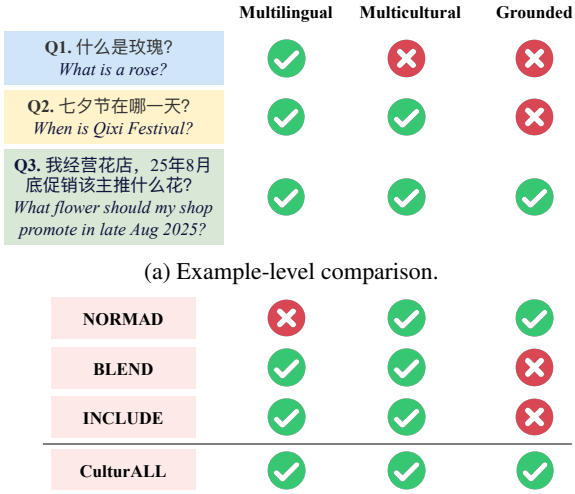
CulturALL: Benchmarking Multilingual and Multicultural Competence of LLMs on Grounded Tasks

Peiqin Lin[1], **Chenyang Lyu**[1], **Wenjiang Luo**[2], **Haotian Ye**[3], **Md Mehrab Hossain**[5], **Chunlan Ma**[3], **Shaoxiong Ji**[4,5], **Younes Samih**[6], **Bo Zeng**[1], **Fan Jiang**[1], **Yuanbin Cao**[1], **Dilda Duisenbek**[2], **Adrian Neo Sau Xun**[2], **Daria Pozdniakova**[2], **Liubou Misevich**[2], **Nevena Marinković**[2], **Ngoc Gia Linh Nguyen**[2], **Thi Khanh Linh Do**[2], **Sarakmatak Sophy**[2], **Baotian Hu**[8], **Guanhua Chen**[9], **Gongbo Tang**[2], **Alham Fikri Aji**[7], **Longyue Wang**[1], **Weihua Luo**[1]

[1]Alibaba Group [2]Beijing Language and Culture University [3]LMU Munich
[4]ELLIS Institute Finland [5]University of Turku [6]IBM Research AI, UAE [7]MBZUAI
[8]Harbin Institute of Technology, Shenzhen [9]Southern University of Science and Technology

Abstract

Large language models (LLMs) are now deployed worldwide, inspiring a surge of benchmarks that measure their multilingual and multicultural abilities. However, these benchmarks prioritize generic language understanding or superficial cultural trivia, leaving the evaluation of grounded tasks—where models must reason within real-world, context-rich scenarios—largely unaddressed. To fill this gap, we present CulturALL, a comprehensive and challenging benchmark to assess LLMs' multilingual and multicultural competence on grounded tasks. CulturALL is built via a human–AI collaborative framework: expert annotators ensure appropriate difficulty and factual accuracy, while LLMs lighten the manual workload. By incorporating diverse sources, CulturALL ensures comprehensive scenario coverage. Each item is carefully designed to present a high level of difficulty, making CulturALL challenging. CulturALL contains 2,610 samples in 14 languages of 51 regions, distributed across 16 topics to capture the full breadth of grounded tasks. Experiments show the best LLM achieves only 44.48% accuracy, leaving substantial room for improvement.[1]



(a) Example-level comparison.

(b) Benchmark-level comparison.

Figure 1: (a) Example-level: Q1 is multilingual only; Q2 adds cultural knowledge; Q3 requires all three, posing the hardest challenge. (b) Benchmark-level: existing representative benchmarks test at most two axes, while CulturALL spans all three.

1 Introduction

As LLMs are adopted across the globe, it is imperative to evaluate how well they perform in diverse languages and cultures. Existing multilingual and multicultural benchmarks, e.g., BLEND (Myung et al., 2024), INCLUDE (Romanou et al., 2024), and Global MMLU (Singh et al., 2025), cover a wide range of languages and cultures, but their content is dominated by encyclopedic trivia. Consequently, they say little about how LLMs perform on the everyday tasks people actually care about, e.g., planning a trip or making an online purchase. Recent efforts have started to introduce grounded evaluations, e.g., CultureBank (Shi et al., 2024) and NORMAD (Rao et al., 2025). However, they are mostly in English and cover only a narrow band of grounded tasks, mainly social interactions. This gap prompts a key question: *How effectively can LLMs tackle culturally grounded tasks across different languages and cultures?*

A truly capable LLM must solve grounded tasks across diverse linguistic and cultural contexts, because these tasks reflect what users actually need. These tasks are particularly challenging because they probe three complementary capacities of an LLM: **(1) language comprehension (multilingual)**: the capacity to accurately parse and interpret a user's native tongue; **(2) cultural knowledge acquisition (multicultural)**: the ability to access and recall long-tail, domain-specific cultural facts; and

[1]Code and data are publicly available at https://github.com/AIDC-AI/Marco-LLM.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4208–4224
November 6-10, 2026 ©2026 Association for Computational Linguistics

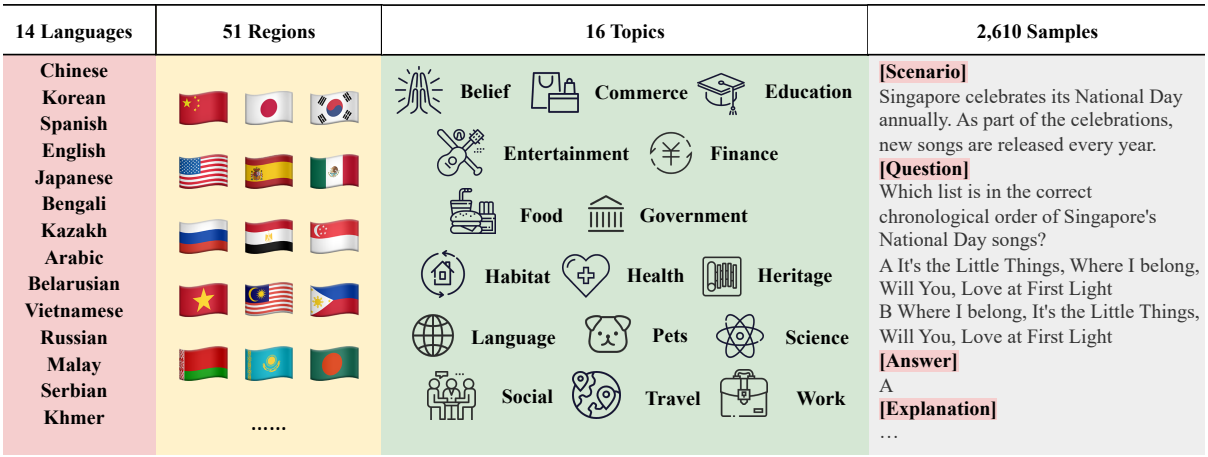


Figure 2: CulturALL is a comprehensive and challenging benchmark. It contains 2,610 samples in 14 languages across 51 regions, distributed among 16 topics to capture the full breadth of grounded tasks. As the given example illustrates, each item presents a grounded scenario followed by its question. Successfully solving each item requires an LLM to fuse these cues with its stored knowledge and reason to the correct answer.

(3) contextual reasoning (grounded): the skill of integrating that information and synthesizing it into an accurate response. As illustrated in Fig. 1a, Q1 merely tests a LLM's multilingual ability, and Q2 adds a cultural fact. In contrast, Q3 requires the full chain of multilingual, multicultural, and grounded reasoning: the LLM must first interpret the Chinese query, identify the relevant late-August festival in China and its customs, recall the symbolic meanings of different flowers, and finally synthesize this information into a concise, culturally appropriate reply via reasoning. Coordinating this chain of culturally grounded reasoning is anything but trivial. CulturALL operationalizes this evaluation through objective-answer formats (multiple-choice and short-form answers), while drawing its scenarios from real-world cultural situations.

In response, we introduce CulturALL, the first benchmark to assess LLM performance in grounded scenarios across diverse languages and cultures (Fig. 1b). CulturALL is constructed using a novel human-LLM collaborative framework that leverages expert annotators for factual accuracy and elevated difficulty, while LLMs assist in generating and enriching diverse scenarios, ensuring comprehensive coverage and challenging samples. Fig. 2 shows CulturALL's extensive language and cultural coverage, as well as its challenging nature. The characteristics of CulturALL are as follows: 1) coverage: the 2,610 samples in CulturALL span 16 topics that encompass diverse facets of daily life and society, covering cultures from 51 regions across 14 languages; 2) challenging: answering each scenario-based question is difficult because it requires LLMs to integrate nuanced cultural knowledge with strong reasoning skills.

Using CulturALL, we find that existing LLMs struggle with culturally grounded tasks, and improving their performance requires effective web search and strong reasoning capabilities. In summary, our contributions are multifold.

- We design a unified human-LLM framework, which can be applied to create benchmarks with wide coverage and high difficulty.
- We present CulturALL, the first benchmark explicitly designed to assess LLMs' multilingual and multicultural competence across a wide spectrum of realistic tasks.
- We benchmark state-of-the-art LLMs on CulturALL and deliver an in-depth analysis, highlighting key strengths and failure modes.

2 CulturALL: Construction and Statistics

Robust evaluation of LLMs on multilingual and multicultural tasks requires datasets that are both diverse and challenging. To achieve this at scale, we introduce a unified human-LLM framework that combines human expertise with the generative power of LLMs, resulting in CulturALL, which offers broad coverage and high difficulty.

An overview of this framework is shown in Fig. 3. The framework begins with cultural topic sourcing (§2.1), compiling an extensive list of

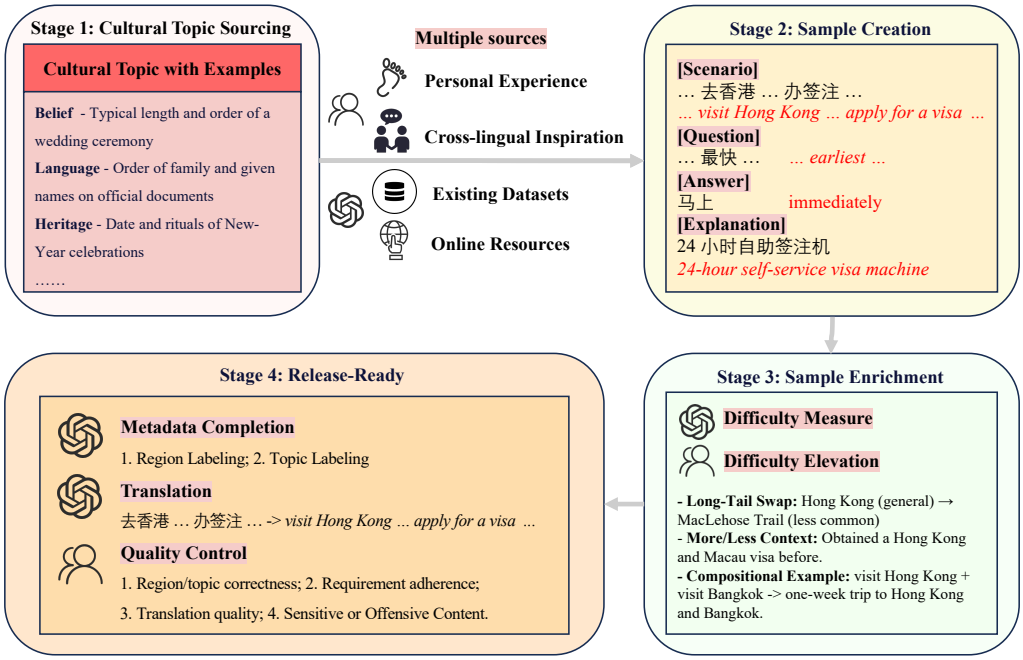


Figure 3: The data construction framework of CulturALL: 1) Cultural Topic Sourcing: assemble a list of cultural topics; 2) Sample Creation: craft original items for each topic; 3) Sample Enrichment: enhance realism and increase difficulty; 4) Release-Ready: complete sample information and conduct quality validation.

cultural topics and illustrative examples. Next is sample creation (§2.2), where we draft seed instances for these topics, drawing on sources such as personal experience and online materials. These drafts are then refined during sample enrichment (§2.3) to increase their difficulty and better mirror grounded scenarios. The final stage—Release-Ready (§2.4)—completes each sample with topic/region labels and English translation, and then conducts thorough quality checks.

We emphasise that *LLMs are never used as knowledge sources* in our framework. All factual grounding derives from human annotators' first-hand cultural expertise, curated existing benchmarks, and human-verified online resources. LLMs serve solely as productivity tools for mechanical subtasks, while human annotators perform every step where outputs may be unreliable—including authoring original scenarios, verifying topics, elevating difficulty, and conducting final quality control. A detailed breakdown is in §G.

We define a culture group as the population of a single country or region. To capture broad cultural expertise, we collaborate with annotators from a wide range of countries, regions, and linguistic backgrounds. Real-world queries seldom state their cultural origin explicitly, so LLMs must infer it from implicit cues—vocabulary, idioms, institutions, and other context signals. For this reason, each annotator composes samples in the dominant language of their locale—e.g., English in the United States and Mandarin Chinese in mainland China—embedding authentic local references that models must recognize and interpret. The annotation guidelines and annotator profiles are in §A.

2.1 Stage 1: Cultural Topic Sourcing

To spur the creation of grounded tasks across a broad spectrum of cultural topics, the cultural-topic-sourcing stage aims to generate a comprehensive list that covers nearly every facet of daily life through human–LLM collaboration. We first compile a preliminary set of topics with concise scope descriptions, drawing on prior research (Yin et al., 2022; Romanou et al., 2024; Chiu et al., 2025) and heuristics, and then engage *gpt-4o-2024-11-20* in several iterative rounds to merge, refine, and expand both the topics and their accompanying descriptions. With the final list in place, we craft seed examples from personal experience and instruct *gpt-4o-2024-11-20* to expand them, yielding a pool of 160 language-agnostic illustrative instances (10 per topic) that serve as scaffolding for the subsequent sample-creation stage. The topic list accompanied by descriptions and three representative culture-related scenarios is in Tab. 4 (§B).

Field	Description
language	ISO-639-1 code, e.g. en, zh
region	ISO-3166-1 alpha-2 code, e.g. US, CN
topic	One of the 16 predefined topics
scenario	Narrative context in which the question is asked
question	The question itself
answer	Correct answer or answer key
explanation	Concise justification of the answer

Table 1: Metadata for each item.

2.2 Stage 2: Sample Creation

2.2.1 Sample Format

Tab. 1 outlines the schema that each sample must adhere to. During annotation, the language field is predefined based on the source data or annotator's information, while region and topic are automatically generated by the LLM (see §2.4.1). All remaining fields are reviewed and completed by the annotators. Below, we detail the requirements for the overall sample and each field that must be verified or completed by annotators.

Sample Craft a culturally grounded item that evaluates an LLM's ability to employ cultural knowledge. Cultural knowledge includes but not limited to local vocabulary, social norms, cultural commonsense, regulations, and domain-specific knowledge. Generic trivia (e.g., math puzzles or textbook facts) is out of scope. Two items are considered distinct only if they probe different knowledge or reasoning steps, not if they are merely paraphrases of each other. Every sample is designed to have a single objective, verifiable answer (either a discrete choice or a short-form answer), ensuring unambiguous evaluation.

Scenario Construct a grounded scenario, withholding any explicit hints that would let a model solve the task without relevant cultural knowledge.

Question Ensure the query arises from the scenario and cannot be answered correctly without an understanding of the relevant cultural knowledge.

Answer To facilitate automatic evaluation, answers should be objective and as brief as possible. If an objective free-form answer is impractical, convert the question to a four-option multiple-choice format (A–D) and return only the chosen letter.

Explanation When appropriate, supply the cultural or domain knowledge that supports the answer. These explanations make CulturALL more transparent for readers and pave the way for using CulturALL in future free-text evaluation tasks.

2.2.2 Cultural Knowledge Sourcing

Personal Experience (Human) To capture unwritten social cues, emerging slang, and region-specific practices, we ask annotators to draw from their personal experiences. These first-hand contributions result in scenarios that are both authentic and deeply rooted in context. Annotators receive a detailed list of topics with descriptions and examples (§2.1), which they can adapt, use as inspiration for new culturally relevant instances, or supplement with ideas from local forums.

Cross-lingual Inspiration (Human) An example in one language often sparks analogous ideas in annotators who speak other languages. For instance, a Chinese query about obtaining a visa for Hong Kong may inspire a French annotator to create a comparable scenario involving a French employee applying for a Belgian work permit. To facilitate this transfer, we translate existing samples into English (see §2.4.2 for details), serving as a shared pivot that enables native speakers of other languages to more easily create parallel data.

Existing Datasets (LLM) Many prior cultural benchmarks contain culture-relevant items yet lack explicit grounding in grounded contexts. We refine these items through rewriting using *gpt-4o-2024-11-20*, anchoring each item in a concrete scenario while preserving its original knowledge requirements. Details are provided in §C.

Online Resources (LLM) We collect culture-rich materials from online resources, focusing primarily on mining posts from Xiaohongshu, guided by our cultural topic example list. For each target country/region (Tab. 5, §C), we combine the region name with topic seeds generated during the cultural-topic–sourcing stage (in Chinese) as search queries to efficiently surface relevant local content. This crawl returns 3,518 pages. Each page is translated into the country/region's dominant language with `gpt-4o-2024-11-20` (Fig. 6, §D). The retrieved content is then supplied to *gpt-4o-2024-11-20* as raw input for drafting candidate items (Fig. 8, §D).

2.3 Stage 3: Sample Enrichment

To ensure CulturALL better reflects challenging grounded demands, every draft item undergoes a process of "up-leveling" as illustrated in Fig. 3. Specifically, we assess the difficulty of the original samples and categorize them into hard and easy examples. Hard examples are forwarded to Stage 4,

while easy examples undergo a difficulty elevation process to increase its difficulty if possible.

2.3.1 Difficulty Measure (LLM)

Inspired by Phan et al. (2025); Fabbri et al. (2025), we utilize three LLMs—*gpt-4o-2024-11-20*, *claude-3.5-sonnet-1022*, and *qwen-max-2024-09-19*—to quantify the difficulty of items produced by Stage 2. An item is classified as challenging if at most one of the three LLMs provides the correct answer. Such items are forwarded to Stage 4. Otherwise, they proceed to the difficulty elevation process for further complexity enhancement.

2.3.2 Difficulty Elevation (Human)

We introduce three complementary enrichment strategies to guide human annotators in elevating the difficulty of existing samples:

Long-Tail Swap Common entities are replaced with rarer ones, e.g., substituting the general location "Hong Kong" with "MacLehose Trail," a lesser-known hiking route within the region.

More/Less Context Additional situational details are introduced, requiring the answer to hinge on conditional, multi-step reasoning (e.g., determining if a traveler has a prior visa). Conversely, unnecessary context that provides hints to LLMs can be removed to increase the challenge.

Compositional Example Two independent knowledge points are combined into a single query—for example, merging the entry requirements for both Hong Kong and Bangkok—forcing the model to engage in compositional reasoning.

Annotators are encouraged to apply one or more of these techniques to enhance sample difficulty. If no opportunities for improvement, annotators may leave it unchanged. These refinements amplify the task complexity, moving beyond superficial matching and compelling models to demonstrate deeper understanding and advanced reasoning capabilities.

2.4 Stage 4: Release-Ready

2.4.1 Metadata Completion (LLM)

The Metadata Completion step utilizes *gpt-4o-2024-11-20* to fill in the `region` and `topic` fields. For `region`, the prompt shown in Fig. 9 (§D) is used to generate the corresponding ISO 3166-1 alpha-2 code. For the `topic`, we prompt the LLM to select the most suitable topic from a predefined list (§2.1) based on the created sample, using the prompts provided in Fig. 10 (§D).

2.4.2 Translation (LLM)

In this sub-step, *gpt-4o-2024-11-20* translates each sample into English, using the translation prompt in Fig. 6 (§D). This process provides a unified reference language for readers while also inspiring annotators to develop new cross-regional samples.

2.4.3 Quality Control (Human)

Data quality is maintained via a peer-review process wherein each annotator cross-checks samples created by others/LLMs. During the review, annotators are asked with checking the following aspects:

Region/Topic Correctness The assigned region should be a valid ISO 3166-1 alpha-2 code and the topic belongs to the predefined list. Both must accurately align with the content of the sample.

Requirement Adherence Each sample must comply with the requirements outlined in §2.2.1.

Translation Quality LLM's translation output should be correctly align with the source input.

Sensitive or Offensive Content The output should not include personally identifiable information or harmful, offensive, or inappropriate content.

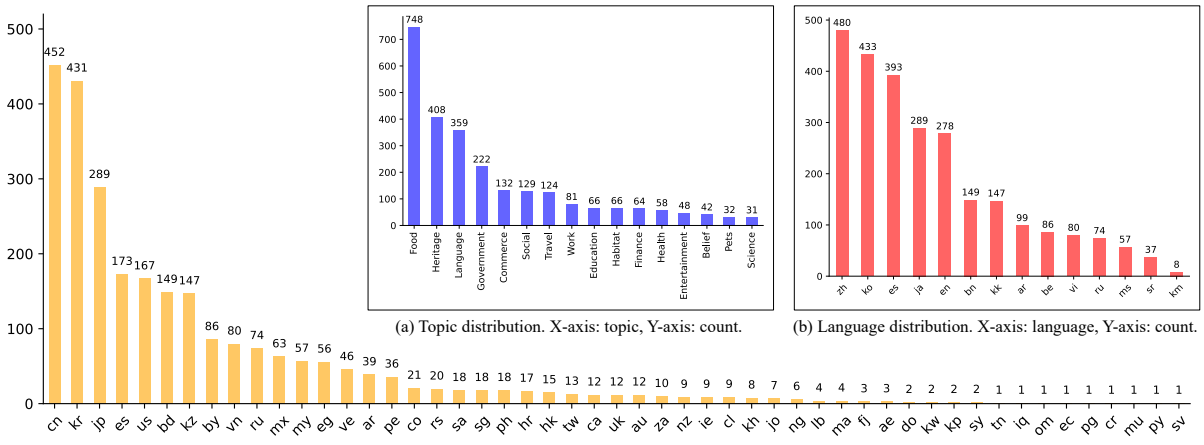
Based on these criteria, annotators have three options: **accept**, **revise**, or **reject**. A sample should be marked as **accept** if it meets all four criteria. If issues are identified, the annotator should attempt to **revise** the sample to ensure it fully satisfies the requirements. However, if the sample cannot be revised to meet the criteria, it should be marked as **reject**. Based on these criteria, 75.0% of the samples were marked as **accept**, 8.1% were marked as **revise**, and 16.9% were marked as **reject**.

To measure annotation reliability, every sample created by one annotator was reviewed by a different annotator (Task D, §A.2.4), providing full intra-language coverage. In addition, we randomly selected 100 samples for a second independent quality check; all 100 samples were judged consistent with the original decisions, confirming the overall reliability of the annotation process.

2.5 Statistics

The resulting CulturALL comprises 2,610 samples, spanning 14 languages and 51 regions. The distributions of samples across topics, languages, and regions are illustrated in Fig. 4.

To ensure that CulturALL remains genuinely challenging, it excludes any item that is correctly solved by all model settings (§3). We categorize



(a) Topic distribution. X-axis: topic, Y-axis: count.

(b) Language distribution. X-axis: language, Y-axis: count.

(c) Region distribution. X-axis: region, Y-axis: count.

Figure 4: Distributions across topics, languages, and regions. The first row includes: (a) topic distribution and (b) language distribution, and the second row shows (c) region distribution.

the 2,610 examples in CulturALL based on how many of the 15 settings answered them correctly. Items solved by 10–14 settings are labeled as **Easy** (470 items, 18.01%), while those solved by 5–9 settings form the **Medium** subset (700 items, 26.82%). Finally, items solved by at most four settings are grouped into the **Hard** subset (1,440 items, 55.17%). Fig. 13 in §E further illustrates the language distribution of examples according to the number of settings that answered them correctly.

3 Setup

Model Selection Tab. 2 presents the 15 experiments conducted to benchmark top-performing LLMs on CulturALL. Our analysis includes 8 leading LLMs from the Text Arena leaderboard as of 18 August 2025,[2] spanning both open-source and proprietary models. The experiments feature 15 distinct configurations, achieved by varying reasoning capabilities and the inclusion of web search. This comprehensive experimental setup facilitates a systematic evaluation of key factors, such as reasoning performance, web search integration, model size, and other critical characteristics.

Prompt Design To benchmark LLMs with CulturALL, we conduct zero-shot prompting. In addition, we also require the LLMs to answer the question in as few words as possible to ease the follow-up evaluation. The concrete prompt is provided in Fig. 11 of §D.

Metric Since all reference answers in CulturALL are strictly objective, automatic assessment can be applied. During evaluation, the judge LLM, *gpt-4o-2024-11-20*, is provided with the full item (scenario, question, gold answer, and optional explanation) alongside the prediction from the evaluated LLM. Leveraging the prompt provided in Fig. 12, the judge assesses whether the prediction aligns with the reference answer. Each item yields a binary outcome (correct or incorrect), and the overall performance is measured as accuracy, defined as the proportion of correctly judged items.

To validate the LLM-as-a-judge approach, we compared its judgments against traditional metrics including exact-match and substring-overlap on a 100-item sample. For multiple-choice questions, LLM-as-a-judge scores are largely consistent with direct accuracy (letter match). In the small number of divergent cases, LLM-as-a-judge typically performs better because it handles minor formatting differences (e.g., lowercase vs. uppercase, punctuation variants) more robustly.

4 Results

4.1 Performance Across LLMs

The performance of evaluated LLMs across various configurations on the complete CulturALL dataset is detailed in Tab. 2. The resulting scores support several key observations:

Best Setting Still Falls Short Among all experiments, gemini-2.5-pro_auto_true, which utilizes gemini-2.5-pro with its strongest reasoning capability and web search integration, achieves the highest accuracy at 44.48%. However, this performance remains far from ideal, indicating significant room

[2] https://lmarena.ai/leaderboard/text

ID	Experiment Name	Model Name	Open	Reasoning	Web	All	Hard	Med.	Easy
1	gemini-2.5-pro_auto_true	gemini-2.5-pro	No	auto	Yes	**44.48**	**18.47**	**65.71**	**92.55**
2	gemini-2.5-pro_auto_false	gemini-2.5-pro	No	auto	No	37.89	10.07	59.57	90.85
3	gemini-2.5-pro_128_true	gemini-2.5-pro	No	128 tokens	Yes	39.27	15.49	55.71	87.66
4	gemini-2.5-flash_auto_true	gemini-2.5-flash	No	auto	Yes	33.68	12.78	46.71	78.30
5	gpt-5-20250807_high_false	gpt-5	No	high	No	37.59	10.28	57.71	91.28
6	gpt-5-20250807_medium_false	gpt-5	No	medium	No	37.20	9.31	58.29	91.28
7	gpt-5-20250807_low_false	gpt-5	No	low	No	37.24	9.38	59.00	90.21
8	claude-opus-4-20250514_high_false	claude-opus-4	No	1024 tokens	No	36.70	9.44	56.00	91.49
9	claude-opus-4-20250514_low_false	claude-opus-4	No	disabled	No	36.48	9.03	56.43	90.85
10	claude-sonnet-4-20250514_high_false	claude-sonnet-4	No	1024 tokens	No	32.76	8.82	46.43	85.74
11	qwen-max_auto_true	qwen-max	No	hybrid	Yes	19.31	5.14	24.71	54.68
12	qwen-max_auto_false	qwen-max	No	hybrid	No	18.97	5.69	24.00	52.13
13	qwen3-235b-a22b_high_true	qwen3-235b-a22b	Yes	hybrid	Yes	22.49	4.24	31.43	65.11
14	qwen3-235b-a22b_high_false	qwen3-235b-a22b	Yes	hybrid	No	23.68	4.65	33.57	67.23
15	qwen3-30b-a3b_high_true	qwen3-30b-a3b	Yes	hybrid	Yes	17.62	4.72	23.86	47.87

Table 2: Performance of evaluated LLMs with diverse settings on the complete CulturALL dataset and its three subsets, categorized by difficulty level. All results are reported as accuracy (%). Reasoning: reasoning capability. Web: the use of web search. Open: open-source availability. Experiment Name: Model Name_Open_Web.

for LLM improvement in handling the challenging scenarios presented by CulturALL.

Open-Source LLMs Lag Behind As shown in the results, gemini-2.5-pro (ID 2), gpt-5 (ID 5), and claude-opus-4 (ID 8) achieve accuracy rates of 37.89%, 37.59%, and 36.70%, respectively, demonstrating comparable performance among proprietary models. In contrast, the open-source qwen series shows a significant performance gap, with its top-performing setting, qwen3-235b-a22b_high_false (ID 14), achieving only 23.68%. This disparity highlights the challenges faced by open-source models in addressing multilingual and culturally grounded tasks, particularly when competing with advanced proprietary alternatives.

Higher Variants Consistently Outperform Their Counterparts The performance comparisons align with the publicly stated performance hierarchies from model providers (e.g., Gemini-2.5-Pro is positioned above Gemini-2.5-Flash; Claude Opus is positioned above Claude Sonnet). Gemini-2.5-pro (ID 1) outperforms gemini-2.5-flash (ID 4) by 10.80%, supporting its position as the more powerful variant within the Gemini series. Similarly, claude-opus-4 (ID 8) achieves a 3.94% higher accuracy than claude-sonnet-4 (ID 10), reinforcing its superior performance.

Reasoning Effort Affects LLMs Unevenly Reasoning capabilities play a crucial role in determining the performance of LLMs, but the impact varies across models. For example, gemini-2.5-pro with the most advanced reasoning setting of "auto" (ID 1) outperforms the same model configured with minimal reasoning (128 tokens, ID 3) by 5.21%, highlighting the positive effect of enhanced reasoning efforts. However, increasing reasoning capabilities for gpt-5 (ID 5 vs. 6 vs. 7) and claude-opus-4 (ID 8 vs. 9) fails to raise their scores by a considerable margin. This discrepancy could stem from the fact that these LLMs lack sufficiently robust multi-step reasoning abilities to navigate complex cultural scenarios effectively.

Web Search Plays a Critical Role Equipping gemini-2.5-pro with a web-search tool raises its score by 6.59% (ID 1 vs. 2), demonstrating the benefit of external retrieval. In contrast, the qwen models gain little advantage from the web search tool (ID 11 vs. 12, ID 13 vs. 14), suggesting they are not yet sufficiently trained to leverage web-search results for solving grounded tasks.

4.2 Performance Across Difficulty Levels

As shown in Tab. 2, the performance of LLMs across the three subsets highlights several key patterns. First, the relative ranking of different experimental settings remains consistent across Easy, Medium, and Hard tasks, indicating stable differences in their capabilities. Second, the performance gap between commercial and open-source models becomes more pronounced as tasks grow easier: on the Easy subset, gemini-2.5-pro_auto_true (ID 1) achieves an impressive accuracy of 92.55%, while the best-performing open-source model (ID 14) achieves only 67.23%, trailing by a substantial margin. Finally, all systems continue to struggle with

the Hard subset; even the top-performing setting (ID 1) achieves 18.47%, highlighting the need for further advancements.

4.3 Performance Across Languages

We further investigate the performance of different experimental settings across languages, as depicted in the heatmap in Fig. 5. Two important caveats apply when interpreting this figure. First, because CulturALL is *not* a parallel benchmark—each language subset tests culturally specific knowledge rather than translations of the same questions—scores across languages **should not be interpreted as direct comparisons of model competence**; they instead reflect the difficulty of the language-specific items. Second, we verify robustness by excluding languages with fewer than 100 samples and recomputing all accuracies; results (§F) confirm that overall conclusions are consistent.

With these caveats in mind, we observe that certain languages yield comparatively higher scores (e.g., Serbian), which likely reflects that the items collected for those languages are less challenging for current LLMs, rather than indicating superior model capability in those languages. This interpretation is further supported by the fact that the easy/medium/hard distribution of items differs substantially across languages (see Fig. 13 in Appendix E), which influences the aggregate scores reported here. Conversely, languages such as Japanese and Arabic consistently receive lower scores across all settings. Notably, the performance rankings of languages vary significantly across different experimental settings. For example, gemini-2.5-flash_auto_true achieves an accuracy of 69.80% in Bengali, demonstrating competitive results against state-of-the-art settings. However, in other languages, e.g., Arabic and Chinese, its performance lags behind alternative configurations.

4.4 English vs. In-Language Prompts

Previous work (Yin et al., 2022) revealed that samples written in English tend to achieve superior performance compared to those written in native languages, as LLMs typically exhibit greater proficiency in English. In our study, we evaluated the performance using native prompts and their English translations under the experimental setting of ID 1. The results show that employing English prompts yields an accuracy of 36.40%, which is 8.08% lower than the accuracy achieved with the original native prompts. We conducted a case study to identify the causes of this drop; two main factors emerged. First, when culture-specific information (e.g., a region name or local institution) is absent from the English context (see Q3 in Fig. 1), translation prevents the LLM from accessing the culturally grounded knowledge needed for a correct answer. Second, even when such information is present in the translated text, translating it into English biases the model toward an English-centric perspective, leading it to retrieve inappropriate world knowledge and produce wrong answers. We note that translation quality itself could also be a contributing factor; however, we used a strong translation model (*gpt-4o-2024-11-20*) and verified output quality during Stage 4. The consistent performance drop across languages suggests that the primary cause is the loss of culturally implicit cues rather than translation errors per se.

These findings highlight the urgent need to enhance LLMs' capabilities to adapt to diverse cultural and linguistic contexts.

5 Related Work

A wide range of benchmarks have been have been introduced to assess the general capabilities of LLMs (Hendrycks et al., 2021; Wang et al., 2024b). As LLMs become ubiquitous across the world, researchers are paying growing attention to their performance across diverse languages (Wu et al., 2025) and cultures (Hershcovich et al., 2022; Adilazuarda et al., 2024; Pawar et al., 2024; Liu et al., 2025).

Early efforts to assess LLMs under multilingual and multicultural settings have curated knowledge bases and probing tasks to test the LLMs' capacity on culture-specific knowledge acquisition (Yin et al., 2022; Wang et al., 2024a; Myung et al., 2024; Zhou et al., 2025; Hasan et al., 2025; Arora et al., 2025). Findings from these studies show that LLMs still display pronounced cultural biases and uneven performance across different regions of the world (Naous et al., 2024; Mitchell et al., 2025). While these benchmarks shed light on what LLMs know about diverse cultures in different languages, they do not fully assess multilingual and multicultural competence, which requires LLMs not only to store cultural knowledge but also to apply it flexibly in grounded scenarios (Rao et al., 2025).

To obtain a clearer picture of LLMs' multilingual and multicultural competence, recent benchmarks have shifted from decontextualized trivia to grounded scenarios, covering social interaction

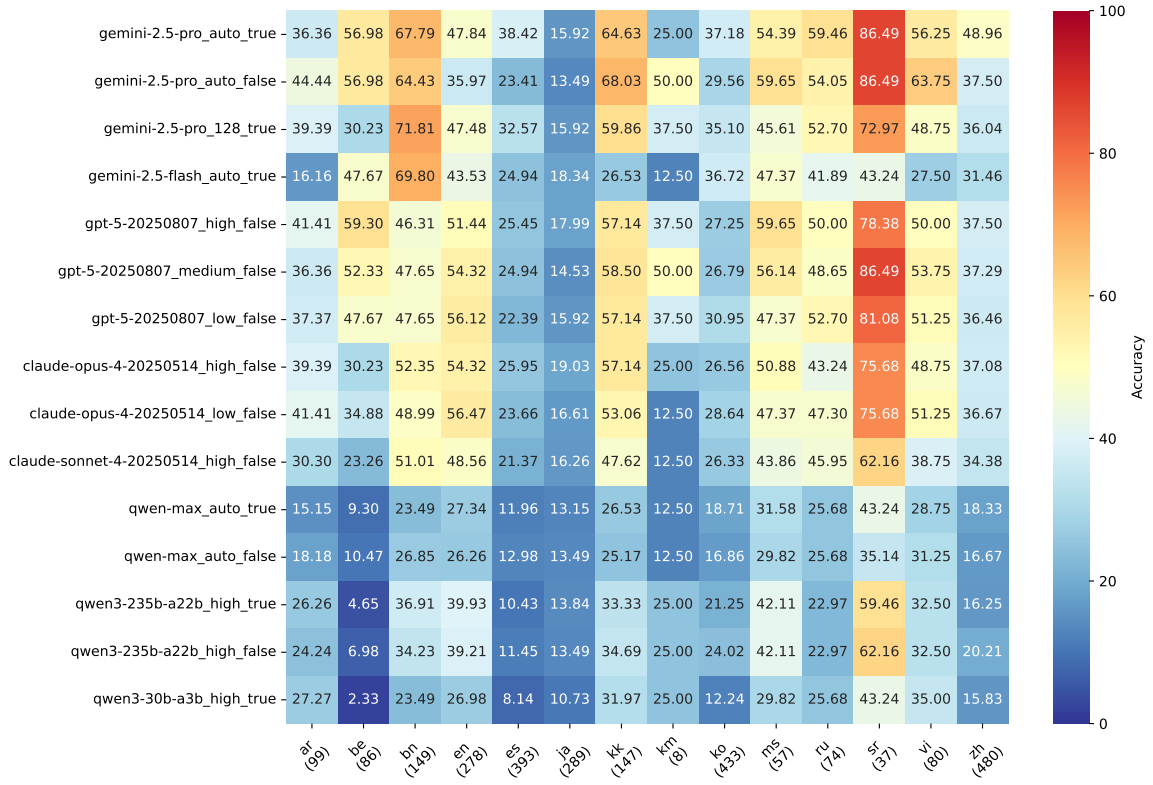


Figure 5: Performance of various experimental settings across 14 languages. X-axis: languages (along with their sample counts), Y-axis: different experimental settings.

(Rao et al., 2025; Yin et al., 2024; Qiu et al., 2025), psychology (Jin et al., 2025), and cultural proverbs (Liu et al., 2024). However, they are usually monolingual or confined to a narrow domain, leaving them a long way from the breadth and diversity of situations that arise in grounded applications.

6 Conclusion

In this paper, we introduce CulturALL, the first benchmark designed to evaluate the multilingual and multicultural capabilities of LLMs across grounded tasks. CulturALL is built using a human-LLM collaboration framework, ensuring both comprehensiveness and a high level of challenge. Through an in-depth analysis of LLMs on CulturALL, we highlight the critical need to enhance their information retrieval and reasoning skills.

Limitations

The scope and design of this study inevitably come with certain limitations, which we outline in this section. Addressing these limitations in future work can help provide a more nuanced and comprehensive understanding of LLMs' performance across cultural contexts.

Coverage Bias As illustrated in Fig. 4, certain topics, languages, and regions are underrepresented in the dataset. Our proposed unified data construction framework provides a practical foundation for expanding CulturALL to improve coverage in the future.

Focus on Regional Cultural Groupings In this work, we define cultural groups predominantly based on geographic regions. While this approach provides a high-level understanding of general cultural differences, it does not fully capture more nuanced or cross-cutting aspects of culture. Factors such as religion, age, socio-economic status, gender, and education significantly influence cultural perspectives and may intersect in ways that are not represented by regional groupings alone. Future studies should explore these more fine-grained cultural dimensions to offer a holistic assessment of LLMs' cultural-grounded reasoning capabilities.

Reliance on Objective Answer Evaluation To ensure consistency and reproducibility in our experimental setup, we focus exclusively on tasks with objective, verifiable answers to enable automatic evaluation. While these tasks serve as a robust

benchmark for model performance, they do not account for the complexities of free-text generation, which is a key feature of LLMs in multilingual and culturally nuanced applications. Investigating free-text generation compared to objective reasoning tasks is an important avenue for future exploration to better understand LLMs' ability to engage with subjective, open-ended questions influenced by cultural relativism.

Exclusion of Multimodal Inputs Our study focuses entirely on text-based inputs without considering multimodal contexts, such as the integration of visual, auditory, or other non-textual signals. However, cultural understanding often extends beyond textual communication to include visual symbolism, nonverbal gestures, and audio cues, all of which hold significant meaning in cultural interactions. Future research should explore the impact of multimodality on LLM performance when tackling culturally grounded tasks to better model the complexities of human communication.

Ethical Considerations

Annotation Process and Annotator Profile Prior to starting the annotation process, annotators undergo a comprehensive briefing on the guidelines detailed in §A.2. Each annotator must first label a pilot batch of 10 randomly selected examples. Only those who complete this trial accurately—demonstrating full comprehension of the guidelines—may advance to the main annotation phase. During production we run continuous spot-checks and feedback rounds to keep quality high.

The annotation team consists of native speakers or individuals with extensive immersion in the target language (see §A.1 for demographics). Annotators were informed about the purpose of the data collection, its intended use, and storage policies through detailed instructions and a privacy agreement.

Reproducibility Challenges and Mitigation Strategies To facilitate reproducibility, we will make publicly available: (i) the key code components used for data collection, processing, and evaluation; (ii) the finalized CulturALL dataset accompanied by detailed documentation of its construction and evaluation workflow; and (iii) the full set of model outputs, performance scores, and all experimental configurations required to replicate our results.

All artifacts will be made publicly available at the time of publication, enabling anyone to fully reproduce the entire workflow, from raw inputs to the final results presented in our tables.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 15763–15784. Association for Computational Linguistics.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. Calmqa: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11772–11817. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. Culturalbench: A robust, diverse and challenging benchmark for measuring lms' cultural knowledge through human-ai red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 25663–25701. Association for Computational Linguistics.

Alexander R Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. 2025. Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms. *arXiv preprint arXiv:2507.17476*.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N. Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. Nativqa: Multilingual culturally-aligned natural query for llms. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14886–14909. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Daniel Hershcovich, Stella Frank, Heather C. Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6997–7013. Association for Computational Linguistics.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. Language model alignment in multilingual trolley problems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2016–2039. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Trans. Assoc. Comput. Linguistics*, 13:652–689.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L. Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11995–12041. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *CoRR*, abs/2406.09948.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16366–16393. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *CoRR*, abs/2411.00860.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Haoyi Qiu, Alexander R. Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3978–4005. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2373–2403. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. INCLUDE: evaluating multilingual language understanding with regional knowledge. *CoRR*, abs/2411.19799.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4996–5025. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui,

Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18761–18799. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 370–390. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *CoRR*, abs/2504.15521.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *CoRR*, abs/2205.12247.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Safeworld: Geo-diverse safety alignment. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 9840–9867. Association for Computational Linguistics.

ID	Nationality	Target Lang.	Background
A	Chinese	Chinese	Native speaker
B	Chinese	Chinese	Native speaker
C	Bangladeshi	Bengali	Native speaker
D	Kazakh	Kazakh	Native speaker
E	Moroccan	Arabic	Native speaker
F	Belarusian	Belarusian	Native speaker
G	Vietnamese	Vietnamese	Native speaker
H	Russian	Russian	Native speaker
I	Malaysian	Malay	Native speaker
J	Serbian	Serbian	Native speaker
K	Cambodian	Khmer	Native speaker
L	Chinese	Korean	Ten years' study
M	Chinese	Spanish	Eight years' study
N	Chinese	English	Twenty years' study; five years' residence
O	Chinese	Japanese	Ten years' study; one year's residence

Table 3: Annotator profiles: nationality, target language, and background with the target language.

A Annotation

A.1 Annotators

We relied on 15 volunteer annotators from universities and industrial research labs. All volunteers are native speaker of the target language, or near-native with more than 5 consecutive years of residence and study in the language community. Table 3 gives an overview of their backgrounds.

The 51 regions covered by CulturALL are achieved through two complementary mechanisms. First, several annotators cover multiple regions within their language (e.g., the Chinese annotators contribute samples from both mainland China and Hong Kong/Taiwan, and the Arabic annotator covers several Arabic-speaking countries). Second, for regions without a dedicated local annotator, we draw on samples adapted from existing benchmarks (Section C), which were originally created by or with input from local experts. In such cases, we assume the original knowledge sources are reliable, and all adapted samples undergo the same quality-control review (Task D) as annotator-created items.

A.2 Guideline

As illustrated in Fig. 3, human annotators are assigned four distinct tasks, which are detailed below.

A.2.1 Task A: Sample Creation (Personal Experience)

1. Read §2.2.1 to fully understand the sample requirements.
2. Browse the full topic list—including descriptions and seed examples (§2.1)—to select a topic that interests you.
3. Craft an original sample based on your personal experience whenever possible, using seed items and local forums as inspiration.

A.2.2 Task B: Sample Creation (Cross-lingual Inspiration)

1. Read §2.2.1 to fully understand the sample requirements.
2. Read the English translations of examples originally created in other languages.
3. Whenever possible, write a culturally plausible example in your native language that is similar to the provided ones.

A.2.3 Task C: Difficulty Elevation

1. Review §2.2.1 to fully understand the sample requirements and §2.3.2 to learn strategies for increasing difficulty.
2. Retrieve an easy sample for your language and region.
3. If possible, enhance its difficulty using the elevation techniques described in §2.3.2.

A.2.4 Task D: Quality Control

1. Verify the sample against three criteria: Region/Topic Correctness, Requirement Adherence, Translation Quality, and Sensitive or Offensive Content, as outlined in §2.4.3.
2. **Accept** the sample if it fully satisfies all three criteria without any issues.
3. If issues are identified, **revise** the sample to ensure it meets all requirements.
4. If the sample cannot be revised to meet the criteria, it should be marked as **reject**.

B Topics

Tab. 4 presents the full topic list, each entry paired with a short description and three illustrative culture-specific scenarios; the complete set of seed examples will be released publicly.

C Adapting Existing Datasets

Data Sources We repurpose six public benchmarks:

- **GEOMLAMA** (Yin et al., 2022): QA pairs in five language–region pairs.
- **SeaEval** (Wang et al., 2024a): multiple-choice questions in four language–region pairs.
- **INCLUDE** (Romanou et al., 2024): 44 languages. We retain items whose `regional_feature` is *region implicit*, *region explicit*, or *culture*.
- **CulturalBench** (Chiu et al., 2025): 45 countries/regions, English only. Each item is translated into the dominant local language (Tab. 6).
- **Global-MMLU** (Singh et al., 2025): 42 languages. We keep only culture-sensitive questions.
- **MultiNRC** (Fabbri et al., 2025): QA pairs in three language–region pairs.

Translation Whenever an item is not already in the target language, we translate it with `gpt-4o-2024-11-20` using the prompt shown in Fig. 6 (§D).

Grounding The translated (or original) text is then converted into our sample format with the prompt in Fig. 7, ensuring that each item remains original cultural knowledge while satisfying CulturALL's annotation schema.

D Prompts

This section lists the prompts employed in our study, including translation (Fig. 6), grounded sample generation based on existing datasets (Fig. 7), grounded sample generation based on online resources (Fig. 8), region labeling (Fig. 9), topic labeling (Fig. 10), CulturALL evaluation (Fig. 11), and prediction judgment (Fig. 12).

E Difficulty Distribution Across Languages

F Results Excluding Low-Resource Languages

Seven of the 14 languages in CulturALL have fewer than 100 samples (Arabic: 99, Belarusian: 86, Viet-

Topic	Description	Example 1	Example 2	Example 3
Belief	Systems of conviction that shape values, rituals, institutions, life-cycle events, and views on existence—covering religious faith, spiritual practice, secular ethics, and cultural traditions (e.g., funerary customs and ideas of an afterlife).	Typical length and order of a wedding ceremony	Dietary restrictions during major religious holidays	Whether to pull the lever in the classic trolley-problem dilemma
Commerce	Buying, selling, marketing, and payment of goods and services—from daily necessities to luxury fashion—across bricks-and-mortar shops, e-commerce sites, and mobile wallets.	Typical opening hours for supermarkets	Return policy for online purchases	Legal limits on alcohol sales in retail stores
Education	Formal and informal learning, teaching, research, and skill-building for all ages, settings, and disciplines.	Courses normally taken in middle school	National university-entrance-exam format	Grading scale used in secondary schools
Entertainment	Media, arts, sports, games, performances, hobbies, and events created for leisure and enjoyment.	Popular sport clubs	National mascots or iconic cartoon characters	Gambling age and casino legality
Finance	Earning, saving, budgeting, investing, insuring, transferring, and distributing wealth during life and after death.	Color that signals a stock-price rise or fall on trading screens	Common payment methods in everyday shopping	Typical tax-filing deadline for individuals
Food	Agriculture, sourcing, processing, cooking, nutrition, beverages, and dining culture from farm to table.	Typical breakfast foods	Is tipping expected in restaurants?	Common allergens that must be listed on packaged food
Government	Public policy, legislation, courts, law enforcement, defense, emergency response, and civic administration.	Highway speed limits	Emergency number to call when lost in the mountains	Length of mandatory military or civil service
Habitat	Homes, buildings, infrastructure, utilities, urban planning, ecosystems, weather patterns, and sustainability practices.	Typical home-heating system	Floor-numbering convention in multi-story buildings	Recycling rules for household waste
Health	Physical, mental, and emotional well-being—prevention, treatment, fitness, wellness, palliative, and end-of-life care.	Standard childhood-vaccination schedule	Prescription vs. over-the-counter drug availability	Legal age of consent for medical decisions
Heritage	Past events, living traditions, festivals, monuments, and other cultural inheritances—and their study, preservation, and commemoration.	Date and rituals of New-Year celebrations	Historic event marked by a public holiday	Customs from a particular historical period
Language	Official and minority languages, scripts, dialects, idioms, emotional nuance, politeness levels, sign language, literacy, and translation norms.	Order of family and given names on official documents	Appropriate greetings and honorifics in business	Meaning and proper use of a common proverb
Pets	Care, health, training, companionship, and welfare of domesticated animals.	Rules for bringing pets on public transport	Mandatory rabies vaccination for dogs	Cultural status of certain animals
Science	Systematic inquiry into the natural world and its applications—research, engineering, technology, and innovation.	Unit used to state distance between two cities	Standard format for writing dates	Whether smartphones support dual-SIM use
Social	Family, friendships, romance, community networks, demographics, and social issues.	Table etiquette at family gatherings	Meaning of two women holding hands in public	Typical blind-dating process
Travel	Planning, transport, logistics, accommodation, tourism, and movement of people or goods.	Information needed before booking a city trip	Visa rules for a 90-day tourist stay	Cost of popular tourist attractions
Work	Careers, labor markets, workplaces, productivity tools, and professional development.	Statutory length of paid annual leave	Legal steps for ending an employment contract	Region-specific unique occupations

Table 4: Cultural topics with concise descriptions and illustrative examples.

Translation

Translate the following text to {target_language}:
{source}
Do not output anything else.

Figure 6: Prompt used for the translation task, where {target_language} is the target language name and {source} is the input source text for translation.

Grounded Sample Generation (Existing Datasets)

You are given a sample describing some cultural knowledge:
{source_excerpt}
Generate a grounded item (scenario + question + answer + explanation) that assesses the consultant's grasp of the cultural knowledge.
Ensure the generated item preserves the same cultural knowledge as the example. Do not modify the choices or the correct answer.
The generated sample should be in the same language as the given sample.
The output format should be:
[Scenario]
XXX
[Question]
XXX
[Answer]
XXX
[Explanation]
XXX
Do not output any other things.

Figure 7: Prompt used for grounded-sample creation. Placeholders {source_topic}, {source_excerpt}, and {topic_list} are replaced with the corresponding inputs.

Grounded Sample Generation (Online Resources)

You are given a sample describing some cultural knowledge:
{source_excerpt}
Generate a grounded item (scenario + question + answer + explanation) that assess the consultant's grasp of the cultural knowledge.
Ensure the generated sample preserves the same cultural knowledge as the provided example.
The answer should be objective and as brief as possible. If an objective free-form answer is impractical, convert the question to a four-option multiple-choice format (A–D) and return only the chosen letter.
The generated sample should be in the same language as the given sample.
The output format should be:
[Scenario]
XXX
[Question]
XXX
[Answer]
XXX
[Explanation]
XXX
Do not output any other things.

Figure 8: Prompt used for grounded-sample creation from online resources. The placeholder {source_excerpt} is replaced with the source text.

Region Labeling

Based on the following text and its written language, retrieve the corresponding ISO 3166-1 alpha-2 country code (only one lowercase two-letter country code).
Scenario: {scenario}
Question: {question}
Do not output any other things.

Figure 9: Prompt used for region classification, where {scenario} and {question} are the provided fields of the given sample.

Topic Labeling

Based on the following text, select the most appropriate topic from the topic list `{topic_list}`.
Scenario: `{scenario}`
Question: `{question}`
Do not output any other things.

Figure 10: Prompt used for topic classification. `{scenario}` and `{question}` are the provided fields of the given sample. `{topic_list}` is the predefined list.

CulturALL Evaluation

Using the scenario as context, answer the question in as few words as possible.
Scenario: `{scenario}`
Question: `{question}`
Answer:

Figure 11: Prompt used for CulturALL evaluation, where `{scenario}` and `{question}` are the provided fields of the given sample.

Judgement

Please evaluate whether the model prediction is correct based on the given scenario, question, answer, and explanation.
Scenario: `{scenario}`
Question: `{question}`
Answer: `{answer}`
Explanation: `{explanation}`

Model prediction: `{prediction}`
Only output `1` or `0` with no additional text. `1` means correct, `0` means incorrect.

Figure 12: Prompt used to evaluate the model's prediction based on the given scenario, question, answer, and explanation. The prompt ensures binary evaluation for correctness.

Lang	Language	Reg	Region
ar	Arabic	sa	Saudi Arabia
be	Belarusian	by	Belarus
bn	Bengali	bd	Bangladesh
en	English	us	United States
es	Spanish	es	Spain
ja	Japanese	jp	Japan
kk	Kazakh	kz	Kazakhstan
km	Khmer	kh	Cambodia
ko	Korean	kr	South Korea
ms	Malay	my	Malaysia
ru	Russian	ru	Russia
sr	Serbian	rs	Serbia
vi	Vietnamese	vn	Vietnam
zh	Chinese	cn	China

Table 5: ISO 639-1 language codes (Lang) with language names and their representative ISO 3166-1 alpha-2 country/region codes (Reg) and region names, sorted by Lang.

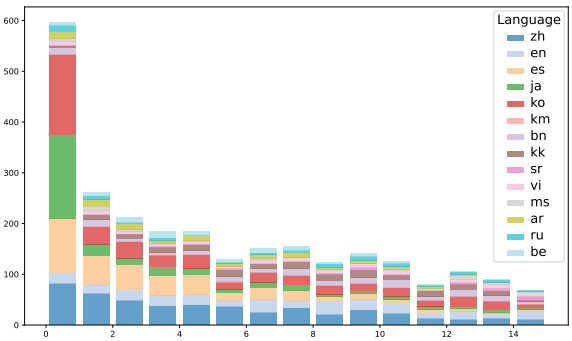


Figure 13: Language distribution of examples based on the number of settings that answered them correctly. X-axis: Number of settings that answered correctly, Y-axis: Count of examples.

namese: 80, Russian: 74, Malay: 57, Serbian: 37, Khmer: 8). To ensure that the main results are not driven by these low-resource subsets, we recompute all accuracies after excluding these seven languages. The filtered dataset contains 2,169 samples across the remaining seven languages (Bengali, English, Spanish, Japanese, Kazakh, Korean, Chinese).

Table 7 compares the original and filtered accuracies. The differences are modest (within 4 percentage points), and the overall ranking of models is preserved. This confirms that the main conclusions are not skewed by the low-resource language subsets.

Reg	Region	Lang	Language	Reg	Region	Lang	Language	Reg	Region	Lang	Language
ar	Argentina	es	Spanish	au	Australia	en	English	bd	Bangladesh	bn	Bengali
by	Belarus	be	Belarusian	ca	Canada	en	English	cl	Chile	es	Spanish
cn	China	zh	Chinese	co	Colombia	es	Spanish	eg	Egypt	ar	Arabic
es	Spain	es	Spanish	fj	Fiji	en	English	hr	Croatia	hr	Croatian
hk	Hong Kong	zh	Chinese	ie	Ireland	en	English	iq	Iraq	ar	Arabic
jp	Japan	ja	Japanese	kh	Cambodia	km	Khmer	kp	North Korea	ko	Korean
kr	South Korea	ko	Korean	kz	Kazakhstan	kk	Kazakh	lb	Lebanon	ar	Arabic
ma	Morocco	ar	Arabic	mu	Mauritius	en	English	mx	Mexico	es	Spanish
my	Malaysia	ms	Malay	ng	Nigeria	en	English	nz	New Zealand	en	English
pe	Peru	es	Spanish	pg	Papua N. Guinea	en	English	ph	Philippines	tl	Tagalog
rs	Serbia	sr	Serbian	ru	Russia	ru	Russian	sa	Saudi Arabia	ar	Arabic
sg	Singapore	en	English	sy	Syria	ar	Arabic	tw	Taiwan	zh	Chinese
uk	United Kingdom	en	English	us	United States	en	English	ve	Venezuela	es	Spanish
vn	Vietnam	vi	Vietnamese	za	South Africa	en	English				

Table 6: ISO 3166-1 alpha-2 country/region codes (Reg) and region names with their representative ISO 639-1 language codes (Lang) and language names, sorted alphabetically by Reg.

	All		Hard		Med.		Easy	
Model	**Orig**	**Filt**	**Orig**	**Filt**	**Orig**	**Filt**	**Orig**	**Filt**
gemini-2.5-pro_auto_true	44.48	42.51	18.47	17.91	65.71	64.87	92.55	93.33
gemini-2.5-pro_128_true	39.27	38.08	15.49	15.59	55.71	56.09	87.66	88.33
gemini-2.5-flash_auto_true	33.68	33.43	12.78	11.91	46.71	49.82	78.30	82.78
gpt-5-high_false	37.59	34.39	10.28	8.95	57.71	55.20	91.28	90.56
claude-opus-4-high_false	36.70	35.18	9.44	9.59	56.00	56.63	91.49	90.83
claude-sonnet-4-high_false	32.76	31.86	8.82	9.11	46.43	48.92	85.74	84.44
qwen3-235b-a22b_high_false	23.68	22.87	4.65	4.96	33.57	34.95	67.23	66.39

Table 7: Accuracy (%) before (Orig) and after (Filt) excluding low-resource languages (<100 samples). Differences are within 4 points; model rankings are preserved.

G LLM–Human Role Division

A key design principle of our framework is a clear separation of roles between human annotators and LLMs. *LLMs are never used as knowledge sources.* The factual grounding of all samples comes from (1) human annotators drawing on personal cultural expertise, (2) curated existing benchmarks, and (3) human-verified online resources. LLMs serve only as productivity tools for mechanical subtasks: expanding a seed topic list (Stage 1), rewriting existing benchmark items into grounded format (Stage 2), measuring relative item difficulty (Stage 3), and filling metadata fields and translating samples (Stage 4). In every stage, *human annotators perform the steps where LLM outputs may be unreliable*: they author original scenarios from first-hand experience, select and verify topics, elevate sample difficulty, and conduct the final quality-control review that accepts, revises, or rejects every item. Table 8 summarises this division of responsibilities.

Stage	**LLM role**	**Human role**
1. Topic Sourcing	Expand seed topic list	Select and verify topics
2. Sample Creation	Rewrite existing items; draft from web posts	Author original scenarios; verify all instances
3. Sample Enrichment	Measure relative difficulty	Select, revise, and approve samples
4. Release-Ready	Fill metadata; translate	Quality-control every item (accept/revise/reject)

Table 8: Division of responsibilities between LLMs and human annotators at each pipeline stage.