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, most existing benchmarks lack comprehensiveness and are challenging to scale and adapt across different cultural contexts, as their frameworks often lack guidance from well-established cultural theories and tend to rely on expert-driven manual annotations. To address these issues, we propose *CultureScope*, the most comprehensive evaluation framework to date for assessing cultural understanding in LLMs. Inspired by existing cultural studies, we design a novel dimensional schema for cultural knowledge classification, comprising 3 layers and 140 dimensions, which guides the automated extraction of culture knowledge and corresponding evaluation datasets for any given cultures. Experimental results demonstrate that our method can effectively evaluate cultural understanding. They also reveal that existing large language models lack comprehensive cultural competence, and merely incorporating multilingual data does not necessarily enhance cultural understanding. All code and data files are available at `https://github.com/AnonymousUserForSubmission/Culture`

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide spectrum of natural language processing tasks, and are now widely adopted in applications ranging from virtual assistants to educational tools (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 real-world contexts (Benkler et al., 2023; AlKhamissi et al., 2024; Shen et al., 2024). For example, if a healthcare chatbot powered by an LLM recommended placing an elderly parent in a nursing facility to a Chinese user, it fails to account for the cultural significance of filial piety (Xiaoshun) in traditional Chinese society. While the suggestion aligns with Western notions of efficiency, it conflicts with deeply rooted Chinese values, leading to user dissatisfaction and negative public response. Therefore, how to comprehensively and accurately evaluate LLMs' cultural understanding abilities in specific cultural contexts has become an urgent issue that must be addressed for their broader adoption.

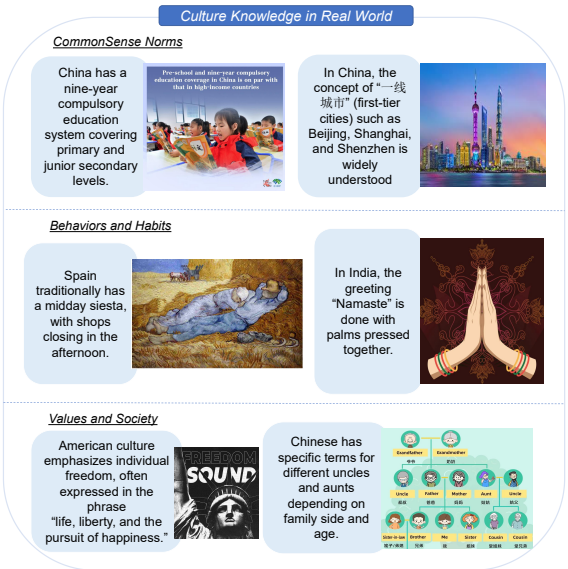


Figure 1: The classification of real-world cultural knowledge is inherently complex and diverse, encompassing a wide range of aspects such as institutions, geography, behavioral habits, etiquette, values, and social structures. Existing evaluation datasets often fail to comprehensively and accurately cover these types of knowledge.

Cultural understanding refers to a model's ability to comprehend culturally grounded definitions and concepts, identify cultural stereotypes or biases, and appropriately apply cultural knowledge

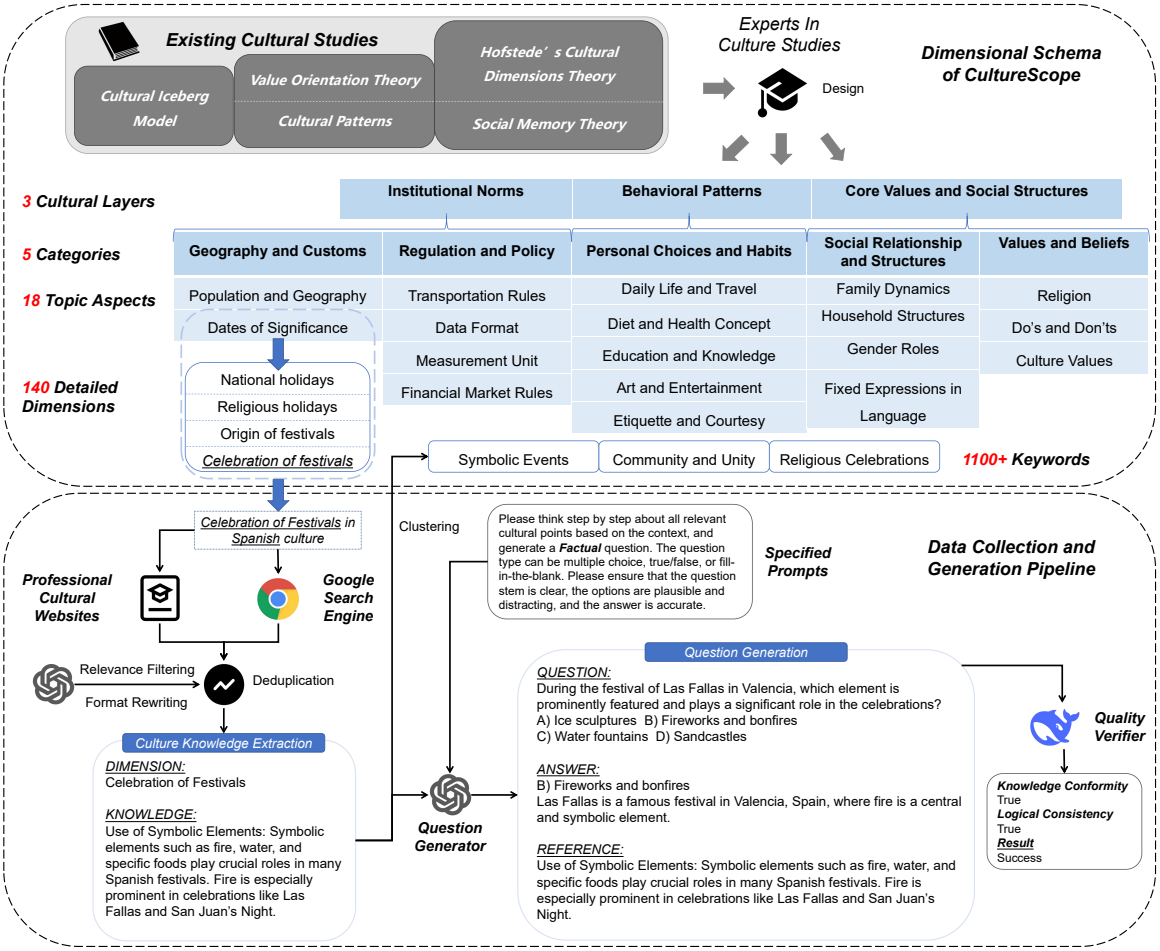


Figure 2: The overall framework of CultureScope. By adjusting the name of the target culture in knowledge extraction, this framework can be readily adapted to assess cultural understanding for any given target culture. We invite experts in culture studies to to develop the most comprehensive knowledge classification schema to date for evaluating the cultural understanding of large language models, based on the most influential existing cross-cultural and mono-cultural theories. This framework enables more fine-grained analysis and interpretation of LLMs' performances on specific cultural dimensions.

in real-world or context-sensitive scenarios (Earley et al., 2002). As is shown in Figure 1, culture is a complex and holistic concept (Tylor, 1871), and cultural knowledge in the real world scenarios encompasses multiple dimensions. Despite growing efforts in the evaluation and enhancement of cultural understanding, existing benchmarks still suffer from significant limitations.

First, existing evaluations tend to lack comprehensiveness, which limits their practical utility, and the assessed capability scores may be misaligned with real-world scenarios. Previous studies either directly rewrite existing surveys (Jiang et al., 2024; Zhao et al., 2024; Li et al., 2024), or generate and manually annotate data by selecting a small number of keywords (Chiu et al., 2024; Lee et al., 2024; Myung et al., 2024). Their classification frameworks for cultural knowledge dimensions lack rigorous theoretical guidance and tend to rely on intuitive and empirical experience. **Second**, existing evaluation benchmarks are difficult to scale and adapt to other cultural and linguistic contexts, as they heavily rely on manual annotation (Myung et al., 2024; Palta and Rudinger, 2023; Mousi et al., 2024). The high cost, labor intensity, and cultural specificity of manual labeling limit their generalizability across diverse cultural settings.

In this paper, we propose CultureScope, an automated, scalable, and currently the most comprehensive evaluation framework for LLMs' cultural understanding capabilities. We argue that the evaluation framework for cultural understanding should meet four key criteria: *comprehensiveness, operability, interpretability, and scalability*. To address the issues above, we develop the most comprehensive classification schema of cultural knowledge to

Benchmark	Theoretical Guidance	Commonsense Norms	Behaviors & Habits	Values & Society	Transferability to other cultures	Number of dimensions
BLEnD (2024)	✗	✗	✓	✗	✗	6
CultureLLM (2024)	✗	✗	✗	✓	✗	7
WorldValuesBench (2024)	✗	✗	✗	✓	✗	12
KorNAT (2024)	✗	✗	✓	✓	✗	/
ARADICE (2024)	✗	✓	✓	✓	✗	30
Ours(CultureScope)	✓	✓	✓	✓	✓	**140**

Table 1: Existing benchmarks lack guidance from well-established cultural theories, with dimension systems largely designed based on empirical heuristics. As a result, they provide limited coverage of cultural knowledge, hindering comprehensive evaluation of culture understanding. Moreover, the reliance on manual annotation in many existing methods leads to limited automation and poses significant challenges in extending them to diverse cultural contexts.

date based on existing cultural theories, propose an evaluation framework applicable to any culture, and construct several evaluation datasets.

We conduct extensive evaluations of existing LLMs. Our findings indicate that cultural capability fundamentally involves the mastery and application of cultural knowledge. Relying exclusively on linguistic knowledge or deeper reasoning is insufficient for improving a model's cultural alignment. Furthermore, existing models perform inconsistently across different cultural dimensions and languages, which provides clear guidance for targeted enhancements of the models. Our contributions are as follows:

- Building on previous monocultural and cross-cultural research, we propose the most comprehensive framework to date for assessing LLMs' capabilities in specific cultural contexts. This framework captures both surface-level practices and deep-level assumptions in a certain culture, covering a wide range of cultural knowledge, including cultural facts, behavioral norms, values, and beliefs, enabling automatic data generation and can be flexibly extended to any given culture.

- We extract a substantial volume of cultural knowledge embedded in specific cultural contexts from various data sources, which can facilitate subsequent evaluation and training of LLMs. These cultural knowledge possess high quality, and the accuracy of knowledge instances retrieved and generated in different languages remain comparable.

- We construct several comprehensive evaluation datasets and conduct a series of extensive experiments on commonly used LLMs to evaluate their cultural understanding abilities. The performances of these LLMs vary significantly across different cultural dimensions and languages. Enhancing models solely through multilingual data and deep reasoning is insufficient to achieve stable improvements in cultural understanding.

2 Related Work

In recent years, the evaluation and improvement of LLMs' cultural understanding capabilities have received significant attention within the research community and industry. These capability assessments can broadly be divided into two categories. (1) Question answering on factual knowledge within specific cultural contexts (Onohara et al., 2024; Singh et al., 2024; Nayak et al., 2024), and (2) Evaluation of cultural values and moral reasoning (Hadar-Shoval et al., 2024; Karinshak et al., 2024; Aksoy, 2025; Ji et al., 2025). Common data construction methods include manual annotation and LLM generation. However, current research on the evaluation of LLMs' cultural understanding capabilities still faces notable limitations. Table 1 provides a detailed comparison between our evaluation benchmark and previous works. On one hand, most existing evaluation benchmarks are not grounded in robust theoretical frameworks, and their dimensional structures are typically based on **intuitive, experience-driven approaches**. As a result, their evaluation frameworks fail to cover a broad spectrum of cultural knowledge and cannot fully reveal the limitations and areas for improvement of LLMs in culturally rich scenarios. On the other hand, existing methods often rely on **expert manual annotation**, which incurs high data generation costs and makes it difficult to scale to other cultures. Some datasets even involve hundreds of annotators, which is evidently costly and difficult to replicate (Myung et al., 2024).

(a) Spanish Cultural Knowledge

(b) Chinese Cultural Knowledge

Figure 3: Distribution of Retrieved Cultural Knowledge in Spanish and Chinese. 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.

3 CultureScope Framework

3.1 Dimensional Schema Design

Current cultural theories emphasize that culture is an iceberg-like concept, whose deeper layers are closely related to people's behaviors and the surrounding environment (Hall, 1976; Baker, 2009). We construct a hierarchical classification of cultural knowledge that captures both surface-level practices and deep-level assumptions, dividing it into three layers: Institutional Norms, Behavioral Patterns, and Core Values and Social Structures. The three layers respectively cover cultural knowledge of common norms, behaviors and habits, and underlying values/social structures. As is shown in Figure 2, the schema is divided into four levels: *Culture Layer, Category, Topic Aspect, and Detailed Dimension*, which offers systematic dimensions to assess LLMs' cultural understanding. The detailed content of the dimensional schema and its theoretical foundations are shown in Appendix C.

3.2 Cultural Knowledge Extraction

Some studies have found that even datasets widely regarded as reliable, such as Wikipedia, still exhibit notable cultural biases (Naous et al., 2023). Therefore, in our data construction process, we incorporate complementary data sources to ensure diversity and to mitigate the risk of bias from relying on a single source. The first is manually selected professional cultural websites[1]. These sites offer highly reliable content, but the amount of data, the coverage across dimensions is limited, and data extraction often requires specialized processing. The other is data retrieved through Google search. This method allows us to cover the full range of dimensions, and provides a large and scalable volume of data, but the quality of the content requires further verification. We convert the detailed dimensions in the dimensional schema into search keywords and combine them with the target culture to construct retrieval queries(e.g., Celebration of Festivals in Spanish culture). In this step, we apply GPT-4o to filter the page content, summarize and extract information from the retrieved texts.

Size	Knowledge Instances		Evaluation Dataset	
	En	Local	En	Local
Spanish	2767	2483	11.09k	10.68k
Chinese	3177	3535	11.2k	11.2k

Table 2: Statistics of CultureScope. We report the number of knowledge instances in each language-specific contexts. The number of data samples is the same across different types of question content(Factual, Conceptual, Misleading, Multi-hop) across languages.

3.3 Evaluation Dataset Generation

We draw on cognitive science theories (Wilberschied and Lee, 2015; Avgousti, 2018; Feng et al.,

[1]For example, https://culturalatlas.sbs.com.au/

Culture	Annotator	Knowledge	Evaluation Dataset
Spanish	Annotator1	98.57%	99.46%
	Annotator2	97.85%	99.12%
Chinese	Annotator1	97.14%	100.0%
	Annotator2	99.29%	99.11%

Table 3: The quality assessment results of the cultural knowledge and the evaluation dataset via manual sampling inspection. The validation of the cultural knowledge instances focuses on identifying factual or expression-related errors, and the quality of the evaluation set emphasizes the logical consistency among the question, the answer, and the reference knowledge.

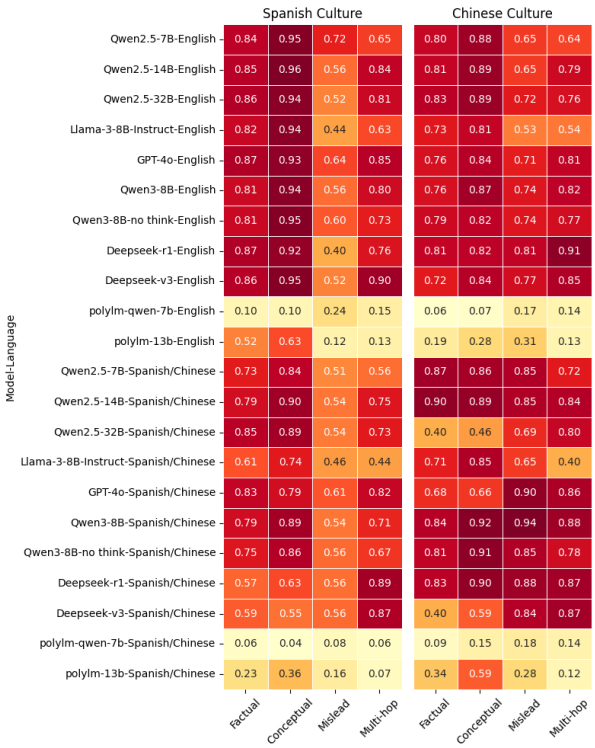


Figure 4: Models' performances across different question content types. Many models demonstrate culture-specific weaknesses in identifying cultural bias.

2025) to define four content types for our questions: Factual, Conceptual, Misleading and Multi-hop, as this design ensures that the evaluation not only probes different facets of cultural knowledge but also reflects the practical challenges models face in real-world cultural understanding and interaction. These four question types respectively assess whether the model knows cultural facts, understands the underlying meaning of cultural phenomena, can identify cultural biases, and is able to synthesize multiple cultural elements while grasping the deeper logic or internal connections among cultural phenomena. The question formats include

Culture	Model	Acc	Alpaca	MMLU
Spanish	Base	82.14%	81.80%	85.40%
	Trained	**87.32%**	82.00%	85.20%
Chinese	Base	73.21%	81.80%	85.40%
	Trained	**74.29%**	81.95%	85.55%

Table 4: We report the training results of Qwen2.5-7B-Base. Training LLMs on *extracted cultural knowledge* can enhance their cultural understanding abilities while preserving their general-purpose capabilities.

multiple-choice, true/false, short answer, and essay questions. For each type of question, we apply specific expert-designed prompts to clearly define the task requirements. Question generation is based on a Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) approach, randomly selecting a set of cultural rules for the LLM to reference during generation. Detailed prompt formulations are provided in the Appendix E.

3.4 Quality Control

For cultural knowledge and generation-related tasks, the framework provides LLM-based quality evaluation and filtering. For the validation of cultural knowledge, we primarily focus on the consistency between the rewritten content and the source webpages, since we found that when using queries such as "eating habits in Spanish culture", the top five results returned by Google are consistently high-quality webpages that are highly relevant to the target dimension of the query. In addition to widely recognized platforms such as Wikipedia, these results also include official releases from relevant government agencies of the target country, authoritative media outlets, long-established authoritative tourism information portals and so on.

On the other hand, the evaluation of the generated questions primarily focuses on examining the logical consistency among the question itself, the provided answer, and the underlying reference cultural knowledge. Specifically, it assesses whether the question is formulated in a way that is coherent with the reference knowledge, whether the corresponding answer logically follows from the question, and whether any contradictions or reasoning errors arise across these three components.

4 Evaluation Benchmark

4.1 Evaluation Metric

We evaluate model performance using two metrics. First, **Accuracy** measures the proportion of correct

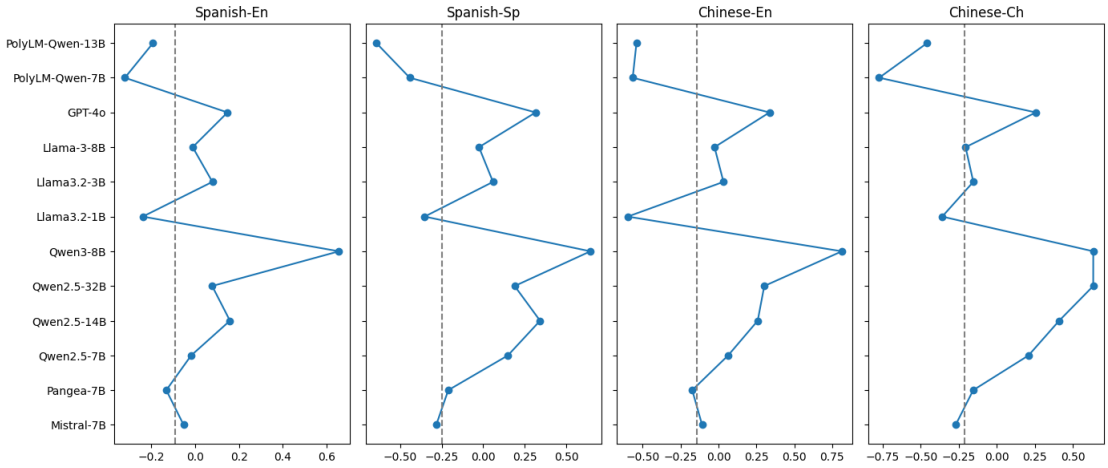


Figure 5: BT-Scores of the target models and the baseline models, estimated from pairwise comparisons against two baselines(Mistral-7B and Pangea-7B). The BT model fits global ability parameters from the comparison outcomes; higher scores indicate stronger relative performance within this fitted ability space.

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}.

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

The second is the **Bradley–Terry Ability Score (BT-Score)**, a metric designed to capture relative model capability by fitting pairwise win–loss outcomes between a target model and baseline models.

Given pairwise comparison results between a target model M and a baseline model B, each instance records one of three outcomes: M wins, B wins, or a tie. For a pair of models (i, j), let w_{ij} denote the number of times i is preferred over j. The Bradley–Terry model (Bradley and Terry, 1952) assumes each model i has a latent ability $\theta_i \in R$, and the probability that i is preferred over j is

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

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

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

and estimate the ability parameters $\{\theta\}$ by maximizing this log-likelihood. Finally, we define the **BT-Score** of the target model M 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 We independently construct evaluation datasets grounded in Chinese culture and Spanish-speaking cultural contexts. The constructed benchmark consists of **11962** cultural knowledge instances and **44.1k** questions in total. Table 2 shows the detailed statistics of each culture and language. The distribution of the English version of the cultural knowledge instances is illustrated in Figure 3. The distribution of cultural knowledge instances exhibits a trend similar to that of the dimensions in the schema, a pattern that holds consistently across different languages.

Data Quality We invite human experts to manually evaluate the quality of the culture knowledge and the correctness of the evaluation dataset. Table 3 shows the quality evaluation results of the constructed benchmark. The validation of cultural knowledge focuses on identifying factual or expression-related errors, whereas the evaluation of the dataset questions emphasizes the logical consistency among the question, the answer, and the reference knowledge. The results show that the generated cultural knowledge and evaluation questions are of high quality and can be used to assess the cultural understanding capabilities of LLMs,

Models	Spanish Culture							Chinese Culture						
	Lang	Acc	G & C	PC & H	R& P	SR & S	V & B	Lang	Acc	G & C	PC & H	R& P	SR & S	V & B
General-purpose models														
Qwen2.5-7B	en	0.832	0.850	0.808	0.800	0.958	0.868	en	0.741	0.883	0.681	0.740	0.771	0.829
	sp	0.659	0.750	0.620	0.640	0.688	0.737	cn	0.827	0.800	0.815	0.810	0.896	0.868
Qwen2.5-14B	en	0.870	0.883	0.851	**0.840**	**0.958**	0.908	en	0.783	**0.917**	0.761	0.717	0.750	0.867
	sp	0.745	0.767	**0.728**	0.730	0.771	0.790	cn	**0.871**	**0.917**	**0.844**	**0.880**	**0.896**	**0.908**
Qwen2.5-32B	en	0.861	**0.900**	0.848	0.820	0.896	0.908	en	**0.798**	0.917	**0.772**	**0.740**	**0.854**	0.842
	sp	0.753	0.817	0.714	0.710	0.792	**0.880**	cn	0.586	0.667	0.583	0.550	0.563	0.592
Llama3-1B	en	0.800	0.783	0.786	0.760	0.938	0.829	en	0.651	0.683	0.627	0.576	0.792	0.720
	sp	0.563	0.567	0.554	0.480	0.625	0.658	cn	0.655	0.610	0.634	0.600	0.729	0.790
Llama3-3B	en	0.800	0.783	0.786	0.760	0.938	0.829	en	0.651	0.683	0.627	0.576	0.792	0.720
	sp	0.563	0.567	0.554	0.480	0.625	0.658	cn	0.655	0.610	0.634	0.600	0.729	0.790
Llama3-8B	en	0.800	0.783	0.786	0.760	0.938	0.829	en	0.651	0.683	0.627	0.576	0.792	0.720
	sp	0.563	0.567	0.554	0.480	0.625	0.658	cn	0.655	0.610	0.634	0.600	0.729	0.790
GPT-4o	en	**0.875**	0.850	**0.870**	0.840	0.938	**0.921**	en	0.780	0.850	0.765	0.690	0.750	**0.921**
	sp	**0.761**	**0.850**	0.710	**0.770**	**0.813**	0.829	cn	0.773	0.800	0.765	0.750	0.771	0.816
Deep-reasoning models														
Deepseek-r1	en	0.845	0.733	0.851	0.810	0.938	**0.895**	en	**0.837**	0.883	**0.801**	**0.840**	**0.854**	**0.921**
	sp	0.664	0.810	0.629	0.586	0.729	0.737	cn	0.870	**0.917**	0.848	0.840	**0.896**	**0.934**
Deepseek-v3	en	**0.879**	**0.933**	**0.859**	0.860	0.938	0.895	en	0.794	**0.900**	0.743	0.790	0.851	0.868
	sp	0.644	0.683	0.623	0.660	0.542	0.733	cn	0.675	0.683	0.659	0.640	0.688	0.763
Qwen3-8B	en	0.834	0.833	0.786	**0.870**	**0.979**	0.868	en	0.796	0.867	0.757	0.770	0.833	0.895
	sp	**0.734**	**0.817**	**0.703**	**0.720**	**0.792**	0.763	cn	**0.893**	0.917	**0.877**	**0.890**	0.896	0.934
Qwen3-8B-f	en	0.827	0.833	0.804	0.800	0.958	0.855	en	0.779	0.900	0.750	0.740	0.792	0.829
	sp	0.713	0.817	0.688	0.680	0.688	**0.776**	cn	0.838	0.867	0.823	0.810	0.875	0.882
Models Continually-trained on Multilingual Corpus														
PolyLM-7B	en	0.113	0.133	0.130	0.090	0.042	0.105	en	0.111	0.133	0.091	0.120	0.167	0.118
	sp	0.057	0.083	0.044	0.040	0.083	0.092	cn	0.141	0.117	0.145	0.140	0.125	0.158
PolyLM-13B	en	0.458	0.367	0.491	0.390	0.500	0.474	en	0.225	0.150	0.243	0.172	0.292	0.250
	sp	0.204	0.200	0.207	0.250	0.063	0.224	cn	0.330	0.283	0.330	0.330	0.250	0.421

Table 5: Main evaluation results on Spanish culture and Chinese culture. We evaluate the question-answer accuracy of several types of LLMs on different languages and cultural categories mentioned above. Lang refers to the language of the questions. Qwen3-8B-f refers to Qwen3-8B without deep reasoning enabled. Abbreviations follow a consistent pattern: for example, "G&C" stands for Geography and Customs; other abbreviations follow the same convention.

which also validates the effectiveness of our automatic quality control method.

5 Experiment Results

We evaluate different models across all dimensions, languages, and question types, and analyzed potential influencing factors. The detailed setups are shown in Appendix A. The overall performances of all LLMs are listed in Table 5 and Figure 5. The two metrics exhibit consistency in characterizing model capability.

Observation 1: The cultural understanding ability of LLMs is language-dependent. Overall, as is shown in Table 5, the accuracy of model responses to cultural questions is directly related to the language of the questions. For Spanish-speaking cultures, When the language switch from English to Spanish, all models show a trend of accuracy decline.

On the other hand, the models demonstrate **culture-specific weaknesses in identifying cultural bias**. As is shown in Figure 4, the models' performances on handling Mislead-type questions are significantly lower than on other types within Spanish culture, while in Chinese culture, such a disparity is less pronounced and exhibits lower consistency across models. This suggests that the models may be more susceptible to misleading content or less robust in reasoning under deception in specific cultural contexts.

Meanwhile, all models exhibit **varying performances across different *Topic Aspects***. We evaluate the finer-grained *Categories* within each *Layer* and present a subset of the results in the Figure 7, and the complete results can be found in the Appendix B. In the context of Spanish culture, as the cultural knowledge evaluation progresses from shallow to deep layers, the models exhibit increasing performance divergence within the same dimension when questioned in different languages. At the **Institutional Norm Layer**, all models perform poorly on questions related to *Measurement Units*, while demonstrating better mastery of knowledge concerning *Dates of Significance*.

Observation 2: Larger models typically exhibit stronger performance on cultural under-

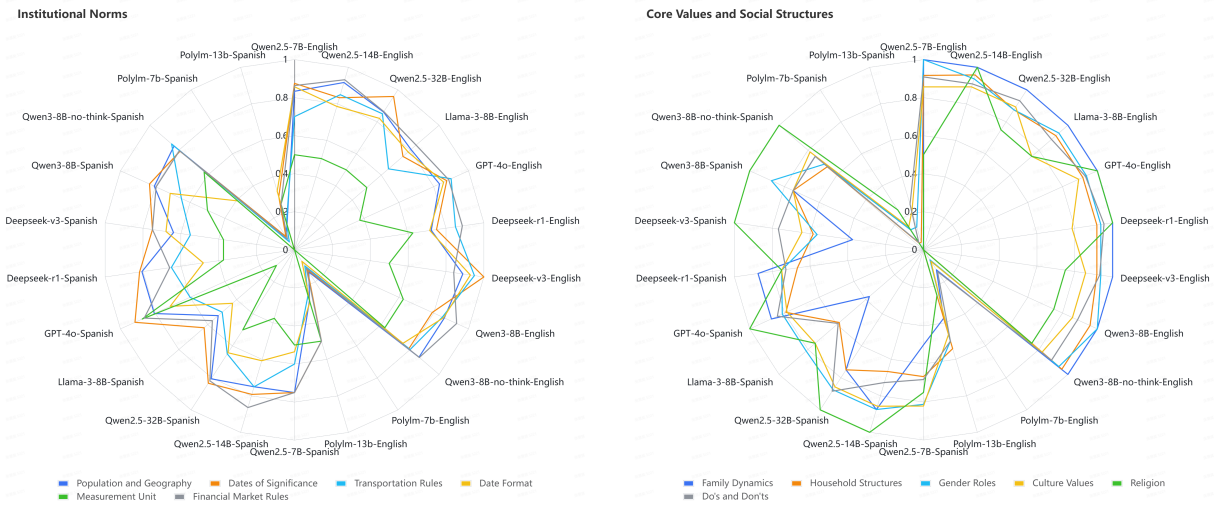


Figure 6: Accuracy on the Institutional Norms and Core Values and Social Structures layers within the Spanish cultural context. Performances on institutional norms are comparable across English and Spanish, whereas values and social structures reveal notable cross-linguistic differences, particularly in *Family Dynamics* and *Religion*.

standing tasks, as they encode a broader range of knowledge. Models from the same series exhibit a generally positive correlation between performance and **model size**. On one hand, for Spanish culture, increasing model size consistently leads to notable performance gains on Spanish-language questions. On the other hand, for Chinese culture, as the scale increases, nearly all models show improved performance on both English and Chinese questions.

Observation 3: Deep reasoning does not inherently yield substantial gains on cultural understanding capability. Although models with deep thinking enabled outperform those without it in the context of Chinese culture, a similar advantage is not observed for Spanish-speaking cultures. Deep thinking models we used(Deepseek-r1 and Qwen3) were trained on larger Chinese corpora and thus possess stronger Chinese language capabilities. This suggests that deep thinking yields greater benefits only when the model has been trained on sufficiently rich corpora in the target language. Without access to relevant knowledge, no amount of reasoning can fully compensate for the limitations in cultural understanding.

Observation 4: Language≠Culture. As shown in the Figure 7, compared to the similarly sized Qwen models, PolyLM (Wei et al., 2023), which incorporates additional multilingual data, exhibits a significant performance drop for both Spanish and Chinese cultures, rather than an improvement. 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 capability evaluation.

Observation 5: Knowledge enhances cultural understanding. 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 trained model shows notable improvements, supporting our hypothesis that cultural understanding is strongly correlated with the acquisition of cultural knowledge, and that enhancing a model's knowledge of culture facilitates the advancement of its cultural understanding.

6 Conclusion

In this study, we address the critical challenge of systematically and comprehensively evaluating the cultural understanding capability of LLMs. We propose a theoretically grounded and extensible framework for cultural understanding evaluation. Our experimental results on Spanish and Chinese cultures reveal that current LLMs exhibit significant differences in cultural capability across languages and dimensions. Cultural understanding fundamentally depends on a model's cultural knowledge rather than on language knowledge alone, and that deep reasoning does not reliably compensate for cultural knowledge gaps. These findings highlight the limitations of conventional multilingual training for achieving true cultural alignment, which offers valuable insights for the future training, evaluation, and deployment of culturally aligned LLMs.

7 Limitations

This study has two main limitations. First, due to cost constraints, the evaluation was conducted only on Spanish and Chinese cultural contexts, although the proposed framework is highly reproducible and could be extended to other cultures. Second, biases in the datasets remain, which is a common issue in related research; despite efforts to minimize them, they cannot be entirely eliminated.

8 Ethical Considerations

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 was collected exclusively from publicly available internet sources, including educational materials, open-access cultural databases, and publicly shared questionnaires or forums. 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. 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.

References

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

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *ArXiv*, abs/2402.13231.

M. Avgousti. 2018. Intercultural communicative competence and online exchanges: a systematic review. *Computer Assisted Language Learning*, 31:819 – 853.

W. Baker. 2009. The cultural iceberg.

Noam Benkler, Drisana Mosaphir, Scott E. Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *ArXiv*, abs/2312.10075.

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.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms.

Michael Desmond, Zahra Ashktorab, Werner Geyer, Elizabeth M. Daly, Martín Santillán Cooper, Qian Pan, Rahul Nair, Nico Wagner, and Tejaswini Pedapati. 2025. Evalassist: Llm-as-a-judge simplified. pages 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 e-prints*, pages arXiv–2407.

E. Earley, Chris Earley, and Pharmacia Quotes. 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*.

Dong Guo, Daming Yang, Haoyang Zhang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Maurice Halbwachs. 1992. *On Collective Memory*. University of Chicago Press, Chicago, IL. Original work published 1950.

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

Geert Hofstede. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA.

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*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71.

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

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*.

Florence R. Kluckhohn and Fred L. Strodtbeck. 1961. *Variations in Value Orientations*. Row, Peterson and Company, Evanston, IL.

Alfred L. Kroeber and Clyde Kluckhohn. 1952. *Culture: A Critical Review of Concepts and Definitions*. Harvard University Press, Cambridge, MA.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Red Hook, NY. Curran Associates, Inc. ArXiv:2005.11401.

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

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

Aohan Liu, Baolin Feng, Bin Xue, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

G. Mohan, R. P. Kumar, P. V. 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. *Knowl. Inf. Syst.*, 66:5047–5070.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*.

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

Sanjay Nadhavajhala and Yingbei Tong. 2024. Rubra-mistral-7b-instruct-v0.3.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2024. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. *arXiv preprint arXiv:2410.17250*.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

An Yang Qwen Team, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

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

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. pages 5668–5680.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

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

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

Wilberschied and F. Lee. 2015. Intercultural communicative competence: Literature review. 2:4.

A Yang, A Li, B Yang, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

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

A Experimental Setups

A.1 Evaluation Setups

Our evaluation include three types of LLMs: General-purpose models, including Qwen-2.5-series (Qwen Team 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 trained specifically on multilingual data, including Polylm-Qwen-7B and Polylm-Qwen-13B (Wei et al., 2023). The baselines used for computing the BT-Score are Mistral-7B-Instruct-v0.3 (Nadhavajhala and Tong, 2024) and Pangea-7B (Yue et al., 2024). All local models are run with FP16 precision on NVIDIA A800 GPUs.

A.2 Training Setups

We constructed a training dataset based on the mined 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 2e6, using the CPUAdamW optimizer with gradient offloading enabled to reduce GPU memory usage. The model was trained in bfloat16 precision with gradient checkpointing to further optimize memory efficiency. The total number of training steps was determined dynamically based on the dataset size.

B Supplementary Data

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.

We evaluate the finer-grained *Categories* within each *Layer* and the results are shown in Figure 7. 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.

In the context of Spanish culture, as the cultural 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 English, all models perform poorly, whereas on Spanish questions, Art and Entertainment outperforms other secondary dimensions. 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. In contrast, for Chinese culture, as the evaluation deepens, the models' performance across different languages tend to stabilize within the same dimension.

Statistics In addition to the statistics mentioned in the main technical content, we also analyze the data distribution and sources of the extracted cultural knowledge instances.

For each culture, we conduct knowledge extraction and evaluation dataset generation in both English and the corresponding native language (i.e., Spanish or Chinese). We conduct a categorical analysis of the source websites of the cultural knowledge instances, identifying eight distinct types of knowledge sources: Government/official websites, Educational institutions/research, News/media, Language learning/translation, Business/commerce, Social media/forums, Culture/tourism, E-commerce platforms. We extract cultural knowledge of Chinese and Spanish cultures using both English and the corresponding target languages. Each instances are build upon CultureScope framework. Figure 8 shows the statistics of data source for Spanish and Chinese culture bases retrieved in corresponding languages. The data sources of the two knowledge bases exhibit distinct distributions: approximately 20% of the Spanish knowledge base is derived from cultural and tourism websites, whereas this source category accounts for only a small proportion in the Chinese knowledge base. This highlights certain differences in the distribution of internet data across languages within cultural contexts.

Figure 9 illustrates the distribution of cultural knowledge across dimensions in different cultural knowledge bases retrieved using different languages. Despite notable differences in data sources among the knowledge bases, the knowledge distributions appear to be largely consistent, following similar patterns that align with the number of dimensions defined in the taxonomy. This suggests that the density of culturally relevant textual content on the Internet is relatively uniform across languages, and that through retrieval and rewriting, comparable amounts of cultural knowledge can be obtained for each dimension.

C Culture Knowledge Classification Schema

C.1 Theoretical Foundations

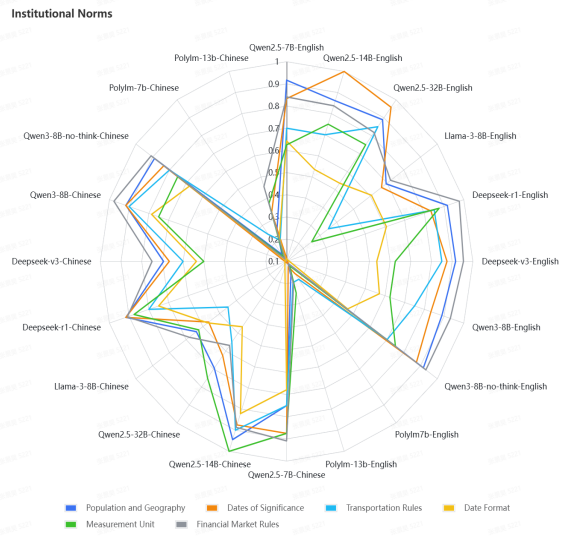
The *Cultural Iceberg Model* (Hall, 1976) posits that most aspects of culture lie beneath the surface—such as beliefs, values, and thought patterns—while visible behaviors and customs represent only a small, superficial portion. *Value Orientation Theory* (Kluckhohn and Strodtbeck, 1961) further explains that cultures differ systematically in how they resolve fundamental human problems, such as relations with nature, time, and social hierarchy. Building on this, the concept of *Cultural Patterns* (Kroeber and Kluckhohn, 1952) emphasizes the recurring configurations of values and practices that shape collective behavior. *Hofstede's Cultural Dimensions Theory* (Hofstede, 1980) provides a quantitative framework for comparing cultures along key dimensions such as power distance, individualism, and uncertainty avoidance, highlighting the structured variability of cultural systems. Meanwhile, *Social Memory Theory* (Halbwachs, 1992) underscores the role of shared narratives and collective memory in maintaining cultural continuity across generations.

Together, these frameworks suggest that culture operates on multiple interrelated Cultural Layers—Institutional Norms (formalized social rules), Behavioral Patterns (observable practices), and Core Values and Social Structures (deep-rooted ideational and relational systems), and each layer includes a series of hierarchical subcomponents—thus inspiring the dimensional schema proposed in this study.

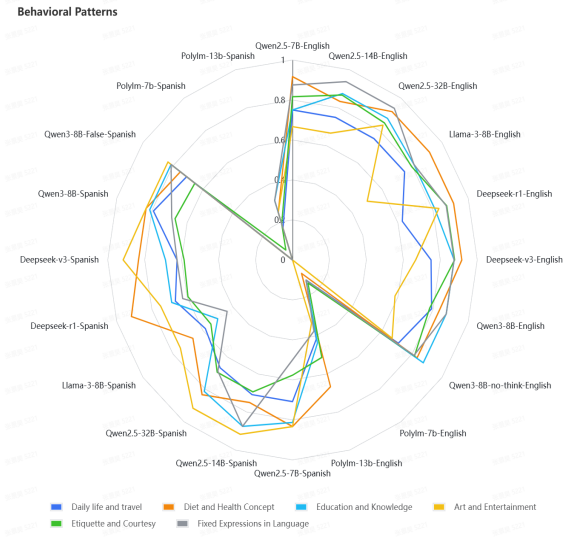
C.2 Classification Principles

To comprehensively and scientifically evaluate the cultural understanding capability of LLMs, the dimensional schema for classifying cultural knowledge should satisfy four core principles. **1) Comprehensiveness**: The dimensions should cover cultural knowledge as fully as possible. **2) Opera-**

(a) Institutional Norms of Spanish culture

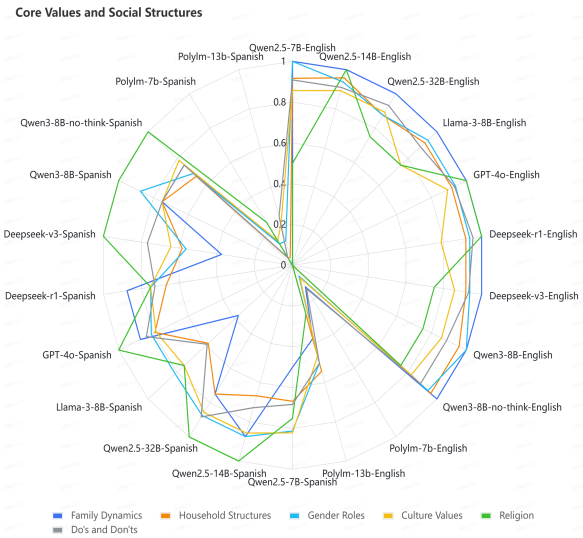


(b) Institutional Norms of Chinese culture

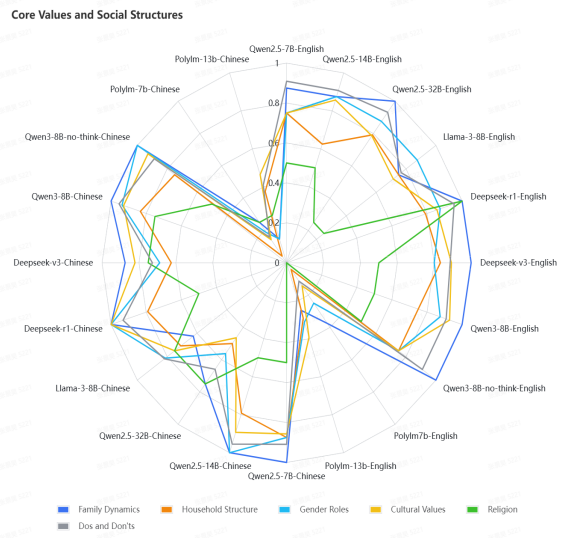


(c) Behavioral Patterns of Spanish culture

(d) Behavioral Patterns of Chinese culture



(e) Core Values and Social Structures of Spanish culture



(f) Core Values and Social Structures of Chinese culture

Figure 7: Performance of Models across Layers and Languages

tionalizability: The framework should effectively guide data collection and the practical construction of knowledge bases. **3) Interpretability**: It should enable clear attribution analysis of model behaviors, revealing preferences and performance variations across dimensions. **4) Extensibility**: The system should be designed for high scalability through data-driven expansion, supporting increasingly fine-grained evaluation in the future.

C.3 Detailed Dimensions

Our proposed dimension schema contains 4 levels, *Cultural Layers, Category, Topic Aspect and Detailed Dimensions*. There are in total 3 Cultural Layers, 5 Categories, 18 Topic Aspects and 140 Detailed Dimensions, encompassing a wide range of cultural knowledge from explicit commonsense facts, behavioral patterns to implicit values. We transfer the most detailed dimensions into keywords for Google search, thereby effectively guiding the construction and evaluation of the cultural knowledge base. The top *Cultural Layer* is the **Institutional Norms Layer**, which includes explicit commonsense knowledge such as geography, customs, rules, and policies within specific cultural contexts; the middle layer is **Behavioral Patterns Layer**, which encompasses personal choices and habits, including aspects such as daily life and travel, education, and concepts of knowledge; the bottom layer is the **Core Values and Social Structures Layer**, which includes deep cultural knowledge such as social relationships and structures, values, and beliefs. Each *Cultural Layer* include a series of *Categories, Topic Aspects and Detailed Dimensions.* A detailed description of the dimensions is provided in Table 7.

D Data Instances

This section provides representative samples from the constructed dataset, covering both cultural knowledge instances and question examples in evaluation dataset.

D.1 Cultural Knowledge

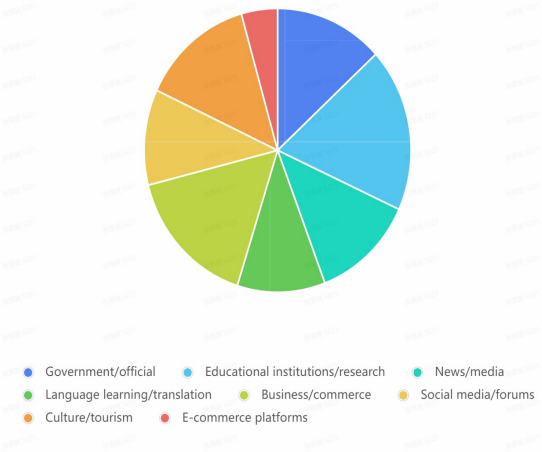
The extracted cultural knowledge instances are organized according to the evaluation dimensions. Each cultural knowledge instance provides a characteristic description of the corresponding cultural dimension as manifested in a specific cultural context, encompassing relevant commonsense knowledge, shared 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 6.

D.2 Evaluation Dataset

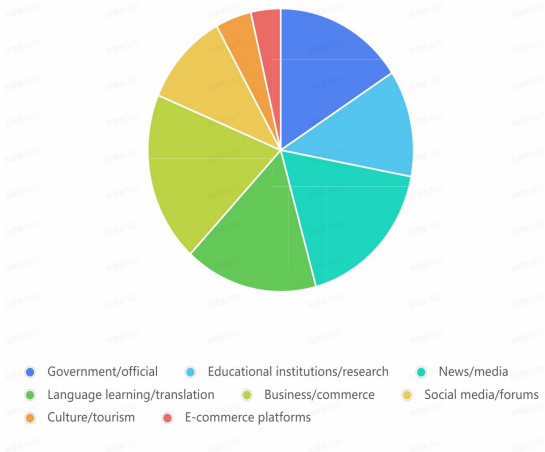
As is mentioned in the main context, the questions in our evaluation dataset are categorized into four types based on their content: Factual, Conceptual, Misleading, and Multi-hop reasoning. Each instance in the evaluation set consists of a question, its corresponding answer, and the associated cultural knowledge. The questions are categorized into subjective types (short answer and essay questions) and objective types (single-choice, multiple-choice, and true/false questions). Here, we present the examples organized by their content types in Table 8.

E Prompts

During the dataset construction process, we leveraged large language models (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. These prompts are language-specific and correspond to the language of the target data. Here, we present the prompts used across different languages.

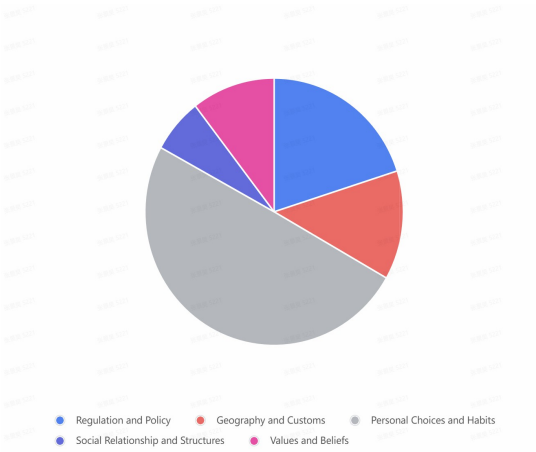


(a) Spanish Culture

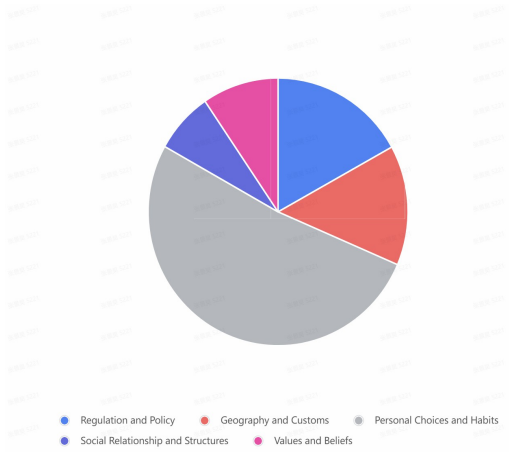


(b) Chinese Culture

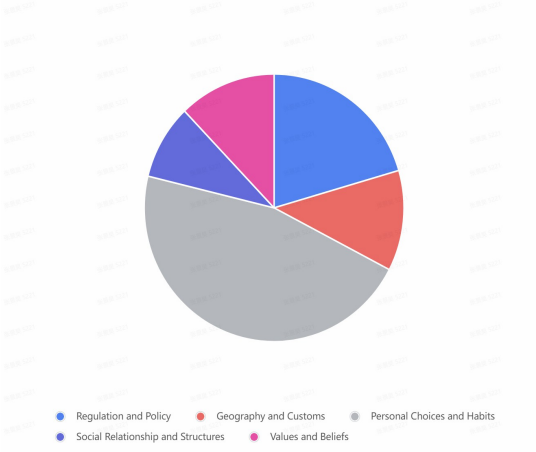
Figure 8: Data Source Statistics for Different Extracted Cultural Knowledge Instances



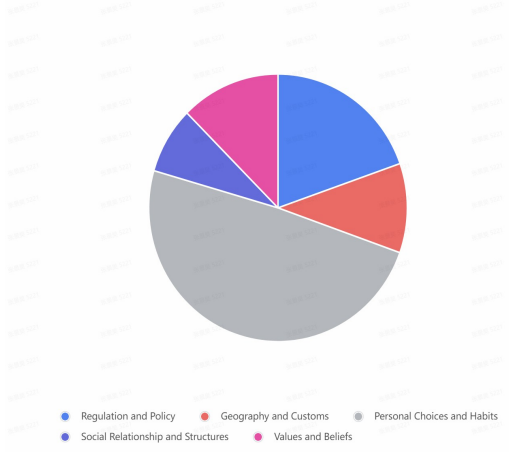
(a) Spanish Cultural Knowledge in English



(b) Spanish Cultural Knowledge in Spanish



(c) Chinese Cultural Knowledge in English



(d) Chinese Cultural Knowledge in Chinese

Figure 9: The distribution of knowledge instances retrieved through searches in different languages.

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

Table 6: 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.

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

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

Layer	Category	Topic Aspect	Detailed Dimension
			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
Behavioral Patterns	Personal Choices & Habits	Etiquette and Courtesy	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

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 7: The complete dimensional schema of CultureScope

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	Use of Symbolic Elements: Symbolic elements such as fire, water, and specific foods play crucial roles in many Spanish festivals. Fire is especially prominent in celebrations like Las Fallas and San Juan's Night.
Conceptual	In a Spanish business context, you notice that after a formal meeting, colleagues often suggest coffee or drinks. What is the primary purpose? A) Finalize decisions B) Build personal trust C) Discuss new opportunities D) Reduce stress	B) To build personal relationships and trust outside of the formal setting	Informal gatherings (after-work events, coffee breaks) build trust and relationships outside formal settings. Personal bonds are often foundational before formal business discussions in Spain.
Mislead	You are a foreign exchange student in Spain, and you notice that in your host family, the mother is a full-time working professional while the father stays at home to take care of the children. You also observe that both the son and daughter share household chores equally. Your friend from back home, who has never been to Spain, makes the following comment: Ẅow, that's so unusual! I always thought Spanish women were expected to stay at home and do all the housework, and the men didn't help much. Also, I heard that in Hispanic households, daughters are supposed to do more chores than sons. Question: Based on your observations and the context provided, which of the following statements best reflects a critical understanding of Spanish culture? A) My friend's comment is accurate; Spanish women are traditionally expected to stay at home and do all the housework, and daughters do more chores than sons. B) My friend's comment reflects a common stereotype, but it does not accurately represent the reality of modern Spanish families, where roles are more flexible and shared. C) My friend's comment is partially correct; while Spanish women traditionally did more housework, this is no longer the case, and now only daughters do more chores than sons. D) My friend's comment is entirely incorrect; Spanish women have always worked outside the home, and there is no expectation for daughters to do more chores than sons.	B) My friend's comment reflects a common stereotype, but it does not accurately represent the reality of modern Spanish families, where roles are more flexible and shared.	Increasing numbers of Spanish women now pursue careers while balancing family duties. In Hispanic households, daughters are expected to help with chores, while sons participate less. In Spanish culture, women traditionally take primary responsibility for household chores and child-rearing.

Type	Question	Answer	Associated Knowledge
Multi-hop	Imagine you are a Spanish student studying in Madrid. You have been invited to a family dinner at your classmate's home. The dinner is scheduled for the 15th of the month, and your classmate mentions that this particular month has a special significance in Spanish culture. During the dinner, you notice that the family shares stories and memories related to the month, and there is a lot of laughter and bonding over the meal. 1. Contextual Understanding: - What could be the possible month of the dinner, given its cultural significance? - How does the name of this month relate to its historical or mythological origins? 2. Cultural Practices: - Why do you think shared meals are so important in Spanish culture? - How does the practice of sharing stories during meals contribute to the family bond? 3. Language and Grammar: - If you were to write a note to your classmate thanking them for the dinner, how would you correctly abbreviate the name of the month in Spanish? - What grammatical rules should you follow when writing the name of the month in a sentence?	1. Contextual Understanding: - The month of the dinner is likely to be July (Julio) or August (Agosto), as these months are named after significant historical figures (Julius Caesar and Augustus, respectively). Both Julius Caesar and Augustus were influential Roman leaders, and their names are still celebrated in Spanish culture. - Julio is named after Julius Caesar, who reformed the Roman calendar, while Agosto is named after Augustus, who was the first Roman emperor and continued the legacy of his adoptive father, Julius Caesar. 2. Cultural Practices: - Shared meals in Spanish culture are deeply rooted in the tradition of strengthening family ties and fostering a sense of community. These meals are not just about eating but also about spending quality time together, sharing experiences, and creating lasting memories. - The practice of storytelling during meals is a way to pass down family history, values, and traditions from one generation to the next. It helps to create a strong sense of identity and belonging within the family, making the meal a more meaningful and bonding experience. 3. Language and Grammar: - To correctly abbreviate the name of the month in Spanish, you would use j̈ul.ïor julio and ägo.ïor agosto. - When writing the name of the month in a sentence, it should be in lowercase unless it starts the sentence or is part of a proper noun. For example, Ël 15 de julio fuimos a una cena en casa de mi compañero.̈(On July 15th, we went to a dinner at my classmate's house.)	Cultural significance of months: Julio (Julius Caesar), Agosto (Augustus); Sept–Dec = Latin numbers. Food as bonding ritual in Spanish culture. Grammar: months in lowercase unless at sentence start; masculine nouns; accepted abbreviations like "en." for enero.

Table 8: 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.

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 behind 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*
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:

Knowledge Injection(English)

Reference:
reference knowledge
Question:
question