

CultureScope: A Dimensional Lens for Probing Cultural Understanding in LLMs

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 suffers from incomplete and insufficiently rich cultural modeling due to the absence of a comprehensive and scientifically grounded approach to cultural modeling, and many benchmarks remain difficult to scale because they rely heavily on manual data construction. Building on and synthesizing insights from established cultural theories, and in close collaboration with domain experts, we propose the most comprehensive cultural evaluation framework to date. We construct a novel stratified hierarchical schema comprising 140 dimensions, which provides rigorous guidance for the automated extraction of cultural knowledge and the construction of corresponding evaluation datasets for any cultural context. Experimental results on constructed datasets demonstrate that our method can effectively evaluate cultural understanding. They also reveal that existing LLMs lack comprehensive cultural competence, which 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.,

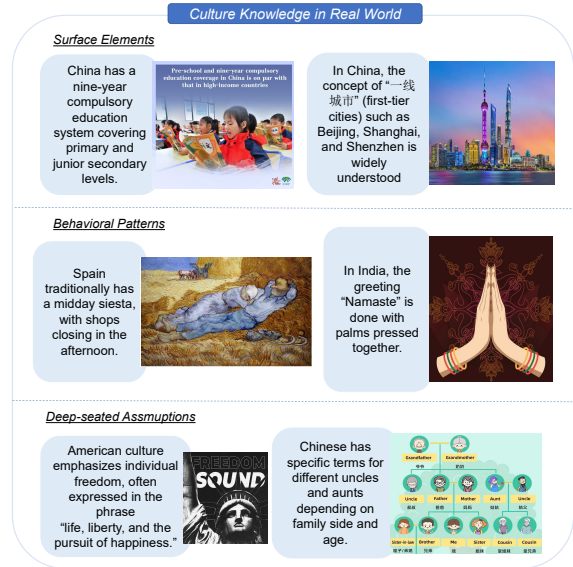


Figure 1: In the real world, the composition of an individual culture is inherently stratified, comprising observable surface elements (e.g., artifacts, concepts), intermediate behavioral patterns (e.g., etiquette, habits), and deep-seated assumptions (e.g., values, society). Existing evaluation datasets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

2025; Ramezani and Xu, 2025). For example, if a social assistant powered by an LLM fails to distinguish between politeness conventions and genuine intentions in Chinese culture context, it may generate inappropriate suggestions that negatively affect 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.

Cultural understanding refers to a model's ability to **comprehend** culturally grounded definitions and 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

Benchmark	Design of Taxonomy			Knowledge Coverage		
	Theory Informed	Abstract Construct	Operationalizable Dimensions	Norms	Habits	Values
AraDiCE-Culture (2025)	✗	✗	✗	✓	✓	✗
CulturalBench (2024)	✗	✗	✗	✗	✗	✓
KorNAT (2024)	✗	✗	✓	✓	✗	✓
BLEnD (2024)	✗	✓	✗	✗	✓	✗
WorldValuesBench (2024)	✓	✓	✗	✗	✗	✓
CDEval (2024)	✓	✓	✗	✗	✗	✓
CultureSynth (2025)	✓	✓	✗	✓	✓	✓
Ours(CultureScope)	✓	✓	✓	✓	✓	✓

Table 1: In existing evaluation efforts, many benchmarks for cultural understanding are primarily intuition-driven, and those informed by **cultural theories** are often derived from theoretical frameworks that do not aim to capture the stratified nature of culture and provide coarse and abstract constructs. Taxonomies providing comprehensive modeling of the **operationalizable dimensions** remain scarce, leading to insufficient **coverage of cultural knowledge** in current work.

is an inherently stratified concept (Tylor, 1871), comprising observable surface elements (e.g., artifacts, concepts), intermediate behavioral patterns (e.g., etiquette, habits), and deep-seated assumptions (e.g., values, beliefs). Existing benchmarks, however, fail to capture this stratified structure.

Current benchmarks address only a single stratum of culture, leaving other essential aspects unexamined. As is presented in Table 1, many benchmarks (Myung et al., 2024; Lee et al., 2024; Chiu et al., 2024; Mousi et al., 2025) construct taxonomies based on researchers’ intuition without systematic theoretical grounding, resulting in fragmented and incomplete coverage of cultural knowledge. Some benchmarks (Wang et al., 2024; Zhao et al., 2024; Zhang et al., 2025) adopt relative theories such as Hofstede’s dimensions (Hofstede and Bond, 1984; Hofstede, 1984), World Values Survey (Inglehart et al., 2000) or library classification systems (Satija, 2000). However, these frameworks were originally designed to characterize *cultural values and attitudes* or *academic knowledge*, and they provide no guidance for organizing factual cultural knowledge (e.g., measurement units, date formats, festival customs) or behavioral norms (e.g., dining etiquette, greeting conventions), resulting in limitations in capturing the stratified depth from surface norms to deep values that characterizes how culture actually operates.

On the other hand, 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 issues above, we propose Cul-

tureScope, a novel evaluation **framework** for LLMs’ cultural understanding capabilities and construct a corresponding multi-culture benchmark. **Our goal is to** rigorously evaluate LLMs’ alignment with the cultural understanding of **specific** countries or regions. Achieving this requires a thorough clarification of the various aspects of knowledge that constitute “culture”. In light of this, we integrate and extend representative cultural theories to develop a stratified and operational classification schema for cultural evaluation, propose a culture-agnostic evaluation framework, and construct multiple corresponding evaluation datasets to evaluate LLMs’ abilities to understand, master, and apply cultural knowledge. CultureScope is **labor-efficient, scalable**, and currently the most **comprehensive** in terms of cultural knowledge coverage, enabling the most thorough evaluation of understanding of a specific culture.

Our notion of “comprehensiveness” is defined with respect to the operational coverage of evaluable cultural knowledge within a given culture. Through expert design informed by multiple theoretical frameworks, our schema is, in this sense, the most comprehensive to date. We conduct an empirical study to quantify the coverage of existing cultural benchmarks with respect to our schema, the detailed results are reported in Appendix D. When mapping existing cultural benchmarks to our schema, their maximum 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:

- By integrating and extending prior cultural theory research, we provide an comprehensive cultural modeling for cultural evaluation

133 and introduce the most complete classification
134 schema to date, which guides the collection of
135 cultural knowledge and evaluation datasets.

- 136 • We extract a substantial volume of cultural
137 knowledge embedded in specific cultural con-
138 texts from various data sources and construct
139 the corresponding evaluation datasets. We
140 conduct quality control through source filter-
141 ing and expert annotation.
- 142 • We conduct extensive experiments on com-
143 monly used LLMs to evaluate their abilities to
144 understand, master, and apply cultural knowl-
145 edge. We observe that developing cultural
146 capability requires specialized training, partic-
147 ularly for the practical applicational tasks.

148 2 Related Work

149 Current studies on the evaluation of LLMs’ cul-
150 tural understanding mainly include: (1) Factual
151 knowledge mastery (Onohara et al., 2025; Singh
152 et al., 2025; Nayak et al., 2024), and (2) Cultural
153 values and moral alignment (Hadar-Shoval et al.,
154 2024; Karinshak et al., 2024; Ji et al., 2024; Ak-
155 soy, 2025; Jiang et al., 2025). However, current
156 research on the evaluation of LLMs’ cultural under-
157 standing capabilities still faces notable limitations.
158 As is shown in Table 1, while some existing stud-
159 ies (Wang et al., 2024; Yadav et al., 2025; Zhou
160 et al., 2025) draw inspirations from relevant cul-
161 tural theories, these theoretical frameworks gener-
162 ally do not aim for the comprehensive model-
163 ing of cultural knowledge, resulting in incomplete
164 taxonomies that cannot fully reveal the stratified
165 structure of culture. As such, these frameworks
166 fail to provide sufficient foundations for a thorough
167 and operational assessment of cultural understand-
168 ing. Many benchmarks instead rely on **intuitive,**
169 **experience-driven** approaches in their taxonomy
170 construction, resulting in evaluation frameworks
171 that fail to cover the full spectrum of cultural knowl-
172 edge and cannot fully reveal the limitations or im-
173 provement needs of LLMs in culturally rich scen-
174 arios. On the other hand, existing methods often
175 **heavily** rely on **expert manual data construction,**
176 and some datasets even involve hundreds of an-
177 notators (Palta and Rudinger, 2023; Myung et al.,
178 2024; Mousi et al., 2025), which incurs high data
179 generation costs and is difficult to scale.

3 CultureScope Framework 180

3.1 Dimensional Schema Design 181

182 While many studies recognize the complexity and
183 the **stratified** nature of culture (Hall, 1976; Sasu,
184 2016), the need for comprehensive and accurate
185 modeling of cultural knowledge for evaluating
186 LLMs’ cultural capabilities has received little at-
187 tention. To address this, we collaborated with mul-
188 tiple experts in cultural and cognitive studies to
189 construct the most comprehensive knowledge clas-
190 sification schema to date for cultural evaluation.
191 Guided by the principles of *Comprehensiveness,*
192 *Operationalizability, Interpretability,* and *Extensi-*
193 *bility,* the schema provides a rigorous foundation
194 for assessing LLMs’ cultural understanding. As
195 presented in Figure 2, it organizes cultural knowl-
196 edge in a **stratified** hierarchical tree structure, with
197 the top level encompassing knowledge ranging
198 from surface-level Institutional Norms, through
199 intermediate Behavioral Patterns, to deeply em-
200 bedded Core Values and Social Structures. We
201 demonstrate, from both theoretical and empirical
202 perspectives, that our schema provides more com-
203 plete coverage of cultural knowledge in specific
204 cultural contexts. The detailed content of the di-
205 mensional schema, principles and its **advantages**
206 are shown in Appendix D.

3.2 Cultural Knowledge Extraction 207

208 To ensure both reliability and comprehensiveness,
209 our culture knowledge is derived from two types
210 of sources. 1) *reliability*: manually selected profes-
211 sional cultural websites. 2) *comprehensiveness*:
212 data dynamically retrieved through Google and
213 Bing search engines. The detailed sources of cul-
214 tural knowledge are provided in Appendix F. Some
215 studies have found that even datasets widely re-
216 garded as reliable, such as Wikipedia, still exhibit
217 notable cultural biases (Naous et al., 2024). More-
218 over, as some aspects of culture are constantly
219 evolving, dependence on static databases inevitably
220 results in incomplete or outdated data. By com-
221 bining static, high-quality data sources with dy-
222 namic retrieval, we can cross-validate information
223 to reduce biases while effectively handling rapidly
224 changing aspects of cultural knowledge, such as
225 expressions of fashion. Regarding the dynamics
226 of culture, we provide further elaboration in the
227 Appendix E. We convert the leaf nodes of the strat-
228 ified hierarchical dimensional schema into search
229 queries(e.g., Celebration of Festivals in Spanish

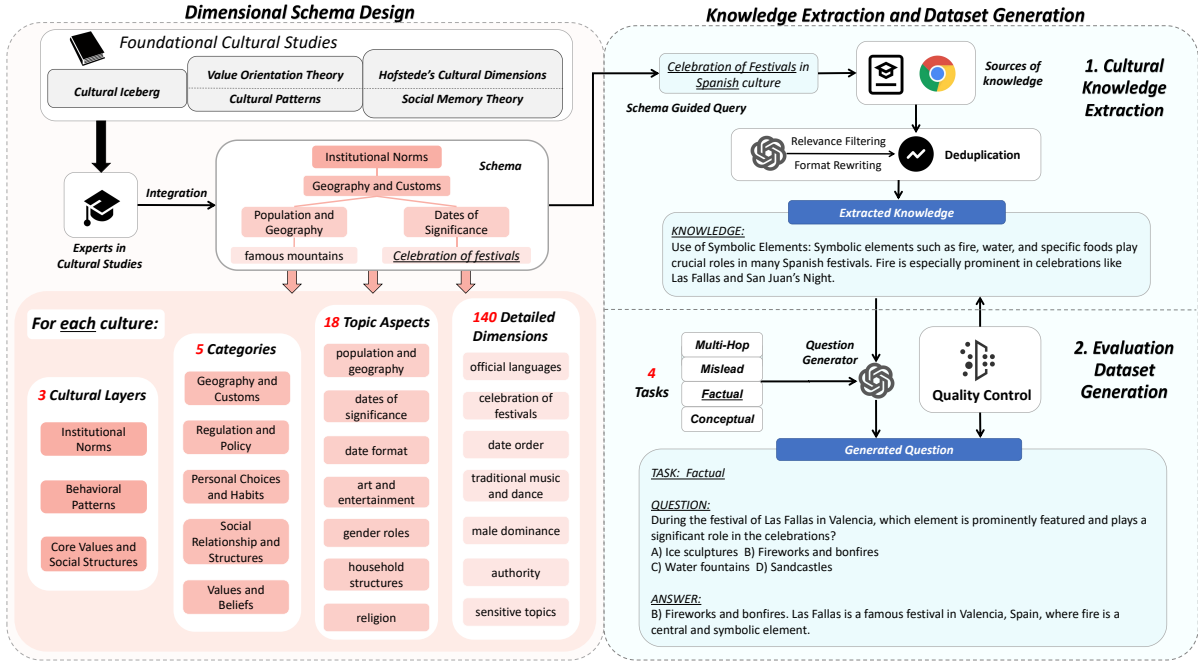


Figure 2: The overall framework of CultureScope. We engage cultural studies experts to develop a schema that models cultural knowledge at a fine-grained level in specific cultural contexts, addressing aspects largely overlooked by previous work, and providing a framework for evaluating LLMs’ cultural 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.

230 culture). For each query, we select the top five web
 231 pages, retain their textual content, and apply GPT-
 232 4o to summarize, and extract relevant information
 233 from the retrieved texts. Examples of extracted cul-
 234 tural knowledge are provided in the Appendix H.

235 3.3 Evaluation Dataset Generation

236 We adapted cognitive science (Wilberschied Ph D,
 237 2015; Avgousti, 2018; Feng et al., 2025) theo-
 238 ries and define four types of tasks for our ques-
 239 tions, each assessing whether the model: (1)Fac-
 240 tual: **knows** cultural facts, (2)Conceptual: **under-**
 241 **stands** the underlying meaning of cultural phenom-
 242 ena, (3)Misleading: can **identify** cultural biases in
 243 **practical** scenarios, (4)Multi-hop: **apply** its knowl-
 244 edge to reason and act appropriately in complex set-
 245 tings. This ensures the evaluation focuses not only
 246 the **understanding and mastery** of cultural knowl-
 247 edge but also its **practical application**, providing a
 248 comprehensive assessment of the model’s cultural
 249 understanding. Question generation is based on a
 250 RAG (Lewis et al., 2020) approach. We employ
 251 Qwen3-Max (Yang et al., 2025) as the generator.
 252 For each type of question, we apply specific expert-
 253 designed prompts to clearly define the task require-

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.

254 All relative prompts are in Appendix I.

255 3.4 Quality Control

256 We conducted quality control from two perspec-
 257 tives: source selection and manual annotation.
 258 When using queries such as “eating habits in Span-
 259 ish 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. 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. The annotation results are presented in the Table 2. We report the annotators’ average annotation results and their agreement. We further annotated the judgments produced by the LLM Judge to assess 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 subjective 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.

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

The second metric is the **Bradley–Terry Ability Score (BT-Score)**, which quantifies a model’s relative capability by fitting pairwise win–loss outcomes against baseline models. For each pairwise comparison 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 $w_{i,j}$ denote the number of times model i is preferred over model j . The Bradley-Terry model (Bradley

and Terry, 1952) assumes that each model i has a latent ability parameter $\theta_i \in \mathcal{R}$, such that the probability of i being preferred over j is given by

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

Using all pairwise outcomes between M and the baseline models B_1, B_2 , we construct the log-likelihood

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

and estimate the ability parameters θ by maximizing this log-likelihood. Finally, the **BT-Score** of the target model M is defined as

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

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 We provide full examples of all data instances in Appendix H. The complete dimensional schema 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 10. 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.

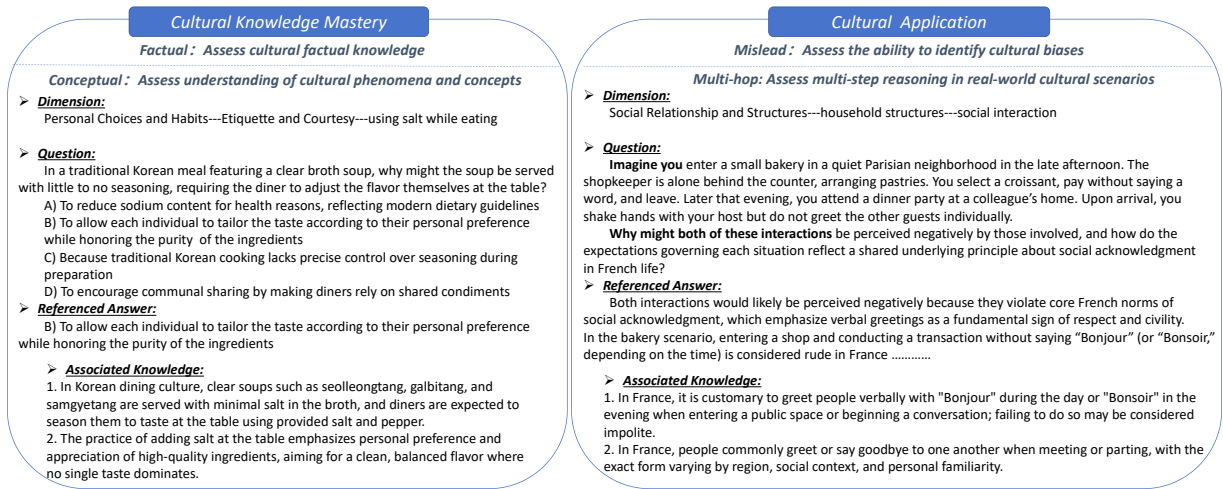


Figure 3: Examples of the generated evaluation dataset. We designed four types of tasks to evaluate LLMs' mastery of cultural knowledge and their ability to apply such knowledge. Mastery of cultural knowledge includes understanding cultural **facts** and grasping cultural **concepts**, while cultural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

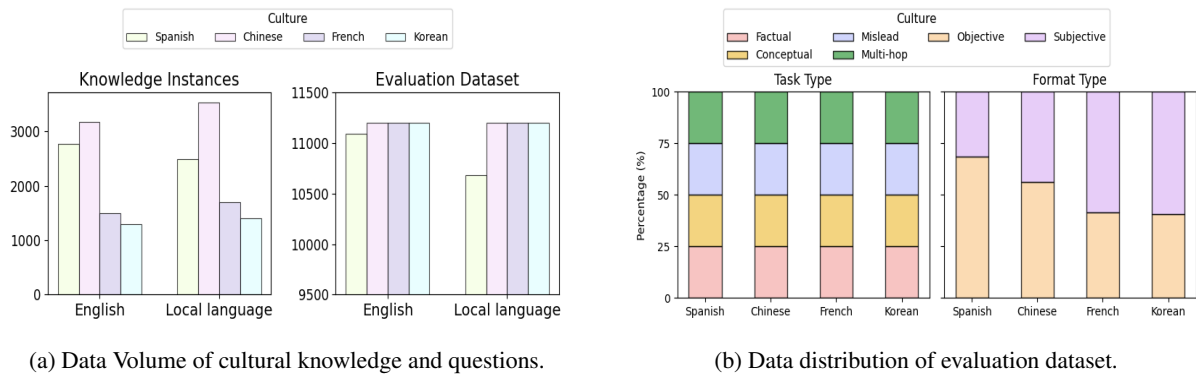


Figure 4: Statistics of constructed benchmarks. The generated questions are balanced in terms of tasks, with the four types (Factual, Conceptual, Mislead, Multi-hop) represented in equal proportions. In terms of format, the ratio of subjective to objective questions is also reasonably proportioned.

5 Experiment Results

We evaluate different models across all dimensions, languages, and question types, and analyzed potential influencing factors. All detailed setups are shown in Appendix A. The overall performances of all LLMs are listed in Table 3 and Figure 5. The statistical reliability of our results is illustrated 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 a large part of this gap is attributable to insufficient language proficiency. Many models produce substantial repetition and

show inadequate analysis when asked and answering in local language, regardless of whether they are online or local models. Further analyses of model outputs are presented in the Appendix C.1.

Observation 2: Weaker in cultural application. As is shown in Figure 7, most models perform worse on Mislead and Multi-hop questions, which assess cultural application, compared to other question types. This aligns with intuitive cognition: cultural 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

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 datasets 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 Institutional 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.

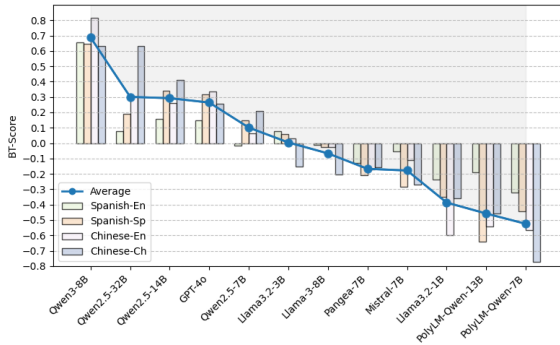


Figure 5: BT-Scores of the tested models, estimated from pairwise comparisons against two baselines(Mistral-7B and Pangea-7B). The BT model fits global ability params from the comparison outcomes; higher scores indicate stronger relative performance within this ability space. The abilities estimated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

an average margin of **26.7%** on tested cultures.

Observation 4: Knowledge injection does not promise better cultural application. We employ supervised fine-tuning (SFT) to train the model on the **extracted cultural knowledge**, with the results presented in the Table 4. The performance of the

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.

trained model shows notable improvements, 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 regarding 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

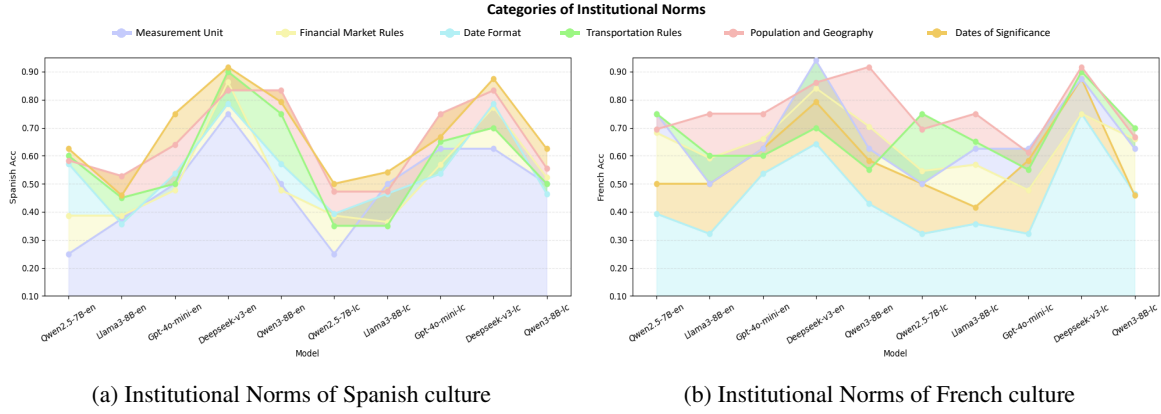


Figure 6: Accuracy on the Institutional Norms within the Spanish cultural context and French cultural context. Many categories of cultural knowledge are areas where models typically exhibit limited capability—for example, Measurement Unit in Spanish culture and Date Format in French culture.

Model	Spanish Culture				Chinese Culture			
	Factual	Conceptual	Mislead	Multi-hop	Factual	Conceptual	Mislead	Multi-hop
Qwen2.5-7B-En	0.61	0.49	0.39	0.46	0.67	0.59	0.56	0.52
Qwen2.5-14B-En	0.76	0.62	0.56	0.59	0.71	0.71	0.65	0.62
Qwen2.5-32B-En	0.74	0.67	0.49	0.64	0.76	0.74	0.68	0.62
Llama3-1B-En	0.31	0.24	0.11	0.12	0.25	0.25	0.11	0.12
Llama3-3B-En	0.45	0.34	0.22	0.36	0.46	0.37	0.36	0.28
Llama3-8B-En	0.57	0.54	0.42	0.44	0.49	0.43	0.49	0.48
GPT-4o-mini-En	0.74	0.57	0.41	0.52	0.73	0.66	0.51	0.45
Qwen3-8B-En	0.62	0.69	0.64	0.65	0.64	0.76	0.71	0.74
Qwen3-8B-En-wo-think	0.69	0.77	0.73	0.69	0.69	0.78	0.76	0.72
PolyLM-7B-En	0.11	0.02	0.11	0.01	0.11	0.01	0.10	0.03
Qwen2.5-7B-Loc	0.56	0.39	0.43	0.44	0.46	0.31	0.39	0.50
Qwen2.5-14B-Loc	0.66	0.52	0.49	0.46	0.58	0.38	0.59	0.51
Qwen2.5-32B-Loc	0.72	0.62	0.56	0.63	0.68	0.46	0.53	0.49
Llama3-1B-Loc	0.17	0.06	0.24	0.14	0.14	0.05	0.02	0.06
Llama3-3B-Loc	0.46	0.28	0.28	0.26	0.19	0.08	0.21	0.24
Llama3-8B-Loc	0.53	0.36	0.43	0.43	0.41	0.38	0.36	0.40
GPT-4o-mini-Loc	0.76	0.55	0.52	0.54	0.55	0.53	0.46	0.49
Qwen3-8B-Loc	0.61	0.61	0.61	0.57	0.49	0.72	0.81	0.61
Qwen2.5-7B-En	0.86	0.79	0.71	0.47	0.89	0.74	0.73	0.34
Qwen2.5-14B-En	0.88	0.81	0.84	0.61	0.92	0.74	0.84	0.49
Qwen2.5-32B-En	0.90	0.81	0.83	0.66	0.93	0.73	0.86	0.61
Llama3-1B-En	0.75	0.60	0.10	0.09	0.79	0.60	0.17	0.11
Llama3-3B-En	0.82	0.69	0.53	0.26	0.86	0.66	0.59	0.25
Llama3-8B-En	0.79	0.76	0.66	0.42	0.86	0.67	0.64	0.34
GPT-4o-mini-En	0.56	0.62	0.81	0.65	0.59	0.54	0.74	0.51
Qwen3-8B-En	0.81	0.78	0.73	0.62	0.88	0.76	0.76	0.56
Qwen3-8B-En-wo-think	0.86	0.75	0.74	0.66	0.89	0.74	0.84	0.54
PolyLM-7B-En	0.41	0.35	0.02	0.02	0.44	0.34	0.04	0.01
Qwen2.5-7B-Loc	0.86	0.73	0.56	0.38	0.59	0.43	0.45	0.17
Qwen2.5-14B-Loc	0.88	0.77	0.77	0.55	0.65	0.47	0.64	0.28
Qwen2.5-32B-Loc	0.89	0.79	0.79	0.58	0.65	0.51	0.69	0.29
Llama3-1B-Loc	0.74	0.54	0.08	0.01	0.50	0.36	0.06	0.01
Llama3-3B-Loc	0.82	0.65	0.21	0.15	0.56	0.41	0.04	0.02
Llama3-8B-Loc	0.85	0.70	0.59	0.29	0.59	0.44	0.51	0.30
GPT-4o-mini-Loc	0.56	0.56	0.77	0.61	0.45	0.43	0.71	0.49
Qwen3-8B-Loc	0.78	0.71	0.69	0.59	0.54	0.46	0.62	0.46

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.

reasoning outperforms non-deep reasoning in only 56% of the experimental results. We observe cases where deep reasoning underperforms non-deep reasoning, and a detailed analysis is provided in the Appendix C.2. Deep reasoning is not a univer-

sal key; its effectiveness depends on the model’s underlying knowledge and reasoning capabilities. When the model’s scale is insufficient to support the knowledge and capabilities required for reasoning, each step of reasoning may further propagate errors, leading to overall inappropriate responses.

Observation 6: Language≠Culture. Compared to the similarly sized Qwen models, PolyLM (Wei et al., 2023), which incorporates additional multilingual data, exhibits a significant performance drop for all tested cultures. This indicates that multilingualism does not equate to multicultural understanding, and increasing the proportion of data in a particular language during training does not necessarily enhance cultural capabilities.

Observation 7: Weakness in Certain Cultural Categories. Many categories of cultural knowledge are areas where models typically exhibit limited capability. As is shown in Figure 6, for *Measurement Unit* in Spanish culture and *Date Format* in French culture, nearly all models perform worse than the other categories. The complete results of the Categories can be found in the Appendix B.

6 Conclusion

By proposing a theoretically grounded and extensible schema, 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.

7 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 schema 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.

8 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.

543
544
545
546
547

548
549
550
551

552
553
554
555
556
557

558
559
560
561

562
563
564
565

566
567
568
569
570
571

572

573
574
575
576
577

578
579
580
581
582

583
584
585
586

587
588
589
590
591

592
593

594
595

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.

Lewis A Coser. 1992. *On collective memory*.

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.

Clifford Geertz. 2017. *The interpretation of cultures*. Basic books. 596
597

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. 598
599
600
601
602
603

Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrahi, 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. 604
605
606
607
608
609

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. 610
611
612
613
614

Edward T. Hall. 1976. *Beyond Culture*. Anchor Press / Doubleday, Garden City, New York. 615
616

Geert Hofstede. 1984. *Culture’s consequences: International differences in work-related values*, volume 5. sage. 617
618
619

Geert Hofstede and Michael H Bond. 1984. Hofstede’s culture dimensions: An independent validation using rokeach’s value survey. *Journal of cross-cultural psychology*, 15(4):417–433. 620
621
622
623

Geert Hofstede and Culture’s Consequences. 2001. Comparing values, behaviors, institutions and organizations across nations. *Tilburg University, Netherlands: SAGE Publishing*. 624
625
626
627

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*. 628
629
630
631
632

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*. 633
634
635
636
637
638

Ronald Inglehart and Christian Welzel. 2005. Cultural change and democracy: the human development sequence. *New York: Cambridge University*. 639
640
641

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenye Hua, and Yongfeng Zhang. 2024. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*. 642
643
644
645

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*. 646
647
648
649
650

763	and linguistic biases in multilingual evaluation. In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 18761–18799.	
764		
765		
766		
767	Edward Burnett Tylor. 1871. <i>Primitive culture: researches into the development of mythology, philosophy, religion, art, and custom</i> , volume 2. J. Murray.	
768		
769		
770	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 <i>Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP</i> , pages 1–16.	
771		
772		
773		
774		
775		
776	Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, and 1 others. 2023. PolyIm: An open source polyglot large language model. <i>arXiv preprint arXiv:2307.06018</i> .	
777		
778		
779		
780		
781	Lee F Wilberschied Ph D. 2015. Intercultural communicative competence: Literature review.	
782		
783	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. <i>arXiv preprint arXiv:2505.22793</i> .	
784		
785		
786		
787		
788		
789		
790	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. <i>arXiv preprint arXiv:2505.09388</i> .	
791		
792		
793		
794		
795	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. <i>arXiv preprint arXiv:2412.15115</i> .	
796		
797		
798		
799	Taha Yasserli, Robert Sumi, and János Kertész. 2012. Circadian patterns of wikipedia editorial activity: A demographic analysis. <i>PloS one</i> , 7(1):e30091.	
800		
801		
802	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 <i>The Thirteenth International Conference on Learning Representations</i> .	
803		
804		
805		
806		
807		
808		
809	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. <i>arXiv preprint arXiv:2509.10886</i> .	
810		
811		
812		
813		
814	Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. World-valuesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. <i>arXiv preprint arXiv:2404.16308</i> .	
815		
816		
817		
818		
	Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. <i>arXiv preprint arXiv:2502.12057</i> .	819
		820
		821
	Hao Zhu, Bodhisattwa Prasad Majumder, Dirk Hovy, and Diyi Yang. 2025. Social intelligence in the age of llms. In <i>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)</i> , pages 51–55.	822
		823
		824
		825
		826
		827
		828
	A Experimental Setups	829
	A.1 Evaluation Setups	830
	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 PolyIm-Qwen-7B and PolyIm-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.	831
		832
		833
		834
		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
	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.	848
		849
		850
		851
		852
		853
		854
	A.2 Training Setups	855
	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	856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868

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.

B Supplementary Experimental Results

Our experiments aim to examine: 1) To what extent can current LLMs align with specific cultural knowledge. 2) What are the factors that influence the performances of LLMs on culture-specific knowledge-related tasks. 3) Whether there is a noticeable difference in the performances of LLMs across different cultural dimensions, languages, and types of question.

B.1 Result robustness

To ensure the reliability of our evaluation results, we performed repeated sampling for a subset of models and computed their BT confidence intervals, as shown in Table 8. Specifically, each model was independently sampled six times, and BT confidence was estimated using 1,000 bootstrap samples, along 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 evaluate the finer-grained *Categories* within each *Layer* and the results are shown in Figure 8. The use of different languages leads to notable differences in cultural understanding capability across various dimensions. We find that the models perform differently across different Layers depending on the cultural context. Here we provide some analysis of models’ performances of Spanish culture. In the context of Spanish culture, as the cul-

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

Table 5: 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 6: Details annotation results of all annotators. Two annotators are invited for each culture.

tural knowledge evaluation progresses from shallow to deep layers, the model exhibits increasing performance divergence within the same dimension when questioned in different languages. At the Institutional Norm layer, evaluation results across different languages show a certain degree of consistency. All models perform poorly on questions related to Measurement Units, while demonstrating better mastery of knowledge concerning Dates of Significance. At the Behavioral Habits layer, models exhibit different preferences across languages. For Art and Entertainment questions in Spanish, all models perform poorly, whereas on English questions, Art and Entertainment performs relatively well. At the Cultural Values layer, language-dependent preference differences become more pronounced. Nearly all models perform well on Religion-related Spanish questions but poorly on the same questions in English. Conversely, for Family Dynamics and Household Structures, almost all models answer English questions very well but perform poorly on Spanish ones.

C 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.

C.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. The English response is more structured, detailed, and focused on the underlying

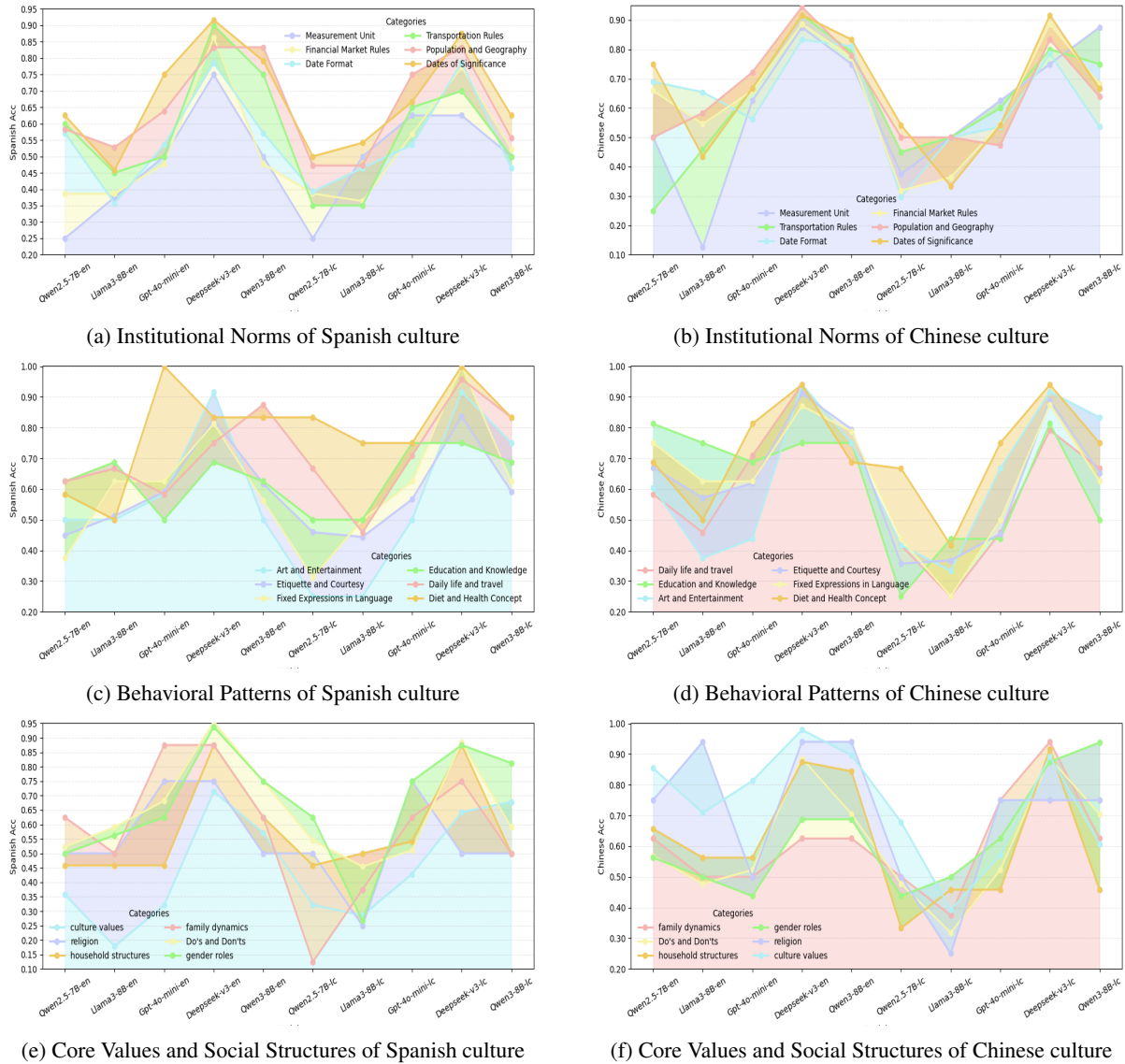


Figure 8: Performance of Models across Layers and Languages (a-f).

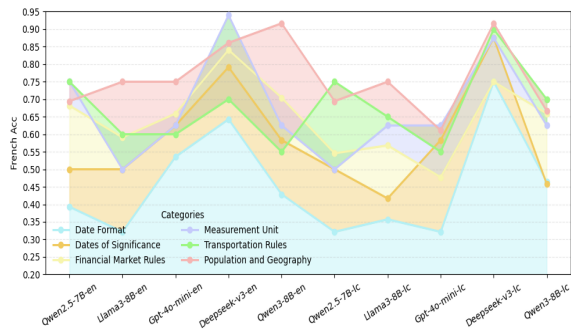
Culture	Annotator	Age	Gender	Education Level	Cultural Background
Spanish	Annotator1	-29	Female	Master	Spanish
	Annotator2	30-39	Female	Doctorate	Spanish & Chinese
Chinese	Annotator1	-29	Male	Master	Chinese
	Annotator2	40-49	Female	Bachelor	Chinese
French	Annotator1	-29	Female	Master	French
	Annotator2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

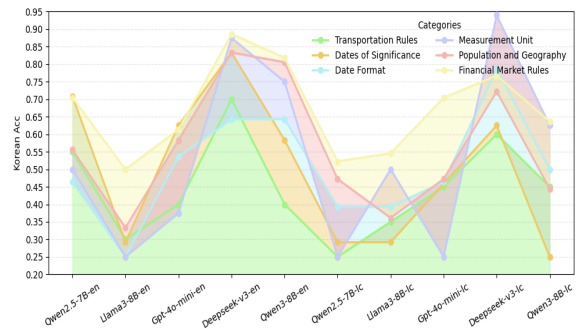
cultural mechanisms, whereas the Korean response is more generalized and repetitive, emphasizing social norms without fully articulating the causal logic behind the observed behaviors. Such phenom-

ena were consistently observed across all tested models.

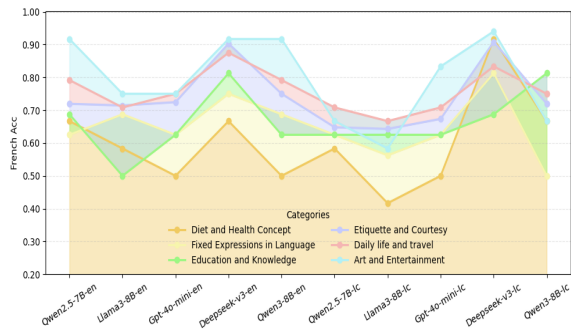
Question At a formal dinner in Korea, a young employee arrives with their department head and a



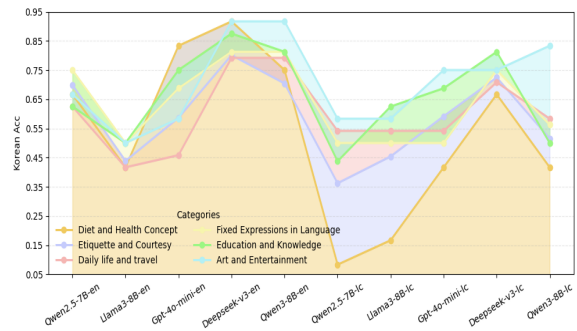
(g) Institutional Norms of French culture



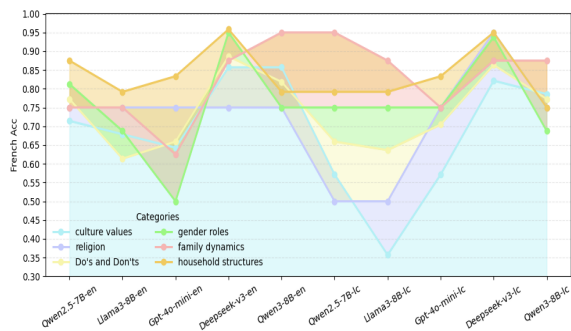
(h) Institutional Norms of Korean culture



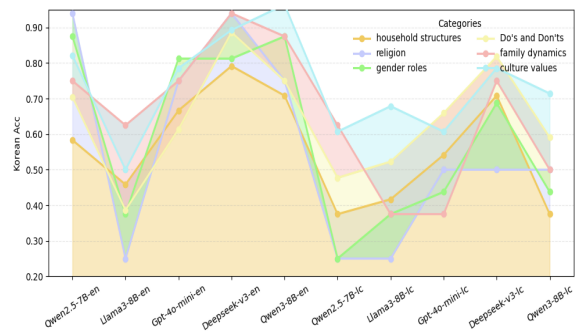
(i) Behavioral Patterns of French culture



(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).

senior executive from another division. The dining room has a traditional layout with seating arranged around a low table near the inner wall and additional chairs closer to the entrance. Upon entering, the young employee instinctively moves toward a seat near the door but pauses before sitting. Moments later, after the senior executive takes a seat furthest from the entrance, the department head selects a position between the executive and the door. Only after both have seated does the young employee sit down—and even then, they wait with their hands in their lap until someone begins eating.

Why did the young employee delay both sitting and starting to eat, and how do these actions reflect unspoken expectations about interpersonal dynamics in this setting?

English response The young employee’s actions reflect a deep understanding and respect for the unspoken expectations and social norms in the Korean dining etiquette setting. Here’s a detailed explanation of why they delayed both sitting and starting to eat:

1. ****Respect for Hierarchical Order**:** In Korean culture, there is a strong emphasis on respecting hierarchical relationships. The young employee waited for the senior executive and the department head to take their seats before sitting down. This shows that they recognize the senior executive as the most important person at the table and the department head as the next in line. By waiting for them to be seated, the young employee is demonstrating deference to their positions and ensuring that the most senior individuals are comfortably

986 seated first.

987 2. ****Observation of Seating Arrangements****:

988 The young employee observed the seating arrange-

989 ment and waited for the senior executive to take the

990 seat furthest from the entrance. This is a common

991 practice in Korean dining etiquette, where the most

992 senior or important person is often seated at the end

993 of the table, facing the entrance. By waiting for

994 the senior executive to take this seat, the young em-

995 ployee is showing that they understand and respect

996 the seating hierarchy.

997 3. ****Waiting for the Host to Begin****: After both

998 the senior executive and the department head were

999 seated, the young employee waited for someone

1000 to start eating before they began. This is a sign of

1001 respect for the host or the person who is leading

1002 the meal. In Korean culture, it is customary to wait

1003 for the host or the most senior person to start eating

1004 before others join in. This shows that the young

1005 employee is aware of the social dynamics and is

1006 being considerate of the group's expectations.

1007 4. ****Maintaining Proper Posture and Man-**

1008 **ners****: The young employee waited with their

1009 hands in their lap until someone began eating. This

1010 demonstrates good manners and a respect for the

1011 dining etiquette. In Korean culture, it is considered

1012 polite to maintain proper posture and to wait for

1013 the appropriate time to start eating. By waiting

1014 with their hands in their lap, the young employee is

1015 showing that they are attentive to the social norms

1016 and are being respectful of the dining etiquette.

1017 In summary, the young employee's actions re-

1018 flect a deep understanding of the unspoken expect-

1019 ations and social norms in the Korean dining eti-

1020 quette setting. They demonstrated respect for hi-

1021 erarchical relationships, observed proper seating

1022 arrangements, waited for the appropriate time to

1023 start eating, and maintained good manners. These

1024 actions are crucial in maintaining positive interper-

1025 sonal dynamics and showing respect in a formal

1026 dining setting in Korea. The young employee de-

1027 layed sitting and starting to eat to show respect for

1028 hierarchical relationships, observe proper seating

1029 arrangements, wait for the appropriate time to start

1030 eating, and maintain good manners.

1031 **Korean response(Translated into English)** The

1032 reason this young employee delayed both sitting

1033 down and starting the meal is to follow social norms

1034 and etiquette in Korean society. In Korea, behavior

1035 varies according to the status and role of others,

1036 and this young employee delayed sitting in order

1037 to respect the positions of the department head

1038 and senior executives and to follow their actions.

1039 These behaviors reflect the unspoken expectations

1040 regarding interpersonal relationships in this setting.

1041 1. **Mutual Respect**: The young employee's act

1042 of following the department head and senior execu-

1043 tives' behavior is an expression of mutual respect.

1044 It shows that the employee respects supervisors and

1045 colleagues and acknowledges their status and roles.

1046 2. **Social Norms**: In Korean society, mirroring

1047 the actions of others is considered a social norm.

1048 The young employee's behavior demonstrates ad-

1049 herence to these norms.

1050 3. **Teamwork**: By following the department head

1051 and senior executives' actions, the young employee

1052 emphasizes teamwork. This shows that they respect

1053 and cooperate with others as members of the team.

1054 4. **Cultural Etiquette**: The young employee's be-

1055 havior demonstrates adherence to cultural etiquette

1056 in Korean society. It shows that they understand

1057 and follow the values and norms of Korean society.

1058 Therefore, the young employee's delaying of

1059 both sitting and starting the meal is a behavior that

1060 reflects following social norms and etiquette in Ko-

1061 rean society, encompassing mutual respect, social

1062 norms, teamwork, and cultural etiquette in interper-

1063 sonal relationships. Such behaviors play an impor-

1064 tant role in maintaining and strengthening social

1065 relationships in Korea.

1066 These behaviors clearly illustrate social norms

1067 and etiquette in Korean society, reflecting mutual

1068 respect, social norms, teamwork, and cultural eti-

1069 quette in interpersonal relationships. Such behav-

1070 iors play an important role in maintaining and

1071 strengthening social relationships in Korea. The

1072 young employee's actions show that they under-

1073 stand and adhere to the values and norms of Korean

1074 society. This plays a key role in maintaining and

1075 strengthening social relationships in Korea.

1076 The young employee's actions demonstrate so-

1077 cial norms and etiquette in Korean society, re-

1078 flecting mutual respect, social norms, teamwork,

1079 and cultural etiquette in interpersonal relationships.

1080 Such behaviors play an important role in maintain-

1081 ing and strengthening social relationships in Korea.

1082 The young employee's actions indicate that they

1083 understand and comply with the values and norms

1084 of Korean society. This is essential for maintaining

1085 and strengthening social relationships in Korea.

1086 The young employee's actions demonstrate so-

1087 cial norms and etiquette in Korean society, re-

1088 flecting mutual respect, social norms, teamwork,

1089	and cultural etiquette in interpersonal relationships.	1140
1090	Such behaviors play an important role in maintain-	1141
1091	ing and strengthening social relationships in Korea.	1142
1092	The young employee’s actions indicate that they	1143
1093	understand and comply with the values and norms	1144
1094	of Korean society. This is essential for maintaining	1145
1095	and strengthening social relationships in Korea.	1146
1096	The young employee’s actions demonstrate so-	1147
1097	cial norms and etiquette in Korean society, re-	1148
1098	fecting mutual respect, social norms, teamwork,	1149
1099	and cultural etiquette in interpersonal relationships.	1150
1100	Such behaviors play an important role in maintain-	1151
1101	ing and strengthening social relationships in Korea.	1152
1102	The young employee’s actions indicate that they	1153
1103	understand and comply with the values and norms	1154
1104	of Korean society. This is essential for maintaining	1155
1105	and strengthening social relationships in Korea.	1156
1106	The young employee’s actions demonstrate so-	1157
1107	cial norms and etiquette in Korean society, re-	1158
1108	fecting mutual respect, social norms, teamwork,	
1109	and cultural etiquette in interpersonal relationships.	
1110	Such behaviors play an important role in maintain-	
1111	ing and strengthening social relationships in Korea.	
1112	The young employee’s actions indicate that they	
1113	understand and comply with the values and norms	
1114	of Korean society.	
1115	Analysis Compared to the Korean response, the	
1116	English answer demonstrates superior performance	
1117	for several reasons. First, it establishes a clear	
1118	causal and logical structure: each observed behav-	
1119	ior of the young employee—delaying sitting and	
1120	waiting to start eating—is explicitly linked to a	
1121	corresponding social norm or expectation. This	
1122	“behavior → cultural rule → social meaning” map-	
1123	ping allows the reader to understand not only what	
1124	the employee does, but why these actions carry	
1125	significance within Korean dining etiquette.	
1126	Second, the English response provides concrete	
1127	contextual details, such as the order of seating, the	
1128	position relative to the entrance, and the mainte-	
1129	nance of posture, thereby situating abstract norms	
1130	within observable actions. In contrast, the Korean	
1131	answer relies heavily on generalized concepts such	
1132	as “respect,” “social norms,” and “teamwork,” with	
1133	substantial repetition and little reference to specific	
1134	behavioral cues, reducing both analytical depth and	
1135	clarity.	
1136	Third, the English version employs academic	
1137	and analytical language (unspoken expectations,	
1138	hierarchical order, interpersonal dynamics), which	
1139	effectively conveys both conceptual understand-	
	ing and cultural nuance. Even though the English	
	answer contains minor factual inaccuracies, these	
	are easily correctable. In contrast, the Korean ver-	
	sion suffers from both conceptual vagueness and	
	structural redundancy, making it less amenable to	
	revision for scholarly precision.	
	The difference between the English and Korean	
	responses primarily arises from training data dis-	
	tribution and task familiarity. The model has been	
	exposed to far more English-language examples	
	of academic analysis and culturally specific ex-	
	planations, allowing it to generate structured, de-	
	tailed, and mechanism-focused answers. In con-	
	trast, Korean-language training data on this type of	
	hierarchical etiquette is more limited, causing the	
	model to produce more generalized and repeti-	
	tive explanations that emphasize respect and so-	
	cial norms without fully articulating the underlying	
	cultural mechanisms.	
	C.2 With vs. Without Deep Thinking	
	We present the results of Qwen3-8B with and with-	
	out deep reasoning, as well as Qwen3-max with	
	deep reasoning enabled. 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 high-	
	lights the crucial role of knowledge capacity and	
	reasoning ability in multi-step inference.	
	Chain-of-thought reasoning relies on the model’s	
	ability to sequentially link premises, apply relevant	
	knowledge, and infer causal or contextual relation-	
	ships. If the model’s knowledge base is incomplete	
	or imprecise—particularly regarding culturally spe-	
	cific norms or social conventions—each step in	
	the reasoning chain may propagate errors, lead-	
	ing to an overall inaccurate response. In contrast,	
	larger models typically possess broader training	
	data coverage and more robust internal representa-	
	tions, which allow them to generate correct interme-	
	diate inferences. Consequently, deep reasoning am-	
	plifies the model’s strengths when sufficient knowl-	
	edge is available but can exacerbate errors when	
	the model’s understanding is limited. This demon-	
	strates that the effectiveness of chain-of-thought	
	reasoning is contingent on the model’s underlying	
	knowledge and reasoning capacity, rather than be-	
	ing inherently superior in all contexts.	
	Question During a dinner party in Lyon, a guest	
	presents the host with a brand-new vacuum cleaner	
	as a housewarming gift, saying it will make clean-	

ing easier after all the cooking. Later that evening, another guest arrives bearing a bouquet of red roses and a bottle of Bordeaux. The host accepts both gifts politely but seems visibly uncomfortable. A third guest, noticing this, offers a small box of chocolates instead and is warmly thanked.

Why might the host have reacted with discomfort to the first two gifts, despite the guests' good intentions?

Qwen3-8B with thinking The host might have felt uncomfortable because the first two gifts were not appropriate for a housewarming. 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. The red roses and Bordeaux are also not appropriate for a housewarming, as they are more suitable for a romantic or celebratory occasion. The third gift, a box of chocolates, is a more appropriate and thoughtful gift for a housewarming, which is why the host was grateful.

Qwen3-8B wo thinking The host likely felt uncomfortable with the first two gifts because they were **not appropriate for a housewarming**.

- 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.

- A **bouquet of red roses and a bottle of Bordeaux** might be interpreted as a **romantic or romanticized gesture**, which could be inappropriate if the host is single, not interested in romance, or if the gift was intended for someone else (like a spouse or partner).

contrast, a **box of chocolates** is a **neutral, universally appreciated gift** that fits well with a housewarming celebration, making it more appropriate and less potentially awkward.

So, the host's discomfort likely stemmed from the **inappropriateness or unintended meaning** behind the first two gifts, rather than any ill will from the guests.

Qwen3-max with thinking The host's discomfort with the first two gifts stems from cultural and social nuances in French etiquette, particularly in a refined city like Lyon, known for its strong traditions around hospitality and gift-giving. Here's a breakdown of why each gift caused discomfort, despite good intentions:

1. **The Vacuum Cleaner (First Gift)**: - **Implied Criticism**: 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. The guest's comment about "making cleaning easier after all the cooking" inadvertently suggests the host's efforts are messy or inadequate. This is particularly sensitive in Lyon, a city proud of its culinary heritage and domestic hospitality (e.g., the tradition of *les mères lyonnaises*, where women historically ran renowned home-style restaurants). A practical gift like this may feel utilitarian and impersonal, undermining the host's pride in their home and the evening's ambiance. - **Cultural Context**: Housewarming gifts in France typically lean toward decorative, consumable, or symbolic items (e.g., wine, art, or gourmet foods) that celebrate the home rather than "fix" it. A vacuum cleaner breaches this norm, evoking discomfort by highlighting chores over conviviality.

2. **Red Roses and Bordeaux (Second Gift)**: - **Romantic Connotations of Red Roses**: In France, red roses are heavily associated with romantic love and courtship. Presenting them to a host—especially in a group setting—can be misinterpreted as a romantic advance, causing embarrassment. This is amplified if the host is in a relationship, as it may imply impropriety or overstep social boundaries. Even among friends, such a bold floral choice is typically reserved for intimate partners, not casual acquaintances. - **Bordeaux as a Double-Edged Gift**: While a bottle of Bordeaux is a classic, well-intentioned gift in wine-loving Lyon, pairing it with red roses creates a jarring juxtaposition. The roses overshadow the wine, making the entire gift feel like a romantic gesture rather than a token of appreciation for hospitality. This could put the host in an awkward position, especially if other guests misinterpret the giver's intentions.

Why the Chocolates Were Well-Received: - The third guest's small box of chocolates is a culturally safe, neutral gift in France. Chocolates are: - **Appropriate for the Occasion**: They align with French *l'art de recevoir* (the art of hosting), as they can be shared immediately with all guests, reinforcing communal enjoyment. - **Non-Controversial**: Unlike the vacuum (critical) or roses (romantic), chocolates carry no hidden implications. They are a traditional, elegant gesture

1292	that respects boundaries. - **Lyon-Specific Relevance** : Lyon is famous for its chocolatiers (e.g.,	1343
1293	master <i>*chocolatiers*</i> like Bernachon), making this	1344
1294	gift locally meaningful and flattering.	1345
1295		
1296	Broader Cultural Insight: French etiquette em-	
1297	phasizes <i>*la politesse*</i> (politeness) and avoiding	
1298	<i>*faux pas*</i> that disrupt social harmony. Gifts	
1299	should uplift the recipient without implying judg-	
1300	ment or overfamiliarity. The first two gifts, though	
1301	well-meant, violated these principles: - The vac-	
1302	uum cleaner focused on labor (cleaning) rather than	
1303	pleasure (enjoying the home). - The red roses	
1304	risked misinterpretation in a context where sub-	
1305	tlety and restraint are valued. The host’s polite	
1306	acceptance but visible discomfort reflects a desire	
1307	to maintain grace while inwardly recoiling from	
1308	these breaches. The chocolates, in contrast, exem-	
1309	plify the ideal French gift: thoughtful, shareable,	
1310	and free of unintended messages.	
1311	This scenario underscores how cultural context	
1312	transforms seemingly generous gestures—what	
1313	might be practical or romantic elsewhere can be-	
1314	come a social misstep in Lyon’s nuanced setting.	
1315	The third guest’s awareness of these norms allowed	
1316	them to offer a gift that truly honored the host.	
1317		
1318	Analysis The first response, generated with	
1319	chain-of-thought reasoning by a smaller model, is	
1320	detailed but ultimately incorrect: it interprets the	
1321	host’s discomfort superficially as “inappropriate for	
1322	a housewarming,” failing to account for the cultural	
1323	and social implications of each gift.	
1324	The second response, produced without deep	
1325	reasoning, is more concise and accurate: it cor-	
1326	rectly identifies the implicit social and cultural	
1327	meanings—such as the vacuum cleaner’s utili-	
1328	tarian connotation and the romantic implications	
1329	of red roses—demonstrating that simpler, high-	
1330	probability reasoning can sometimes outperform	
1331	complex reasoning when the model’s knowledge is	
1332	limited.	
1333	The third response, generated by a larger model	
1334	with chain-of-thought reasoning enabled, produces	
1335	the correct output with detailed justification.	
1336	This contrast illustrates that chain-of-thought	
1337	reasoning amplifies the model’s existing knowl-	
1338	edge: with insufficient knowledge, deeper rea-	
1339	soning can propagate errors, whereas with suf-	
1340	ficient knowledge and representational capacity,	
1341	deep reasoning enables accurate, nuanced explana-	
1342	tions. Collectively, these examples highlight that	
	both the model’s knowledge base and reasoning	
	strategy critically determine output quality, and that	1343
	deeper reasoning is not inherently superior without	1344
	adequate supporting knowledge.	1345
	D Culture Knowledge Classification	1346
	Schema	1347
	In this section, we provide a detailed introduction	1348
	to the cultural knowledge modeling schema we pro-	1349
	pose. We demonstrate its advantages over existing	1350
	approaches from both Theoretical Foundations	1351
	and Statistical Results . In addition, we present the	1352
	design principles and detailed dimensions of the	1353
	schema.	1354
	D.1 Theoretical Foundations	1355
	Regarding the design of the dimensional schema,	1356
	we invited multiple experts for discussion, includ-	1357
	ing two professors specializing in cognition and	1358
	culture, one professor in computer science, and	1359
	three PhD students. This schema integrates nu-	1360
	merous existing cross-cultural and single-culture	1361
	research theories. The specific design is as follows:	1362
	At the explicit layer, we follow the “surface”	1363
	part of the cultural iceberg model (Hall, 1976)	1364
	and focus on institutionalised, directly observable	1365
	aspects of culture, such as traffic rules, date formats,	1366
	measurement units, financial regulations, popula-	1367
	tion and geography, and public holidays. These	1368
	categories capture what visitors typically first en-	1369
	counter when interacting with a new society.	1370
	At the intermediate layer, we model everyday	1371
	practices and habitual choices that are less for-	1372
	malised but still observable: daily routines and	1373
	mobility, food and health beliefs, views on edu-	1374
	cation and knowledge, artistic and entertainment	1375
	preferences, etiquette and rituals, and conventional	1376
	linguistic expressions. This layer is informed by	1377
	Hall’s work on high-context and low-context com-	1378
	munication theory (Hall, 1976), Cultural Pat-	1379
	terns (Kroeber and Kluckhohn, 1952) and Kluck-	1380
	hohn & Strodtbeck’s Value Orientation The-	1381
	ory (Kluckhohn and Strodtbeck, 1961), which em-	1382
	phasise how values are enacted through recurrent	1383
	patterns of Behaviour and communication.	1384
	At the implicit layer, we move to deep struc-	1385
	tures of culture: family relations and social hi-	1386
	erarchy, gender roles and equality, broader “cul-	1387
	tural dimensions” of value preferences, religious	1388
	beliefs, and taboos (Do’s and Don’ts). Here	1389
	we draw inspiration from Social Memory The-	1390
	ory (Coser, 1992) and explicitly incorporate Hof-	1391

1392	stede’s value dimensions (Hofstede, 1984)(e.g.,	Extensibility: The system should be designed for	1442
1393	power distance, individualism–collectivism, uncer-	high scalability through data-driven expansion, sup-	1443
1394	tainty avoidance, masculinity–femininity) and re-	porting increasingly fine-grained evaluation in the	1444
1395	lated value-orientation work to guide which aspects	future.	1445
1396	of hierarchy, authority, risk, and role expectations	These principles also serve as the guiding tenets	1446
1397	are encoded in our “social relations and structures”	for our schema design:	1447
1398	and “values and beliefs” categories.		
1399	In summary, Hofstede’s dimensions are used at	• Regarding Comprehensiveness: As described	1448
1400	the deep layer of our taxonomy and are comple-	in the previous section, we have achieved this	1449
1401	mented by the cultural iceberg perspective, cul-	goal by maximizing the comprehensive cover-	1450
1402	tural patterns, value-orientation theories, high/low-	age of cultural knowledge.	1451
1403	context communication, and social memory theory	• Regarding Operationalizability: By refining	1452
1404	to cover both visible institutional rules and hid-	the schema into detailed dimensions at the	1453
1405	den value systems. Whether in works that design	keyword level, we can directly convert the con-	1454
1406	schemas purely based on intuition or those that di-	tent within the schema into queries for knowl-	1455
1407	rectly adopt theories like Hofstede’s dimensions,	edge extraction, effectively supporting knowl-	1456
1408	our schema design is more comprehensive.	edge collection and dataset construction.	1457
1409	D.2 Empirical Study of Knowledge Coverage	• Regarding Interpretability: We have designed	1458
1410	To more intuitively demonstrate the advantages	the schema into a tree structure, which fa-	1459
1411	of our framework in terms of cultural knowledge	facilitates a coarse-to-fine hierarchical analysis	1460
1412	coverage and completeness, we selected several	of cultural evaluation results and yields more	1461
1413	representative existing works that are grounded in	granular assessment outcomes.	1462
1414	culture studies and conducted a empirical analysis	• Regarding Extensibility: The tree struc-	1463
1415	of their data to identify their categorical coverage.	ture also contributes to achieving scalability.	1464
1416	The results are shown in the Table 9. It can be	Whether supplementing dimension nodes for	1465
1417	observed that the data used in these studies can	specific cultures in a top-down manner or ex-	1466
1418	all be categorized within our framework, and their	panding dimension nodes based on data clus-	1467
1419	cultural knowledge coverage represents only a sub-	tering in a bottom-up approach, implementa-	1468
1420	set of ours. Even the most comprehensive prior	tion is straightforward.	1469
1421	work covers no more than 40% of the dimensions		
1422	in our 140-dimensional taxonomy. This demon-	D.4 Detailed Dimensions	1470
1423	strates that, without a well-designed, scientific-	Our proposed dimension schema contains 4 levels,	1471
1424	ally grounded, fine-grained knowledge taxonomy, cul-	<i>Cultural Layers, Category, Topic Aspect and De-</i>	1472
1425	tural competence evaluations based solely on in-	<i>tailed Dimensions</i> . There are in total 3 Cultural	1473
1426	tuition or coarse-grained cultural theories struggle	Layers, 5 Categories, 18 Topic Aspects and 140	1474
1427	to achieve high coverage of knowledge within a	Detailed Dimensions, encompassing a wide range	1475
1428	specific culture.	of cultural knowledge from explicit commonsense	1476
1429	D.3 Design Principles for Schema	facts, Behavioral patterns to implicit values.	1477
1430	To comprehensively and scientifically evaluate the	The surface-level <i>Cultural Layer</i> is the Insti-	1478
1431	cultural understanding capability of LLMs, we ar-	tutional Norms Layer , which includes explicit	1479
1432	gue that the dimensional schema for classifying	commonsense knowledge such as geography, cus-	1480
1433	cultural knowledge should satisfy four core prin-	tomms, rules, and policies within specific cultural	1481
1434	ciples. 1) Comprehensiveness: The dimensions	contexts; the intermediate layer is Behavioral Pat-	1482
1435	should cover cultural knowledge as fully as pos-	terns Layer , which encompasses personal choices	1483
1436	sible. 2) Operationalizability: The framework	and habits, including aspects such as daily life and	1484
1437	should effectively guide data collection and the	travel, education, and concepts of knowledge; the	1485
1438	practical construction of knowledge bases. 3) In-	deep-level layer is the Core Values and Social	1486
1439	terpretability: It should enable clear attribution	Structures Layer , which includes deep cultural	1487
1440	analysis of model behaviors, revealing preferences	knowledge such as social relationships and struc-	1488
1441	and performance variations across dimensions. 4)	tures, values, and beliefs. Each <i>Cultural Layer</i>	1489

1490 include a series of *Categories, Topic Aspects and*
1491 *Detailed Dimensions*. A detailed description of the
1492 dimensions is provided in Table 13.

1493 E The Dynamics of Culture

1494 Numerous studies have pointed out that culture
1495 itself is dynamically evolving; therefore, relying
1496 solely on static data sources for evaluation is in-
1497 sufficient. Although dynamism is not the issue
1498 addressed in this particular study, the dynamic na-
1499 ture of our data sources and collection methods
1500 ensures that the quality of our evaluation is not sig-
1501 nificantly compromised. We elaborate on this point
1502 below.

1503 Efficient and Incremental Knowledge Updating

1504 When certain cultural knowledge undergoes sig-
1505 nificant modifications or changes, we only need
1506 to reconstruct **queries** for the affected dimensions
1507 and crawl the web. Google’s and Bing’s search
1508 mechanism ensures that we can obtain the most
1509 up-to-date results, thereby naturally enabling the
1510 corresponding updates to cultural knowledge.

1511 **Integration of Dynamic Sources** On one hand,
1512 our evaluation dataset is built on **dynamic data**
1513 **sources**, and the resulting knowledge has been
1514 validated to be of high quality. Although high-
1515 quality internet-sourced data, such as information
1516 from authoritative websites, is relatively static be-
1517 cause each individual page is fixed, these sources
1518 have limited coverage. To address this, our frame-
1519 work integrates dynamic web search to expand cul-
1520 tural coverage. All web-search-based knowledge
1521 included in the benchmark undergoes human verifi-
1522 cation, and the statistics demonstrate that the frame-
1523 work reliably acquires accurate and comprehensive
1524 representations of dynamic culture.

1525 Low Update Frequency of High-Quality Infor-

1526 **mation Sources** On the other hand, high-quality
1527 cultural knowledge sources (Wikipedia, profes-
1528 sional cultural portals, official governmental or
1529 tourism sites, etc.) tend to have relatively **low fre-**
1530 **quency** with **small incremental** changes about cul-
1531 tural knowledge. Prior empirical work shows that
1532 high-quality cultural sources such as Wikipedia
1533 and governmental cultural portals are character-
1534 ized by low-frequency, incremental updates once en-
1535 tries reach maturity (Halfaker et al., 2013; Yasseri
1536 et al., 2012). Anthropological and cross-cultural re-
1537 search further demonstrates that core cultural struc-
1538 tures—values, festivals, etiquette systems—are

1539 slow-moving variables that remain stable across
1540 decades (Hofstede and Consequences, 2001; Ingle-
1541 hart and Welzel, 2005; Geertz, 2017). Only periph-
1542 eral, policy-related aspects of culture (e.g., festival
1543 regulations) undergo periodic adjustments, forming
1544 a minor proportion of cultural knowledge (Ferdi-
1545 nand and Kitchin, 2012). And this portion covers
1546 only 25 dimensions in our schema, accounting for
1547 no more than 18% of all 140 dimensions. This
1548 update scale is entirely acceptable for evaluation
1549 purposes.

1550 To conclude, only a subset of cultural dimen-
1551 sions undergo periodic updates—e.g., festival
1552 regulations, and certain institution-related prac-
1553 tices—and their update proportions are relatively
1554 small compared to the stable core of cultural con-
1555 tent. Hence, using these sources does not compro-
1556 mise the validity or timeliness of our benchmark.

1557 F Data Sources and Data Distribution

1558 **Data Sources** We conduct a categorical analy-
1559 sis of the source websites of the cultural knowl-
1560 edge instances, identifying eight distinct types
1561 of knowledge sources: Government/official web-
1562 sites, Educational institutions/research, News/me-
1563 dia, Language learning/translation, Business/com-
1564 merce, Social media/forums, Culture/tourism, E-
1565 commerce platforms. Figure 9 shows the statistics
1566 of data source for cultural knowledge instances
1567 retrieved in corresponding languages. The data
1568 sources of the two knowledge sets exhibit distinct
1569 distributions: approximately 20% of the Spanish
1570 knowledge instances are derived from cultural and
1571 tourism websites, whereas this source category ac-
1572 counts for only a small proportion in the Chinese
1573 knowledge base. This highlights certain differences
1574 in the distribution of internet data across languages
1575 within cultural contexts.

1576 Table 11 presents the URLs, categories, and fre-
1577 quencies of source webpages retrieved during the
1578 cultural knowledge extraction process for the four
1579 culture. Due to space limits, we display the top
1580 12 URLs with the highest frequencies. It can be
1581 observed that the most frequently occurring source
1582 webpages belong to categories considered rela-
1583 tively high-quality for cultural data collection, such
1584 as educational institutions/research organizations,
1585 government/official websites, culture/tourism, and
1586 language learning/translation.

1587 **Cultural Knowledge Distribution** Figure 10 il-
1588 lustrates the distribution of cultural knowledge

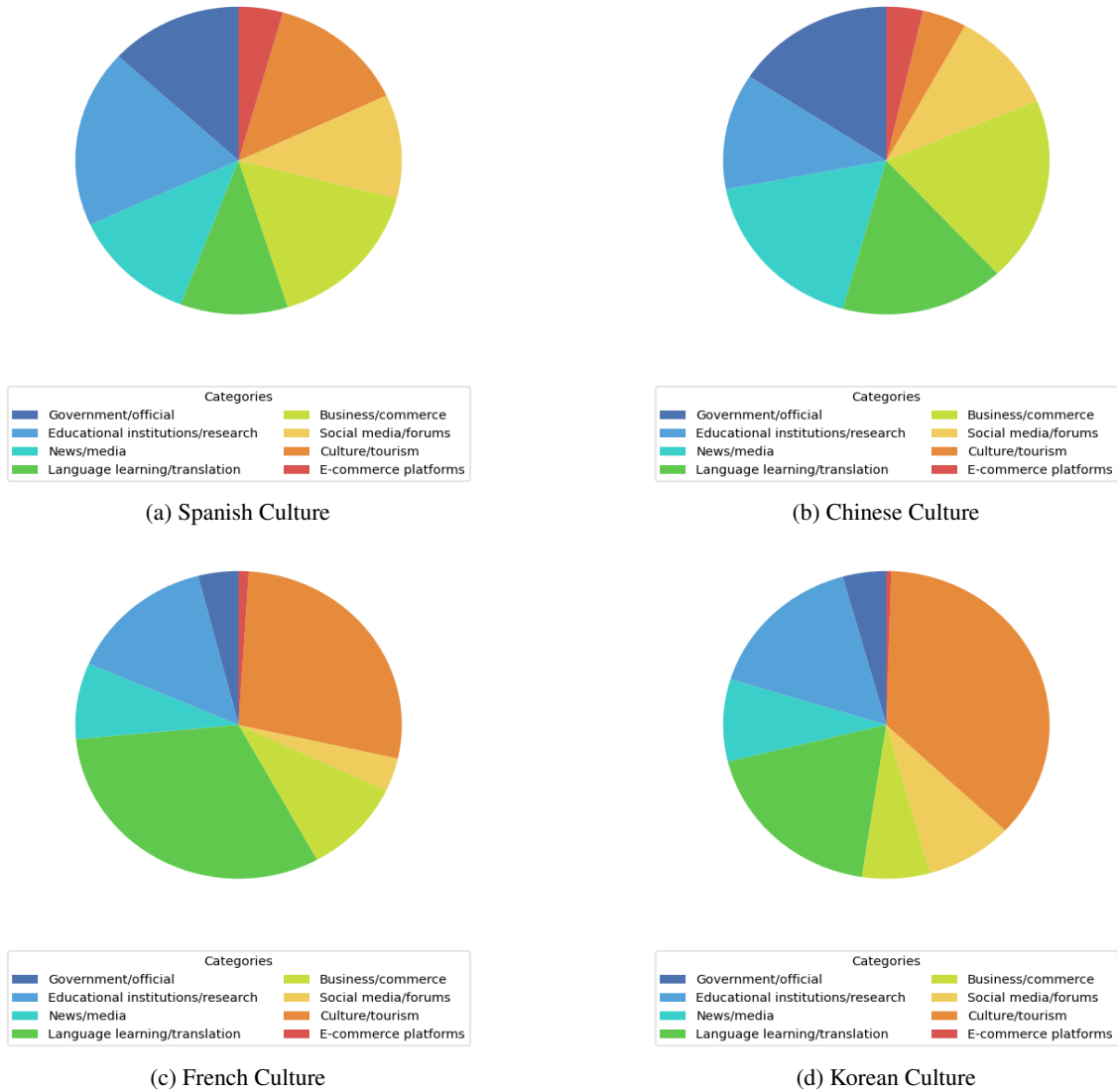


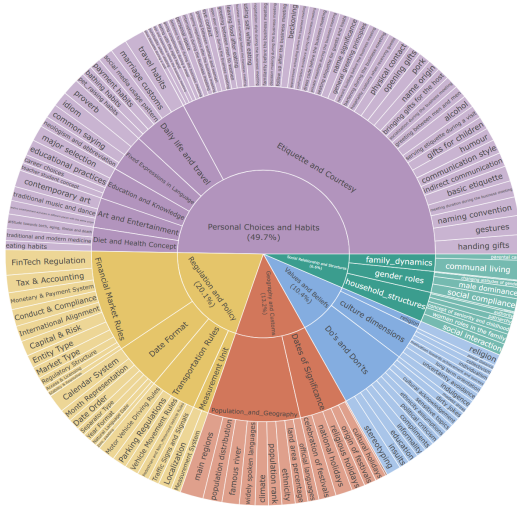
Figure 9: Data Source Statistics for Different Extracted Cultural Knowledge Instances

1589 across dimensions in different cultural knowledge
 1590 sets retrieved using different languages. Despite no-
 1591 table differences in data sources among the knowl-
 1592 edge sets, the knowledge distributions appear to be
 1593 largely consistent, following similar patterns that
 1594 align with the number of dimensions defined in the
 1595 taxonomy. This suggests that the density of cultur-
 1596 ally relevant textual content on the Internet is rela-
 1597 tively uniform across languages, and that through
 1598 retrieval and rewriting, comparable amounts of cul-
 1599 tural knowledge can be obtained for each dimen-
 1600 sion.

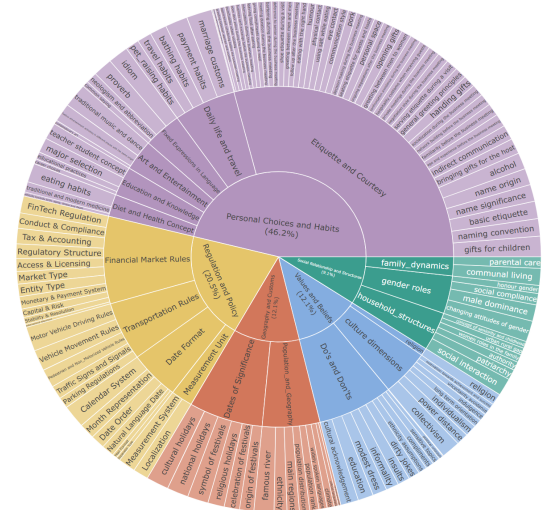
1601 G Annotation Details

1602 **Annotation of Knowledge and Questions** We
 1603 invite experts to annotate the key components of
 1604 the entire evaluation framework for reliability and

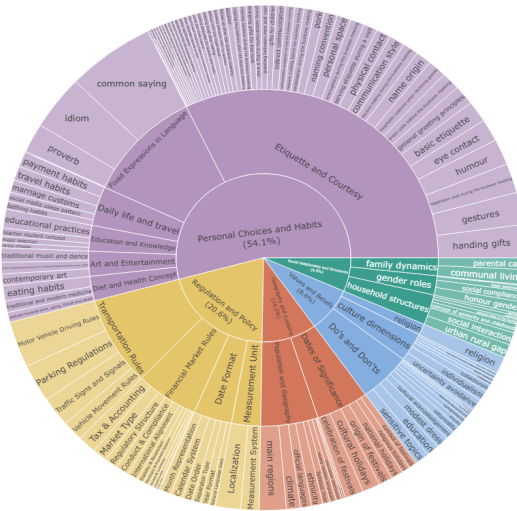
1605 consistency. For each culture, two experts with
 1606 the corresponding cultural background are asked
 1607 to perform sampled annotations on both the cul-
 1608 tural knowledge and the evaluation set. For cultural
 1609 knowledge, the annotation focuses on whether each
 1610 entry contains serious factual errors or biases;
 1611 for the evaluation dataset, the focus is on whether the
 1612 questions, answers, and related cultural knowledge
 1613 are logically consistent. All annotators are col-
 1614 leagues and students from our research institution,
 1615 and they have been fairly compensated for their
 1616 participation. Table 12 presents the instructions.
 1617 Table 7 presents information of the invited experts.
 1618 For each culture, the agreement between the two
 1619 annotators is measured by their simple agreement
 1620 rate. The detailed annotation result of each culture
 1621 is shown in Table 6.



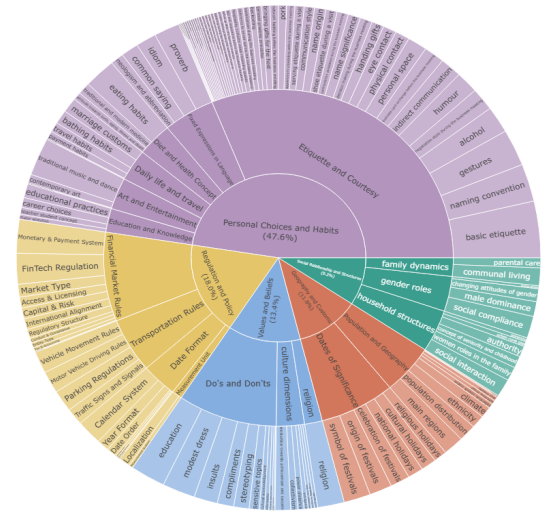
(a) Spanish Cultural Knowledge



(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 schema, and their distribution is consistent with the dimensions defined in the schema.

Annotation of LLM Judges In order to validate the reliability of using LLMs as judges for assessing the accuracy of subjective questions, 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 5.

H Data Instances

H.1 Cultural Knowledge

The extracted cultural knowledge instances are organized according to the evaluation dimensions. Each cultural knowledge instance provides a characteristic description of the corresponding cultural dimension as manifested in a specific cultural context, encompassing relevant commonsense knowledge, shared beliefs, values, and more. During dataset construction, we used the aforementioned detailed evaluation dimensions as cues for retrieval and rewriting. Cultural knowledge associated with the same dimension is grouped together. Furthermore, we clustered the knowledge instances under each dimension to derive even more fine-grained sub-dimensions, which can serve as references for future dataset construction and evaluation. We provide a detailed illustration of examples from the constructed knowledge bases in Table 10.

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 14.

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.

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 8: 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.

Dataset	Theory Ground	Covered Dimensions	Coverage
CultureSynth	multinational library classification systems	<ul style="list-style-type: none"> main regions religious holidays origin of festivals symbol of festivals eating habits attitude towards birth, aging, illness and death traditional and modern medicine bringing gifts for the host educational practices teacher student concept major selection career choices proverb religion communal living urban rural gap 	56 / 140 = <u>40%</u>
CDEval	Hofstede's Cultural Dimensions	<ul style="list-style-type: none"> celebration of festivals travel habits social media usage pattern eating habits attitude towards birth, aging, illness and death traditional and modern medicine educational practices teacher student concept uncertainty avoidance contemporary art decision making during the business meeting communal living parental care authority concept of seniority and childhood power distance 	22 / 140 = <u>15.71%</u>
WorldValuesBench	World Value Survey	<ul style="list-style-type: none"> social interaction social compliance collectivism power distance motivation towards achievement and success uncertainty avoidance long term orientation indulgence modest dress religion changing attitudes of gender communication style indirect communication humour gestures eye contact 	31 / 140 = <u>22.14%</u>

Table 9: Coverage of evaluation dimensions for several representative benchmark studies. The dimensional taxonomies of these datasets are all subsets of ours, with the highest coverage not exceeding 40%. Due to space limits, we only present a subset of the covered dimensions.

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

Table 10: 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.

Culture	URL	Type
Spanish	https://culturalatlas.sbs.com.au	Educational institutions/research
	https://businessculture.org	Educational institutions/research
	https://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
	https://www.quora.com	Social media/forums
	https://kummuni.com	Business/commerce
	https://www.spanishpod101.com	Language learning/translation
	https://www.enforex.com	Language learning/translation
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
	http://www.npc.gov.cn	Government/official websites
	https://www.topchinatravel.com	Culture/tourism
	https://lcchineseschool.com	Language learning/translation
https://www.chinese101.com	Language learning/translation	
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.cafs-sorbonne.fr	Educational institutions/research
	https://www.justlanded.com	Culture/Tourism
	https://businessculture.org	Educational institutions/research
	https://francevibes.com	Culture/Tourism
	https://welcome.fr	Culture/Tourism
	https://www.mondly.com	Language learning/translation
https://www.statista.com	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
	https://www.korea.net	Government/official websites
	https://www.koreatimes.co.kr	News/media
	https://bulletpointchristian.com	Culture/Tourism
	https://www.learnkoreantv.com	Language learning/translation

Table 11: 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. Due to space limitations, we list the top 12 URLs with the highest counts.

Target	Instruction
Knowledge	<p>You are asked to evaluate individual cultural knowledge entries. Your task is to determine whether each entry contains serious factual errors or cultural biases.</p> <p>Specifically, please assess:</p> <p>Factual correctness: Whether the described cultural facts, practices, or norms are accurate and not misleading.</p> <p>Bias and stereotyping: Whether the entry contains overgeneralizations, stereotypes, or culturally biased statements that could lead to misunderstanding or unfair judgments.</p> <p>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.</p> <p>Please base your judgment on widely accepted cultural knowledge and avoid overly strict or subjective interpretations.</p>
Evaluation Dataset	<p>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.</p> <p>Specifically, please assess:</p> <p>Question–answer alignment: Whether the answer directly and appropriately addresses the question.</p> <p>Consistency with cultural knowledge: Whether the answer is logically supported by the provided cultural knowledge and does not contradict it.</p> <p>Internal coherence: Whether there are logical inconsistencies, contradictions, or reasoning gaps within or across the question, answer, and cultural context.</p> <p>Focus on logical coherence rather than stylistic quality or linguistic fluency.</p>
LLM Judges	<p>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.</p> <p>Specifically, please assess:</p> <p>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.</p> <p>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.</p> <p>Consistency with evaluation criteria: Whether the Judge applies the stated evaluation standards appropriately and consistently, rather than making arbitrary or superficial judgments.</p> <p>Focus on whether the Judge’s output can be considered a reliable proxy for human evaluation, rather than evaluating the underlying model response itself.</p>

Table 12: Instructions of the annotators

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 schema of CultureScope(Part 1/4)

Layer	Category	Topic Aspect	Detailed Dimension
Behavioral Patterns	Personal Choices & Habits	Daily Life and Travel	<ul style="list-style-type: none"> payment habits travel habits bathing habits pet-raising habits social media usage pattern marriage customs
		Diet and Health Concept	<ul style="list-style-type: none"> eating habits attitudes towards birth, aging, illness and death traditional and modern medicine
		Education and Knowledge	<ul style="list-style-type: none"> educational practices teacher student concept major selection career choices
		Art and Entertainment	<ul style="list-style-type: none"> festive entertainment activities in different places with the same origin traditional music and dance contemporary art
		Etiquette and Courtesy	<ul style="list-style-type: none"> 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 schema 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 schema 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 schema of CultureScope(Part 4/4)

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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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. Analysis: China's education system does emphasize standardized tests and unified curriculum, which reflects the pursuit of certainty and consistency in knowledge mastery. - However in fact, in recent years, the Chinese government and educational institutions have begun to attach importance to cultivating students' innovation, critical thinking and practical skills.....	<ol style="list-style-type: none"> 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 14: (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.	<ol style="list-style-type: none"> 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 14: (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.

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

===

Knowledge Summary(Spanish)

Eres un investigador especializado en la cultura española. Se te proporcionará un texto de una página web (puede estar en cualquier idioma).

Tu tarea es identificar las características culturales que estén relacionadas con el siguiente aspecto específico de la cultura española:

Dimensión cultural:dimension

Por favor, extrae del texto solo los elementos relevantes que estén claramente relacionados con esta dimensión cultural.

Escribe las características en español, siguiendo este formato:

[Título breve de la característica]

[Descripción clara y concisa en español]

Fuente de información: [Frase, palabra clave o idea tomada directamente del texto original]

Escribe siempre en español. .

Texto de la página web:

""

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 inside 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