively from publicly available sources, including educational materials, open-access cultural databases, and publicly shared questionnaires or forums, in compliance with the policies and terms of service of the respective platforms. No private, proprietary, or identifiable personal data was included.

We acknowledge that cultural data can reflect deeply held values and potentially contentious viewpoints. In constructing the dataset, we took care to avoid harmful stereotypes and minimizing cultural bias. When selecting and annotating the data, we consulted domain experts in cultural studies to ensure contextual fidelity and respect toward the represented communities.

Intended use Furthermore, the dataset is intended solely for research purposes, such as evaluating the capabilities of language models in understanding cultural knowledge. It is not designed to make normative claims about cultures, nor should it be used for profiling, ranking, or stereotyping individuals or groups. To mitigate risks of misuse, we restrict the release of the dataset under a research-use license and require users to agree not to employ it for downstream applications that could cause harm, including but not limited to cultural discrimination, misinformation propagation, or automated decision-making in sensitive contexts (e.g., hiring, immigration, education). We are committed to ongoing ethical reflection and welcome feedback from cultural communities, stakeholders, and the wider research community.

Use of AI Assistants We use ChatGPT to correct our grammatical errors and polish our texts.

References

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

Maria Iosifina Avgousti. 2018. Intercultural communicative competence and online exchanges: A systematic review. *Computer Assisted Language Learning*, 31(8):819–853.

G Bharathi Mohan, R Prasanna Kumar, P Vishal Krishh, A Keerthinathan, G Lavanya, Meka Kavya Uma Meghana, Sheba Sulthana, and Srinath Doss. 2024. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, 66(9):5047–5070.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Fan Bu, Zheng Wang, Siyi Wang, and Ziyao Liu. 2025. An investigation into value misalignment in llm-generated texts for cultural heritage. *arXiv preprint arXiv:2501.02039*.

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 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms.

Michael Desmond, Zahra Ashktorab, Werner Geyer, Elizabeth M Daly, Martin Santillan Cooper, Qian Pan, Rahul Nair, Nico Wagner, and Tejaswini Pedapati. 2025. Evalassist: Llm-as-a-judge simplified. 39(28):29637–29639.

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 preprint arXiv:2407.21783*.

P Christopher Earley. 2002. Redefining interactions across cultures and organizations: Moving forward with cultural intelligence. *Research in organizational behavior*, 24:271–299.

Meina Feng, Muhammad Noor Bin Abdul Aziz, and Syarizan Dalib. 2025. Exploring intercultural communicative competence among chinese students: Insights from a qualitative study. *SAGE Open*, 15(2):21582440251336634.

Nicole Ferdinand and Paul Kitchin. 2012. *Events management: An international approach*. Sage.

Clifford Geertz. 1973. The interpretation of cultures: selected essays.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*, 11:e55988.

Maurice Halbwachs. 2020. *On collective memory*. University of Chicago press.

Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How wikipedia's reaction to popularity is causing its decline. *American behavioral scientist*, 57(5):664–688.

Edward T. Hall. 1976. *Beyond Culture*. Anchor Press / Doubleday, Garden City, New York.

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

Geert Hofstede and Culture‘s Consequences. 2001. Comparing values, behaviors, institutions and organizations across nations. *Tilburg University, Netherlands: SAGE Publishing*.

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*.

Ronald Inglehart and Christian Welzel. 2005. Cultural change and democracy: the human development sequence. *New York: Cambridge University*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating

the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*.

Florence R Kluckhohn and Fred L Strodtbeck. 1961. Variations in value orientations.

Alfred Louis Kroeber and Clyde Kluckhohn. 1952. Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. Kornat: Llm alignment benchmark for korean social values and common knowledge. *arXiv preprint arXiv:2402.13605*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jiaqing Liang. 2025. Lsrl: Memory efficient large model training framework. https://github.com/lsdefine/lsrl.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. Aradice: Benchmarks for dialectal and cultural capabilities in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218.

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.

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

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. 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 932–950.

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.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

Aida Ramezani and Yang Xu. 2025. The discordance between embedded ethics and cultural inference in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14726–14747.

Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. 2025. Industrial applications of large language models. *Scientific Reports*, 15(1):13755.

Anamaria Sasu. 2016. The cultural iceberg and other cultural models. *Buletinul Științific al Universității Tehnice de Construcții București Seria: Limbi Străine și Comunicare*, 9(2):79–87.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Edward Burnett Tylor. 1871. *Primitive culture: researches into the development of mythology, philosophy, religion, art, and custom*, volume 2. J. Murray.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. Cdeval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, and 1 others. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Lee F Wilberschied Ph D. 2015. Intercultural communicative competence: Literature review.

Srishti Yadav, Lauren Tilton, Maria Antoniak, Taylor Arnold, Jiaang Li, Siddhesh Milind Pawar, Antonia Karamolegkou, Stella Frank, Zhaochong An, Negar Rostamzadeh, and 1 others. 2025. Cultural evaluations of vision-language models have a lot to learn from cultural theory. *arXiv preprint arXiv:2505.22793*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Taha Yasseri, Robert Sumi, and János Kertész. 2012. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1):e30091.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Xinyu Zhang, Pei Zhang, Shuang Luo, Jialong Tang, Yu Wan, Baosong Yang, and Fei Huang. 2025. Culturesynth: A hierarchical taxonomy-guided and retrieval-augmented framework for cultural question-answer synthesis. *arXiv preprint arXiv:2509.10886*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

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

Hao Zhu, Bodhisattwa Prasad Majumder, Dirk Hovy, and Diyi Yang. 2025. Social intelligence in the age of llms. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 51–55.

A Experimental Setups

A.1 Evaluation Setups

Models Our evaluation include three types of LLMs: General-purpose models, including the Instruct models of Qwen-2.5-series (Yang et al., 2024), Llama-3-series (Dubey et al., 2024) and GPT-4o (Hurst et al., 2024); Deep-reasoning models and its non-deep-reasoning version, including Qwen-3 (Yang et al., 2025), Deepseek-r1 (Guo et al., 2025)and Deepseek-v3 (Liu et al., 2024); Models continually-trained specifically on multilingual data, including Polylm-Qwen-7B and Polylm-Qwen-13B (Wei et al., 2023). The baselines used for computing the BT-Score are Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and Pangea-7B (Yue et al., 2024). We adopt DeepSeek-R1 as the judge model due to its strong reasoning capabilities, which enable reliable reference-based judgment.

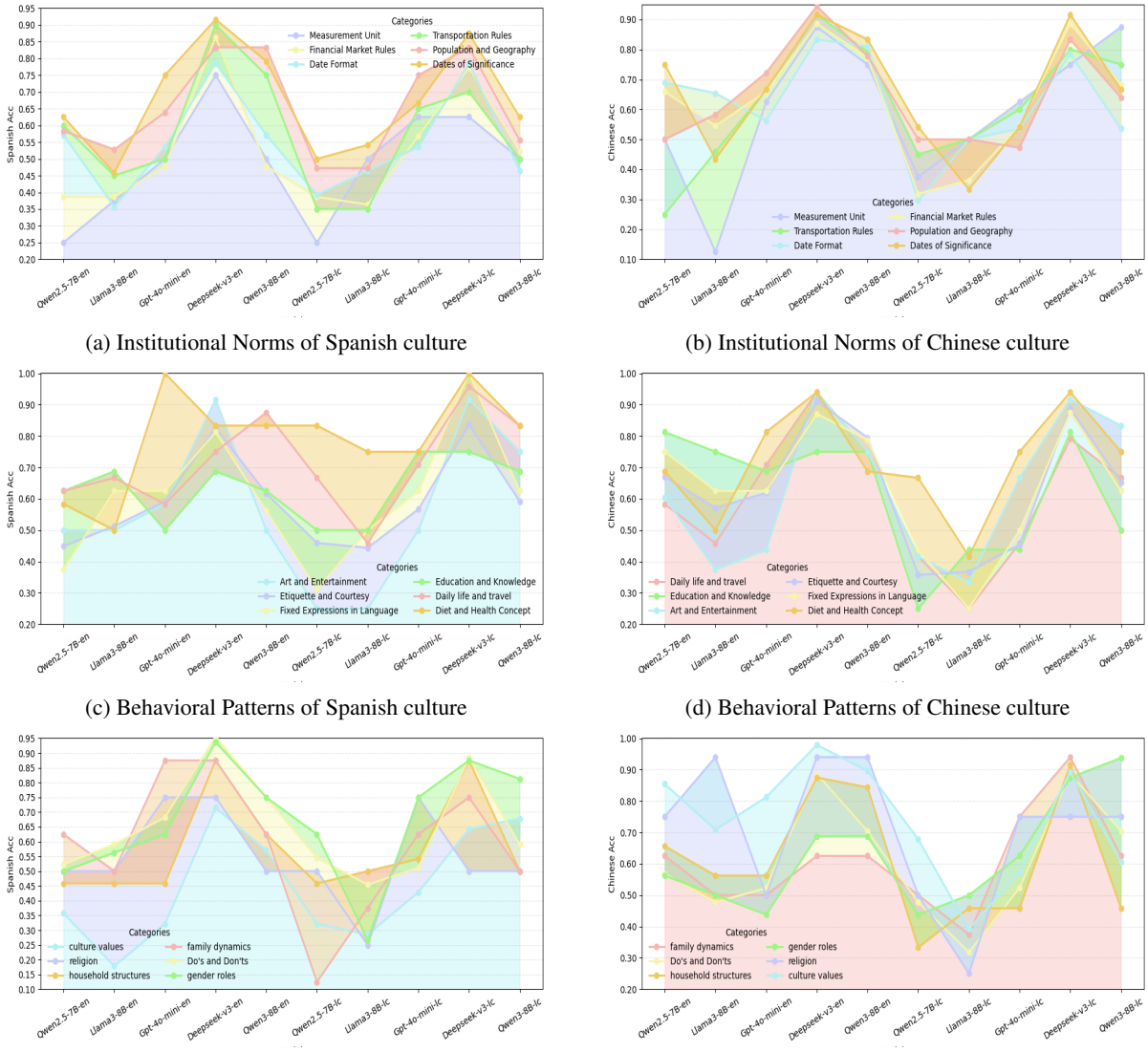
Parameters All local models are run with FP16 precision on NVIDIA A800 GPUs. Inference is conducted with VLLM, with the model loaded in float16 precision. The SamplingParams are set as follows: temperature=0.0, n=1, max_tokens=1024, and stop_token_ids=[2]; all other parameters retain their default values.

A.2 Training Setups

The training dataset and the evaluation dataset are completely disjoint. We constructed a novel training dataset based on the extracted cultural knowledge and performed SFT training on Qwen2.5-7B-base, aiming to enable it to memorize cultural knowledge that it was previously unfamiliar with. The training process was conducted using LSRL Framework (Liang, 2025). The cultural knowledge was formatted as question–answer pairs, with questions generated by Qwen2.5-72B-Instruct, and tokenized with a maximum sequence length of 12000. The training was conducted for 1 epoch with a batch size of 2 and a learning rate of 2e-6, using the CPUAdamW optimizer with gradient offloading enabled to reduce GPU memory usage. The model was trained in bfloat16 precision with gradient checkpointing to further optimize memory efficiency. The total number of training steps was determined dynamically based on the dataset size. To mitigate overfitting, we also incorporated some general-purpose datasets into the training data, including IFEval, Alpaca, and GSM8k.

Culture	Model	Language	Mean	95% CI	Reported Acc
Spanish	Qwen2.5-7B-Instruct	en	0.492	(0.481, 0.506)	0.489
		sp	0.429	(0.416, 0.440)	0.425
	LLama3-8B-Instruct	en	0.482	(0.453, 0.510)	0.486
		sp	0.406	(0.387, 0.424)	0.400
	Deepseek-r1	en	0.860	(0.840, 0.885)	0.857
		sp	0.845	(0.830, 0.862)	0.846
	GPT-4o-mini	en	0.580	(0.573, 0.586)	0.582
		sp	0.566	(0.535, 0.594)	0.579
	PolyLM-7B	en	0.061	(0.054, 0.070)	0.062
		sp	0.030	(0.027, 0.033)	0.030
Chinese	Qwen2.5-7B-Instruct	en	0.581	(0.553, 0.612)	0.580
		ch	0.417	(0.393, 0.450)	0.411
	LLama3-8B-Instruct	en	0.477	(0.448, 0.517)	0.474
		ch	0.386	(0.365, 0.411)	0.382
	Deepseek-r1	en	0.880	(0.867, 0.895)	0.874
		ch	0.892	(0.884, 0.901)	0.894
	GPT-4o-mini	en	0.568	(0.555, 0.584)	0.567
		ch	0.519	(0.498, 0.540)	0.510
	PolyLM-7B	en	0.064	(0.057, 0.072)	0.060
		ch	0.062	(0.056, 0.067)	0.064
French	Qwen2.5-7B-Instruct	en	0.694	(0.645, 0.734)	0.696
		fr	0.618	(0.575, 0.657)	0.616
	LLama3-8B-Instruct	en	0.645	(0.602, 0.678)	0.650
		fr	0.592	(0.566, 0.614)	0.598
	Deepseek-r1	en	0.537	(0.525, 0.548)	0.545
		fr	0.517	(0.492, 0.537)	0.516
	GPT-4o-mini	en	0.406	(0.395, 0.422)	0.410
		fr	0.790	(0.750, 0.824)	0.761
	PolyLM-7B	en	0.203	(0.181, 0.222)	0.201
		fr	0.117	(0.114, 0.121)	0.117
Korean	Qwen2.5-7B-Instruct	en	0.668	(0.628, 0.705)	0.669
		kr	0.410	(0.396, 0.431)	0.408
	LLama3-8B-Instruct	en	0.628	(0.585, 0.660)	0.625
		kr	0.458	(0.435, 0.479)	0.458
	Deepseek-r1	en	0.546	(0.525, 0.566)	0.539
		kr	0.707	(0.688, 0.730)	0.701
	GPT-4o-mini	en	0.343	(0.333, 0.355)	0.341
		kr	0.558	(0.535, 0.575)	0.561
	PolyLM-7B	en	0.210	(0.187, 0.229)	0.209
		kr	0.071	(0.068, 0.074)	0.071

Table 5: We report the results from six independent runs and compute the BT confidence intervals using 1,000 bootstrap samples. **All** reported results fall within the 95% confidence intervals, demonstrating the credibility of our evaluation.



(a) Institutional Norms of Spanish culture

(b) Institutional Norms of Chinese culture

(c) Behavioral Patterns of Spanish culture

(d) Behavioral Patterns of Chinese culture

(e) Core Values and Social Structures of Spanish culture

(f) Core Values and Social Structures of Chinese culture

Figure 8: Performance of Models across Layers and Languages (a–f).

B Supplementary Experimental Results

B.1 Result robustness

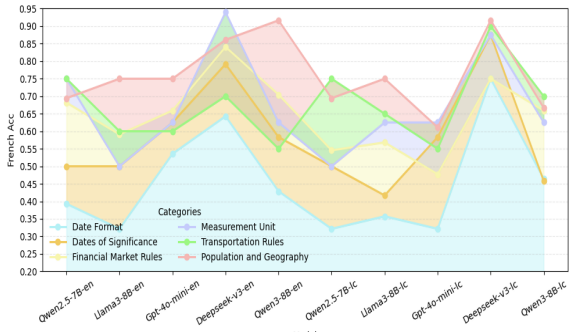
We performed repeated sampling for a subset of models and computed their BT confidence intervals, as shown in Table 5. Specifically, each model was independently sampled six times, and BT confidence was estimated using 1,000 bootstrap samples, with the corresponding mean values. **All** reported results fall within the 95% confidence intervals, demonstrating the credibility of our evaluation.

B.2 Fine-Grained Results

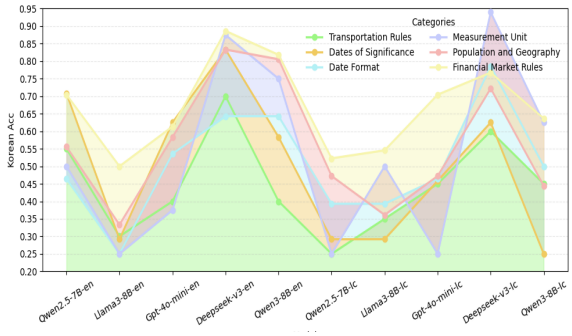
We evaluated models across stratified Layers and finer-grained Categories (Figure 8). In Spanish culture, performance divergence across languages intensifies as evaluation depth increases: **Institutional Norms:** Models show relative cross-lingual consistency. Performance is universally poor on Measurement Units but consistently strong on Dates of Significance; **Behavioral Habits:** Language preferences emerge. In Art and Entertainment, models struggle with Spanish queries but perform better with English prompts; **Cultural Values:** Divergence is most pronounced. In Religion, models excel in Spanish but fail in English. Conversely, for Family Dynamics and Household Structures, English queries yield higher performance than Spanish ones.

C Culture Knowledge Classification Taxonomy

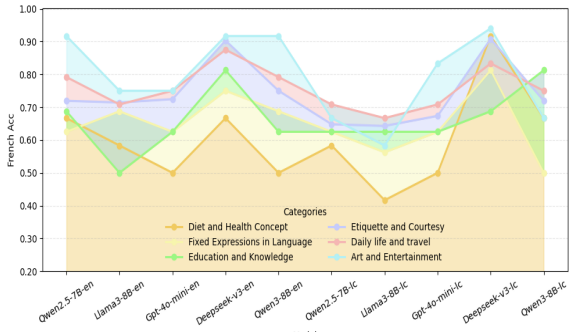
In this section, we provide a detailed introduction to the cultural knowledge modeling taxonomy we pro-



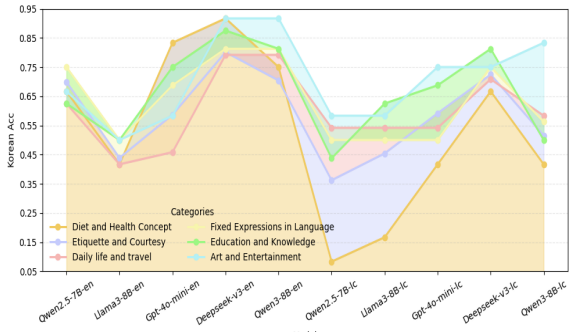
(g) Institutional Norms of French culture



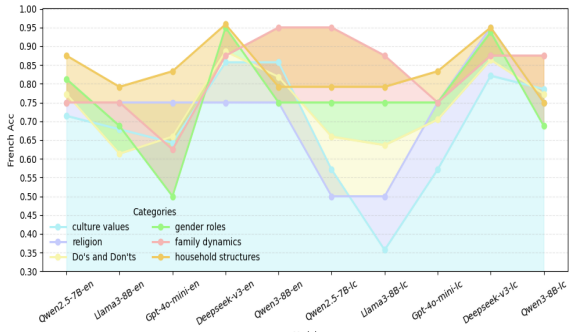
(h) Institutional Norms of Korean culture



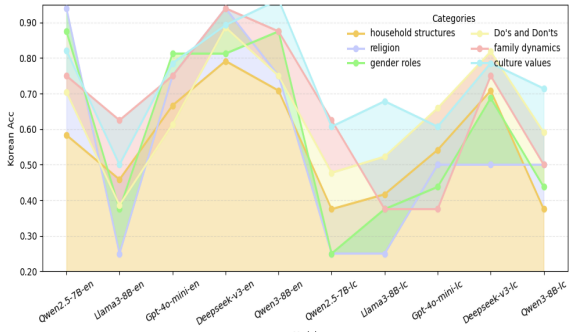
(i) Behavioral Patterns of French culture



(j) Behavioral Patterns of Korean culture



(k) Core Values and Social Structures of French culture



(l) Core Values and Social Structures of Korean culture

Figure 8: Performance of Models across Layers and Languages (g–l).

pose. We demonstrate its advantages over existing approaches from both **Theoretical Foundations** and **Statistical Results**. In addition, we present the design principles and detailed dimensions of the taxonomy.

C.1 Theoretical Foundations

Regarding the design of the dimensional taxonomy, we invited multiple experts for discussion, including two professors specializing in cognition and culture, one professor in computer science, and three PhD students. This taxonomy integrates numerous existing cross-cultural and single-culture research theories. The specific design is as follows:

At the explicit layer, we follow the "surface" part of the **cultural iceberg model** (Hall, 1976) and focus on institutionalised, directly observable aspects of culture, such as traffic rules, date formats, measurement units, financial regulations, population and geography, and public holidays. These categories capture what visitors typically first encounter when interacting with a new society.

At the intermediate layer, we model everyday practices and habitual choices that are less formalised but still observable: daily routines and mobility, food and health beliefs, views on education and knowledge, artistic and entertainment preferences, etiquette and rituals, and conventional linguistic expressions. This layer is informed by Hall's work on **High-Context and Low-Context Culture** (Hall, 1976), **Cultural Patterns** (Kroeber and Kluckhohn, 1952) and Kluckhohn & Strodtbeck's **Value Orientation Theory** (Kluckhohn and Strodtbeck, 1961), which emphasise how values

are enacted through recurrent patterns of Behaviour and communication.

At the implicit layer, we move to deep structures of culture: family relations and social hierarchy, gender roles and equality, broader "cultural dimensions" of value preferences, religious beliefs, and taboos (Do's and Don'ts). Here we draw inspiration from **Social Memory Theory** (Halbwachs, 2020) and explicitly incorporate **Hofstede's value dimensions** (Hofstede, 1984)(e.g., power distance, individualism–collectivism, uncertainty avoidance, masculinity–femininity) and related value-orientation work to guide which aspects of hierarchy, authority, risk, and role expectations are encoded in our "social relations and structures" and "values and beliefs" categories.

In summary, Hofstede's dimensions are used at the deep layer of our taxonomy and are complemented by the cultural iceberg perspective, cultural patterns, value-orientation theories, high/low-context communication, and social memory theory to cover both visible institutional rules and hidden value systems. Whether in works that design taxonomys purely based on intuition or those that directly adopt theories like Hofstede's dimensions, our taxonomy design is more comprehensive.

C.2 Design Principles for Taxonomy

To evaluate LLM cultural capability, our taxonomy follows four core principles:

- Comprehensiveness: Maximizing coverage across diverse cultural domains.
- Operability: Refining dimensions to the keyword level to directly guide data extraction and construction.
- Interpretability: Utilizing a tree structure for coarse-to-fine hierarchical analysis and granular performance attribution.
- Extensibility: Supporting both top-down node supplementation and bottom-up data-driven expansion within the scalable tree framework.

C.3 Detailed Dimensions

Our proposed dimension taxonomy contains 4 levels, *Cultural Layers, Category, Topic Aspect and Detailed Dimensions*. There are in total 3 Cultural Layers, 5 Categories, 18 Topic Aspects and 140 Detailed Dimensions, encompassing a wide range of cultural knowledge from explicit commonsense facts, Behavioral patterns to implicit values.

The surface-level *Cultural Layer* is the **Institutional Norms Layer**, which includes explicit commonsense knowledge such as geography, customs, rules, and policies within specific cultural contexts; the intermediate layer is **Behavioral Patterns Layer**, which encompasses personal choices and habits, including aspects such as daily life and travel, education, and concepts of knowledge; the deep-level layer is the **Core Values and Social Structures Layer**, which includes deep cultural knowledge such as social relationships and structures, values, and beliefs. Each *Cultural Layer* include a series of *Categories, Topic Aspects and Detailed Dimensions.* A detailed description of the dimensions is provided in Table 13.

D Case study of Failure modes

This section presents case analyses of selected experimental results to provide deeper insight into our findings, comprising two parts: (1) instances in which English succeeds but the local language fails, and (2) instances in which non-deep reasoning succeeds whereas deep reasoning fails. Each part includes examples of different responses, along with an analytical commentary.

D.1 English vs. Local Language

We presented the responses of Qwen2.5-7B-Instruct to the same question in different languages and observed a clear difference in quality between the two outputs. Such phenomena were consistently observed across all tested models.

Question At a formal dinner, a young employee pauses before taking a seat near the door, waiting until a senior executive and department head are seated in their respective hierarchical positions. After sitting, the employee waits for others to begin eating before starting their own meal. Why did the employee delay these actions, and how do they reflect unspoken expectations regarding interpersonal dynamics and social hierarchy in Korea?

English response The young employee's actions demonstrate a rigorous adherence to Korean dining hierarchy through four key behaviors:

1. Hierarchical Deference: By waiting for the senior executive and department head to sit first, the employee acknowledges the established social order and ensures superiors are accommodated.

2. Seating Protocol: Their observation of the seating arrangement respects the tradition where

the most senior individual occupies the shangzuo (seat furthest from the entrance).

3. Synchronized Initiation: Waiting for the most senior person or host to begin eating recognizes their role as the leader of the meal, a core sign of respect in collective dining.

4. Disciplined Etiquette: Maintaining a formal posture with hands in their lap signifies attentiveness to social norms and self-restraint until the meal officially commences.

In summary, these actions reflect a sophisticated understanding of how physical gestures operationalize unspoken hierarchical dynamics.

Korean response(Translated into English) The employee delayed sitting and starting the meal to follow social norms and etiquette in Korean society. Their behavior reflects the following principles:

1. Mutual Respect: Following the supervisors' and executives' behavior is an expression of mutual respect, showing they respect colleagues and acknowledge their roles.

2. Social Norms: In Korean society, mirroring the actions of others is considered a social norm, and the employee's behavior demonstrates adherence to these norms.

3. Teamwork: By following the seniors' actions, the employee emphasizes teamwork, showing they respect and cooperate with others as members of the team.

4. Cultural Etiquette: The employee's behavior demonstrates adherence to cultural etiquette, showing they understand and follow the values and norms of Korean society.

The young employee's actions demonstrate social norms and etiquette in Korean society.....
The young employee's actions demonstrate social norms and etiquette in Korean society.....

Analysis The English response provides a mechanism-focused analysis, mapping specific physical actions to their underlying cultural logic. In contrast, the Korean response is characterized by high redundancy and conceptual vagueness, repeatedly using generalized terms like "respect" and "social norms" without explaining the actual cultural rules. While the first version offers taxonomic depth, the second relies on circular reasoning, offering little information beyond the surface-level observation.

D.2 With vs. Without Deep Thinking

We present the results of Qwen3-8B with and without deep reasoning, as well as Qwen3-max with deep reasoning enabled below. We find that when deep reasoning is enabled, the 8B model makes errors in core explanations, whereas the larger Qwen3-max produces correct reasoning. This observation highlights the crucial role of **knowledge capacity** and **reasoning ability** in multi-step inference.

Question A host in Lyon reacts with visible discomfort after receiving a vacuum cleaner. However, he warmly thank a third guest who offers chocolates. Why did the first gift cause discomfort despite the guest' good intention?

Qwen3-8B with thinking A vacuum cleaner is typically a practical gift, but it might not be suitable for a housewarming, as it is more of a household appliance rather than a gift for the occasion.....

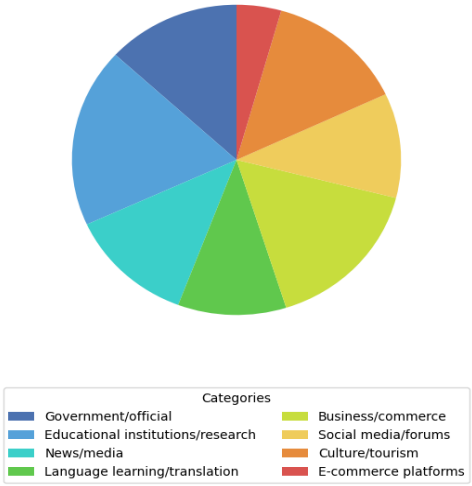
Qwen3-8B wo thinking - A vacuum cleaner is a practical gift, but it may be seen as overly utilitarian or even impersonal, especially in a social setting like a dinner party. **It could also imply that** the host is focused on cleaning rather than enjoying the occasion.....

Qwen3-max with thinking In French culture, especially during a dinner party centered around cooking and hosting, a vacuum cleaner can be interpreted as a **subtle critique of the host's home cleanliness or housekeeping skills**.....

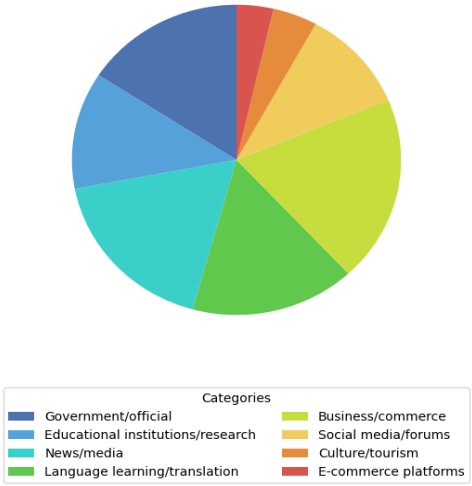
Analysis While the smaller model's standard response correctly identifies the hidden cultural metaphors, enabling CoT paradoxically leads it to a superficial, incorrect interpretation. In contrast, the larger model succeeds with CoT, providing rigorous justification. This demonstrates that reasoning amplifies internal knowledge: when capacity is limited, deeper reasoning does not refine the answer but instead propagates and amplifies errors. Ultimately, CoT's efficacy is strictly contingent upon the model's underlying knowledge density.

E The Dynamics of Culture

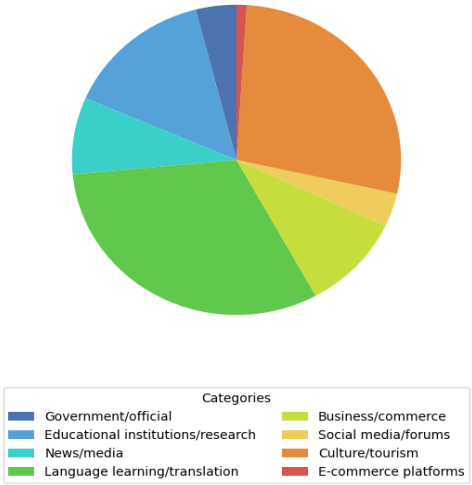
To accommodate the evolving nature of culture, our framework utilizes dynamic data sources and collection methods. While this study does not focus on cultural dynamism directly, our adaptive data pipeline ensures that the evaluation remains current



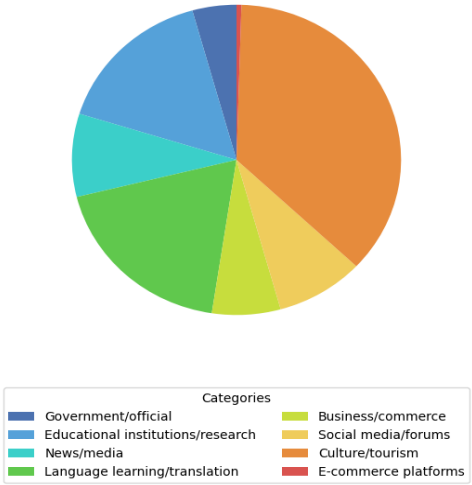
(a) Spanish Culture



(b) Chinese Culture



(c) French Culture



(d) Korean Culture

Figure 9: Data Source Statistics for Different Extracted Cultural Knowledge Instances

and robust, as elaborated below. We elaborate on this point below.

Efficient and Incremental Knowledge Updating When cultural knowledge evolves, we simply reconstruct queries for the affected dimensions and re-crawl the web. By leveraging Google's search mechanisms, we naturally obtain the most current results, ensuring our knowledge base remains updated and synchronized with real-world shifts.

Integration of Dynamic Sources Our framework combines authoritative data with dynamic web search to balance quality and coverage. While static authoritative websites provide high-quality but limited information, our integration of web-search results—all human-verified—ensures a more comprehensive representation of evolving culture. Statistics confirm that this approach reliably captures accurate and up-to-date cultural nuances.

Stability of High-Quality Sources High-quality cultural sources, such as Wikipedia and official portals, exhibit low-frequency, incremental updates once entries mature (Halfaker et al., 2013; Yasseri et al., 2012). Core cultural structures—including values, festivals, and etiquette—are stable, slow-moving variables (Hofstede and Consequences, 2001; Inglehart and Welzel, 2005). Only peripheral or policy-related aspects (e.g., festival regulations) undergo periodic adjustments (Ferdinand and Kitchin, 2012). These volatile elements account for only 25 dimensions (less than 18%) of our 140-dimensional taxonomy, a scale that is entirely manageable for evaluation purposes.

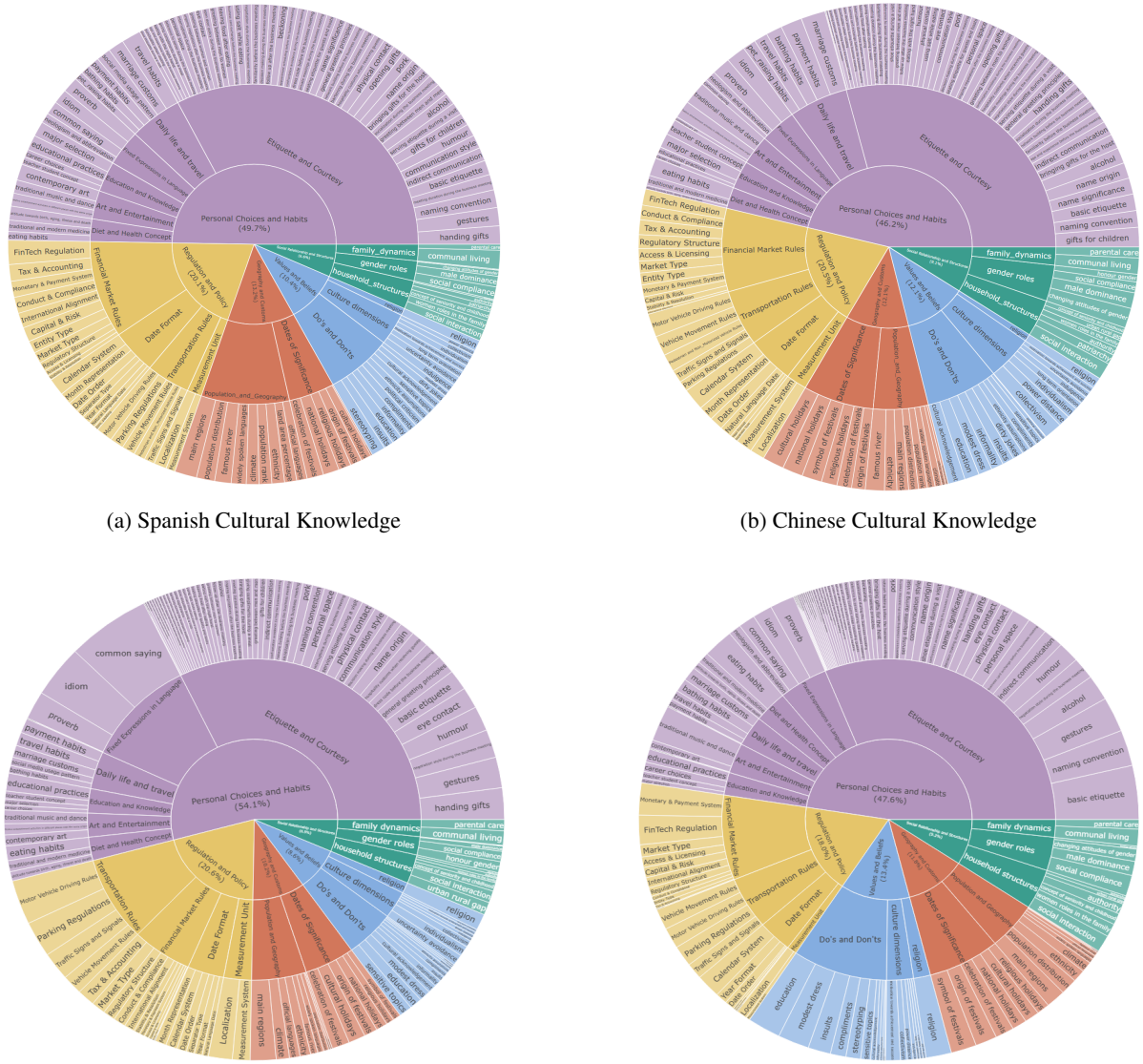
Culture	URL	Type
Spanish	https://culturalatlas.sbs.com.au	Educational institutions/research
	https://businessculture.org	Educational institutions/research
	ttps://www.expatica.com	Culture/tourism
	https://en.wikipedia.org	Educational institutions/research
	https://www.donquijote.org	Language learning/translation
	https://www.lingoda.com	Language learning/translation
	https://baselang.com	Language learning/translation
	https://cibtvisas.com	Business/commerce
Chinese	https://en.wikipedia.org	Educational institutions/research
	https://www.chinahighlights.com	Culture/tourism
	https://culturalatlas.sbs.com.au	Educational institutions/research
	https://www.reddit.com	Social media/forums
	https://lansonplace.com	Business/commerce
	https://travelchinawith.me	Culture/tourism
	https://pmc.ncbi.nlm.nih.gov	Educational institutions/research
	https://commons.ln.edu.hk	Educational institutions/research
French	https://www.fluentu.com	Language learning/translation
	https://www.polyglottistlanguageacademy.com	Language learning/translation
	https://culturalatlas.sbs.com.au	Educational institutions/research
	https://www.learnfrenchwithsonal.com	Language learning/translation
	https://commisceo-global.com	Business/commerce
	https://www.ccfs-sorbonne.fr	Educational institutions/research
	https://www.justlanded.com	Culture/Tourism
	https://businessculture.org	Educational institutions/research
Korean	https://everythingkorea.blog	Culture/Tourism
	https://link.springer.com	Educational institutions/research
	https://www.koreanhabits.com	Culture/Tourism
	https://www.90daykorean.com	Language learning/translation
	https://culturalatlas.sbs.com.au	Educational institutions/research
	https://www.namhansouthkorea.com	Culture/Tourism
	https://pmc.ncbi.nlm.nih.gov	Educational institutions/research
	https://ed-money.tistory.com	Social media/forums

Table 6: This table presents the URLs, categories, and counts of source webpages retrieved in the cultural knowledge extraction processes for the four cultures, using queries in English.

F Data Sources and Data Distribution

Data Sources We identified eight knowledge source categories: government/official, educational/research, news/media, language/translation, business, social media, culture/tourism, and e-commerce. As shown in Figure 9, source distributions vary by language; for instance, culture/-tourism sites contribute 20% of Spanish instances but significantly less for Chinese, highlighting cross-cultural variations in internet data distribution. Table 6 lists the top 8 most frequent source URLs across the four cultures. The most frequent sources consistently belong to high-quality categories, such as educational institutions, government portals, and culture/tourism sites, validating the reliability of our extraction process.

Cultural Knowledge Distribution Figure 10 shows the distribution of cultural knowledge across dimensions. Despite varying data sources, the distributions remain remarkably consistent across languages, aligning closely with our taxonomy's structure. This uniformity suggests that culturally relevant web content density is relatively balanced across different linguistic contexts. Consequently, our retrieval and rewriting pipeline can extract comparable volumes of knowledge for each dimension, regardless of the target language.



(a) Spanish Cultural Knowledge

(b) Chinese Cultural Knowledge

(c) French Cultural Knowledge

(d) Korean Cultural Knowledge

Figure 10: Distribution of Retrieved Cultural Knowledge Instances. These cultural knowledge instances are collected from multiple sources under the guidance of the dimensional taxonomy, and their distribution is consistent with the dimensions defined in the taxonomy.

Culture	Annotator	Age	Gender	Education Level	Cultural Background
Spanish	Annotator1	-29	Female	Master	Spanish
	Annotator2	30-39	Female	Doctorate	Spanish & Chinese
Chinese	Annotator1	-29	Male	Master	Chinese
	Annotator2	40-49	Female	Bachelor	Chinese
French	Annotator1	-29	Female	Master	French
	Annotator2	30-39	Male	Master	French
Korean	Annotator1	-29	Female	Master	Korean
	Annotator2	-29	Male	Bachelor	Korean & Chinese

Table 7: Detailed information for culture annotators. Two annotators with the corresponding cultural background are invited for each culture.

G Annotation Details

Annotation of Knowledge and Questions Two experts per culture (native/background speakers) performed sampled annotations of the cultural knowledge and evaluation sets. Experts assessed knowledge entries for factual errors or biases and evaluated the logical consistency between questions, answers, and source knowledge. Participants—colleagues and students from our institution—were fairly compensated. Expert profiles, instructions, and inter-annotator agreement rates (measured by simple agreement) are detailed in Tables 7, 10, and 9, respectively.

Annotation of LLM Judges We invited human annotators to label the judgment results of LLMs. We collect 1000 samples for each culture. The annotation results are shown in Table 8.

H Data Instances

H.1 Cultural Knowledge

Extracted instances are organized by evaluation dimensions, providing characteristic descriptions of commonsense, beliefs, and values within specific contexts. During construction, these dimensions served as cues for targeted retrieval and rewriting. Within each dimension, instances were further clustered into fine-grained sub-dimensions to guide future evaluations. Table 11 illustrates representative examples from the resulting knowledge instances.

Culture	Agreement of Judges
Spanish	97.80%
Chinese	99.10%
French	98.30%
Korean	97.60%

Table 8: Simple agreement rate of judge results of subjective questions on llm judges and human annotators. We collect 1000 judge samples for each culture.

Culture	Metric	Knowledge	Questions
Spanish	Annotator1	95.71%	96.61%
	Annotator2	92.86%	98.21%
Chinese	Annotator1	97.84%	96.07%
	Annotator2	97.84%	98.93%
French	Annotator1	90.00%	98.21%
	Annotator2	95.00%	98.93%
Korean	Annotator1	92.86%	97.86%
	Agreement	93.57%	98.75%

Table 9: Details annotation results of all annotators. Two annotators are invited for each culture.

H.2 Evaluation Dataset

As is mentioned in the main context, the questions in our evaluation dataset are categorized into four tasks based on their content: Factual, Conceptual, Misleading, and Multi-hop reasoning. Each instance in the evaluation set consists of a question, its corresponding answer, and the associated cultural knowledge. The questions are categorized into subjective types (short answer and essay questions) and objective types (single-choice, multiple-choice, and true/false questions). The complete examples of these types of questions are provided in Table 12.

I Prompts

We leveraged LLMs for text rewriting, question generation, and quality assessment. In these steps, we employed carefully expert-designed prompts to guide the LLMs in performing the corresponding tasks. Here, we present the prompts used across different languages in the last pages. All operations are conducted using prompts written in the corresponding target language; the examples shown here are the English versions.

Target	Instruction
Knowledge	You are asked to evaluate individual cultural knowledge entries. Your task is to determine whether each entry contains serious factual errors or cultural biases. Specifically, please assess: Factual correctness: Whether the described cultural facts, practices, or norms are accurate and not misleading. Bias and stereotyping: Whether the entry contains overgeneralizations, stereotypes, or culturally biased statements that could lead to misunderstanding or unfair judgments. Severity: Only mark an entry as problematic if the error or bias is substantial (i.e., likely to misinform or distort cultural understanding), rather than minor imprecision or simplification. Please base your judgment on widely accepted cultural knowledge and avoid overly strict or subjective interpretations.
Evaluation Dataset	You are asked to evaluate entries in the evaluation dataset, which consist of questions, answers, and associated cultural knowledge. Your task is to determine whether these components are logically consistent with each other. Specifically, please assess: Question–answer alignment: Whether the answer directly and appropriately addresses the question. Consistency with cultural knowledge: Whether the answer is logically supported by the provided cultural knowledge and does not contradict it. Internal coherence: Whether there are logical inconsistencies, contradictions, or reasoning gaps within or across the question, answer, and cultural context. Focus on logical coherence rather than stylistic quality or linguistic fluency.
LLM Judges	You are asked to evaluate the outputs of an LLM-based Judge, which assesses model responses in cultural tasks. Your task is to determine whether the Judge's decision and justification accurately reflect human expert judgment. Specifically, please assess: Judgment correctness: Whether the Judge's final decision (e.g., correct/incorrect, better/worse) aligns with what a human expert would reasonably conclude given the task and cultural context. Reasoning validity: Whether the Judge's explanation correctly identifies the key cultural or logical factors that justify its decision, without relying on irrelevant, hallucinated, or misleading considerations. Consistency with evaluation criteria: Whether the Judge applies the stated evaluation standards appropriately and consistently, rather than making arbitrary or superficial judgments. Focus on whether the Judge's output can be considered a reliable proxy for human evaluation, rather than evaluating the underlying model response itself.

Table 10: Instructions of the annotators

Dimension	Knowledge	Clustered category
Alcohol	[1] Alcohol is deeply integrated into daily life and social activities in Spain, such as drinking beer with friends at bars or enjoying wine with meals.	Cultural Integration
	[2] The legal drinking age in Spain is 18, with strict drink-driving laws (blood alcohol limit: 0.5 g/L).	Legal and Social Norms
	[3] Spaniards practice moderate drinking, with binge drinking being uncommon.	Drinking Habits and Moderation
	[4] Beer and wine are the most popular alcoholic beverages, with craft beer gaining popularity.	Regional and Traditional Beverages
Physical Contact	[1] Spaniards emphasize physical contact as a tactile and emotional expression to convey friendliness, intimacy, or support.	Frequency and Context of Physical Contact
	[2] Public displays of affection (e.g., holding hands) between couples are widely accepted as normal in Spanish culture.	Cultural Acceptance and Boundaries
	[3] Friends use gestures like nudging arms, linking arms, or holding shoulders to show appreciation and support.	Forms of Physical Contact
	[4] Casual touches (e.g., adjusting clothes or lightly touching collars) signal approachability and friendliness.	Forms of Physical Contact
Widely Spoken Languages	[1] Spanish is the official language in 20 countries and one U.S. territory (Puerto Rico).	Official Status and Recognition
	[2] Spanish is widely spoken in regions like the United States, parts of Africa, Asia, and Oceania.,	Geographical Distribution
	[3] The majority of Spanish speakers reside in Hispanic America, particularly Mexico (largest population of native speakers).	Geographical Distribution
	[4] Spanish is recognized as an official/co-official language in international organizations (UN, EU, OAS, African Union).	Official Status and Recognition

Table 11: Examples of extracted knowledge instances. Dimension refers to the finest-grained level used for retrieval and rewriting; these dimensions serve as keywords for Google search. Cluster category denotes the result of clustering the cultural knowledge, with each piece of knowledge corresponding to a more fine-grained category.

Type	Question	Answer	Associated Knowledge
Factual	During the festival of Las Fallas in Valencia, which element is prominently featured and plays a significant role in the celebrations? A) Ice sculptures B) Fireworks and bonfires C) Water fountains D) Sandcastles	B) Fireworks and bonfires	1. Use of Symbolic Elements: Symbolic elements such as fire, water, and specific foods play crucial roles in many Spanish festivals. 2. Fire is especially prominent in celebrations like Las Fallas and San Juan's Night.
Conceptual	In a traditional Korean meal featuring a clear broth soup, why might the soup be served with little to no seasoning, requiring the diner to adjust the flavor themselves at the table? A) To reduce sodium content for health reasons, reflecting modern dietary guidelines B) To allow each individual to tailor the taste according to their personal preference while honoring the purity of the ingredients C) Because traditional Korean cooking lacks precise control over seasoning during preparation D) To encourage communal sharing by making diners rely on shared condiments	B) To allow each individual to tailor the taste according to their personal preference while honoring the purity of the ingredients	1. In Korean dining culture, clear soups such as seolleongtang, galbitang, and samgyetang are served with minimal salt in the broth, and diners are expected to season them to taste at the table using provided salt and pepper. 2. The practice of adding salt at the table emphasizes personal preference and appreciation of high-quality ingredients, aiming for a clean, balanced flavor where no single taste dominates.
Mislead	When a teacher at an international school was communicating with Chinese students, he found that Chinese students were generally unwilling to participate in class discussions and almost never questioned the teacher's explanations. The teacher therefore believed that the Chinese education system does not encourage critical thinking and innovation at all. Do you think the teacher's point of view is correct? A. True B. False	B. False.	1. In cultures with high uncertainty avoidance, formal rules and procedures are emphasized, and organizations often focus on structured decision-making and task orientation. 2. China's education system traditionally emphasizes standardized exams and a unified curriculum, but recent reforms increasingly promote critical thinking and innovation. 3. Traditional Chinese cultural values often favor gradual, incremental change rather than radical shifts.

Table 12: (Part1) Examples of generated evaluation data set. Based on content classification, the evaluation dataset comprises four distinct question types. Each question includes a prompt, a corresponding answer, and relevant cultural knowledge. Factual and Conceptual questions assess the model's grasp of factual cultural knowledge and conceptual understanding, respectively, focusing on the evaluation of **knowledge** itself; whereas Mislead and Multi-hop questions evaluate the model's ability to identify biases in real-world scenarios and perform multi-step reasoning, respectively, emphasizing the **application** of cultural knowledge.

Type	Question	Answer	Associated Knowledge
Multi-hop	Imagine you enter a small bakery in a quiet Parisian neighborhood in the late afternoon. The shopkeeper is alone behind the counter, arranging pastries. You select a croissant, pay without saying a word, and leave. Later that evening, you attend a dinner party at a colleague's home. Upon arrival, you shake hands with your host but do not greet the other guests individually. Why might both of these interactions be perceived negatively by those involved, and how do the expectations governing each situation reflect a shared underlying principle about social acknowledgment in French life?	Both interactions would likely be perceived negatively because they violate core French norms of social acknowledgment, which emphasize verbal greetings as a fundamental sign of respect and civility. In the bakery scenario, entering a shop and conducting a transaction without saying "Bonjour" (or "Bonsoir," depending on the time) is considered rude in France. Even in brief, transactional encounters, it is customary—and expected—to greet the shopkeeper upon entry. This small act acknowledges their presence and humanity, transforming an impersonal exchange into a moment of mutual recognition. Failing to do so can come across as dismissive or arrogant. Similarly, at the dinner party, while shaking hands with the host is appropriate, neglecting to greet other guests—typically with a "Bonjour," a handshake, or cheek kisses depending on familiarity—is seen as impolite. In French social settings, acknowledging each person present.....	1. In France, it is customary to greet people verbally with "Bonjour" during the day or "Bonsoir" in the evening when entering a public space or beginning a conversation; failing to do so may be considered impolite. 2. In France, people commonly greet or say goodbye to one another when meeting or parting, with the exact form varying by region, social context, and personal familiarity. 3. Acknowledging others with a verbal greeting such as "Bonjour" or "Bonsoir" when entering shared spaces like shops or elevators is essential, as it recognizes the other's presence and shows respect.

Table 12: (Part2) Examples of generated evaluation data set. Based on content classification, the evaluation dataset comprises four distinct question types. Each question includes a prompt, a corresponding answer, and relevant cultural knowledge. Factual and Conceptual questions assess the model's grasp of factual cultural knowledge and conceptual understanding, respectively, focusing on the evaluation of **knowledge** itself; whereas Mislead and Multi-hop questions evaluate the model's ability to identify biases in real-world scenarios and perform multi-step reasoning, respectively, emphasizing the **application** of cultural knowledge.

Layer	Category	Topic Aspect	Detailed Dimension
Institutional Norms	Geography & Customs	Population and Geography	population rank population distribution land area percentage main regions ethnicity official languages widely spoken languages famous river climate
		Dates of Significance	national holidays religious holidays cultural holidays origin of festivals celebration of festivals symbol of festivals
	Regulation & Policy	Transportation Rules	vehicle movement rules traffic signs and signals pedestrian and non-motorized vehicles motor vehicle driving rules parking regulations
		Data Format	data order separator type year format month representation zero-padding natural language date calendar system
		Measurement Unit	measurement system localization
		Financial Market Rules	regulation structure market type entity type access & licensing conduct & compliance capital & risk monetary & payment system finTech regulation tax & accounting international alignment stability & resolution

Table 13: The complete dimensional taxonomy of CultureScope(Part 1/4)

Layer	Category	Topic Aspect	Detailed Dimension
Behavioral Patterns	Personal Choices & Habits	Daily Life and Travel	payment habits travel habits bathing habits pet-raising habits social media usage pattern marriage customs
		Diet and Health Concept	eating habits attitudes towards birth, aging, illness and death traditional and modern medicine
		Education and Knowledge	educational practices teacher student concept major selection career choices
		Art and Entertainment	festive entertainment activities in different places with the same origin traditional music and dance contemporary art
		Etiquette and Courtesy	basic etiquette naming convention name origin name significance punctuality when visiting shoe etiquette during a visit hospitality customs when receiving guests bringing gifts for the host seating etiquette for guests and hosts serving etiquette during a visit leaving food after eating using salt while eating giving compliments during a meal eating with the right hand alcohol pork handing gifts gifts for children opening gifts general greeting principles greeting between men and men greeting between women and women

Table 13: The complete dimensional taxonomy of CultureScope(Part 2/4)

Layer	Category	Topic Aspect	Detailed Dimension
Behavioral Patterns	Personal Choices & Habits	Etiquette and Courtesy	greeting between men to women appointment scheduling before the business meeting dress code before the business meeting business card exchange before the business meeting network building before the business meeting age and experience before the business meeting familiarity before the business meeting socialization during the business meeting meeting duration during the business meeting open door policy during the business meeting interruptions during the business meeting deference to senior during the business meeting negotiation style during the business meeting decision making during the business meeting bartering during the business meeting private meetings during the business meeting confrontation avoidance during the business meeting follow up after the business meeting ongoing negotiations after the business meeting communication style indirect communication humour physical contact personal space gestures beckoning eye contact

Table 13: The complete dimensional taxonomy of CultureScope(Part 3/4)

Layer	Category	Topic Aspect	Detailed Dimension
Core Values and Social Structures		Fixed Expressions in Language	idiom common saying proverb neologism and abbreviation
	Social Relationship and Structures	Family Dynamics	communal living parental care
		Household Structures	authority patriarchy women roles in the family social interaction urban rural gap concept of seniority and childhood
		Gender Roles	male dominance social compliance honour gender changing attitudes of gender
	Values and Beliefs	Culture Values	power distance collectivism individualism motivation towards achievement and success uncertainty avoidance long term orientation indulgence
		Religion	religion
		Do's and Don'ts	modest dress informality compliments cultural acknowledgement education insults dirty jokes political criticism sensitive topics ethnicity assumptions stereotyping

Table 13: The complete dimensional taxonomy of CultureScope(Part 4/4)

Knowledge Summary(English)

I will provide a web article. Please extract the key characteristics and content related to the cultural dimension in Chinese culture from it. Present these features clearly under distinct headings.
Each feature should begin with a section titled:
Title Then list the following points below:
Description of the feature:
Source of information: (Quote the original text and indicate the paragraph if possible.)
The content should be well-structured and logically coherent. If the information is insufficient to support a certain feature, do not fabricate content.
The article is as follows:
===
Input Text
===

Question Instruction(Factual)

Based on the context, think through all relevant cultural points step by step and generate a **factual** question. The question type can include single-choice, true/false, or fill-in-the-blank. Ensure that the question stem is clear, the options are plausible but misleading (distractors), and the answer is accurate.

Question Instruction(Conceptual)

Based on the context, think through all relevant cultural points step by step and generate a **conceptual** explanation question. The question should focus on the learner's understanding of the concepts, structures, or values Insind cultural phenomena, rather than simple memorization. Suitable formats include multiple-choice or true/false questions. Ensure the question is thought-provoking and the answer is well-justified.

Question Instruction(Mislead)

Based on the context, think through all relevant cultural points step by step and generate a **misleading** question to assess whether learners can identify cultural misunderstandings, stereotypes, or biases. The question should focus on learners' critical thinking about culture, identifying which statements or Behaviors reflect misunderstandings, oversimplifications, biases, or stereotypes, and guide them toward more accurate or respectful understandings. Possible formats include multiple-choice, true/false, case analysis, or short-answer questions.

Question Instruction(Multi-hop)

Based on the context, think through all relevant cultural points step by step and generate a **multi-hop** reasoning question to assess whether the learner can synthesize multiple cultural elements and understand the deeper logic or internal connections among cultural phenomena. The question should prompt learners to start from multiple information points, integrate cultural knowledge, and perform logical analysis, comparison, or generalization. Scenario-based, integrated analysis, or comparative reasoning questions are recommended.

Question Generation(English)

Task: Answer in English.
Instruction:
Instruction
Note:
1. The question should avoid explicitly mentioning cultural concepts, terminology, or characteristics, in order to effectively assess the student's understanding of cultural traits.
2. A reference answer should be provided after the question.
Context:
"'
Context
"'
Question:
Reference Answer:

LLM Judge(English)

Given the following question, reference answer, and student's answer, please evaluate if the student's answer is accurate compared to the reference answer. If the answer is correct, 'return "success". If the answer is incorrect or **contradicts with** the reference answer, return "failed"
Question:
question
Reference Answer:
reference answer
Student's Answer:
test response
Evaluation Result:
[Provide the result: "success" or "failed"]')

LLM Judge for Pairwise Comparison(English)

```
Given the following question, a set of reference answer points, and two candidate
answers, please evaluate which candidate answer covers more of the key points
from the reference.
If Candidate A covers more, return "A". If Candidate B covers more, return "B".
If they cover roughly the same amount, return "Tie".
Question: question
Reference Points: reference answer
Candidate A: candidate a
Candidate B: candidate b
Evaluation Result: [Provide the result: "Candidate A", "Candidate B", or "Tie"]
```

Knowledge Injection(English)

Reference:
reference knowledge
Question:
question