Does Mapo Tofu Contain Coffee?
Probing LLMs for Food-related Cultural Knowledge

Anonymous ACL submission

Abstract

Recent studies have highlighted the presence of cultural biases in Large Language Models (LLMs), yet often lack a robust methodology to dissect these phenomena comprehensively. Our work aims to bridge this gap by delving into the FOOD domain—a universally relevant yet culturally diverse aspect of human life. We introduce FMLAMA, a multilingual dataset centered on food-related cultural facts and variations in food practices. We analyze LLMs across various architectures and configurations, evaluating their performance in both monolingual and multilingual settings. By leveraging templates in six different languages, we investigate how LLMs interact with language-specific and cultural knowledge. Our findings reveal that (1) LLMs demonstrate a pronounced bias towards food knowledge prevalent in the United States; (2) Incorporating relevant cultural context significantly improves LLMs' ability to access cultural knowledge; (3) The efficacy of LLMs in capturing cultural nuances is highly dependent on the interplay between the probing language, the specific model architecture, and the cultural context in question. This research underscores the complexity of integrating cultural understanding into LLMs and emphasizes the importance of culturally diverse datasets to mitigate biases and enhance model performance across different cultural domains.

1 Introduction

Asking a French person for the recipe of *Beef Bourguignon* might yield an immediate and precise response, while the same query might pose challenges to a Chinese individual unless posed as 勃艮第牛肉" (its literal translation). In China, the dish is commonly referred to by its broader description, 红酒炖牛肉" (*Red Wine Stewed Beef*), highlighting the main ingredients and cooking technique, albeit without specifying a regional origin. Employing "法式红酒炖牛肉" (*French-style Red Wine Stewed Beef*) with an adjectival description can indicate adherence to French culinary traditions, illustrating how cultural and linguistic nuances influence knowledge transmission. This variability underscores the challenges language models face in navigating cross-cultural culinary contexts.

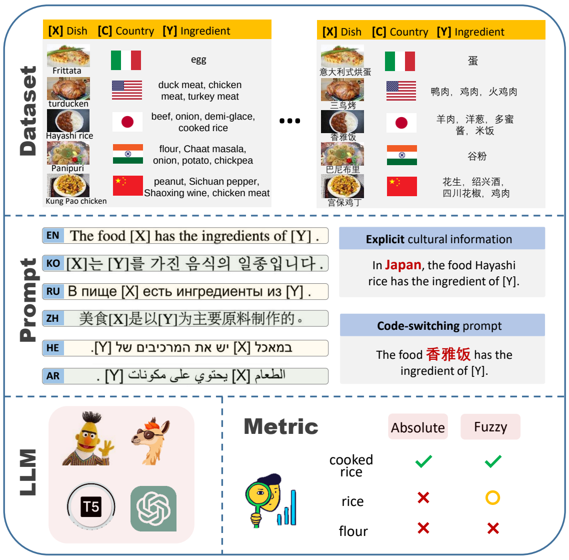


Figure 1: Summary of the various aspects of our work.

Trained on vast datasets, Large Language Models (LLMs) encode a wide array of knowledge, from general facts to domain-specific insights (Kassner et al., 2021; Elazar et al., 2021; AlKhamissi et al., 2022; Wu et al., 2023; Petroni et al., 2019a; Meng et al., 2022; Roberts et al., 2023). This diversity is crucial for their adaptability across various linguistic tasks. However, it also predisposes them to biases, including gender biases (Savoldi et al., 2021; Kaneko et al., 2022), and belief biases (Søgaard, 2021; González et al., 2021; Lent and Søgaard, 2021), potentially leading to the propagation of misinformation and distorting information retrieval. Particularly, *cultural* bias presents a significant challenge, as LLMs tend to favor certain cultures, perpetuate stereotypes, and

Datasets	Format	Topic	Construction method	Size
GeoMLama (Yin et al., 2022)	Manual Template	Geo-Diverse Concept	Manually curated	3125
FORK (Palta and Rudinger, 2023)	CommonsenseQA	Culinary culture	Manually curated	184
StereoKG (Deshpande et al., 2022)	Triplet knowledge	Stereotypes about religion and ethnicity	Automatically constructed	4722
Candle (Nguyen et al., 2023)	Sentences	Several cultural facets (food, drinks, clothing, traditions, rituals, behaviors)	Automatically constructed	47360
FmLAMA (ours)	Triplet knowledge	Food domain	Automatically constructed	33601

Table 1: Comparison of cultural knowledge datasets. Size is in number of instances.

vary in cultural knowledge familiarity (Cao et al., 2023; Deshpande et al., 2022; Yin et al., 2022). Our research scrutinizes the LLM's capacity to access cultural knowledge, guided by research questions on the direction of cultural bias, the impact of cultural context, and the role of language in accessing culturally relevant information.

Contributions. This study introduces several key advancements in the understanding and evaluation of the LLM's ability to access cultural knowledge. Figure 1 provides an overview of our work's various aspects.

- We present FmLAMA, a pioneering dataset focused on the food domain, which is inherently rich in cultural diversity (Cao et al., 2024). This dataset is a multifaceted tool for probing LLMs across cultures and languages.
- We propose novel metrics designed to assess LLMs' ability to accurately and sensitively probe for cultural knowledge, incorporating both absolute- and fuzzy-match techniques.
- We analyze the impact of integrating cultural context and language specificity in prompts, offering insights to optimize LLMs for equitable cross-cultural knowledge retrieval.

Our methodology for automated collection of cultural knowledge corpora extends the analysis potential in other domains, broadening the scope of research on cultural biases in LLMs.

2 Related Work

Cultural knowledge datasets. Cultural knowledge, encompassing the customs, beliefs, traditions, and practices of a culture, is crucial yet challenging to encapsulate. While some researchers focus on manually curating cultural knowledge datasets, others evaluate LLMs' performance on culturally related tasks. Yin et al. (2022) and Palta and Rudinger (2023) have developed benchmarks such as geo-diverse prompts and food-custom datasets (FORK) to probe cultural biases in commonsense reasoning systems. However, manual dataset construction is inefficient and hard to scale, prompting a shift towards automated methods. For instance, StereoKG (Deshpande et al., 2022) offers a scalable knowledge graph that blends cultural knowledge with stereotypes, and Candle (Nguyen et al., 2023) extracts cultural commonsense knowledge from the web, organizing it into clusters. Despite these advances, the variability in data representation—from sentences to triplets using `OpenIE`—poses challenges for consistency and noise control in knowledge probing. Keleg and Magdy (2023) aims to mitigate this by selecting culturally diverse factual triples from Wikidata, focusing mainly on explicit country information. In contrast, our work proposes an automated, efficient approach to constructing a cultural knowledge dataset in a uniform triplet format, addressing the limitations of existing methods and focusing on implicit cultural knowledge. Table 1 contrasts our dataset with prior cultural knowledge collections.

Knowledge probing. Deciphering the knowledge encoded by LLMs poses significant challenges due to their opaque nature , early benchmarks like LAMA (Petroni et al., 2019b) sought to quantify the factual knowledge in English LLMs, while ParaRel (Elazar et al., 2021) highlighted their consistency issues. Subsequent efforts as mLAMA (Kassner et al., 2021) and mParaRel (Fierro and Søgaard, 2022) expanded these benchmarks multilingually, though such methods often focus on single-word entities, limiting their depth of assessment. To address these shortcomings, newer studies (Shin et al., 2020; Zhong et al., 2021; Meng et al., 2022) have evolved towards eliciting more comprehensive factual knowledge, including multi-word entities, with Jiang et al. (2020a) developing algorithms for multi-token predictions. LPAQA (Jiang et al., 2020b) further refines this by optimizing prompt discovery for more accurate knowledge probing. Our work builds on this foundation, targeting multi-token probing within the food domain, characterized by complex expressions like *Trigonella foenum-graecum* and *almond paste*. We also introduce absolute- and fuzzy-match metrics for a nuanced evaluation of LLMs' cultural knowledge.

3 FmLAMA Construction

To assess whether LLMs encode and access cultural information, we develop FmLAMA, a multicultural, multilingual dataset focusing on culinary knowledge. The designed framework can be adapted to other cultural domains.

Step #1: Obtain countries set. Following Zhou et al. (2023), we use countries of food origin to delineate cultural groups. This method leverages countries as proxies for cultural identity, encapsulating diverse traditions, values, and norms that reflect the breadth of human civilizations across geographical boundaries (Minkov and Hofstede, 2012; Peterson et al., 2018).

Step #2: Acquire food instances. We utilize SPARQL to query Wikidata, extracting a vast array of food-related data. This approach exploits Wikidata's RDF triple structure to gather detailed information on food instances, offering a rich source of comprehensive food knowledge.

i. Class. For our food-focused dataset, we concentrate on the dish class and employ two approaches to find food instances:

- Explicit instance of dish, e.g., *bouillabaisse*.
- Inferred through a hierarchy, e.g., *Blanquette de veau* $\xrightarrow{\text{subclass of}}$ stew $\xrightarrow{\text{subclass of}}$ *dish*.

This enables comprehensive inclusion of food instances, represented as $I \xrightarrow{(\text{instance of}|\text{subclass of})+}$ *dish*, where '|' denotes "or", and '+' is "one or more".

ii. Cultural group. We organize food instances by their origin, applying these strategies:

- Directly specified in Wikidata, e.g., *bouillabaisse* $\xrightarrow{\text{country of origin}}$ France.
- Through the associated cuisine category, e.g., *mapo doufu* $\xrightarrow{\text{cuisine}}$ Chinese cuisine $\xrightarrow{\text{country}}$ China.

We exclude dishes with multiple origin countries to maintain cultural specificity.

iii. Properties included. We prioritize the property "has part(s)" to identify food ingredients for each dish. Additional properties like "made from material" and "image" are collected to support future research (e.g., multimodal), though they are not utilized in this study. Language consistency for property descriptions ensures uniformity across the dataset. Our dataset, FmLAMA, comprises 33,601 dish instances, detailed by name, origin, ingredients, and optionally, materials and images. Examples are provided in Appendix A.

Step #3: Filter by language. To explore various language settings, we further create sub-datasets by applying language filtering to the created FmLAMA dataset. This results in FmLAMA-*la*, where '*la*' designates the specific language of the current sub-dataset. In this paper, We focus on a typologically diverse set of languages, namely, English (*en*), Chinese (*zh*), Arabic (*ar*), Korean (*ko*), Russian (*ru*), and Hebrew (*he*), with a filtered sub-dataset of 2590, 815, 571, 807, 961, 462 dishes each. These languages span 4 different language families – Indo-European (English, Russian), Semitic (Hebrew, Arabic), Altaic (Korean), and Sino-Tibetian (Chinese), and are spoken by more than 2.356 billion speakers. Furthermore, these languages represent cultural diversity, being spoken on different continents by groups with rich and distinct cultural backgrounds.

4 Cultural Knowledge Probing

4.1 Probing templates

Following most knowledge-probing methods, we adopt Word Predictions (WP) as knowledge-probing tasks. Specifically, we manually design prompt templates focused on the core attribute "has part(s)", illustrating a connection between a dish (subject) and its ingredient(s) (object). Considering LLMs produce varied predictions based on prompt framing (Elazar et al., 2021; Wang et al., 2023), we craft five templates conveying identical meanings in each prompt language. These templates, in various languages, are depicted in Figure 7 in Appendix B. To explore the impact of introducing cultural context on LLMs' ability to access cultural knowledge, we enhance the basic templates by integrating location adverbials. For instance, "In [C], the food [X] includes the ingredients of [Y]", where [C] denotes the country of origin for the dish [X], [Y] indicates the ingredient object.

4.2 Probing task

The probing task is defined as a candidate object retrieval problem. The candidate set consists of all objects in the filtered sub-dataset FmLAMA-*la*.[1] Our primary objective is to utilize LLMs to obtain the probability of each candidate C and subsequently rank the predicted objects based on these probabilities.

MASK operation. We use the corresponding subject-object tuples ([X], [Y]) as the query and probe LLMs by replacing the subject and masking the object. Considering that each candidate object is tokenized into k subtokens $\{c_1, \cdots, c_k\}$ by LLMs correspondingly, we apply [MASK] token of varying lengths to the objects within each query. So we construct K queries for each food case based on the same template t, K is the maximum number of tokens, each query Q_k is defined as:

$$Q_k = t\left([\mathrm{MASK}] * k\right). \tag{1}$$

Probability acquisition. We use Mean Pooling[2] method to obtain the prediction probability of each candidate. Specifically, for candidate object $C = \{c_1, \cdots, c_k\}$ of length k, we obtain its probability from the likelihoods associated with the [MASK] tokens in Q_k, and the probability of C is calculated as the average of the probabilities of composing its subtokens:

$$P\left(C\right) \;=\; \frac{1}{k}\sum_{i=1}^{k} p\left([\mathrm{MASK}]_i = c_i\right), \tag{2}$$

where $p\left(\cdot\right)$ is obtained after the log softmax operation.

4.3 Probing Metric

Despite the considerable amount of research dedicated to knowledge probing, even studies employing a similar LAMA-style approach lack a standardized evaluation criterion. Although our experiments solely focus on a single relationship, that is, the ingredients of a food item, our probing task poses greater challenges for LLMs: (1) The number of objects in each instance is not fixed. (2) The number of food instances contained in each cultural group varies. (3) The expression of certain ingredients is not always absolute. For instance, in Chinese, both 盐 and 食盐 can denote *salt*. (4) There is flexibility in specifying cooking ingredients. Consider the example of *frico*, the Italian dish known as a cheese crisp in English. Although the knowledge base specifies *cheese* as the ingredient, there is an option to choose a specific type, such as *mozzarella cheese* or *feta*. Considering these constraints, we introduce both an `absolute-match metric`, **Mean Average Precision (mAP)**, and a `fuzzy-match metric`, **Mean Word Similarity (mWS)**.

Mean Average Precision. mAP is widely used in information retrieval settings, assessing the relevance of predicted objects (in our case ingredients) only when they precisely match the golden ones. The precision at rank k (P@k) for a given food instance i is defined as follows:

$$\mathrm{P@k} = \frac{|\mathrm{ing}_i \cap \mathrm{topk}_i|}{k}, \tag{3}$$

where ing_i is the golden ingredients set, topk_i signifies the set of top-k objects with the highest predicted probability of belonging to food item i by LLMs. Then the average precision of food item i is computed as follows:

$$\mathrm{AP}_i = \frac{1}{|\mathrm{ing}_i|}\sum_{k}^{n} \mathrm{P@k} \times \mathrm{rel@k}, \tag{4}$$

where n refers to the size of the candidate object set and rel@k is a relevance indicator function, which equals 1 if the object at rank k is relevant to food item i and equals to 0 otherwise. Finally, we compute the mAP in the following way:

$$\mathrm{mAP} = \frac{1}{|G|}\sum_{i \in G} \mathrm{AP}_i, \tag{5}$$

where G represents a food group we are focusing on (i.e. a subset of FmLAMA).

Mean Word Similarity. mWS is defined based on the semantic similarity between predicted and golden objects. First, we define the similarity score $S\left(i, g\right)$ for each ingredient g within each food instance i. Only the predicted objects in the top-l rankings that are most similar to g contribute to the evaluation score, where l is the size of ing_i. Mathematically, $S\left(i, g\right)$ is defined as follows:

$$S@l\left(i, g\right) = \max_{p \in \mathrm{topl}_i}\left[\cos\left(w_g, w_p\right)\right], g \in \mathrm{ing}_i, \tag{6}$$

where w_g and w_p are the continuous word representations for the ingredient g and the predicted

[1] As the size of the filtered sub-dataset increases, the candidate object set also expands, leading to greater difficulty in probing. Consequently, in this paper, results obtained by probing across different filtered sub-datasets cannot be used for horizontal comparison.

[2] Mean Pooling is usually used in multi-token probing and much better than max-pooling and first-pooling methods, this is demonstrated by the experimental results of different pooling methods from Wu et al. (2023).

Origin	Count	Encoder-only LLMs						Encoder-Decoder LLMs		Avg.
		Bb-c	Bb-u	Bl-c	Bl-u	mB-c	mB-u	mT5	T5	
Italy	215 (8.3%)	3.93±2.29	5.43±2.26	4.39±2.44	5.32±3.14	5.25±2.40	5.65±4.03	4.49±1.37	5.09±1.82	4.94
U.S.	285 (11.0%)	10.26±4.94	**14.58±6.29**	11.60±4.67	**12.72±6.25**	**13.79±5.16**	**13.10±7.62**	**12.37±3.37**	**13.16±5.32**	**12.70**
Turkey	98 (3.8%)	7.80±5.13	10.53±5.09	6.72±3.97	8.02±5.66	11.55±6.37	9.12±6.28	6.29±2.77	7.84±3.98	8.48
Japan	186 (7.2%)	7.99±3.75	7.53±4.09	8.68±4.03	6.64±3.75	7.18±3.41	7.59±4.59	5.57±2.63	5.64±1.40	7.10
France	175 (6.8%)	5.01±2.99	6.57±2.90	6.15±2.58	5.55±2.29	5.64±2.62	6.02±3.59	3.63±1.90	4.05±1.67	5.33
U.K.	83 (3.2%)	8.40±5.34	8.84±3.39	11.17±6.01	9.29±5.83	10.48±4.16	8.52±6.48	6.29±3.94	7.71±3.55	8.83
Mexico	57 (2.2%)	6.41±3.12	6.93±2.95	7.72±2.58	7.16±4.55	8.08±2.64	7.92±2.26	3.61±0.97	4.18±2.72	6.50
India	132 (5.1%)	**12.63±6.56**	14.18±5.99	**13.37±7.32**	12.10±6.99	9.78±4.34	10.56±5.67	8.64±3.41	5.41±2.55	10.83
Germany	57 (2.2%)	4.94±3.21	5.42±3.39	5.64±3.43	5.78±2.69	7.40±2.65	7.81±4.92	4.20±1.61	5.79±1.26	5.87
China	97 (3.8%)	10.43±4.17	12.35±3.13	12.36±3.34	11.27±4.69	12.54±3.47	11.79±5.54	7.93±1.97	9.36±3.18	11.00
Iran	21 (0.8%)	7.00±5.17	5.96±4.26	6.32±4.56	6.33±3.82	8.31±4.73	9.42±5.85	10.37±3.09	5.67±0.42	7.42
Greece	21 (0.8%)	3.15±2.93	3.99±2.46	2.78±1.84	2.73±1.63	3.72±3.24	3.12±1.42	2.12±0.92	0.73±0.27	2.79
Others	1031 (40.0%)	5.92±3.04	6.30±2.81	6.41±2.88	6.37±3.17	6.06±2.51	5.62±2.42	3.82±1.21	3.61±1.06	5.51
ALL	2580 (100.0%)	6.87±3.46	8.09±3.40	7.59±3.36	7.51±3.77	7.70±2.91	7.43±3.73	5.51±1.75	5.58±1.95	7.04

(a) Performance results evaluated using **mAP** (%).

Origin	Encoder-only LLMs						Encoder-Decoder LLMs		Avg.
	Bb-c	Bb-u	Bl-c	Bl-u	mB-c	mB-u	mT5	T5	
Italy	0.3107±0.07	0.3391±0.05	0.3206±0.07	0.3246±0.05	0.2992±0.05	0.3279±0.04	0.3287±0.02	0.3097±0.01	0.3201
U.S.	0.3593±0.08	**0.3957±0.07**	0.3793±0.07	**0.3813±0.07**	**0.3730±0.05**	**0.3760±0.06**	**0.3796±0.03**	**0.3844±0.04**	**0.3786**
Turkey	0.3431±0.07	0.3592±0.08	0.3542±0.07	0.3521±0.06	0.3478±0.08	0.3480±0.05	0.3501±0.04	0.3243±0.03	0.3474
Japan	0.3247±0.07	0.3321±0.06	0.3312±0.08	0.3072±0.05	0.2902±0.05	0.2965±0.05	0.3072±0.03	0.2815±0.01	0.3088
France	0.3280±0.07	0.3420±0.04	0.3348±0.05	0.3372±0.03	0.3261±0.05	0.3278±0.03	0.3342±0.02	0.3268±0.02	0.3321
U.K.	0.3132±0.08	0.3343±0.03	0.3401±0.06	0.3362±0.05	0.3177±0.04	0.3255±0.05	0.3126±0.04	0.3079±0.02	0.3234
Mexico	0.3123±0.05	0.3418±0.03	0.3473±0.06	0.3254±0.05	0.3087±0.05	0.3152±0.02	0.3276±0.01	0.3130±0.01	0.3239
India	**0.3638±0.09**	0.3741±0.08	**0.3796±0.09**	0.3531±0.08	0.3215±0.06	0.3367±0.07	0.3352±0.05	0.2920±0.02	0.3445
Germany	0.3127±0.10	0.3268±0.08	0.3362±0.08	0.3191±0.06	0.3118±0.05	0.3473±0.06	0.3189±0.05	0.3263±0.03	0.3249
China	0.3210±0.04	0.3493±0.02	0.3476±0.02	0.3354±0.03	0.3398±0.02	0.3309±0.04	0.3569±0.02	0.3156±0.03	0.3371
Iran	0.3195±0.07	0.3145±0.09	0.3399±0.08	0.3145±0.08	0.3252±0.07	0.3476±0.04	0.3555±0.03	0.3233±0.04	0.3300
Greece	0.3113±0.08	0.3399±0.06	0.3359±0.08	0.3178±0.06	0.2822±0.08	0.3180±0.02	0.3530±0.05	0.3048±0.01	0.3204
Others	0.3235±0.06	0.3378±0.04	0.3376±0.06	0.3263±0.04	0.3088±0.05	0.3150±0.02	0.3266±0.02	0.2821±0.02	0.3197
ALL	0.3278±0.06	0.3468±0.05	0.3428±0.06	0.3339±0.05	0.3177±0.05	0.3264±0.03	0.3338±0.02	0.3050±0.02	0.3293

(b) Performance results evaluated using **mWS**.

Table 2: Probing performance comparison with English prompts and FMLAMA-*en* sub-dataset. "B/mB" respectively represent abbreviations for BERT and mBERT. "b/l" stands for base/large and "c/u" stands for cased/uncased. **Bold** represents the best-performing cultural group within the same model (each column). Red indicates the best-performing LLMs in each specific cultural group (each row). We find that LLMs typically score higher on U.S. cultural knowledge, and monolingual English LLMs perform better in English-speaking countries. The *Pearson correlation coefficient* between mAP and mWS is 0.66 based on the results above.

object p, respectively, and are obtained using `Fasttext` (Bojanowski et al., 2017; Joulin et al., 2016), $\cos(\cdot)$ is the cosine similarity function. Then we can compute the probing similarity WS_i for each food instance i and mWS for the targeted food group as follows:

$$\mathrm{WS}_i = \frac{1}{|\mathrm{ing}_i|} \sum_{g \in \mathrm{ing}_i} S@l\,(i,g) \qquad (7)$$

$$\mathrm{mWS} = \frac{1}{|G|} \sum_{i \in G} \mathrm{WS}_i \qquad (8)$$

Both mAP and mWS metrics range from [0,1].

5 Experiments

5.1 Access evaluation

We explore the masked language models, including the encoder-only language model BERT (Devlin et al., 2019) and its multilingual version, mBERT, along with the encoder-decoder language model T5 (Raffel et al., 2020) and its multilingual counterpart, mT5 (Xue et al., 2020). Table 2 showcase the probing results based on English prompt on the filtered dataset FMLAMA-*en*.[3]

Metric comparison Regarding the adopted two metrics, we observe the following: 1) Across diverse cultural groups, evaluating probing results with the absolute-match metric, mAP, reveals a significantly greater disparity compared to assessments using the fuzzy-match metric, mWS. This highlights the challenge of absolute-match evaluation and the importance of introducing the fuzzy-match metric. 2) We evaluate the correlation between the two metrics, mAP, and mWS, utilizing all probing results from Table 2, resulting in a Pearson correlation coefficient of 0.66. This signifies a moderate to strong positive correlation between

[3]The probing results with prompts in the other five languages on the corresponding filtered sub-datasets can be found in the Appendix C.

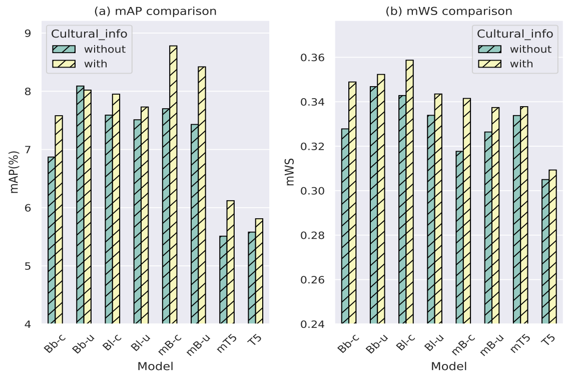


Figure 2: Comparative probing results on FmLAMA-*en*: incorporating cultural information about the origin of dishes into English prompts can enhance the probing of cultural knowledge.

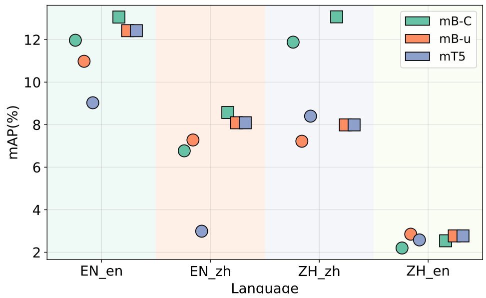


Figure 3: Probing results with monolingual prompts and code-switching prompts, ○ and □ signify the absence or presence of cultural background information in the prompt, respectively.

the two metrics.

Group comparison Regarding the various cultural groups in Table 2, we observe the following: (1) Regardless of absolute or fuzzy evaluation, the U.S. group almost consistently achieves the highest probing results on FmLAMA-*en* with English prompts. This suggests that evaluated LLMs possess a greater familiarity with food-related knowledge specific to the U.S. context. (2) The quantity of knowledge within cultural groups in the knowledge base may not fully reflect the potential knowledge within LLMs. For example, despite the Italy group having more dishes than the U.K. and India groups, LLMs achieve lower probing result scores for the Italy group. We speculate that this difference may be due to the official languages of the U.K. and India including English, aligning with the prompt language in the experiments. However, it is notable that LLMs do not exhibit lower probing result scores in China, despite its official language not including English. (3) Monolingual English LLMs excel in accessing cultural knowledge within English-speaking countries, while multilingual LLMs may not demonstrate superior performance in non-English-speaking countries. This discrepancy may arise from the fact that the pre-training data for monolingual models is solely in English, whereas for multilingual models, it encompasses languages beyond English.

5.2 Prompt analysis

In this part, we examine the impact of different prompt settings on cross-cultural knowledge exploration. This involves integrating references to cultural backgrounds, incorporating code-switching settings, and considering language choices within the prompts.

Cultural background analysis Using the English prompts from Figure 7 as a basis, we incorporate information about the country of origin into the probing prompts for each specific dish (as described in §4.1). Figure 2 presents a comparative analysis of probing results on FmLAMA-*en* across different LLMs, taking into account the inclusion of cultural information mentions in the prompts. We find that, whether through absolute or fuzzy evaluation, English prompts with cultural information achieve higher probing scores in capturing LLMs' knowledge within the food domain. This suggests the importance of emphasizing the cultural background when utilizing LLMs, especially in the exploration of culture-related topics. Moreover, in the comparison of the absolute-match metric mAP, cultural information contributes more in multilingual LLMs compared to monolingual LLMs. This seems to align with real-world social dynamics, indicating that in multicultural settings, introducing cultural backgrounds enhances communication.

Code-switching analysis Code-switching (CS) is the linguistic phenomenon of incorporating multiple languages within a single sentence or conversation. It occurs naturally in conversational speech among multilingual speakers (Aguilar and Solorio, 2020; Doğruöz et al., 2021). To assess the cultural knowledge probing ability of multilingual LLMs under code-switching settings, we configure prompt variations with English-Chinese and Chinese-English CS settings, specifically based on purely English and Chinese prompts, as follows:

EN_en: The food beef bourguignon has the ingredients of [Y].

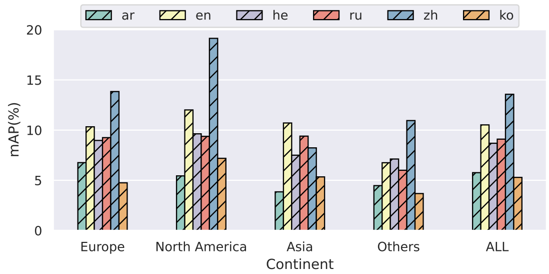


Figure 4: Average probing results of all multilingual LLMs on the filtered sub-dataset with prompts in different languages. The ability of LLMs to access cultural knowledge may not consistently outperform in the cultural group where that language is spoken for each language prompt.

Prompt language	EN		ZH	
Subject language	en	zh	en	zh
Llama2-7b-chat-hf	27.36	17.91	7.28	5.91
Llama2-13b-chat-hf	35.04	22.64	10.43	7.87
vicuna-7b	36.22	19.69	12.01	14.37
vicuna-13b	28.15	20.87	17.72	16.54
gpt-3.5-turbo	47.64	34.06	20.28	25.00

Table 3: Accuracy (%) of decoder-only LLMs' results.

EN_zh: The food 红酒炖牛肉 has the ingredients of [Y].

ZH_zh: 美食红酒炖牛肉是以[Y]为主要原料制作的。

ZH_en: 美食beef bourguignon是以[Y]为主要原料制作的。

where EN and ZH signify the main language of the prompt, while en and zh indicate the language of the subject. We aim to predict the ingredients using the same language as the prompt. Probing results on each multilingual LLMs display for Figure 3[4]: (1) In code-switching settings, the probing abilities of LLMs for EN_zh relative to EN_en and ZH_en relative to ZH_zh both notably decrease, with the latter showing a more pronounced decline in the mB-c language model. (2) Whether in code-switching mode or not, the introduction of relevant cultural backgrounds in the prompts generally facilitates cultural knowledge probing.

Language analysis We adopt prompts in six different languages and conduct knowledge probing on multilingual LLMs. To ensure a fair comparison of the probing results across different language prompts, we filter a sub-dataset consisting of only 175 food instances that have labels in all of the languages that are involved. This underscores the lack of aligned knowledge across multiple languages. Considering the relatively limited size of the filtered dataset, we opt for a broader categorization based on continents rather than individual countries to differentiate cultural groups. Appendix E illustrates the distribution of cultural groups within this filtered sub-dataset. Figure 4 displays probing results on the filtered sub-dataset with prompts in different languages. The probing performance varies with prompts in different languages. The ability of LLMs to access cultural knowledge may not consistently outperform in the cultural group where that language is spoken for each language prompt. For instance, Chinese prompts yield better results in English-speaking countries. Our experimental results suggest that determining the most suitable language for knowledge probing in a specific cultural context proves challenging.

5.3 Decoder-only LLMs probing

LLMs, particularly decoder-only models, offer robust capabilities in both natural language understanding and generation. We specifically conduct cultural knowledge probing on instruction-tuned models, including Llama2-chat (Touvron et al., 2023), Vicuna (Chiang et al., 2023), and gpt-3.5-turbo (OpenAI, 2023). Considering the characteristics of these LLMs, we employ a naive approach to probe and evaluate them.[5] Table 3 shows the overall probing results using accuracy for both the monolingual and code-switching prompt settings. We observe that gpt-3.5-turbo demonstrates the highest extent of cultural knowledge. Additionally, all decoder-only LLMs demonstrate better performance with monolingual English prompts compared to monolingual Chinese prompts, and the performance with code-switching prompts is inferior to that with their corresponding monolingual prompts.

5.4 Case study – Ingredients analysis

We conduct a more fine-grained analysis of the model's predictions to better understand its behavior, particularly to discern whether it is relying on "educated" guesses or leveraging its cultural knowledge when providing answers. We focus on two multilingual models, mT5 and mbert-base-uncased, and three languages: Korean, Chinese,

[4]Code-switching probing focused on dishes common in both English and Chinese in a filtered subset of FmLAMA. Only cultural groups with dish counts exceeding 20 are retained. The probing results of each cultural group can be seen in the Appendix D.

[5]The experimental setup details, evaluation method, cultural knowledge probing results for each cultural group, and additional analysis are provided in Appendix F.

Most common ingredients			Gold label ingredients		
en	zh	ko	en	zh	ko
yogurt	water	tangerine	egg	grain flour	egg
avocado	sugar	cumin seeds	flour	egg	wheat flour
vegetable	glucose	hot stone	sugar	sugar	sugar
olive oil	oil	lime juice	potato	rice	potato
rice	salt	pork floss	almond	chicken	salt
bread	egg	rice	meat	tomato	chicken meat
baking powder	quince	drinking water	onion	onion	butter
ice cream	milk	apple	tomato	butter	meat
extra virgin olive oil	sugar	cheese	chicken meat	pork	milk
vegetable oil	cilantro	tofu	butter	potato	bread

Table 4: Ingredients analysis. The first three columns display the 10 most common ingredients for mBERT-base-uncased in English (en), Chinese (zh), and Korean (ko). The remaining three columns show the 10 most common gold standard ingredients, sourced from Wikipedia.

and English[6]. For every language and dish in the multilingual dataset (see §3), we extract the top 5 ingredients with the highest probability as predicted by the model and calculate the set of distinct ingredients. Table 4 shows mbert's 10 most common ingredients in each language. In Korean, Chinese, and English, these ingredients cover 0.59%, 0.91%, and 0.97% of the total predictions, respectively.[7] This indicates that the model tends to depend on repetitive predictions, consistently selecting the same subset of ingredients regardless of the context. [8] For example, in Korean, mbert often predicts ingredients specific to Korean dishes, such as *hot stone*, which refers to various Korean dishes served in a hot stone pot, and *tofu*. However, in Chinese/English, the model tends to predict more widely used ingredients, like *flour* and *sugar*, which are common across various cultural cuisines. This observation might imply that Chinese/English models perform better because the ingredients they predict have broader applicability and are more likely to be found in a variety of dishes.

6 Human evaluation

To check the validity of the two proposed metrics, we conduct a human evaluation of the probing results, including 372 food instances. Specifically, we extract the probing results for all dishes originated in the U.S and China from the best-performing prompt setting in the code-switching EN_en experiments. In each dish instance, only the top-l predicted objects are evaluated, where l is the number of golden labels, similar to the setting of the mWS metric. In addition to assessing the absolute matching between predicted objects and golden objects as measured by mAP, our human evaluation also considers: (1) whether a predicted object can replace a certain golden object in cooking, and (2) whether the predicted object is an ingredient of the dish, even if not listed in the golden label (e.g, *lemon* is an ingredient of the dish "lemon chicken", even though it erroneously has only *chicken meat* in Wikidata). The calculation method for human evaluation scores is the same as mAP. With the participation of four authors, we average their final scores to obtain the ultimate assessment score. Table 5 compares the scores from human evaluation with those from two automated metric evaluations. The *Pearson correlation coefficient* between mAP and Human is 0.88, whereas that between mWS and Human is 0.94 as per the aforementioned results, which indicates the significance of mWS, as it aligns more closely with human evaluation.

Origin	mBERT-uncased			mT5		
	mAP	**mWS**	**Human**	**mAP**	**mWS**	**Human**
U.S.	10.68	37.96	33.61	14.70	38.24	37.56
China	8.40	32.79	25.16	9.40	37.41	32.19

Table 5: Comparison of scores (%) from human evaluation and two automated metric evaluations. The *Pearson correlation coefficient* between mAP and human is 0.88, whereas that between mWS and human is 0.94 as per the aforementioned results.

7 Conclusion

This study presents an automated method for generating extensive cultural knowledge datasets, exemplified by the creation of FmLAMA, a diverse, food-centric dataset that spans multiple cultures and languages. We introduce novel metrics for cultural knowledge evaluation in LLMs, emphasizing the influence of cultural context and language in the probing process. Our findings reveal a predominant bias towards American culture in LLMs when using English prompts, a bias that diminishes with prompts in other languages. Interestingly, incorporating explicit cultural cues in prompts enhances LLMs' cultural knowledge access. The study also highlights the scarcity of culturally diverse knowledge across languages, pointing to a potential root of observed biases in LLMs.

[6]This subset of languages is selected for clarity. The observed trends are consistent across all languages in our dataset.

[7]To compute the coverage we divide the number of times the most common 10 predictions are predicted by the model, by the number of its top 5 predictions. For each dish we take only 5 predictions because more than 0.95% of the dishes contain less than 5 ingredients.

[8]This trend persist for all models, and languages.

8 Limitations

While this study provides valuable insights into cross-cultural knowledge probing in LLMs, it is essential to acknowledge several limitations. Firstly, the food domain knowledge dataset utilized in this research is sourced from Wikidata, which may not offer comprehensive coverage. For example, the dish *soy sauce chicken* may only include the ingredient *chicken meat* while lacking the inclusion of *soy sauce*. Moreover, ingredient descriptions are not always detailed. For instance, the Wikidata gold label might be *oil* when the recipe requires a specific type of oil, such as *sesame oil*. This inconsistency underscores the motivation behind our mWS metric. Furthermore, aside from well-known dishes, certain recipes lack standardization and may vary depending on individual preferences and cooking styles, posing challenges to precise probing. Additionally, the fuzzy-match metric mWS, introduced in this study, relies on Fasttext for obtaining object representation vectors. However, for certain objects in Chinese and Korean, zero vectors may result, rendering similarity calculation impossible. Lastly, we employ manually crafted templates in this paper. However, research has shown that sampling templates from large corpora can also enhance knowledge-probing evaluation. This aspect is deferred to future work. Despite our endeavors to construct comprehensive multilingual and multicultural knowledge repositories, the availability of aligned cross-cultural knowledge remains limited in multilingual settings. This constraint presents challenges in exploring the interaction between language and culture.

References

Gustavo Aguilar and Thamar Solorio. 2020. From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online. Association for Computational Linguistics.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural Adaptation of Recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Heather Lent and Anders Søgaard. 2021. Common sense bias in semantic role labeling. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 114–119, Online. Association for Computational Linguistics.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, Dublin, Ireland. Association for Computational Linguistics.

Michael Minkov and Geert Hofstede. 2012. Is national culture a meaningful concept? cultural values delineate homogeneous national clusters of in-country regions. *Cross-Cultural Research*, 46(2):133–159.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

OpenAI. 2023. ChatGPT. `https://chat.openai.com/`.

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

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation.

Mark F Peterson, Mikael Søndergaard, and Aycan Kara. 2018. Traversing cultural boundaries in ib: The complex relationships between explicit country and implicit cultural group boundaries at multiple levels. *Journal of International Business Studies*, 49:1081–1099.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019a. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019b. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Anders Søgaard. 2021. Locke's holiday: Belief bias in machine reading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8240–8245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. 2023. Assessing the reliability of large language model knowledge. *arXiv preprint arXiv:2310.09820*.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. Do PLMs know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3080–3101, Toronto, Canada. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

A FmLAMA examples

Figure 6 illustrates examples from dataset FmLAMA. Each dish instance in FmLAMA is defined as: $(url, na, cou, la, pa, [ma, im])$, the elements in $[\cdot]$ indicate optional:

- url: the link in Wikidata;
- na: the name of the dish;
- cou: the country of origin of the dish;
- la: the language used in this entry;
- pa: the ingredients of this dish;
- ma: the material used in the dish;
- im: the image of this dish.

Particularly, in each instance, $la = \text{LANG}(na) = \text{LANG}(pa) = \text{LANG}(ma)$. For a dish with the same url, there may be several instances with different languages. There are totally of 33,601 entries in FmLAMA.

B Probing Templates

Probing templates in six involved languages are shown in Figure 7, in which [X] represents the subject (dish) and [Y] indicates the object (ingredient).

C Probing results in each FmLAMA-*la*

Besides the multilingual LLMs discussed in §5.1, we also probe XLM-RoBERTa (Conneau et al., 2020) here.[9] Furthermore, in addition to Hebrew, we configure probing for monolingual LLMs in

[9]Indeed, we probe XLM-RoBERTa on FmLAMA-*en* as well, yielding very poor probing results. Nonetheless, although XLM-RoBERTa's performance remains subpar in other languages, it shows some relatively positive outcomes.

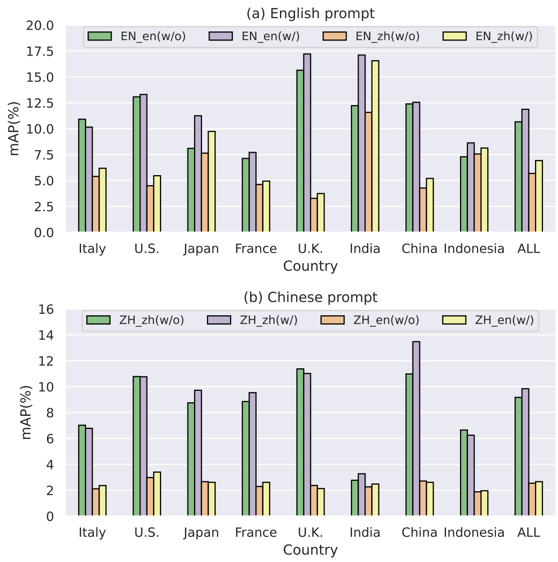


Figure 5: Probing results with code-switching settings in each cultural group, `w/o` and `w/` signify the absence or presence of cultural background information in the prompt, respectively.

Arabic, Russian, Korean, and Chinese, including `asafaya/bert-base-arabic` (Safaya et al., 2020), `DeepPavlov/rubert-base-cased` (Kuratov and Arkhipov, 2019), `kykim/bert-kor-base`, `klue/bert-base` (Park et al., 2021) and `bert-base-chinese`.

The probing results with prompts in the other five languages (Arabic, Hebrew, Russian, Korean, and Chinese) on the corresponding filtered sub-datasets (FmLAMA-*ar*, FmLAMA-*he*, FmLAMA-*ru*, FmLAMA-*ko*, and FmLAMA-*zh*) are depicted in Table 6 through Table 10, respectively. Because certain objects in Chinese and Korean have representation vectors that result in all zeros when obtained through Fasttext, calculating cosine similarity was not feasible. Consequently, mWS evaluation was not conducted for prompts in Chinese and Korean prompts. Overall, irrespective of the language used for probing, LLMs still exhibit a relatively strong familiarity with knowledge in the food domain within American culture. However, they may show a slight preference for certain cultural knowledge; for example, when probing in Arabic, LLMs may show a better capability in probing knowledge related to Iranian groups. Additionally, the probing capability of monolingual LLMs may not necessarily surpass that of multilingual LLMs.

D Cultural results for Code-switching probing

For a more detailed comparison of the knowledge probing abilities of LLMs across various cultural groups, we evaluate the probing results for each cultural group individually. Figure 5 shows the average probing results for the three multilingual LLMs in each cultural group, from which we find: (1) English-dominant prompt settings generally outperform those where Chinese is the primary language. (2) En_en prompt setting excels in probing food knowledge for U.K., U.S., India, and China cultures, while proficiency in the ZH_zh prompt setting is also notable for the above cultural groups except India. (3) In code-switching settings, the detection abilities of EN_zh relative to EN_en and ZH_en relative to ZH_zh both notably decrease, with the latter showing a more pronounced decline. (4) Whether in code-switching mode or not, introducing relevant cultural backgrounds in the prompts aids in the detection of cross-cultural knowledge, manifested in overall probing results and the probing ability within almost every cultural group. The more detailed probing results for each multilingual LLM across various cultural groups can be found in Tables 11 and 12.

E Data distribution of 175 dishes

The figure 8 illustrates the data distribution of the filtered sub-dataset used in the language analysis in §5.2. It encompasses both continent-level and country-level data distributions.

F Decoder-only LLM probing details

In this section, We present the experimental setup for knowledge probing with decoder-only LLMs, and showcase their fine-grained probing results across various cultural groups.

F.1 Experiential setup

Decoder-only LLMs : LlaMa2-chat (Touvron et al., 2023), a collection of pre-trained and fine-tuned chat models; Vicuna (Chiang et al., 2023), a chat assistant trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT; and gpt-3.5-turbo* (OpenAI, 2023), a GPT-3.5 model fine-tuned on human instructions using Reinforcement Learning with Human Feedback (RLHF).

url	dish	origin	language	hasParts	Material (optional)	Image (optional)
https://www.wikidata.org/wiki/Q1022124	beef bourguignon	France	en	red wine, beef, broth	-	
https://www.wikidata.org/wiki/Q1022124	뵈프 부르기뇽	프랑스	ko	맑은국, 적포도주, 쇠고기	-	
https://www.wikidata.org/wiki/Q1022124	红酒炖牛肉	法国	zh	清汤, 红葡萄酒, 牛肉	-	
https://www.wikidata.org/wiki/Q1022124	Bœuf bourguignon	Francia	es	vino tinto, carne de res, caldo	-	
https://www.wikidata.org/wiki/Q7211268	luosifen	China	en	chili pepper, rice vermicelli, peanut, freshwater snail, bamboo shoots, tofu skin	Viviparus quadratus	
https://www.wikidata.org/wiki/Q7211268	螺蛳粉	中国	zh	辣椒，细米粉，笋,淡水蜗牛，腐皮，花生	方形环棱螺	
https://www.wikidata.org/wiki/Q20987994	Шолезард	Иран	ru	рис, шафран, сливочное масло, корица, кардамон, розовая вода	рис, кофе, шафран, корица, сливочное масло, розовая вода, кардамон	
https://www.wikidata.org/wiki/Q1104585	סלט קוב	ארצות הברית	he	תרנגול הבית, ביצה, קותל חזיר, גבינה, עגבנייה, חסה, אבוקדו, ויניגרט'	תרנגול הבית, אבוקדו, גבינה, חסה, ויניגרט, קותל חזיר, ביצה, עגבנייה	
https://www.wikidata.org/wiki/Q997633	أماتريتشانا	إيطاليا	ar	نبيذ أبيض, ملح الطعام, سباغيتي, طماطم, زيت, فلفل حار, غونجالة, صلصة البندورة		

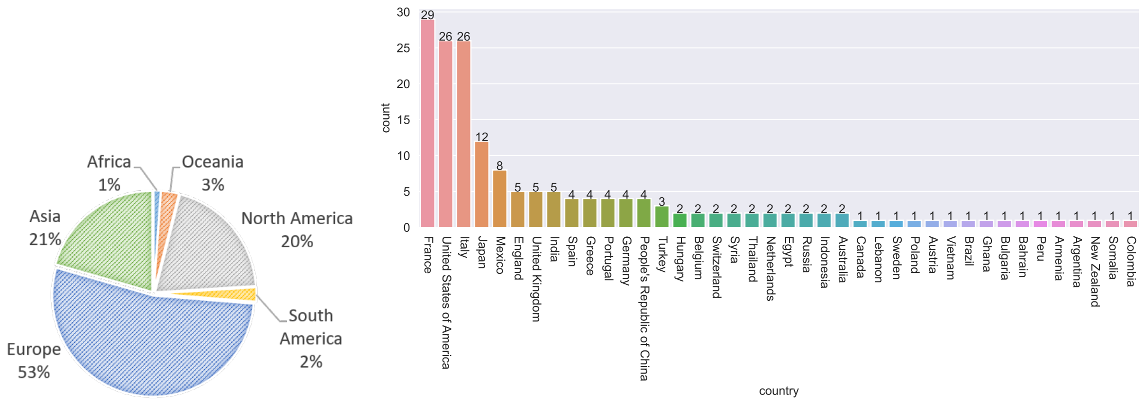
Figure 6: Examples of FmLAMA.

La.	Prompt	La.	Prompt
en	The food [X] has the ingredients of [Y] . [X] is a kind of food with [Y] . [X] is a type of food, comprising [Y] as its main constituents . [Y] is part of the food [X] . The food [X] has part of [Y] .	**ar**	الطعام [X] يحتوي على مكونات [Y] . [X] هو نوع من الطعام مع [Y] . [X] هو نوع من الطعام، يتكون من [Y] كمكوناته الرئيسية . [Y] جزء من الطعام [X] . الطعام [X] يحتوي على جزء من [Y] .
ko	음식 [X]에는 [Y]의 성분이 포함되어 있습니다. [X]는 [Y]를 가진 음식의 일종입니다. [X]는 [Y]를 주성분으로 하는 식품의 일종입니다. [Y]는 음식 [X]의 일부입니다. 음식 [X]에는 [Y]의 일부가 포함되어 있습니다.	**zh**	美食[X]是以[Y]为主要原料制作的。 美食[X]的制作需要[Y]作为主要成分。 在制作美食[X]时，需要使用[Y]作为其中之一的成分。 UTF8gbsn美食[X]中的主要成分之一是[Y]。 [X]是一道美味的食物，它需要使用[Y]制作而成
ru	В пище [X] есть ингредиенты из [Y] . [X] это вид пищи с [Y] . [X] это вид пищи, имеющий [Y] как один из своих основных компонентов . [Y] это часть пищи [X] . Пища [X] содержит часть от [Y] .	**he**	במאכל [X] יש את המרכיבים של [Y]. [X] הוא מאכל עם [Y]. מאכל [X] מכיל את [Y] כמרכיביו העיקריים. [Y] הוא מרכיב במאכל [X]. במאכל [X] יש מרכיבים מ[Y].

Figure 7: Probing templates in six involved languages, with [X] representing the subject and [Y] indicating the object that can be substituted.

Prompt construction Decoder-only LLMs, in contrast to encoder-only LLMs and encoder-decoder LLMs, focus solely on generating text based on contextual information provided to them. Considering these factors, the probing task involving predicting the probability of the [MASK] token during decoder-only LLM probing will no longer be applicable. We extend the initial template with fill-in instructions, such as "The food [X] has the ingredients of []. Please fill in the sentence." We only select the best-performing template for extension and probing of the decoder-only LLMs here. Taking into account the diversity of responses from chat LLMs and the ease of evaluating the probing results, we also consider employing an alternative fill-in instruction: "Please complete the sentence with only one entity". The final results only showcase the probing outcomes under the best-performing settings.

Evaluation For each dish instance, we parse the text generated by the decoder-only LLMs to obtain the final predicted object list. This object list is not derived from a candidate object set, resulting in greater diversity. Therefore, before conducting the matching evaluation, we lemmatize each object in both the golden object list and the predicted object list to obtain the lemmatized format of each word, such as *potatoes* $\rightarrow$ *potato*. We utilize accuracy (ACC) as the metric here, employing a forgiving form of absolute matching to assess result accu-



(a) Continent-level distribution

(b) Country-level distribution

Figure 8: Data distribution for the filtered sub-dataset comprising 175 dishes.

racy. Specifically, for each dish instance, if any predicted object matches any object in the golden object list, we consider the prediction for that dish to be correct.

F.2 Cultural probing results

Table 13 show the probing results in each cultural group, including monolingual prompts and code-switching prompts. In monolingual English prompts, all decoder-only LLMs exhibit superior probing performance for knowledge related to the cultural groups of Italy, the U.S., and France. When using English-Chinese code-switching prompts, the overall probing performance tends to decrease, but the best-performing cultural groups remain largely unchanged. However, there are some improvements in the probing rankings for the China group in `Llama2-7b-chat-hf` and the U.K. group in `vicuna-13b`. In the case of monolingual Chinese prompts, the cultural groups with the best probing performance are primarily concentrated in China, the U.K., and the U.S. Similarly, when using Chinese-English code-switching prompts, the overall probing performance tends to decrease.

Origin	Count	Bb-ar	Bl-ar	mB-c	mB-u	XRb	XRl	mT5	Avg.
Italy	76 (13.3%)	2.16±0.95	2.55±0.97	3.52±1.73	2.55±1.25	1.33±0.63	1.30±0.43	3.35±0.71	2.39
U.S.	54 (9.5%)	3.07±1.45	6.01±3.35	4.74±2.46	4.32±4.47	2.88±2.46	2.26±1.05	6.17±0.97	4.21
Turkey	52 (9.1%)	4.59±2.30	5.97±4.08	4.58±1.37	4.41±3.98	2.50±1.14	1.93±0.11	3.18±0.90	3.88
Japan	44 (7.7%)	3.02±1.11	2.86±1.01	3.76±1.54	2.80±0.56	1.69±0.93	1.30±0.10	2.18±0.77	2.52
France	37 (6.5%)	5.25±2.06	5.87±4.09	4.92±3.30	3.98±3.64	**4.28±4.26**	3.67±2.76	6.17±2.40	4.88
U.K.	26 (4.6%)	6.78±2.88	8.97±6.33	8.53±5.48	6.02±6.67	2.77±3.28	2.18±1.86	**8.84±5.02**	6.30
Mexico	20 (3.5%)	2.95±0.40	5.44±2.22	4.52±2.85	5.74±1.82	2.08±0.34	1.88±0.27	3.24±1.27	3.69
India	18 (3.2%)	3.67±3.17	6.41±3.58	**9.02±7.34**	4.44±5.49	1.90±0.27	1.79±0.13	5.55±2.93	4.68
Germany	13 (2.3%)	6.92±3.83	9.67±5.54	8.03±4.78	**6.05±4.50**	1.45±0.72	1.49±0.65	5.41±3.72	5.57
China	13 (2.3%)	3.25±2.55	2.91±2.51	4.44±3.22	4.33±4.76	1.93±0.16	1.76±0.10	4.04±2.01	3.24
Iran	11 (1.9%)	**7.74±5.56**	**11.09±5.02**	6.61±4.79	4.61±5.41	4.04±4.51	**4.10±4.54**	6.02±4.34	**6.32**
Greece	11 (1.9%)	3.31±0.85	4.98±4.35	4.16±3.98	4.66±4.18	1.79±1.15	1.72±0.89	4.04±1.80	3.52
Spain	10 (1.8%)	6.03±2.68	4.75±3.35	6.16±4.09	3.04±2.87	3.46±1.32	2.79±0.27	5.53±1.39	4.54
Russia	10 (1.8%)	3.36±0.63	2.32±0.49	1.93±0.27	1.82±0.32	2.05±0.41	1.91±0.21	2.47±0.48	2.27
Others	176 (30.8%)	4.03±0.97	3.54±1.57	4.30±1.40	2.83±1.72	2.74±1.19	2.31±0.15	4.56±1.59	3.47
ALL	571 (100.0%)	3.95±1.31	4.66±2.36	4.73±2.03	3.61±2.66	2.48±1.46	2.10±0.59	4.53±1.37	3.72

(a) Performance results evaluated on **mAP** (%).

	Bb-ar	Bl-ar	mB-c	mB-u	XRb	XRl	mT5	Average
Italy	0.3147±0.02	0.2786±0.05	0.2917±0.04	0.2383±0.02	0.1991±0.02	0.1930±0.00	0.3120±0.03	0.2611
U.S.	0.2943±0.03	0.3057±0.05	0.3087±0.02	0.2648±0.05	0.2208±0.05	0.2072±0.02	0.3264±0.02	0.2754
Turkey	0.2996±0.03	0.2754±0.07	0.2787±0.04	0.2432±0.03	0.1941±0.03	0.1866±0.01	0.2867±0.02	0.2520
Japan	0.2977±0.02	0.2416±0.04	0.2685±0.02	0.2409±0.03	0.1795±0.02	0.1729±0.01	0.2791±0.03	0.2400
France	0.3360±0.02	0.3221±0.05	0.3032±0.04	0.2877±0.03	0.2255±0.08	0.2103±0.05	0.3333±0.04	0.2883
U.K.	0.3125±0.05	0.3418±0.06	**0.3632±0.06**	0.2736±0.09	0.2270±0.07	0.2103±0.03	**0.3681±0.04**	0.2995
Mexico	0.3273±0.02	0.2866±0.04	0.3036±0.03	0.2923±0.04	0.2039±0.04	0.2043±0.04	0.3036±0.04	0.2745
India	0.2707±0.03	0.2726±0.07	0.3030±0.04	0.2261±0.06	0.1950±0.02	0.1814±0.01	0.2808±0.03	0.2471
Germany	0.3106±0.03	0.3321±0.06	0.3275±0.05	0.2758±0.05	0.2048±0.03	0.1923±0.00	0.3113±0.02	0.2792
China	0.3006±0.04	0.2658±0.04	0.3366±0.03	0.2854±0.06	0.2035±0.02	0.2000±0.01	0.3037±0.05	0.2708
Iran	0.3714±0.05	**0.3420±0.06**	0.3370±0.05	0.2687±0.05	0.2291±0.07	0.2286±0.07	0.3289±0.04	0.3008
Greece	0.3068±0.01	0.3067±0.06	0.3260±0.05	0.2693±0.03	0.2270±0.01	0.2306±0.02	0.3147±0.04	0.2830
Spain	**0.3930±0.03**	0.3049±0.06	0.3487±0.06	**0.3124±0.03**	**0.2377±0.03**	**0.2395±0.03**	0.3418±0.05	**0.3111**
Russia	0.3353±0.01	0.3068±0.04	0.2885±0.03	0.2784±0.01	0.2264±0.05	0.2183±0.03	0.3157±0.03	0.2813
Others	0.2989±0.01	0.2720±0.05	0.2932±0.04	0.2452±0.01	0.2049±0.04	0.1939±0.01	0.3001±0.02	0.2583
ALL	0.3078±0.02	0.2854±0.05	0.2999±0.03	0.2549±0.02	0.2065±0.04	0.1972±0.02	0.3083±0.02	0.2657

(b) Performance results evaluated on **mWS**.

Table 6: Probing performance comparison with **Arabic** prompts and **FmLAMA-*ar*** sub-dataset. "B/mB" respectively represent abbreviations for BERT and mBERT, "XR" denotes XLM-RoBERTa, "b/l" stands for base/large and "c/u" stands for cased/uncased. **Bold** and underline represent the best-performing and second-performing cultural group within the same model. The *Pearson correlation coefficient* between mAP and mWS is 0.71 based on the results above.

Origin	Count	mB-c	mB-u	XRb	XRl	mT5	Avg.
Italy	59 (12.7%)	5.73±2.00	3.62±1.43	3.60±0.16	4.04±0.85	4.72±0.60	4.34
U.S.	67 (14.5%)	11.59±2.64	6.72±2.59	4.51±0.81	5.05±0.58	11.55±2.31	7.88
Turkey	12 (2.6%)	9.29±7.33	<u>9.31±5.64</u>	**8.83±0.45**	**9.00±0.78**	<u>13.64±7.43</u>	**10.01**
Japan	19 (4.1%)	8.37±4.61	5.81±3.26	<u>7.39±1.97</u>	<u>8.02±0.77</u>	8.56±3.44	7.63
France	65 (14.0%)	9.26±1.15	3.55±1.25	3.61±0.25	4.03±0.54	4.63±1.26	5.02
U.K.	23 (5.0%)	<u>16.17±3.36</u>	**9.89±4.31**	3.54±0.78	3.43±0.84	8.02±1.97	8.21
Mexico	11 (2.4%)	2.38±0.57	2.30±0.86	4.90±1.97	4.93±1.48	2.59±1.09	3.42
India	18 (3.9%)	4.83±3.24	5.68±6.31	5.30±2.41	6.45±0.15	6.30±4.71	5.71
Germany	11 (2.4%)	15.41±5.12	8.46±4.91	2.97±0.33	3.67±1.64	**14.20±4.70**	<u>8.94</u>
China	9 (1.9%)	3.94±3.35	2.63±1.49	6.12±1.36	6.47±0.97	4.19±4.65	4.67
Iran	6 (1.3%)	**16.57±7.01**	7.54±6.06	6.15±2.83	6.27±2.68	6.49±4.44	8.60
Greece	6 (1.3%)	4.33±2.31	1.85±0.60	6.73±2.89	6.82±2.87	4.24±1.18	4.79
Spain	7 (1.5%)	5.17±2.75	2.74±0.72	5.54±2.15	5.77±1.56	7.62±3.32	5.37
Russia	4 (0.9%)	5.22±2.97	4.82±3.18	5.91±0.27	5.44±0.84	3.53±1.26	4.98
Others	146 (31.5%)	5.48±0.57	3.60±1.46	4.94±0.84	5.15±0.43	4.38±1.34	4.71
ALL	463 (100.0%)	7.90±1.15	4.77±1.90	4.70±0.73	5.05±0.08	6.42±1.33	5.77

(a) Performance results evaluated on **mAP** (%).

	mB-c	mB-u	XRb	XRl	mT5	Average
Italy	0.3313±0.01	0.3286±0.03	0.2963±0.00	0.3032±0.02	0.2987±0.07	0.3116
U.S.	0.3706±0.02	0.3442±0.04	0.3090±0.01	0.3130±0.02	0.3565±0.07	0.3387
Turkey	0.3085±0.07	0.3182±0.05	0.3048±0.04	0.2958±0.01	0.2868±0.15	0.3028
Japan	0.2963±0.04	0.2998±0.06	0.2765±0.02	0.2708±0.01	0.2868±0.09	0.2860
France	0.3429±0.01	0.3278±0.04	0.2987±0.02	0.2948±0.01	0.3089±0.05	0.3146
U.K.	<u>0.3832±0.04</u>	**0.3776±0.02**	0.2777±0.03	0.2681±0.01	0.3217±0.06	0.3257
Mexico	0.3585±0.02	0.3534±0.09	0.3444±0.03	0.3260±0.01	0.3605±0.06	0.3486
India	0.3401±0.03	0.3431±0.06	0.2809±0.02	0.2901±0.04	0.2738±0.13	0.3056
Germany	0.3703±0.06	0.3536±0.04	0.2923±0.06	0.2829±0.04	0.3623±0.07	0.3323
China	0.2713±0.02	0.3245±0.05	0.2646±0.04	0.2531±0.02	0.2541±0.12	0.2735
Iran	**0.3856±0.08**	0.3494±0.08	**0.3827±0.00**	**0.3642±0.04**	<u>0.3724±0.06</u>	**0.3709**
Greece	0.3310±0.02	0.3040±0.02	0.3393±0.02	0.3388±0.02	0.3286±0.05	0.3283
Spain	0.3128±0.03	<u>0.3768±0.06</u>	0.3646±0.01	0.3488±0.03	**0.4115±0.10**	<u>0.3629</u>
Russia	0.3431±0.04	0.3713±0.07	<u>0.3803±0.04</u>	<u>0.3596±0.00</u>	0.3268±0.06	0.3562
Others	0.3048±0.01	0.3105±0.05	0.2890±0.02	0.2842±0.01	0.2782±0.08	0.2933
ALL	0.3321±0.01	0.3287±0.04	0.2979±0.02	0.2952±0.01	0.3068±0.07	0.3121

(b) Performance results evaluated on **mWS**.

Table 7: Probing performance comparison with **Hebrew** prompts and **FMLAMA-*he*** sub-dataset. "B/mB" respectively represent abbreviations for BERT and mBERT, "XR" denotes XLM-RoBERTa, "b/l" stands for base/large and "c/u" stands for cased/uncased. **Bold** and <u>underline</u> represent the best-performing and second-performing cultural group within the same model. The *Pearson correlation coefficient* between mAP and mWS is 0.31 based on the results above.

Origin	Count	Bb-ru	mB-c	mB-u	XRb	XRl	mT5	Avg.
Italy	101 (10.5%)	2.97±2.52	6.28±3.53	8.19±3.72	1.79±0.13	1.61±0.29	4.35±1.19	4.20
U.S.	79 (8.2%)	6.86±2.93	8.24±3.97	11.05±3.62	2.29±0.82	1.81±0.15	5.49±1.63	5.96
Turkey	28 (2.9%)	4.12±3.32	4.19±2.62	7.92±0.85	1.98±0.06	1.82±0.50	3.82±3.07	3.98
Japan	68 (7.1%)	4.41±2.32	3.79±1.22	5.37±2.72	1.67±0.22	1.52±0.02	1.91±1.15	3.11
France	93 (9.7%)	4.36±2.40	3.92±2.44	4.82±3.74	1.91±0.67	1.52±0.16	4.86±3.00	3.56
U.K.	33 (3.4%)	6.03±6.40	7.81±3.94	9.74±3.14	**3.14±3.90**	1.37±0.19	**9.79±6.82**	6.31
Mexico	22 (2.3%)	3.84±1.45	1.26±0.72	1.95±0.93	1.30±0.51	1.02±0.08	2.84±0.99	2.04
India	21 (2.2%)	5.86±5.73	**9.58±5.82**	**15.71±6.27**	1.55±0.43	1.49±0.55	5.29±4.03	**6.58**
Germany	45 (4.7%)	4.84±4.05	6.92±2.61	9.73±3.90	2.20±1.63	1.38±0.17	5.09±2.00	5.03
China	30 (3.1%)	**10.09±6.49**	6.93±2.75	10.59±5.73	1.29±0.99	0.83±0.09	5.08±3.75	5.80
Iran	13 (1.4%)	8.25±5.29	6.80±5.72	8.39±7.91	1.53±0.56	1.39±0.53	7.92±6.33	5.71
Greece	18 (1.9%)	3.54±4.18	0.99±0.22	1.15±0.43	1.48±0.09	1.51±0.14	4.09±1.45	2.13
Spain	38 (4.0%)	2.50±1.31	4.83±2.29	6.17±2.48	1.25±0.17	1.25±0.04	4.64±1.45	3.44
Russia	45 (4.7%)	3.61±1.79	3.23±1.01	4.10±2.41	2.18±0.89	**2.16±0.69**	2.90±0.74	3.03
Others	327 (34.0%)	3.36±2.13	4.31±1.80	5.79±1.31	1.77±0.15	1.57±0.29	3.68±1.00	3.41
ALL	961 (100.0%)	4.29±2.63	5.07±1.92	6.89±1.46	1.85±0.49	1.55±0.16	4.29±1.62	3.99

(a) Performance results evaluated on **mAP** (%).

	Bb-ru	mB-c	mB-u	XRb	XRl	mT5	Average
Italy	0.3229±0.06	0.3243±0.06	0.3730±0.06	0.2217±0.05	0.2206±0.05	0.3660±0.03	0.3048
U.S.	**0.3555±0.05**	**0.3370±0.07**	0.3895±0.03	0.2439±0.06	0.2336±0.03	0.3844±0.03	0.3240
Turkey	0.3129±0.05	0.3006±0.06	0.3554±0.05	0.2018±0.07	0.1915±0.04	0.3472±0.05	0.2849
Japan	0.2609±0.04	0.2622±0.02	0.2906±0.01	0.1619±0.04	0.1618±0.04	0.2712±0.01	0.2348
France	0.3233±0.05	0.2967±0.04	0.3378±0.03	0.2126±0.05	0.2099±0.04	0.3825±0.05	0.2938
U.K.	0.3513±0.08	0.3186±0.07	0.3735±0.03	0.2208±0.08	0.2059±0.05	**0.4273±0.07**	0.3162
Mexico	0.3264±0.03	0.3010±0.04	0.3348±0.04	0.2480±0.05	0.2431±0.03	0.3815±0.06	0.3058
India	0.2936±0.07	0.3112±0.09	**0.4053±0.04**	0.1877±0.05	0.1862±0.05	0.3235±0.04	0.2846
Germany	0.3403±0.06	0.3267±0.06	0.3825±0.07	0.2064±0.07	0.1894±0.03	0.3722±0.04	0.3029
China	0.3415±0.08	0.3144±0.04	0.3596±0.04	0.1873±0.06	0.1842±0.05	0.3515±0.04	0.2898
Iran	0.3358±0.07	0.3156±0.05	0.3699±0.06	0.2099±0.06	0.2020±0.04	0.4037±0.05	0.3061
Greece	0.3058±0.11	0.2960±0.05	0.3348±0.02	0.2252±0.04	0.2303±0.05	0.3716±0.02	0.2939
Spain	0.3287±0.04	0.3035±0.06	0.3532±0.03	0.2425±0.04	0.2406±0.03	0.3946±0.04	0.3105
Russia	0.3495±0.03	0.3152±0.05	0.3489±0.04	**0.2590±0.05**	**0.2575±0.05**	0.4177±0.03	**0.3246**
Others	0.3240±0.05	0.3105±0.05	0.3489±0.04	0.2186±0.05	0.2162±0.04	0.3660±0.03	0.2974
ALL	0.3245±0.05	0.3098±0.05	0.3536±0.03	0.2173±0.05	0.2132±0.04	0.3674±0.03	0.2976

(b) Performance results evaluated on **mWS**.

Table 8: Probing performance comparison with **Russian** prompts and **FMLAMA-*ru*** sub-dataset. "B/mB" respectively represent abbreviations for BERT and mBERT, "XR" denotes XLM-RoBERTa, "b/l" stands for base/large and "c/u" stands for cased/uncased. **Bold** and underline represent the best-performing and second-performing cultural group within the same model. The *Pearson correlation coefficient* between mAP and mWS is 0.70 based on the results above.

Origin	**Count**	**Bb-ky**	**Bb-kl**	**mB-c**	**mB-u**	**XRb**	**XRl**	**mT5**	**Avg.**
Italy	88 (10.9%)	6.77±2.85	7.01±2.68	1.19±0.15	3.14±2.45	1.90±1.27	2.96±3.44	6.81±1.50	4.25
U.S.	64 (7.9%)	10.72±5.32	10.49±4.97	2.74±1.35	4.84±4.86	3.07±3.62	3.25±3.75	11.65±2.84	6.68
Turkey	19 (2.4%)	3.23±2.85	2.77±0.81	3.58±2.81	4.53±2.97	1.83±0.07	2.37±0.71	1.75±0.39	2.87
Japan	101 (12.5%)	5.56±3.00	5.73±3.43	2.43±1.00	1.94±1.37	**3.45±0.68**	3.47±0.41	6.77±1.79	4.19
France	67 (8.3%)	2.33±0.74	3.56±1.47	1.40±0.31	2.76±0.17	1.60±0.24	1.62±0.20	2.42±0.12	2.24
U.K.	27 (3.3%)	**12.25±6.50**	**16.39±5.29**	**4.46±4.18**	**5.81±5.41**	2.11±1.30	3.81±4.73	**11.71±4.34**	**8.08**
Mexico	13 (1.6%)	7.40±5.58	4.58±1.71	1.99±0.22	5.58±4.32	1.79±0.17	1.78±0.21	2.28±1.25	3.63
India	32 (4.0%)	4.54±2.64	5.22±1.79	1.20±0.08	2.06±1.10	2.70±3.58	1.68±1.30	5.55±1.09	3.28
Germany	15 (1.9%)	2.26±1.15	3.27±0.71	1.56±0.19	2.74±2.45	2.26±2.37	2.06±1.56	3.51±2.73	2.52
China	54 (6.7%)	6.59±3.41	7.07±3.14	2.10±1.05	2.43±2.01	3.30±3.93	2.80±2.60	6.10±1.59	4.34
Iran	8 (1.0%)	1.74±0.76	8.47±6.64	2.28±0.13	1.50±0.80	1.30±0.48	1.15±0.21	1.56±0.53	2.57
Greece	9 (1.1%)	1.27±0.27	6.00±3.98	2.06±0.09	2.31±0.56	0.97±0.15	1.12±0.11	5.67±4.15	2.77
Spain	21 (2.6%)	2.53±1.12	2.95±1.22	3.84±2.58	2.53±2.12	3.04±2.33	**3.87±3.90**	4.26±2.11	3.29
Russia	8 (1.0%)	2.04±1.88	3.98±2.35	1.28±0.28	1.89±0.78	3.05±4.75	1.71±1.19	1.82±0.33	2.25
Others	281 (34.8%)	4.46±1.86	5.12±1.67	1.84±0.31	1.93±0.99	1.87±1.11	2.00±1.21	4.24±0.63	3.07
ALL	807 (100.0%)	5.42±2.33	6.09±2.16	2.05±0.71	2.68±1.74	2.31±1.52	2.49±1.70	5.56±0.81	3.80

Table 9: Probing performance comparison on **mAP** (%) with **Korean** prompts and **FMLAMA-*ko*** sub-dataset. “B/mB” respectively represent abbreviations for BERT and mBERT, “XR” denotes XLM-RoBERTa, “b/l” stands for base/large and “c/u” stands for cased/uncased. **Bold** and underline represent the best-performing and second-performing cultural groups within the same model.

Origin	**Count**	**Bb-zh**	**mB-c**	**mB-u**	**XRb**	**XRl**	**mT5**	**Avg.**
Italy	49 (5.8%)	13.30±4.45	10.57±2.58	14.37±2.34	10.52±3.21	11.61±3.41	10.15±1.19	11.75
U.S.	101 (11.9%)	15.92±3.83	14.10±2.67	16.08±1.70	9.00±5.25	9.31±4.08	12.67±0.77	12.85
Turkey	12 (1.4%)	8.26±1.41	4.10±1.61	4.91±1.53	2.43±0.85	2.02±0.54	4.94±0.48	4.44
Japan	114 (13.4%)	11.51±4.77	12.29±2.43	12.89±2.23	7.31±4.70	8.32±4.04	7.85±0.48	10.03
France	70 (8.2%)	15.34±4.47	15.31±3.04	15.29±2.19	10.36±5.22	12.79±3.97	7.90±0.97	12.83
U.K.	38 (4.5%)	**23.26±5.66**	**15.85±4.90**	**18.49±4.34**	**14.03±6.94**	**14.95±4.25**	9.00±2.26	**15.93**
Mexico	19 (2.2%)	17.20±3.14	12.52±1.39	12.63±1.61	10.68±3.44	13.99±1.16	**15.19±1.90**	13.70
India	24 (2.8%)	5.69±2.14	4.96±2.12	6.79±2.52	3.50±2.36	5.27±1.84	1.01±0.11	4.54
Germany	16 (1.9%)	6.97±2.09	11.28±1.30	13.08±3.16	5.61±4.45	5.74±2.34	7.55±0.64	8.37
China	87 (10.2%)	16.68±7.05	15.47±4.87	17.70±6.07	9.20±5.50	12.71±6.72	8.33±1.34	13.35
Iran	7 (0.8%)	9.25±1.04	5.60±1.05	11.30±6.38	7.38±6.42	8.71±6.74	3.21±0.09	7.58
Greece	7 (0.8%)	4.29±0.79	2.13±0.70	1.93±0.28	2.70±1.17	5.58±2.37	4.35±0.48	3.50
Spain	12 (1.4%)	11.98±1.11	7.52±2.42	10.18±1.07	7.55±2.97	9.46±3.33	11.65±1.92	9.72
Russia	8 (0.9%)	5.48±0.98	2.82±0.84	3.79±0.69	2.56±0.98	3.65±1.62	2.11±0.43	3.40
Others	251 (30.8%)	10.55±3.73	9.90±2.97	10.28±1.60	6.59±3.07	9.38±3.56	5.70±0.41	8.73
ALL	815 (100.0%)	12.99±3.96	11.78±2.65	13.01±2.06	8.05±4.00	9.98±3.69	7.88±0.55	10.62

Table 10: Probing performance comparison with **Chinese** prompts and **FMLAMA-*zh*** sub-dataset. “B/mB” respectively represent abbreviations for BERT and mBERT, “XR” denotes XLM-RoBERTa, “b/l” stands for base/large and “c/u” stands for cased/uncased. **Bold** and underline represent the best-performing and second-performing cultural groups within the same model.

Origin	mB-c	mB-u	mT5	Average
Italy	9.70±4.24	10.42±6.42	<u>12.61±2.28</u>	10.91
U.S.	<u>14.91±4.71</u>	12.86±5.63	11.45±3.22	<u>13.07</u>
Japan	8.27±3.62	9.04±5.07	6.99±3.13	8.10
France	8.26±3.15	8.13±4.40	5.00±2.18	7.13
U.K.	**19.84±7.17**	**13.97±9.13**	**13.11±7.97**	**15.64**
India	13.87±4.46	<u>13.94±7.00</u>	8.84±3.44	12.22
China	14.89±3.74	13.23±5.08	9.04±2.73	12.39
Indonesia	8.33±3.68	6.75±2.38	6.79±2.12	7.29
ALL	11.97±3.27	10.98±4.65	9.03±2.77	10.66

(a) EN_en: **mAP** (%)

Origin	mB-c	mB-u	mT5	Average
Italy	5.77±3.04	6.18±3.57	**4.21±1.12**	5.39
U.S.	5.11±2.31	5.75±1.41	2.62±0.63	4.49
Japan	10.51±2.87	<u>9.66±4.55</u>	2.75±1.23	<u>7.64</u>
France	5.02±2.41	6.06±1.66	2.76±0.54	4.61
U.K.	3.14±1.06	3.77±0.98	2.93±0.54	3.28
India	**14.43±9.56**	**16.22±7.65**	<u>4.05±1.59</u>	**11.57**
China	3.87±1.39	6.08±3.77	2.85±1.04	4.27
Indonesia	<u>10.60±1.80</u>	8.76±3.04	3.32±0.90	7.56
ALL	6.77±2.26	7.28±1.95	2.99±0.65	5.68

(b) EN_zh: **mAP** (%)

Origin	mB-c	mB-u	mT5	Average
Italy	7.25±3.72	6.66±5.45	7.15±1.16	7.02
U.S.	13.23±3.64	7.69±5.49	**11.42±2.11**	10.78
Japan	12.19±3.70	6.45±4.36	7.61±1.81	8.75
France	11.64±4.46	7.25±5.96	7.66±2.84	8.85
U.K.	<u>13.60±6.01</u>	**9.17±7.54**	<u>11.34±4.36</u>	**11.37**
India	3.70±1.51	3.03±1.66	1.57±0.18	2.77
China	**15.04±3.69**	<u>9.09±8.25</u>	8.80±3.81	<u>10.98</u>
Indonesia	9.65±1.93	4.88±2.55	5.41±2.06	6.65
ALL	11.88±3.49	7.22±5.48	8.40±2.22	9.17

(c) ZH_zh: **mAP** (%)

Origin	mB-c	mB-u	mT5	Average
Italy	1.53±0.20	2.28±0.25	2.52±0.56	2.11
U.S.	2.02±0.21	**3.46±0.49**	**3.45±0.69**	**2.98**
Japan	<u>2.59±0.44</u>	2.74±0.73	<u>2.65±0.15</u>	2.66
France	2.04±0.60	2.41±0.58	2.42±0.44	2.29
U.K.	1.42±0.22	<u>3.40±1.27</u>	2.29±0.35	2.37
India	2.03±0.37	2.53±0.77	2.23±0.91	2.26
China	**2.85±0.44**	2.99±1.34	2.29±0.56	<u>2.71</u>
Indonesia	2.04±0.27	2.36±0.36	1.25±0.20	1.88
ALL	2.20±0.33	2.85±0.64	2.58±0.34	2.54

(d) ZH_en: **mAP** (%)

Table 11: Code-switching analysis: Probing results by using prompts **without** introducing cultural background.

Origin	mB-c	mB-u	mT5	Average
Italy	9.20±4.55	9.97±4.71	11.27±2.42	10.15
U.S.	15.49±4.37	13.05±5.40	11.37±4.10	13.30
Japan	11.81±3.16	11.84±3.99	10.11±3.43	11.25
France	8.29±2.81	8.93±4.29	5.92±1.38	7.71
U.K.	**20.02±9.00**	<u>15.81±7.65</u>	**15.79±6.17**	**17.21**
India	<u>17.95±2.38</u>	**21.88±7.42**	<u>11.51±4.65</u>	<u>17.11</u>
China	14.77±3.49	13.75±6.35	9.13±3.70	12.55
Indonesia	9.02±1.01	8.75±2.36	8.11±3.19	8.63
ALL	13.06±3.26	12.42±4.46	10.12±3.16	11.87

(a) EN_en: **mAP** (%)

Origin	mB-c	mB-u	mT5	Average
Italy	7.57±1.67	6.79±1.71	4.18±1.02	6.18
U.S.	6.01±1.71	7.51±1.08	2.86±0.55	5.46
Japan	<u>13.96±2.05</u>	<u>10.45±3.31</u>	4.77±1.61	<u>9.73</u>
France	5.28±1.72	6.63±0.78	2.92±0.70	4.94
U.K.	3.19±0.91	3.60±0.77	4.44±1.47	3.74
India	**23.58±2.64**	**19.51±6.39**	**6.59±2.59**	**16.56**
China	5.23±0.14	6.39±3.55	3.98±1.08	5.20
Indonesia	10.12±1.01	7.91±2.69	<u>6.35±1.81</u>	8.13
ALL	8.57±1.25	8.09±1.35	4.10±0.98	6.92

(b) EN_zh: **mAP** (%)

Origin	mB-c	mB-u	mT5	Average
Italy	6.35±3.08	7.67±4.89	6.33±1.80	6.78
U.S.	13.60±3.06	8.50±5.41	<u>10.17±1.91</u>	10.76
Japan	13.02±2.43	7.32±4.47	8.83±1.60	9.72
France	12.12±4.46	<u>9.00±5.67</u>	7.48±2.82	9.53
U.K.	<u>13.97±6.02</u>	8.21±7.81	**10.87±4.47**	<u>11.02</u>
India	4.60±1.88	3.38±2.06	1.82±0.30	3.27
China	**20.49±4.29**	**10.22±8.87**	9.71±4.53	**13.47**
Indonesia	9.09±3.50	4.12±2.30	5.53±1.74	6.25
ALL	13.07±3.13	7.99±5.34	8.46±2.24	9.84

(c) ZH_zh: **mAP** (%)

Origin	mB-c	mB-u	mT5	Average
Italy	1.57±0.18	2.73±0.63	<u>2.77±0.61</u>	2.36
U.S.	2.29±0.13	**3.61±1.16**	**4.30±0.73**	**3.40**
Japan	2.79±0.65	2.69±1.35	2.35±0.60	<u>2.61</u>
France	<u>3.02±0.86</u>	2.56±0.70	2.25±0.20	<u>2.61</u>
U.K.	1.77±0.21	2.38±1.43	2.24±0.38	2.13
India	2.30±0.51	<u>2.85±1.29</u>	2.30±0.86	2.48
China	**3.12±0.71**	2.39±1.70	2.31±0.45	<u>2.61</u>
Indonesia	2.39±0.46	2.28±0.77	1.20±0.12	1.96
ALL	2.54±0.42	2.77±1.07	2.68±0.21	2.66

(d) ZH_en: **mAP** (%)

Table 12: Code-switching analysis: Probing results by using prompts **with** introducing cultural background.

Origin	Llama2-7b-chat-hf				Llama2-13b-chat-hf			
	EN_en	**EN_zh**	**ZH_zh**	ZH_en	**EN_en**	**EN_zh**	**ZH_zh**	**ZH_en**
Italy	**42.86**	20.41	2.04	2.04	48.98	**30.61**	6.12	10.20
U.S.	37.62	13.86	4.95	4.95	**49.50**	27.72	5.94	12.87
Japan	14.91	16.67	3.51	4.39	20.18	15.79	4.39	6.14
France	39.71	22.06	5.88	11.76	48.53	26.47	2.94	10.29
U.K.	36.84	18.42	**15.79**	5.26	39.47	21.05	15.79	7.89
India	20.83	12.50	4.17	4.17	33.33	12.50	12.50	**16.67**
China	17.65	**23.53**	10.59	15.29	21.18	24.71	**16.47**	15.29
Indonesia	6.90	10.34	0	6.90	24.14	13.79	3.45	3.45
All	27.36	17.91	5.91	7.28	35.04	22.64	7.87	10.43

(a) Performance results of Llama2-chat.

Origin	vicuna-7b				vicuna-13b			
	EN_en	**EN_zh**	**ZH_zh**	ZH_en	**EN_en**	**EN_zh**	**ZH_zh**	ZH_en
Italy	**51.02**	26.53	8.16	12.24	**42.86**	30.61	18.37	22.45
U.S.	49.50	**27.72**	17.82	13.86	37.62	25.74	20.79	22.77
Japan	21.93	14.04	7.02	7.02	16.67	14.04	9.65	7.02
France	44.12	23.53	11.76	13.24	39.71	20.59	14.71	17.65
U.K.	47.37	21.05	15.79	10.53	34.21	**31.58**	**23.68**	26.32
India	41.67	4.17	12.50	12.50	16.67	16.67	12.50	12.50
China	23.53	18.82	**27.06**	15.29	17.65	18.82	20.00	22.35
Indonesia	20.69	6.90	10.34	13.79	20.69	10.34	13.79	13.79
All	36.22	19.69	14.37	12.01	28.15	20.87	16.54	17.72

(b) Performance results of vicuna.

Origin	gpt-3.5-turbo			
	EN_en	**EN_zh**	**ZH_zh**	**ZH_en**
Italy	**69.39**	**48.98**	20.41	26.53
U.S.	60.40	39.60	22.77	25.74
Japan	28.95	31.58	10.53	16.67
France	63.24	33.82	17.65	27.94
U.K.	57.89	36.84	26.32	26.32
India	54.17	12.50	12.50	16.67
China	30.59	34.12	**31.76**	**37.65**
Indonesia	34.48	13.79	20.69	13.79
All	47.64	34.06	20.28	25.00

(c) Performance results of gpt-3.5-turbo.

Table 13: **Accuracy** (%) of decoder-only LLMs' probing results in each cultural group.