Culture-Wiki: Benchmarking Different Cultures in Knowledge Recalling Tasks

Anonymous ACL submission

Abstract

While the multilingual capacity of Large Language Models (LLMs) is well-documented, their ability to recall culturally specific knowledge remains underexplored. Most culturally aware evaluation frameworks rely on discriminative tasks, failing to capture the complexity, diversity, and evolving nature of cultural knowledge. To address these issues, we introduce CULTURE-WIKI, a multilingual benchmark for culturally-grounded generation in LLMs across four linguistic-cultural settings (German, Hindi, Italian, and Spanish) and nine UNESCO-inspired cultural domains. Built through a scalable Wikipedia-based pipeline, CULTURE-WIKI includes three generative tasks targeting complementary levels of cultural knowledge: concept- & entity-aware MT, encyclopedic definition generation, and open-ended question answering. Our findings highlight critical limitations in current LLMs and underscore the need for more robust, culturally aware evaluation frameworks.

1 Introduction

Large Language Models (LLMs) are increasingly deployed across linguistic and cultural contexts (Xiong et al., 2024; Kasneci et al., 2023; De Angelis et al., 2023). To be truly useful in such settings, these systems must go beyond linguistic competence and capture the cultural knowledge that underpins everyday communication (Pawar et al., 2025). Models performing well on English-centric benchmarks may still fail to recognize culturally specific entities, misrepresent local knowledge, or produce culturally inappropriate outputs despite being fluent (Zhao et al., 2025). Robust evaluation of cultural capabilities is therefore essential for responsible multilingual deployment.

A central challenge is the difficulty of formalizing *culture*. Social sciences define culture as a dynamic system of shared meanings, practices, and symbols that vary across societies (Lévi-Strauss, 1963; Geertz, 1973; UNESCO, 2002). However, despite a growing interest in culturally grounded evaluation (Yao et al., 2024; Hupkes and Bogoychev, 2025a; Hasan et al., 2025), evaluation benchmarks tend to limit to a few cultural aspects such as food or etiquette (Zhou et al., 2025). Moreover, cultural evaluation benchmarks predominantly rely on single-task, discriminative or constrained formulations, such as multiple-choice question answering (Hasan et al., 2025; Chiu et al., 2025), leaving open-ended *generative* abilities underexplored.

To address these limitations, we introduce CULTURE-WIKI: a scalable, multilingual, and multitask benchmark for evaluating cultural knowledge in LLMs, spanning over 30k items, covering three generative tasks: concept- & entity-aware machine translation (MT), encyclopedic definition generation (DG), and open-ended question answering (QA). The benchmark covers four target cultures, each evaluated in its corresponding language[1] (German, Hindi, Italian, and Spanish) and spans nine UNESCO-derived cultural domains (e.g., *Visual Arts & Crafts*, *Books & Press*, *Music*). CULTURE-WIKI is built on a novel, semi-automated annotation pipeline based on Wikipedia that is extensible to other culture-targeted languages and contexts.

The benchmark also enables investigating how culture-specific entities and concepts, expressed in a non-native language, affect downstream task performance.[2]

Our evaluation of several LLMs reveals sharp

[1] For this work, we use the official language of a country as a proxy for its cultural context and shared knowledge. While some of these languages are spoken across multiple countries, our manual validation process ensures that the selected content remains grounded in the culture associated with the reference country. Specifically, we associate German with Germany, Hindi with India, Italian with Italy, and Spanish with Spain.

[2] We use English as the target language given its role as a *lingua franca*.

performance disparities across cultural domains and translation types, proving that no current model is consistently robust and highlighting critical gaps in culturally-specific settings.

In summary, our key contributions are:

- **A scalable Wikipedia-based framework** for constructing culturally grounded benchmarks with minimal manual annotation.
- **CULTURE-WIKI:** a multilingual, multitask benchmark with over 30k items covering 4 culture-associated languages and 9 UNESCO domains, enriched with cross-lingual lexicalization signals.
- **An extensive cross-task LLM evaluation**, complemented by targeted analyses that expose consistent disparities and critical failure modes in modern LLMs' cultural reasoning.

We release our code and data at `omitted.link` under CC BY-SA 4.0.

2 Related Work

Multitask and Multilingual Benchmarking Multitask benchmarks have become a standard tool for evaluating LLMs. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) introduced early task collections for encoder models, while BIG-bench (Srivastava et al., 2023), MMLU (Hendrycks et al., 2021), and HELM (Liang et al., 2023) extended evaluation to broader reasoning and knowledge tasks to modern LLMs. Despite their scale, these benchmarks are predominantly English-centric, implicitly assuming that linguistic competence transfers across cultural contexts.

Multilingual benchmarks such as XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020), and Belebele (Bandarkar et al., 2024) extend coverage across languages to study cross-lingual transfer. However, they often treat languages as interchangeable channels rather than as carriers of distinct cultural knowledge. As a result, localization is frequently approximated via translation of English-centric benchmarks, which inherently preserves Western-focused content, as revealed by Singh et al. (2025). Multilingual generation-oriented benchmarks, such as FLORES (Guzmán et al., 2019), further expand linguistic diversity, however focusing only on the task of machine translation.

Overall, existing multitask and multilingual benchmarks provide broad coverage of linguistic abilities, but tend to overlook how cultural grounding shapes knowledge representation, recall, and reasoning.

Cultural Probing Benchmarks A few works have instead investigated systematic weaknesses of LLMs in culturally grounded settings. XCOPA (Ponti et al., 2020) demonstrated degraded commonsense reasoning under cultural variation, with later work confirming similar gaps across languages and regions (Chiu et al., 2025; Myung et al., 2024; Hasan et al., 2025; Hupkes and Bogoychev, 2025b). However, existing benchmarks often lack strong control over cultural relevance, e.g., some heuristically sample from sources like Wikipedia without explicit cultural filtering (Arora et al., 2025; Zhao et al., 2025). In contrast, we explicitly maximize precision of culture-linked Wikipedia entities and concepts for a target country and apply rigorous filtering with manual verification to ensure high-precision cultural grounding.

A second limitation is the narrow operationalization of cultural competence. Prior work largely relies on discriminative formats such as multiple-choice or binary judgments (Kim et al., 2024; Rao et al., 2025), or treats translation as a proxy for cultural knowledge (Yao et al., 2024; Xu et al., 2026). While useful, these settings mostly capture surface-level recognition. Even generative QA benchmarks like CaLMQA (Arora et al., 2025) remain limited to a single task type. In contrast, we introduce a multi-tier benchmark of cultural competence spanning multiple tasks, enabling a more extensive diagnostic of cultural reasoning beyond prior benchmarks.

3 The CULTURE-WIKI Benchmark

In this section, we describe our creation pipeline of CULTURE-WIKI, which comprises three complementary generative tasks of increasing difficulty: concept- and entity-aware machine translation (**MT**), encyclopedic definition generation (**DG**), and open-ended question answering (**QA**).

The annotation process was carried out by two native-speaker annotators per language, all of whom are PhD students with strong backgrounds in linguistics and NLP. All the annotators are co-authors of the present paper.

3.1 Collection of Culturally Relevant Pages

In the first step, we collect a set of *culturally relevant* Wikipedia pages $\hat{P}^L$ for each culture-targeted language L. Each page $p \in \hat{P}^L$ describes a real-world concept or entity that is specifically relevant to L. This is done in two sub-steps: first, we automatically collect a set of candidate pages P^L_{rel}, then we perform manual filtering to increase data quality, obtaining $\hat{P}^L$.

Candidates Retrieval To identify culturally relevant pages, we leverage Wikipedia's category graph. This knowledge graph consists of a language-specific set of categories C^L, such as *Musica*[3] and *Musica Italiana*, organized hierarchically, and of a language-specific set of pages P^L such as *Luciano Pavarotti*, linked to the categories.

Our goal is to construct $C^L_{\text{rel}} \subseteq C^L$, the set of categories that are culturally *relevant* to L. To do so, we maintain two language-specific lists: a *target-culture list*, containing morphological variants country name and demonyms in the target language L (e.g., *Italia* and *italiano/a/i/e*), and an *other-culture list*, containing references to all other countries and regions (e.g., *Scandinavia*) (see Appendix E). We then perform a depth-first traversal going down the Wikipedia's category graph (starting from category *Culture* in each language wikipedia dumps) extracting all categories and descendant categories whose names contain a substring match from the *target-culture list* (e.g., *Musica italiana*). We halt the traversal when reaching any category whose names contain a substring match from the *other-culture list* (e.g., *Musica francese*).

Filtering steps To ensure the cultural relevance of our retrieved data, we perform automatic and manual filtering steps, involving native-speaker annotators for each of the four culture-targeted languages in our benchmark.

In order to have a feasible number of instances to annotate, we reduce C^L_{rel} obtaining $\bar{C}^L_{rel}$ using two filters. First, we rank the categories in C^L_{rel} by their number of linked pages, and retain the minimum set of top-ranking categories covering at least 90% of the pages. This reduction yields a manageable set of ∼2% for each of our four target languages. Then, we further remove all categories irrelevant to the target culture (e.g., proper names or *stub* categories) by manually checking them (see Appendix A for further details). In this way, we obtain $\bar{C}^L_{\text{rel}}$ and the associated set of pages $\bar{P}^L_{\text{rel}}$.

To ensure data quality, we manually annotate a refined subset of pages, reducing the total number to keep the annotation effort feasible. In order to have a balanced and meaningful subset, we first classify every page leveraging CNER taxonomy (Martinelli et al., 2024), which categorizes pages into a set of semantic classes (e.g., *Location, Event, Measure, Disease, Food*, etc.) and two types (*Named Entities* and *Concepts*). We then exclude the 50 most viewed pages for each CNER class and type to account for popularity bias.[4] Finally, we sample from $\bar{P}^L_{rel}$, ensuring an equal proportion across classes and types. Native speaker annotators then manually inspect the sampled pages, discarding those that do not meet our definition of cultural relevance following the guidelines reported in Appendix C. This process yields the final set of culturally relevant pages $\hat{P}^L$. The inter-annotator agreement is reported in Appendix F.

3.2 Data Enrichment

We now annotate each culturally relevant page $p \in \hat{P}^L$ along two dimensions to support fine-grained downstream analyses: (i) its associated UNESCO category and (ii) the way its title is translated into English.

UNESCO Classification Annotators assign each page to one of 9 *cultural domains* derived from the UNESCO Framework for Cultural Statistics (UNESCO Institute for Statistics, 2025), covering categories such as *Visual Arts & Crafts* (i.e., artistic forms such as painting, sculpture, photography, and traditional crafts) and *Cuisine & Food Culture* (i.e., traditional dishes, culinary techniques, and food-related practices tied to cultural identity); the full taxonomy is in Appendix C.2.

This categorization provides a theoretically grounded and culturally specific framing that complements the CNER classification, which instead is designed for general-purpose concept and named entity recognition rather than cultural analysis. Table 1 reports the distribution of UNESCO categories for each annotated page.

Translation Type Annotation Annotators also assign each page in $\hat{P}^L$ to a *translation type*, that

[3] Examples in this section are given for $L = IT$.

[4] Our manual inspection revealed that highly popular pages are associated to globally known items.

Lang.	Cultural Heritage	Performing Arts	Visual Arts & Crafts	Books & Press	Audio-visual	Design & Creative Svcs.	Music	Cuisine & Food Cult.	Festivals	Other
Italian	33.8	2.1	2.6	7.5	9.1	3.9	13.3	16.4	5.6	5.8
German	45.6	2.1	2.6	7.6	10.5	3.3	4.7	6.3	7.6	9.8
Spanish	43.1	0.9	2.2	7.5	5.3	4.7	3.8	11.5	12.8	8.3
Hindi	33.0	0.8	2.1	6.2	10.0	3.1	2.7	9.0	10.7	22.5

Table 1: Row-wise percentage distribution of instances across UNESCO cultural categories and languages. Percentages in each row sum to 100%.

Translation Type	Italian	German	Spanish	Hindi
LITERAL	470	492	1,052	921
CONCEPTUAL	109	254	99	88
EXPLICITATION	0	1	9	159
LOANWORD	49	7	11	150
NO TRANSLATION	1,298	976	1,029	638
Total	1,926	1,730	2,200	1,956

Table 2: Type translation statistics, across different languages for the extracted and annotated pages.

Language	MT	DG	QA	Total
German	2,472	1,730	2,626	6,828
Hindi	1,009	1,956	3,488	6,453
Italian	3,308	1,926	3,834	9,068
Spanish	1,947	2,200	4,303	8,450
Total	8,736	7,812	14,251	30,799

Table 3: Statistics of the Culture-Wiki benchmark across languages and task categories. MT denotes Machine Translation, DG denotes Definition Generation, and QA denotes Question Answering. The *Total* column reports the number of instances per language, while the final row summarizes the total number of instances per task and across the entire benchmark.

is, a label describing how the title of p^L, natively expressed in language L, is translated into English. The five translation types – LITERAL, CONCEPTUAL, EXPLICITATION, LOANWORD, and NO TRANSLATION – are defined in Appendix C.1. While elicited in the context of translation, these labels capture cross-lingual lexicalization patterns that are relevant to all our three tasks. We reuse them in our analysis to examine how the translatability of a cultural entity or concept affects model performance, beyond the machine translation task. Table 2 shows the distribution of annotated translation types across culturally related Wikipedia pages by language. NO TRANSLATION is the most common category in most culture-targeted languages, indicating that many cultural items lack direct English equivalents and that our pipeline captures culture-specific items effectively. Language-specific patterns also emerge: Hindi has notably more LITERAL translations, likely due to Devanagari transliteration, as well as more LOANWORD and EXPLICITATION cases than the European languages. Further discussion appears in Appendix D.

3.3 Generation of Task Items

Given our set of culturally relevant pages $\hat{P}^L$, we now describe how we generate dataset instances for the three tasks featured in CULTURE-WIKI. Each task is designed so that cultural knowledge is *necessary* to produce a correct output. Items for each task are created in the language of the corresponding culture, using content extracted from the Wikipedia page in that culture's language.

Concept- & Entity-aware (MT) In this task, models are requested to translate sentences from a source language into English, and are evaluated specifically on their ability to correctly translate page titles corresponding to culturally relevant concepts and named entities while preserving meaning and contextual usage. We use English as the target language given its status as the highest-resource language and its central role in the training of all evaluated models.

For each culturally relevant page in $\hat{P}^L$, we extract candidate sentences by first segmenting the page according to Wikipedia's section structure and discarding sections unlikely to contain useful contextual information, such as *See Also* and *Related Notes*, as well as introductory sections, whose definitional content would make the task trivially easy. We split the remaining text into sentences, and discard those shorter than 5 words or not explicitly mentioning the normalized page title.[5] We then randomly sample two sentences per page, yielding the final set of MT items. For each

[5]We lowercase titles and remove parenthetical terms.

page, native-speaker annotators with linguistic expertise provide up to three English translations of the target title, to have a gold references for the evaluation. Annotation details are reported in Appendix C.1.

Encyclopedic Definition Generation (DG) This task evaluates a model's ability to produce concise, Wikipedia-style definitions for culturally grounded concepts and named entities. Given a target page in $\hat{P}^L$, the model is prompted to generate a short, self-contained description identifying what the page is, following Wikipedia's editorial guidelines for lead sections[6]. For each page, we use the first sentence of the Wikipedia introduction as the reference definition for evaluation, as it provides a concise and editorially standardized description of the target concept or entity.

Open-ended Question Answering (QA) This task evaluates whether LLMs can answer factual open-ended questions requiring relational cultural knowledge – that is, knowledge about connections between cultural entities rather than isolated facts about a single item. To construct QA items, for each page in $\hat{P}^L$ we identify hyperlinks within each section that point to pages in the initial collection P^L_{rel}. While pages in P^L_{rel} were not manually verified for cultural relevance, they were retrieved from culturally specific Wikipedia categories and are therefore culturally grounded to some degree (see Section 3.1). Moreover, since they appear as hyperlinks within pages in $\hat{P}^L$, they are by construction related to a culturally relevant item.

For each linked page, we provide an LLM[7] with three sources of context: (i) the introductory section of the page in $\hat{P}^L$, (ii) the section containing the hyperlink, and (iii) the introductory section of the linked page. Based on this information, the model is instructed to generate open-ended question–answer pairs probing the relationship between the two entities, with answers required to be short, specific, and verifiable from the provided context (see Figure 2). Further details about the LLM prompt are reported in Appendix B.

We manually validate a subset of generated QA items and find that approximately $\sim$90% are valid and answerable across all four culture-targeted languages. Annotation guidelines for the qualitative assessment are provided in Appendix G.

[6]Wikipedia Manual of Style

[7]We used `gpt-oss-120b`.

3.4 CULTURE-WIKI Statistics

Our semi-automated pipeline, applied to the Wikipedia dumps[8] of four languages (German, Hindi, Italian, and Spanish) from three different families, results in the CULTURE-WIKI benchmark. The dataset contains more than 30,000 annotated items across 4 languages and 3 tasks. General statistics are reported in Table 3, while more details are discussed in Appendix H.

4 Experimental Setup

4.1 Models

We evaluate a diverse set of instruction-tuned LLMs spanning a wide range of scales, from 1.7B to 70B parameters, to study the effect of model size on cultural knowledge. Our selection is guided by three criteria: scale variation, language coverage, and training data transparency. We include broad-coverage multilingual models, namely Qwen-2.5 (Qwen et al., 2025) (3B, 7B, 14B), Llama-3 (Grattafiori et al., 2024) (3B, 8B, 70B), and Phi-4 (Abdin et al., 2024) (3B, 14B), as strong general-purpose baselines, alongside models trained on documented multilingual corpora with permissive licenses, such as EuroLLM (Martins et al., 2024). To assess the effect of language-specific training, we also include models with multilingual targeted coverage: Minerva (Orlando et al., 2024), an Italian–English bilingual model; Salamandra (Gonzalez-Agirre et al., 2025), a model with dedicated coverage of Spanish and other Iberian languages; the Spanish and German variants of Occiglot (Avramidis et al., 2024), both built on Mistral-7B; and ANITA (Polignano et al., 2024), an Italian adaptation of Llama-3.1-8B. We also include SmolLM2 (Allal et al., 2025) (1.7B) as a lightweight English-centric reference point.

4.2 Prompting and Generation Strategies

All experiments are conducted in a zero-shot setting using the vLLM library (Kwon et al., 2023). To assess robustness to prompt language, we evaluate each model with two prompts: one in English and one in the language associated with the target culture. In the English setting, only the instructions in the system prompt are translated, while culturally grounded content remains in the source language. Full prompts are reported in Appendix I.

[8]We used the Wikipedia dumps of January 2026.

Model	Size	Italian			German			Spanish			Hindi		
		MT	DG	QA	MT	DG	QA	MT	DG	QA	MT	DG	QA
Qwen2.5	1.5B	47.1	15.2	5.6	37.4	16.9	3.4	37.6	20.4	4.5	15.6	0.7	2.6
EuroLLM	1.7B	52.5	15.3	10.1	37.9	15.3	7.6	41.1	21.8	7.2	23.5	1.7	5.3
SmolLM2 (EN)	1.7B	41.5	12.5	3.5	31.4	11.6	2.0	33.8	17.6	2.2	7.7	1.1	1.5
Qwen2.5	3B	56.6	19.0	8.1	45.4	20.2	5.5	46.8	22.9	7.0	20.5	1.8	3.5
Llama-3.2	3B	51.8	18.7	6.5	40.2	19.3	6.3	40.4	24.0	5.2	22.8	1.5	6.8
Phi-4-mini	3.8B	55.3	17.2	9.0	40.7	17.6	9.2	43.5	24.1	8.1	22.9	1.7	4.0
Minerva (IT)	7B	55.3	21.6	23.2	30.2	16.8	3.0	34.8	17.9	4.3	7.8	0.8	2.2
ANITA (IT)	8B	56.3	23.0	15.3	43.6	21.2	14.6	45.3	20.1	10.9	27.8	1.1	8.6
Occiglot-de-en (DE)	7B	58.4	14.7	10.2	46.1	21.5	17.7	48.6	17.2	9.8	17.0	0.9	2.2
Salamandra (ES)	7B	56.6	22.6	15.1	45.9	20.5	14.4	48.7	26.9	21.3	11.0	1.2	1.7
Occiglot-es-en (ES)	7B	53.4	11.0	10.6	41.9	11.4	7.8	49.2	17.0	17.5	15.4	0.5	2.8
Qwen2.5	7B	61.5	21.3	13.0	47.8	22.8	9.7	54.2	23.2	12.5	28.0	3.3	6.0
Llama-3.1	8B	59.4	22.3	15.2	45.0	22.4	14.7	49.1	23.9	11.4	27.9	3.1	13.5
Qwen2.5	14B	63.0	22.9	19.5	47.5	24.4	15.7	54.2	26.5	20.1	33.9	3.2	12.5
Phi-4	14B	65.2	18.0	23.1	50.5	20.4	24.1	55.9	19.2	24.3	36.7	3.2	19.7
Llama-3.1	70B	**67.1**	**29.0**	**28.9**	**54.1**	**27.7**	**29.9**	**59.2**	**32.5**	**28.0**	**37.0**	**5.0**	**27.0**
Average		56.3	19.0	13.6	42.8	19.4	11.6	46.4	22.2	12.1	22.2	1.9	7.5

Table 4: Performance of 16 LLMs across four cultural settings and three tasks. We show the ETA score for MT, ROUGE-L for DG, and sub-EM for QA. Cultural groups are separated into column blocks by language community. Best values are in **bold**, while the highest value per model size is underlined. The target languages of language-specific LLMs are given in brackets, others are multilingual.

4.3 Evaluation Metrics

We evaluate the LLMs on our tasks as follows.

MT We evaluate translation quality using two complementary perspectives: **xCOMET** (Guerreiro et al., 2024) provides a source-conditioned quality estimate without requiring references, while **Entity Translation Accuracy (ETA)** (Conia et al., 2024) measures exact-match accuracy for translated entities and concepts against gold references.

DG Following Periti et al. (2025), we report **ROUGE-L** (Lin, 2004) and **BERTScore** (Zhang et al., 2020) between the generated definition and the ground truth definition to capture lexical and semantic similarity, respectively.

QA Since our questions ask about specific entities and concepts, we evaluate the generated responses using **substring Exact Match (sub-EM)**, which considers a prediction correct if it contains the reference answer as a substring.

5 Results

Table 4 shows the results obtained on CULTURE-WIKI; for readability, we limit the number of metrics to ETA for MT, ROUGE-L for DG and sub-EM for QA; we report the full results in Appendix J. Overall, larger models tend to achieve better performance across the considered culture-targeted languages and tasks. This trend is particularly clear among open-weight multilingual LLMs: models such as Qwen2.5-14B and Llama-3.1-70B consistently outperform smaller ones, especially on QA and MT. The same pattern is also observed within model families, where larger variants generally improve over their smaller counterparts.

However, model size alone does not fully explain performance, as language coverage and pretraining data also appear to play an important role. For instance, Minerva obtains strong QA results for Italian, while Salamandra performs particularly well on Spanish QA and DG, in line with their language distribution of pretraining data. Similarly, language-specific finetuning appears to provide only limited adaptation benefits: ANITA is only marginally better than its backbone model, Llama-3.1, and Occiglot-es-en shows lower performance compared to Salamandra. Moreover, these gains do not transfer to the other culture-targeted languages, indicating that the improvement is modest and remains language-specific.

Language	MT–DG	MT–QA	DG–QA
Italian	0.454	0.622	0.500
German	0.633	0.717	**0.583**
Spanish	0.259	**0.845**	0.317
Hindi	**0.667**	0.828	0.492
Average	0.503	0.753	0.473

Table 5: Correlation between tasks measured through Kendall's τ by languages. Bold indicates highest agreement per column, underline second highest.

Language	Conc.	Lit.	Loan.	No Tr.	Expl.
Machine Translation — ETA					
Italian	19.3	42.8	70.0	63.1	-
German	17.7	38.4	39.7	50.9	-
Spanish	11.8	41.2	57.6	53.1	76.9
Hindi	8.6	29.8	41.8	10.5	42.3
Descriptive Generation — ROUGE-L					
Italian	19.5	21.9	19.4	17.9	-
German	19.3	21.4	13.2	18.4	-
Spanish	22.2	23.4	15.6	21.1	14.1
Hindi	1.8	2.2	0.5	2.2	0.6
Question Answering — Sub-EM					
Italian	12.5	13.4	20.5	13.4	-
German	11.9	13.2	15.6	10.8	-
Spanish	12.1	12.5	24.6	11.6	12.2
Hindi	8.9	7.1	8.7	7.3	8.4

Table 6: Scores of the three tasks in CULTURE-WIKI averaged by the five translation types: LITERAL, CONCEPTUAL, LOANWORD, EXPLICITATION, and NO TRANSLATION.

The results also show substantial variation across culture-targeted languages. Performance is generally higher for German, Italian, and Spanish, while Hindi remains more challenging across all tasks. This gap is especially evident in MT and DG, where Hindi scores are notably lower than those obtained for the European culture-targeted languages. This indicates that low-resource languages – particularly those underrepresented in the pretraining data of these models – still pose significant challenges for current multilingual LLMs.

These cross-task differences point to three factors that jointly affect overall performance: the model scale, the amount and quality of language-specific knowledge acquired through pretraining data, and the culture-targeted language coverage. Among the three tasks, QA appears particularly sensitive to all these factors, highlighting the current limitations of the evaluated models.

5.1 Leaderboard Stability across Tasks

In Table 5, we report Kendall's τ values between model rankings induced by MT, DG, and QA tasks across culture-targeted languages. This measures the consistency of model orderings across tasks probing different aspects of knowledge and reasoning. We observe substantial variability in leaderboard consistency. Overall, MT and QA show higher agreement across all culture-targeted languages, indicating more stable ranking alignment between these tasks, while DG is more variable and less aligned with the others. German and Hindi exhibit relatively stable structures (with τ up to 0.83), whereas Spanish shows strong decoupling between DG and the other tasks ($\tau \leq 0.32$), suggesting that task choice can significantly affect model ordering.

5.2 Results across Translation Types

Table 6 reports the averaged results across the five different types of translation of Wikipedia page titles (cf. Section 3.2).

Of the three tasks, MT is the task where the effect of translation type is most evident. LOANWORDS and NO TRANSLATIONS generally obtain the highest scores, while CONCEPTUAL TRANSLATIONS are consistently the most challenging across all culture-targeted languages. This suggests that models handle surface-level lexical overlap more easily than cases that require a deeper cultural reformulation. For DG and QA, differences across translation types are less pronounced. In DG, LITERAL TRANSLATIONS tend to obtain slightly better results for German, Italian, and Spanish, while LOANWORDS appear more difficult, especially for German and Spanish. In QA, the pattern reverses: LOANWORDS lead to the best results for German, Italian, and Spanish, suggesting that recognizable lexical forms facilitate answer generation. Hindi remains lower across all translation types in both tasks.

QA Results by Translation Type: The Italian Case In Figure 1, we report model-wise deviations from the average QA performance for Italian across translation types. The figure confirms that LOANWORDS tend to be associated with positive deviations for several models, while CONCEPTUAL TRANSLATIONS are more variable and often below the average. Among the Italian-centric models, Minerva-7B shows positive deviations on both LOANWORDS and CONCEPTUAL TRANS-

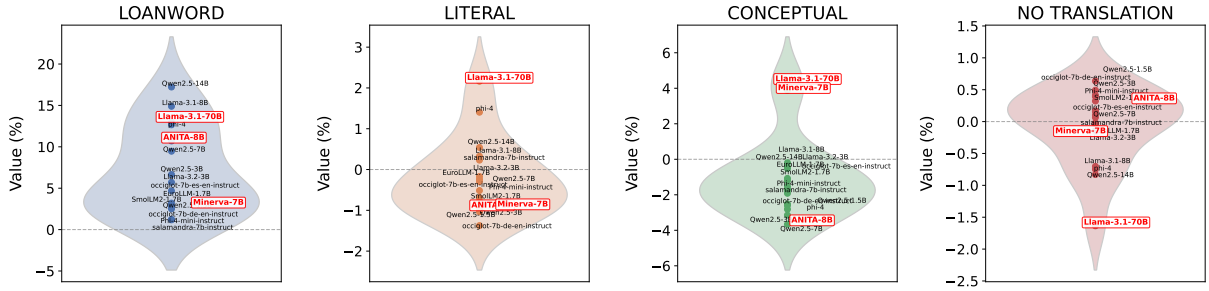

Figure 1: Question answering results across 16 LLMs, reporting Exact Match (EM) scores for Italian conditioned on the translation type of the source pages from which the questions were extracted. Reported values indicate the performance difference for each translation type with respect to the average score across all questions.

LATIONS. This suggests that its Italian-centric pretraining may improve its handling of culture-specific instances that are translated into English as conceptual multi-word expressions. However, this advantage is not uniform across Italian-specific models: ANITA-8B, which is finetuned on Italian data, does not show the same consistent gains, showing impact of pretrained models to a specific language.

6 Qualitative Analysis

We complement quantitative results with a qualitative analysis of model outputs on the Italian and Spanish subsets (further details and qualitative examples in Appendix K).

For MT, we identify three recurrent error types: (i) **source-form retention**, where the source concept or entity is left untranslated despite the existence of an established English form; (ii) **unwarranted translation**, where a concept or entity without a canonical English equivalent is unnecessarily translated; and (iii) **near-equivalent substitution**, where items are replaced with culturally related but incorrect referents.

DG outputs exhibit consistent failure patterns across all model families. Models frequently hallucinate entirely incorrect identities (e.g., describing *La mia storia tra le dita* as a book rather than a song), preserve the correct entity type while altering key details (e.g., *Fontina* → "Fontana"), or over-generalize in ways that erase culturally specific information (e.g., *Viazzi* → "an Italian surname").

QA outputs similarly contain confident but incorrect answers. Common patterns include replacing culturally specific or obscure entities with globally prominent ones (e.g., attributing the term *honorificabilitudinitate* to the *Divina Commedia* instead of *De Vulgari Eloquentia*) and confusing entities belonging to the same category (e.g., misidentifying the Spanish Minister of the Interior who was in office during the El Descanso bombing (*Atentado del restaurante El Descanso*).

7 Conclusion

We introduced CULTURE-WIKI, a multilingual benchmark for evaluating cultural knowledge and generation in LLMs across four linguistic-cultural settings and nine cultural domains, spanning three complementary generative tasks targeting translation, definition and open-ended question answering. Our evaluation shows that culturally grounded performance remains highly uneven. Although larger and language-adapted models generally achieve stronger results, gains are inconsistent across tasks and culture-targeted languages. Low-resource languages (e.g., Hindi) remain challenging across settings, and model rankings vary substantially across tasks and languages, suggesting that current evaluations capture only partial aspects of cultural competence. Results across translation types further show that models handle lexical overlap more reliably than conceptual cultural reformulations. Our qualitative analysis confirms these limitations: models frequently hallucinate culturally related entities, over-generalize culture-specific concepts, or replace local knowledge with globally prominent references. Together, these findings highlight a persistent gap between linguistic fluency and culturally grounded understanding. Beyond benchmarking, we propose a reproducible Wikipedia-based pipeline for scalable construction of culturally aware datasets with minimal annotation effort, easily extensible to new languages and cultural contexts. Our results show that improving culturally aware generation requires moving beyond language-centric paradigms toward frameworks that explicitly capture cultural variation.

Limitations

Despite providing a broad and comprehensive overview, this work presents several limitations that should be acknowledged.

Open-ended Evaluation: Evaluating open-ended generation remains an unresolved challenge in LLM assessment. For the three proposed tasks, we deliberately relied on well-established and easily reproducible automatic metrics. In particular, ROUGE-L and BERTScore, used for the Definition Generation task, mainly capture lexical overlap and semantic similarity, while they do not explicitly assess factual correctness or hallucinations. Similar limitations also affect the substring Exact Match metric adopted for Question Answering evaluation. Although more sophisticated evaluation protocols could provide finer-grained insights, we intentionally avoided introducing task-specific or ad-hoc metrics in order to preserve the simplicity, reproducibility, and usability of the benchmark. We also decided not to employ LLM-as-a-Judge approaches, as they are known to suffer from evaluation biases and instability, particularly in low-resource settings (Zhang et al., 2026). Moreover, culturally grounded content may be especially challenging for judge models that lack sufficient cultural competence.

Reasoning Models: We did not include reasoning-enhanced LLMs in our evaluation, such as recent test-time reasoning models (e.g., DeepSeek-R1 (Guo et al., 2025) and similar architectures). While their inclusion could provide additional insights into the role of explicit reasoning in culturally grounded tasks, we intentionally restricted the experimental setting to standard instruction-tuned models. This choice allows us to keep the analysis focused and avoids introducing an additional experimental dimension that would require a separate and more extensive discussion.

Cross-Language Comparison: All benchmark instances are presented in the native language associated with the target culture. While cross-language analyses could represent an interesting extension – for example, by translating Definition Generation and Question Answering instances into English, or by performing backward translation experiments for Machine Translation – such analyses would require substantial additional annotation and validation efforts. Furthermore, many benchmark items are deeply tied to culture-specific linguistic and contextual nuances that may not transfer faithfully across languages. For these reasons, we leave cross-language evaluation as future work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíek, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

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)*, pages 11772–11817.

Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm, and Kristian Kersting. 2024. Occiglot at wmt24: European open-source large language models evaluated on translation. In *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. 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)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, Volume 11 - 2023.

Clifford Geertz. 1973. *The Interpretation of Cultures: Selected Essays*. Basic Books, New York.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, and Z. F. et al. Wu. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633638.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

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*, pages 14886–14909, Vienna, Austria. 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. *Preprint*, arXiv:2009.03300.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

Dieuwke Hupkes and Nikolay Bogoychev. 2025a. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *Preprint*, arXiv:2504.10356.

Dieuwke Hupkes and Nikolay Bogoychev. 2025b. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *arXiv preprint arXiv:2504.10356*.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Claude Lévi-Strauss. 1963. *Structural Anthropology*. Basic Books, New York. Originally published in French as *Anthropologie Structurale* (1958).

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan,

Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024. CNER: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8336–8351, Mexico City, Mexico. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong 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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Francesco Periti, Roksana Goworek, Haim Dubossarsky, and Nina Tahmasebi. 2025. Definition generation for word meaning modeling: Monolingual, multilingual, and cross-lingual perspectives. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26004–26024, Suzhou, China. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Qwen, An Yang, 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, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403.

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, Andre 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)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali

Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

UNESCO. 2002. UNESCO universal declaration on cultural diversity. In *Records of the General Conference, 31st Session, Paris, 15 October to 3 November 2001, Volume 1: Resolutions*, pages 61–65, Paris. UNESCO. Resolution 25, adopted at the 20th plenary meeting, 2 November 2001.

UNESCO Institute for Statistics. 2025. *2025 UNESCO Framework for Cultural Statistics, Part I: Concepts and Definitions*. UIS/2025/CUL/TD/11. UNESCO Institute for Statistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 17(6):4558–4577.

Lu Xu, Luca Moroni, and Roberto Navigli. 2026. Cultural and knowledge biases in llms through the lens of entity-aware machine translation. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8794–8812, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Hongbin Zhang, Kehai Chen, Xuefen Bai, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. 2026. Mitigating translationese bias in multilingual llm-as-a-judge via disentangled information bottleneck. *Preprint*, arXiv:2603.10351.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

A Category Filtering

In this section, we describe our guidelines for filtering Wikipedia pages based on their internal categorization in order to create our subset of Wikipedia pages that are culturally relevant. Categories are removed if they fall into one or more of the following cases (with examples given in Italian):

- **Proper names**: Categories consisting entirely of proper names – including first names, surnames, and demonyms – are removed, as individual names carry no standalone cultural content (e.g., *Prenomi italiani maschili*, *Cognomi italiani*).

- **Incomplete or unsourced pages**: Categories marked as *Stub* or unsourced are removed, as their pages lack sufficient content for reliable annotation or item generation.

- **Repeated and list categories**: Categories whose pages are structurally identical instances of the same schema – differing only in a year, location, or edition number – are removed as individual pages carry little standalone cultural informativeness. The same applies to categories that aggregate lists rather than individual topic pages (e.g., *Discografie di artisti italiani*). Note that this criterion does not apply to categories whose pages are unique cultural artifacts that happen to share a common attribute: for instance, *Film italiani del 1967* is retained because each page corresponds to a distinct cinematic work.

- **Administrative and bureaucratic categories**: Categories covering administrative

subdivisions, bureaucratic registries, or technical classifications with no direct cultural content are removed (e.g., *Prefissi telefonici italiani*, *Collegi elettorali del Regno d'Italia*, *Armoriali delle famiglie italiane*).

B Task Item Generation

In this section, we provide additional details and list the prompts used throughout our generation pipeline. Note that, while we show the prompts in English for simplicity, we use the relevant target language for prompting the Generator (e.g., Italian prompt for generating Italian task items).

In Figure 2, we show the prompt used to generate the QA task items through `gpt-oss-120b` (we omit the instructions for replying in a JSON format for brevity).

C Annotation Guidelines

We describe the guidelines provided to native-speaker annotators for validating and labeling culturally relevant Wikipedia pages. Given a Wikipedia page, the annotation procedure consists of three stages: (i) deciding whether the page is culturally relevant, (ii) assigning it to a cultural domain, and (iii) producing or validating its English translation. While (i) and (iii) are fully performed manually by native speakers, the cultural-domain assignment is performed automatically using ChatGPT and validated on a subset by native-speaker annotators. The nine UNESCO-derived cultural domains used to define cultural relevance are presented in Appendix C.2.

C.1 Translation of Culturally-relevant Wikipedia Page Titles

This section describes the pipeline and the guidelines provided to annotators for translating and categorizing culturally relevant Wikipedia items. The guidelines were developed to ensure consistency and reproducibility across annotations in different languages.

As a first step, annotators verify that each item is culturally relevant by checking whether it matches at least one of the nine cultural categories defined in Appendix C.2; items that do not match any category are discarded.

For each of the remaining item, annotators consult a set of reference sources and treat their translations as the gold standard, recording up to three different translations per item. Recommended sources include Wikipedia, Wikidata, Oxford English Dictionary (OED), the Encyclopædia Britannica, and Cambridge Encyclopedia, among others.

We define the translations types as the following (here, together with Italian-to-English (IT→EN) examples):

- **LITERAL**: direct lexical translation, i.e., translations word by word.
  - Scuola di Belle di Firenze → Academy of Fine Arts of Florence
  - Abbazia di Thélema → Abbey of Thelema
  - Anni di piombo → Years of Lead
  - Regione ecclesiastica Campania → Ecclesiastical Region of Campania
- **CONCEPTUAL**: meaning-preserving but non-literal rendering, i.e., translations that are not literal.
  - Auto blu → Official state car
  - Albero di maggio → Maypole
  - Enrosadura → Alpenglow
- **EXPLICITATION**: descriptive explanation of the entity, i.e., if the item can include an explanation of what it is as an appropriate translation, you should include that as well.
  - Supplì → Italian fried rice ball
- **LOANWORD**: untranslated borrowing into English, i.e., it does not have an English translation and it is included as-is in reputable sources.
  - Espresso
  - Aperol
  - Campari
- **NO TRANSLATION**: no established English rendering.

During the annotation process, two special cases take precedence over all other translation types: if the item appears unchanged in any reputable English lexical resource, then it must be tagged as LOANWORD, regardless of any other possible translation; conversely, if no translation can be found in any of the listed sources, the item is marked as NO TRANSLATION.

C.2 UNESCO Cultural Categories

Below, we provide definitions and examples for each of the 9 cultural categories used in the annotation process, drawn from the UNESCO Framework for Cultural Statistics. A Wikipedia page was retained in our dataset of culturally relevant pages only if the concept or named entity it describes could be assigned to at least one of the following categories.

Cultural Heritage. Artefacts, monuments, historic sites, built environments, museums, libraries, and archives carrying symbolic, historic, artistic, aesthetic, ethnological or social significance. Heritage is understood as an ongoing social process of meaning-making, not only a static remnant.

Examples include: named monuments, ruins, or archaeological sites (e.g., Colosseum, Pompeii); museums and galleries (e.g., Uffizi Gallery, Galleria Borghese); libraries and archives (e.g., Biblioteca Nazionale Centrale di Roma); historic urban areas or built complexes (e.g., Historic centre of Rome).

Performing Arts. Creative expressions presented live to an audience using voices, bodies or objects. Note: live music is classified under Music, not here.

Examples include: named theatre forms, dance forms, or performance genres (e.g., Commedia dell'arte, Ballet); named performance venues or companies (e.g., Teatro alla Scala); opera, musicals, circus, and puppetry (e.g., Neapolitan puppet theatre).

Visual Arts & Crafts. Creation and commercialisation of artworks and craftwork, including fine arts, photography, illustrations, and handmade craft traditions where manual skill is the primary component.

Examples include: named art movements or styles (e.g., Italian Renaissance painting, Futurism); named artists, sculptors, or photographers (e.g., Amedeo Modigliani); named craft traditions (e.g., Murano glassmaking, Majolica ceramics).

Books & Press. Publishing and dissemination of written content in physical or digital form, including books, newspapers, and magazines. Segmented by format, not subject. Web pages, blogs and wikis are excluded.

Examples include: named literary works or genres (e.g., Divine Comedy, Italian epic poetry); named authors or literary figures (e.g., Italo Calvino); named newspapers, magazines, or publishing houses (e.g., Corriere della Sera).

Audiovisual. Production, dissemination and exhibition of moving images with or without sound, across all formats and channels including streaming.

Examples include: named films, TV programmes, or animation (e.g., Rome, Open City); named film movements or styles (e.g., Italian neorealism); named directors, actors, or broadcasters (e.g., Federico Fellini, Rai).

Design & Creative Services. Activities providing a creative service or intermediary input, including graphic design, fashion design, architectural services, industrial design, interior design, and advertising.

Examples include: named designers, fashion houses, or architects (e.g., Giorgio Armani, Renzo Piano); named design movements or styles (e.g., Italian industrial design); named architectural works where creative practice is the primary subject (e.g., Pantheon as architecture).

Music. Creators, record companies, publishers, live events, radio, instruments, and merchandise. Live music belongs here, not under Performing Arts.

Examples include: named musicians, composers, or singers (e.g., Luciano Pavarotti); named musical genres or traditions (e.g., Italian opera, Canzone napoletana); named festivals, record labels, radio stations, or instrument makers (e.g., Sanremo Music Festival, Stradivari violin).

Cuisine & Food Culture. The practices, traditions, knowledge and symbolic meanings surrounding food preparation, transmission and consumption as a form of cultural identity. Includes named dishes, culinary techniques, traditions and regional culinary practices.

Examples include: named dishes or preparations (e.g., Carbonara, Risotto alla milanese); named food traditions or designations (e.g., Parmigiano Reggiano DOP, Slow Food movement).

Festivals. Themed public celebrations, including feasts, fairs, festivals, and traditional markets.

Examples include: named carnivals, feasts, or civic celebrations (e.g., Carnival of Venice, Palio di Siena); named cultural fairs or trade fairs with cultural identity (e.g., Salone del Mobile Milano);

named traditional markets tied to heritage practices (e.g., Infiorata di Noto).

D Translation Type Discussion

Section 3.2 reports the distribution of translation types across cultures. The largest category across all four languages is NO TRANSLATION, which accounts for the majority of items in Italian, German, and Spanish, and a substantial share in Hindi. This predominance is a direct consequence of how cultural specificity interacts with lexicalization in English: items deeply embedded in a target culture and unfamiliar to international audiences, such as regional dishes, local institutions, vernacular festivals, or minor historical figures, typically lack an established English rendering and are preserved in their original form in encyclopedic and lexicographic sources. The high proportion of NO TRANSLATION items thus serves as an indirect validation of the cultural grounding of CULTURE-WIKI: the more culturally specific an item, the less likely it is to have travelled into the English lexicon, and the more frequently it appears in this category.

In contrast, LOANWORD items are by definition lexicalized in English: terms such as *espresso* and *bruschetta* from Italian, *Bauhaus* and *ablaut* from German, or *yoga* and *biryani* from Hindi, have entered English usage as established borrowings. Their presence in standard English dictionaries and reference works means that models encounter them frequently during pretraining, which explains the consistently higher ETA scores on LOANWORD items reported in Table 6: the correct surface form is already in the model's vocabulary, requiring neither translation nor cultural retrieval.

LITERAL and CONCEPTUAL translations occupy an intermediate position, requiring the model to produce a target English form rather than retrieve a memorised one. The higher proportion of these categories in European languages reflects greater historical exchange between those cultures and the English-speaking world. Hindi, by contrast, shows comparatively more LOANWORD and EXPLICITATION cases, suggesting a different pattern of cultural exportation in which a few widely diffused terms coexist with a larger pool of items that require descriptive rendering when transferred into English.

E Keyword List for Culture Tree Building

The category-graph traversal (Section 3.1) is guided by two manually compiled keyword lists per language: a *target-culture list* containing morphological variants of the country name and demonym, and an *other-culture list* containing country names, demonyms, ethnonyms, and broad region names for all other nations. The lists for Italian are provided below.

Target-culture List

`Italia`, `italiano`, `italiana`, `italiani`, `italiane`, `italico`, `italic`, `italiot`, `Italiae`

Other-culture list

Countries and demonyms `Afghanistan`, `afgano`, `afgana`, `afgani`, `Albania`, `albanese`, `albanesi`, `Algeria`, `algerino`, `algerina`, `algerini`, `Angola`, `angolano`, `angolana`, `angolani`, `Arabia Saudita`, `saudita`, `sauditi`, ... (144 more countries)

Ethnicities `afroamericano`, `afroamericani`, `ispanico`, `ispanici`, `latino`, `latina`, `latini`, `asioamericano`, `asioamericani`, `nativo americano`, `nativi americani`, `indigeno`, `indigeni`, `aborigeno`, `aborigeni`, `maori`, `celtico`, `celtica`, `celtici`, `nordico`, `nordici`, `slavo`, `slava`, `slavi`, `germanico`, `germanica`, `germanici`, `anglosassone`, `anglosassoni`, `scandinavo`, `scandinava`, `scandinavi`, `baltico`, `baltici`, `basco`, `baschi`, `catalano`, `catalani`, `curdo`, `curda`, `curdi`, `rom`, `romani`, `ebreo`, `ebrea`, `ebrei`, `arabo`, `araba`, `arabi`, `persiano`, `persiana`, `persiani`, `turco`, `berbero`, `berberi`, `beduino`, `beduini`, `punjabi`, `bengalese`, `bengalesi`, `tamil`, `han`, `tibetano`, `tibetani`, `uiguro`, `uiguri`, `mongolo`, `pashtun`, `zulu`, `yoruba`, `igbo`, `swahili`, `amarico`, `amarici`, `somalo`

Regions `Scandinavia`, `scandinavo`, `scandinavi`, `Balcani`, `balcanico`, `balcanici`, `Iberia`, `iberico`, `iberici`, `Medio Oriente`, `Vicino Oriente`, `Estremo Oriente`, `Asia orientale`, `Asia sudorientale`, `Asia meridionale`, `Asia centrale`, `Africa occidentale`, `Africa orientale`, `Africa settentrionale`, `Africa subsahariana`, `Sahara`, `Maghreb`, `Arabia`,

Language	Cohen's κ
Italian	0.895
German	0.849
Spanish	0.886
Hindi	0.474

Table 7: Inter-annotator agreement (Cohen's κ) on the `no_culture` discard decision, computed over 80 items per language annotated independently by two expert native-speaker annotators.

Language	Acceptance
Italian	93.5
German	87.8
Spanish	94.0
Hindi	72.5

Table 8: Manual quality validation of generated QA items per culture-targeted language, performed by expert native-speaker annotators.

`arabico`, `arabica`, `Anatolia`, `Levante`, `Mesopotamia`, `Caucaso`, `Caraibi`, `America Latina`, `America meridionale`, `America centrale`, `America settentrionale`, `Mesoamerica`, `Polinesia`, `Melanesia`, `Micronesia`, `Oceania`

F Inter-annotator Agreement on Cultural Relevance

In Table 7, we report inter-annotator agreement using Cohens κ for all four languages. For each language, a random subset of 20 instances was independently annotated by two native-speaker annotators.

The results indicate consistently high agreement across languages ($\kappa \geq 0.85$, corresponding to "almost perfect agreement"), suggesting strong reliability of the annotation scheme. However, Hindi shows comparatively lower agreement than the other languages. This discrepancy may be attributed to the annotators originating from different regions of India, which may reflect systematic regional variation in linguistic interpretation and cultural framing. This finding highlights potential limitations in treating cultural and linguistic phenomena as homogeneous at the national level.

G Quality Validation of Generated QA Items

To validate the QA items generated by `gpt-oss-120b` (see Section 3.3), we manually inspect a random subset of items for each culture-targeted language. The validation is performed by expert native-speaker annotators following a unified set of guidelines designed to assess whether each question–answer pair is well-formed, faithful to the source context, and suitable as an evaluation item.

A question is accepted only if it is grammatically correct, well-posed with no false presuppositions (e.g., the Italian question *"Quale dei due comuni della provincia di Latina ha ricevuto la Bandiera arancione?"* [*Which of the two municipalities in the province of Latina received the Orange Flag?*] is rejected because the province of Latina contains 33 municipalities, not two), independent of external phrasing such as *"according to the text"* (e.g., a question about an orchestra asking which one is *"indicato nel testo"* [*indicated in the text*] is rejected for explicitly referring the model back to the passage), unambiguous within the provided context, does not reveal the answer (e.g., the German question *"Welches Medium bedient hr4 im hessischen Rundfunk?"* [*Which medium does hr4 serve in Hessian broadcasting?*] is rejected because the question already implies the answer *Hörfunk* [*radio*], given that the term *Rundfunk* (broadcasting) together with the station name *hr4* leaves little room for an alternative answer), and asks about a single fact. An answer is accepted only if it is grammatically correct, factually accurate with respect to the provided context, properly scoped to directly address the question and unique. Importantly, expert native-speaker annotators are instructed that items should be answerable by an LLM rather than by a human: a question that is difficult for a human reader is not on its own a reason to discard it.

Table 8 reports the acceptance rates per language. The European languages show consistently high agreement (above 87%), with Italian and Spanish reaching approximately 94%. Hindi shows a notably lower acceptance rate (72.5%), in line with the broader pattern observed throughout our evaluation, where low-resource languages remain more challenging across the entire CULTURE-WIKI pipeline.

H Benchmark Statistics across Tasks

After the extraction of Wikipedia pages and the completion of the cultural annotation process, task generation was applied to the extracted pages (Section 3.1 and Section 3.3). Each page was targeted

to generate benchmark items for the three tasks whenever possible. Table 9 reports the number of pages leveraged to create machine translation (MT) benchmark items. Not all pages could be used to generate MT samples, since some pages do not contain references to the page title outside the introduction section.

The reported statistics show that low-resource languages, such as Hindi, exhibit a lower percentage of usable pages for MT item creation, likely due to the comparatively lower coverage and quality of Wikipedia annotations in those languages. In contrast, languages such as Italian achieve almost complete coverage.

Table 10 reports the number of pages used to create question answering (QA) items. In this case, a high proportion of pages can be leveraged to generate QA samples (with German representing the lowest percentage at 75%). This higher usability is mainly due to the larger number of hyperlinks available in culturally annotated pages and culturally relevant pages, denoted as P_{rel} and defined in Section 3.1.

For both the MT and QA tasks, after item generation we retained a maximum of two benchmark items per page in order to avoid over-representation of some pages.

Regarding the definition generation task, every Wikipedia page contains an introduction section; therefore, we generated exactly one benchmark item for each culturally annotated page.

I Task Prompts

Figures 3, 5, and 4 report the English versions of the prompts used for the machine translation, definition generation tasks, and question answering, respectively. All prompts were also manually translated into the four target languages.

J Further Quantitative Results Across Prompt Types and Metrics

In this section, we report LLM performance across different prompting strategies and evaluation metrics. Tables 12 and 11 present the machine translation results measured with xCOMET and ETA scores. Tables 14 and 13 report the definition generation results evaluated with BERTScore and ROUGE-L, showing broadly consistent rankings across the two metrics. Finally, Table 15 presents substring Exact Match (EM) scores across prompts.

Overall, the results indicate that models tend to perform better when prompted in English, highlighting an interesting cross-lingual transfer effect when combining Englisha high-resource language-with the target languages.

K Qualitative Analysis: Representative Model Outputs

In this section, we further discuss the qualitative analysis of failure modes in CULTURE-WIKI items. We focalize our inspection on four models: *Minerva-7B* (Italian-English model), *ANITA-8B* (Italian-adapted model), *Llama-8B* (English-first model), and *Salamandra-7B* (Spanish-English model). For each task, we identify recurring failure modes and illustrate them with representative examples drawn from Italian and Spanish.

In Table 16 we can see machine translation failure examples, where we see that models leave the source-language form untranslated despite an established English rendering: SALAMANDRA preserves the Italian DOP designation *Peperone di Carmagnola* verbatim rather than producing *Carmagnola pepper*, and ANITA retains the Spanish word *fiesta* in *Fiesta del PCE* instead of translating it. In other cases, the model translates an entity that has no canonical English form and should be preserved as-is: Minerva renders the Calabrian Easter cake *Sguta* as "sausage", while SALAMANDRA decomposes the UNESCO-recognised *Camino Real de Tierra Adentro* into *Royal Road of the Interior Land*. A third pattern is the replacement of a source entity with a culturally adjacent but non-equivalent referent: ANITA renders the Italian basketball league *Divisione Nazionale A* as *Serie A Division*, conflating it with the more internationally visible football league, and Llama substitutes the Aragonese card game *Guiñote* with the unrelated Anglo-American *Gin rummy*.

In Table 17 we show an inspection of DG outputs revealing consistent failure patterns across all four model families. In many cases, models fabricate a wholly incorrect identity for the target entity: ANITA describes the 1994 Grignani song *La mia storia tra le dita* as "a book by Angela BlAusser", and Minerva turns the traditional Spanish dish *Sanjacobo* into "a Spanish city in Asturias". In other cases, the model preserves the correct entity type but assigns the wrong specifics, often through phonetic or geographic drift: SALAMANDRA describes the

Translation Type	Italian	German	Spanish	Hindi
LITERAL	468	363	461	217
CONCEPTUAL	110	173	55	38
EXPLICITATION	0	1	7	52
LOANWORD	49	7	10	81
NO TRANSLATION	1,281	751	611	221
Num. Pages	1,908 (99%)	1,294 (75%)	1,144 (52%)	609 (31%)
Total MT items	3,308	2,472	1,947	1,009

Table 9: Statistics of the pages used to generate machine translation (MT) benchmark items, grouped by translation type and language. Percentages indicate the proportion of culturally annotated pages that were successfully leveraged to create MT items for each language. The final row reports the total number of MT benchmark items generated per language.

Language	Num. Pages	Num. QA Items
Italian	1,917 (99%)	3,834
German	1,313 (75%)	2,626
Spanish	2,152 (97%)	4,303
Hindi	1,752 (89%)	3,488

Table 10: Statistics of the pages used to generate question answering (QA) benchmark items, together with the total number of QA instances included in the final benchmark for each language.

Valle d'Aosta DOP cheese *Fontina* as "Fontana", a different cheese, while Llama turns the 2912-metre Pre-Pyrenean massif *Cotiella* into "a small town in the province of Zaragoza". A third pattern is the generation of outputs so generic that all culturally specific content is lost: SALAMANDRA describes the *Embajada de España en Japón* as "the diplomatic representation of Spain in the country of Japan", and Llama reduces the noble Monferrato family *Viazzi* to "an Italian surname diffused above all in the Piedmont region". Table 17 reports the full set of examples.

Finally, in Table 18, we show QA outputs revealing recurring patterns of confident but factually wrong answers across all four model families. A frequent pattern is the substitution of an obscure correct entity with a more internationally visible one: asked which cyclist nicknamed "Locomotiva umana" won the first professional edition of the *Coppa Caivano*, Minerva confidently names *Eddy Merckx* rather than the Italian rider Learco Guerra; asked in which of Dante's treatises *honorificabilitudinitate* appears, SALAMANDRA names the *Divina Commedia* rather than the less canonical *De Vulgari Eloquentia*. Other failures stem from confident fabrication of plausible-sounding details: Llama reports that the arch along the *Via Postumia* celebrates the Roman family of the *Vitruvi* instead of the *Gavia*, and invents *Carta Sevilla* as the traditional name for the unpromulgated *Constitución española de 1856*, where the canonical expression is *non nata*. A third pattern is named-entity confusion within the correct category: ANITA names *Alfonso Calderón* as the minister of the interior who linked the *Atentado del restaurante El Descanso* to Islamic jihad (the correct answer is José Barrionuevo), and SALAMANDRA attributes the long-running *Copa del Rey de vela* sponsorship to *Perfumerías Avenida* rather than to the Puig family business. Table 18 reports the full set of examples.

Model	Italian		German		Spanish		Hindi	
	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}
Minerva-7B-instruct-v1.0	57.1	53.6	31.7	28.6	34.5	35.1	5.8	9.8
Llama-3.1-8B-Instruct	58.5	60.2	43.2	46.8	49.8	48.4	27.0	28.9
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	56.0	56.6	42.5	44.7	44.9	45.7	27.0	28.6
Qwen2.5-1.5B-Instruct	48.1	46.1	38.3	36.5	39.0	36.3	11.7	19.4
Llama-3.1-70B-Instruct	67.2	67.1	54.2	54.1	59.2	59.2	37.2	36.9
salamandra-7b-instruct	58.5	54.7	46.5	45.3	49.7	47.8	11.9	10.0
Phi-4-mini-instruct	52.8	57.9	38.7	42.7	43.7	43.4	21.8	24.1
Qwen2.5-7B-Instruct	62.0	61.0	47.9	47.8	54.8	53.7	28.4	27.7
Qwen2.5-3B-Instruct	56.7	56.6	46.3	44.5	47.9	45.8	20.4	20.6
Qwen2.5-14B-Instruct	63.7	62.4	47.9	47.1	54.6	53.7	34.2	33.6
phi-4	64.8	65.7	50.6	50.4	55.9	55.8	37.7	35.7
occiglot-7b-es-en-instruct	56.2	50.6	42.4	41.3	51.5	46.9	15.2	15.6
occiglot-7b-de-en-instruct	59.3	57.6	46.9	45.2	49.7	47.6	17.0	16.9
SmolLM2-1.7B-Instruct	42.2	40.9	33.9	28.9	34.5	33.2	7.4	8.0
Llama-3.2-3B-Instruct	52.6	51.0	39.9	40.4	41.1	39.8	24.2	21.5
EuroLLM-1.7B-Instruct	54.8	50.2	39.4	36.3	43.2	39.0	22.6	24.4
Average	56.9	55.8	43.1	42.5	47.1	45.7	21.8	22.6

Table 11: The table reports performance of 16 LLMs across four cultural settings (Italian, German, Spanish, Hindi) for machine translation (MT, ETA score). We report prompts in English and the target culture language. Cultural groups are separated into column blocks by language community.

Model	Italian		German		Spanish		Hindi	
	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}
Minerva-7B-instruct-v1.0	77.7	76.6	74.0	66.6	72.2	71.1	24.1	20.7
Llama-3.1-8B-Instruct	85.7	83.1	88.3	83.7	85.3	84.0	52.0	49.1
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	84.4	85.4	87.0	88.7	82.1	83.9	50.9	53.7
Qwen2.5-1.5B-Instruct	79.6	78.0	86.2	84.7	80.8	79.7	30.3	24.5
Llama-3.1-70B-Instruct	88.4	87.2	92.0	91.7	88.4	87.9	56.0	53.9
salamandra-7b-instruct	84.2	77.2	87.9	84.4	83.7	79.6	28.7	27.6
Phi-4-mini-instruct	82.9	84.6	88.4	88.6	84.5	84.4	44.1	31.2
Qwen2.5-7B-Instruct	87.9	87.7	91.4	91.3	88.0	87.5	55.6	54.7
Qwen2.5-3B-Instruct	85.9	85.6	90.3	89.9	86.2	85.9	47.2	43.7
Qwen2.5-14B-Instruct	88.6	88.4	91.4	90.7	88.2	88.2	61.0	59.4
phi-4	89.1	88.8	92.0	91.7	88.8	89.0	62.1	61.6
occiglot-7b-es-en-instruct	58.8	56.8	57.1	50.2	64.5	62.6	19.7	18.3
occiglot-7b-de-en-instruct	74.2	74.7	79.9	80.4	77.5	76.8	27.6	25.7
SmolLM2-1.7B-Instruct	66.6	57.9	72.0	62.9	71.5	68.3	23.6	19.5
Llama-3.2-3B-Instruct	80.7	77.7	86.0	80.9	80.6	79.0	41.5	39.5
EuroLLM-1.7B-Instruct	78.3	69.4	83.6	76.7	78.5	67.4	42.0	35.2
Average	80.8	78.7	84.2	81.4	81.3	79.7	41.6	38.7

Table 12: The table reports performance of 16 LLMs across four cultural settings (Italian, German, Spanish, Hindi) for machine translation (MT, xComet score). We report prompts in English and the target culture language. Cultural groups are separated into column blocks by language community.

System prompt

You are an expert in target culture. Your task is to generate cultural commonsense questions in target. Questions must test implicit cultural knowledge – that is, what a target citizen would take for granted but a non-target person would not know. They should not test memorisation of encyclopaedic facts or dates, but rather understanding of practices, customs, associations, and meanings that belong to lived target culture.

User prompt

Context: "{root_title}" in relation to "{linked_title}"
{context_block}
=====
TASK: Generate up to number question-answer pairs in target language that satisfy ALL of the following criteria:

WHAT THEY MUST TEST
- Knowledge that a target person (or an expert in the specific local culture) would take for granted
- Cultural associations, practical uses, contexts of use, social origins, symbolic meanings, etc.
- Preferably: facts that contradict what a person unfamiliar with target culture would assume by default (e.g., by analogy with foreign cultures or overgeneralisation)
- Questions must be difficult for those unfamiliar with target culture, but obvious for those who know it

QUESTION FORMAT
- Open-ended: not A/B, not yes/no, not multiple choice
- The answer must not appear lexically in the question
- Only one correct answer: if multiple valid answers exist, discard the question
- The question must directly concern the relationship between "{root_title}" and "{linked_title}" in the target cultural context
- Questions should be concise

ANSWER FORMAT
- Short and specific: an entity, a term, a practice, a social group; not an explanation
- Verifiable in the provided text; do not invent facts not supported by the context!

WHAT TO AVOID
- Trivial questions about dates, numbers, founders' names, generic geographic locations
- Questions answerable without cultural knowledge (e.g., by logical inference)
- Questions whose answer is obvious from the entity name itself
- Overly generic questions that do not require knowledge of the specific cultural context

Return ONLY a JSON with the following structure, with no preamble:
{output_format_block}

Figure 2: Prompt used for open-ended QA item generation. {context_block} contains the root page introduction and sections where the linked item appears, plus the linked item introduction; {output_format_block} contains the JSON schema for the output.

System prompt
You are an expert translator from {target_language} to English. Translate only the sentence without any additional explainations.

User prompt
{source_sentence}

Figure 3: Prompt used in our experiments on the task of sentence-level MT.

System prompt
You are an expert in {target_language}. Given the title of a Wikipedia page, generate a descriptive sentence in {target_language}.
The sentence should introduce the topic, and tell the nonspecialist reader what or who the subject is, and often when or where.

If its subject is definable, then the first sentence should give a concise definition: where possible, one that puts the article in context for the nonspecialist.
Similarly, if the title is a specialized term, provide the context as early as possible.

For topics notable for only one reason, this reason should usually be given in the first sentence.

Do not overload the first sentence by describing everything notable about the subject.
Instead, spread the relevant information out over the entire lead.
Avoid cluttering the first sentence with a long parenthetical containing items like alternative spellings and pronunciations; these can make the sentence difficult to read

User prompt
Page title: {article_title}
Sentence:

Figure 4: Prompt used in our experiments on the task of Encyclopedic-style Definition Generation. In this prompt, we used the same guidelines that Wikipedia gives to the annotators.

System prompt
You are an expert question-answering assistant for {target_language}. Respond concisely and directly.

User prompt
{question}

Figure 5: Prompt used in our experiments on the task of open-ended QA.

Model	Italian		German		Spanish		Hindi	
	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}
Minerva-7B-instruct-v1.0	19.4	23.9	17.6	15.9	20.5	15.4	1.0	0.5
Llama-3.1-8B-Instruct	22.8	21.8	23.1	21.7	29.7	18.1	3.3	2.8
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	23.3	22.7	22.2	20.2	29.6	10.7	1.4	0.9
Qwen2.5-1.5B-Instruct	14.9	15.5	18.8	15.0	19.0	21.9	1.1	0.4
Llama-3.1-70B-Instruct	29.1	28.9	27.6	27.8	34.1	30.8	5.6	4.5
salamandra-7b-instruct	23.4	21.8	18.7	22.2	28.2	25.5	1.5	0.8
Phi-4-mini-instruct	17.5	16.9	19.6	15.7	26.4	21.8	1.8	1.5
Qwen2.5-7B-Instruct	21.9	20.7	23.5	22.1	30.1	16.4	3.4	3.2
Qwen2.5-3B-Instruct	21.0	16.9	22.5	18.0	29.0	16.8	3.0	0.5
Qwen2.5-14B-Instruct	23.3	22.5	24.3	24.4	30.9	22.0	3.2	3.2
phi-4	20.5	15.6	19.8	20.9	27.4	10.9	4.0	2.4
occiglot-7b-es-en-instruct	11.1	11.0	11.3	11.4	20.9	13.1	0.6	0.4
occiglot-7b-de-en-instruct	15.7	13.6	21.2	21.8	21.3	13.0	1.1	0.6
SmolLM2-1.7B-Instruct	14.2	10.8	14.6	8.7	21.8	13.3	1.1	1.1
Llama-3.2-3B-Instruct	18.1	19.3	18.8	19.8	25.6	22.3	1.7	1.4
EuroLLM-1.7B-Instruct	15.6	15.1	17.1	13.5	23.3	20.3	1.3	2.2
Average	19.5	18.6	20.0	18.7	26.1	18.3	2.2	1.6

Table 13: The table reports performance of 16 LLMs across four cultural settings (Italian, German, Spanish, Hindi) for definition generation (MT, RougeL score). We report prompts in English and the target culture language. Cultural groups are separated into column blocks by language community.

Model	Italian		German		Spanish		Hindi	
	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}
Minerva-7B-instruct-v1.0	83.9	84.8	82.2	81.3	83.1	81.7	77.8	90.0
Llama-3.1-8B-Instruct	84.5	84.1	83.3	82.8	84.6	82.6	87.3	91.1
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	84.7	84.5	83.1	82.6	84.7	81.4	71.7	79.4
Qwen2.5-1.5B-Instruct	82.8	82.3	82.5	80.7	83.3	82.4	73.7	87.2
Llama-3.1-70B-Instruct	86.0	85.7	84.4	84.4	86.0	85.2	91.6	91.1
salamandra-7b-instruct	84.8	84.2	83.2	83.2	85.1	84.2	75.3	85.3
Phi-4-mini-instruct	83.2	83.0	82.4	81.5	84.0	83.0	83.9	88.9
Qwen2.5-7B-Instruct	83.9	83.1	82.5	81.9	84.5	82.0	91.1	91.2
Qwen2.5-3B-Instruct	84.2	82.7	82.8	81.5	84.6	81.4	91.2	86.6
Qwen2.5-14B-Instruct	84.7	84.5	83.1	83.0	84.9	83.6	91.3	91.5
phi-4	84.2	83.1	82.7	82.9	84.5	80.9	88.4	79.8
occiglot-7b-es-en-instruct	81.0	80.9	80.0	80.1	82.8	80.6	86.6	87.4
occiglot-7b-de-en-instruct	81.8	81.6	82.3	82.5	82.3	80.4	82.1	87.9
SmolLM2-1.7B-Instruct	81.9	80.7	80.9	79.1	82.4	80.2	82.5	90.1
Llama-3.2-3B-Instruct	83.0	83.2	81.9	82.1	83.3	82.3	80.5	90.0
EuroLLM-1.7B-Instruct	82.9	82.3	82.1	80.6	83.3	82.2	74.9	91.2
Average	83.6	83.2	82.5	81.9	84.0	82.1	83.1	88.0

Table 14: The table reports performance of 16 LLMs across four cultural settings (Italian, German, Spanish, Hindi) for definition generation (DG, BertScore). We report prompts in English and the target culture language. Cultural groups are separated into column blocks by language community.

Model	Italian		German		Spanish		Hindi	
	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}	P_{en}	P_{trg}
Minerva-7B-instruct-v1.0	23.5	22.8	3.2	2.9	4.6	4.0	2.1	2.4
Llama-3.1-8B-Instruct	16.3	14.1	15.1	14.3	16.4	6.4	13.7	13.4
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	15.5	15.1	15.4	13.9	15.7	6.1	8.3	8.9
Qwen2.5-1.5B-Instruct	5.2	5.9	3.4	3.4	4.6	4.3	2.4	2.7
Llama-3.1-70B-Instruct	29.7	28.0	31.1	28.6	32.1	23.9	27.4	26.6
salamandra-7b-instruct	15.8	14.5	15.0	13.9	21.2	21.3	1.6	1.7
Phi-4-mini-instruct	9.0	8.9	9.9	8.4	8.8	7.4	4.5	3.6
Qwen2.5-7B-Instruct	13.2	12.8	10.4	9.0	13.2	11.9	5.8	6.3
Qwen2.5-3B-Instruct	8.0	8.3	5.6	5.3	8.1	5.9	3.4	3.5
Qwen2.5-14B-Instruct	19.7	19.3	15.2	16.3	21.2	19.0	12.4	12.6
phi-4	22.9	23.2	24.1	24.0	24.5	24.1	20.2	19.3
occiglot-7b-es-en-instruct	10.6	10.6	8.6	7.1	17.9	17.2	2.9	2.7
occiglot-7b-de-en-instruct	10.3	10.2	18.5	16.9	9.0	10.7	2.0	2.4
SmolLM2-1.7B-Instruct	3.6	3.4	2.0	1.9	2.2	2.3	1.0	1.9
Llama-3.2-3B-Instruct	7.7	5.3	7.1	5.5	7.6	2.8	7.0	6.6
EuroLLM-1.7B-Instruct	10.9	9.3	7.4	7.7	8.7	5.7	5.2	5.4
Average	13.9	13.2	12.0	11.2	13.5	10.8	7.5	7.5

Table 15: The table reports performance of 16 LLMs across four cultural settings (Italian, German, Spanish, Hindi) for question answering (QA, exact match score). We report prompts in English and the target culture language. Cultural groups are separated into column blocks by language community.

Lang.	Model	Source / Reference	Model output	ETA	xCOMET
IT	SALAMANDRA	*Peperone di Carmagnola* / Carmagnola pepper	"The **'Peperone di Carmagnola'** production area is identified…"	0	1.00
IT	Minerva	*Sguta* / Sguta	"…the **sausage** is ready to be eaten."	0	0.81
IT	ANITA	*Divisione Nazionale A* / Divisione Nazionale A	"With the birth of the **Serie A Division**, ratified by the FIP…"	0	0.95
ES	ANITA	*Fiesta del PCE* / Fiesta of the PCE	"The first **PCE party fiesta** took place…"	0	0.98
ES	SALAMANDRA	*Camino Real de Tierra Adentro* / Camino Real de Tierra Adentro	"…became the **Royal Road of the Interior Land**."	0	0.91
ES	Llama	*Guiñote* / Guiñote	"**Gin rummy** is widely spread in the Spanish communities of Aragon…"	0	0.81

Table 16: Representative MT failure cases for Italian and Spanish; the entity-level output is shown in **bold**. Rows are colour-coded by language.

Lang.	Model	Title / Gold Gist	Model Output	R-L	BS
IT	ANITA	*La mia storia tra le dita* / 1994 song by Gianluca Grignani	"a book by Angela BlAusser, a collection of stories about the life of a woman with extraordinary tactile sensitivity…"	0.03	0.79
IT	SALAMANDRA	*Fontina* / Valle d'Aosta DOP cheese	"**Fontana** is an Italian semi-hard cheese produced with cow's milk."	0.00	0.79
IT	Llama	*Viazzi* / noble Monferrato family	"*Viazzi* is an Italian surname, diffused above all in the Piedmont region."	0.11	0.82
ES	Minerva	*Sanjacobo* / traditional ham-and-cheese fried dish	"*Sanjacobo* is a Spanish city located in the autonomous community of Asturias, noted for its rich cultural heritage…"	0.02	0.78
ES	Llama	*Cotiella* / Pre-Pyrenean limestone massif, 2912 m	"*Cotiella* is a small town and municipality of the province of Zaragoza, in the autonomous community of Aragón…"	0.12	0.81
ES	SALAMANDRA	*Embajada de España en Japón* / Spanish embassy in Japan	"the diplomatic representation of Spain in the country of Japan."	0.00	0.87

Table 17: Representative DG outputs for Italian and Spanish. Rows are colour-coded by language.

Lang.	Model	Question	Gold	Model Output
IT	Minerva	Which cyclist nicknamed "Locomotiva umana" won the first professional edition of the Coppa Caivano?	Learco Guerra	"**Eddy Merckx** won the Coppa Caivano in its first professional edition in 1968."
IT	SALAMANDRA	In which treatise by Dante does *honorificabilitudinitate* appear?	De Vulgari Eloquentia	"In the **Divina Commedia**."
IT	Llama	Which Roman family is celebrated by the arch along the Via Postumia?	Gavia	"The arch celebrated the family of the **Vitruvi**."
ES	ANITA	Which minister of the interior linked the El Descanso attack to Islamic jihad?	José Barrionuevo	"**Alfonso Calderón**."
ES	Llama	What expression is used for the unpromulgated Constitution of 1856?	non nata	"The expression is ***Carta Sevilla***."
ES	SALAMANDRA	Which Spanish family perfumery and fashion business sponsored the Copa del Rey de vela (1985–2006)?	Puig	"**Perfumerías Avenida**."

Table 18: Representative QA outputs for Italian and Spanish. Rows are colour-coded by language.