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

Li Zhou[1,2], Taelin Karidi[3], Wanlong Liu[4], Nicolas Garneau[5], Yong Cao[6], Wenyu Chen[4], Haizhou Li[1,2], Daniel Hershcovich[5]

[1]The Chinese University of Hong Kong, Shenzhen [2]Shenzhen Research Institute of Big Data
[3]Hebrew University of Jerusalem [4]University of Electronic Science and Technology of China,
[5]University of Copenhagen, [6]University of Tübingen, Tübingen AI Center

{lizhou21, haizhouli}@cuhk.edu.cn, taelin.karidi@mail.huji.ac.il, liuwanlong@std.uestc.edu.cn,
{nicolas.garneau, dh}@di.ku.dk, yongcao2018@gmail.com, cwy@uestc.edu.cn

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* in English 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 (bó gěn dì niú ròu) (its literal translation). In China, the dish is also commonly referred to by its broader description, 红酒炖牛肉 (hóng jiǔ dùn niú ròu) (*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 (fǎ shì hóng jiǔ dùn niú ròu) (*French-style Red Wine Stewed Beef*) with an adjectival description can indicate adherence to French culinary traditions. This practice illustrates how cultural and linguistic nuances shape knowledge transmission. In cross-cultural communication, when direct translations are unavailable, speakers often switch between languages—a phenomenon known as code-switching—to better convey meaning (Aguilar and Solorio, 2020; Doğruöz et al., 2021). This variability underscores the challenges language models encounter in navigating cross-cultural culinary contexts and highlights the difficulty of responding with cultural knowledge, which involves using specific language, engaging in code-switching, and integrating cultural context.

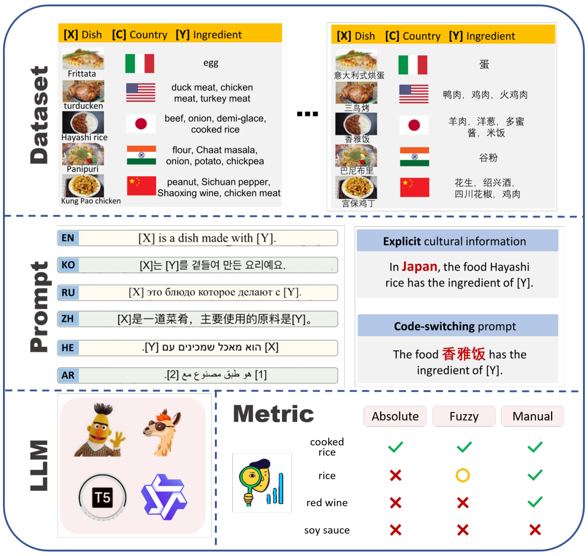


Figure 1: Summary of the various aspects of our work.

Trained on vast datasets, Large Language Models (LLMs) encode a wide array of knowledge, but also face challenges with various biases, such as those related to gender (Savoldi et al., 2021; Kaneko et al., 2022), belief (Søgaard, 2021; González et al., 2021; Lent and Søgaard, 2021), and culture (Cao et al., 2023; Deshpande et al., 2022;

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 9840–9867

Yin et al., 2022; Mukherjee et al., 2024; Singh et al., 2024). In this paper, we adopt food-domain knowledge data as cultural semantic proxy, "country of origin" as cultural demographic proxy (Adilazuarda et al., 2024), and address the following research questions: (1) What cultural biases exist within the cultural knowledge contained in LLMs? (2) How can we best prompt and evaluate LLMs to elicit culture-specific knowledge? Specifically, we first design an automated construction method for the dataset FmLAMA (§3). Next, we introduce the cultural knowledge probing method (§4), encompassing the probing task, template and metric. We then conduct experiments (§5) and analysis (§6) to address our proposed research questions. Finally, we perform an error analysis on the probing results (§7). Figure 1 provides an overview of our work's various aspects, and our **contributions** are:

- We present FmLAMA, a pioneering dataset focused on the food domain, which is inherently rich in cultural diversity. This dataset is a multifaceted tool for probing LLMs across cultures and languages.
- We propose novel metrics aimed at evaluating LLMs' capacity to accurately and sensitively probe cultural knowledge, utilizing a combination of automatic and manual evaluation methods.
- 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. [1]

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. Appendix A.1 shows a comparison of our dataset with existing ones.

Knowledge probing. Deciphering the knowledge encoded by LLMs poses significant challenges due to their opaque nature , early benchmarks like LAMA (Petroni et al., 2019) 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*.

3 FmLAMA Construction

To assess whether LLMs encode and access cultural information, we develop FmLAMA, a mul-

[1] Data and code are available in `https://github.com/lizhou21/FmLAMA-master`.

ticultural, 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 $I \xrightarrow{\text{(instance of|subclass of)+}}$ *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. The property "has part(s)" identifies food ingredients for a 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. In each instance, the descriptive language for all attributes remains consistent.

Ultimately, our constructed dataset, FmLAMA, comprises 33,600 dish instances, detailed by name, origin, ingredients, and optionally, materials and images, encompassing 128 cultural country groups and 250 languages. The average number of ingredients in FmLAMA dataset is 2.04. Examples and statistics are provided in Appendix A.

4 Cultural Knowledge Probing

4.1 Probing task

We adopt Word Prediction (WP) as the knowledge-probing task, following previous work (Fierro and Søgaard, 2022; Wu et al., 2023; Fierro et al., 2024b). Firstly, we manually design a prompt template t focused on the core attribute "has part(s)", illustrating a connection between a dish (subject) and its ingredient(s) (object). WP is usually implemented as a candidate retrieval problem. The candidate set consists of all objects in the focused, filtered sub-dataset of FmLAMA.[2] The 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. Using subject-object tuples ([X], [Y]) as queries, we probe LLMs by replacing the subject and masking the object. Considering that each candidate object is tokenized into d subtokens $\{c_1, \cdots, c_d\}$ by LLMs correspondingly, we apply [MASK] token of varying lengths to the objects within each query. D queries are constructed for each dish subject X based on the same template t, D is the maximum number of object tokens, each query Q_d^X is defined as:

$$Q_d^X = t\left(X, [\text{MASK}] * d\right), d \in [1, D]. \quad (1)$$

Probability acquisition. We use Mean Pooling[3] method to obtain the prediction probability of each candidate. Specifically, for candidate object $C = \{c_1, \cdots, c_d\}$ of length d, its probability is obtained from the likelihoods associated with the [MASK] tokens in Q_d^X, and the predicted probability of C is calculated as the average of the probabilities of composing its subtokens:

$$P\left(Q_d^X, C\right) = \frac{1}{d}\sum_{i=1}^{d} p\left(Q_d^X, [\text{MASK}]_i = c_i\right), \quad (2)$$

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

[2] 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.

[3] Mean Pooling has been shown to be superior to other pooling methods for multi-token probing (Wu et al., 2023).

4.2 Probing template

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. For example, "[X] is a dish made with [Y]", "[X] is a type of food that includes [Y]".

Cultural information. To explore the impact of introducing explicit cultural context on LLMs' ability to access cultural knowledge, we enhance the basic templates by integrating country information. For instance, "In [C], [X] is a dish made with [Y]", where [C] denotes the country of origin for the dish [X], [Y] indicates the ingredient object.

Multilingual prompts. To explore how different prompt language settings affect LLMs' cultural knowledge probing abilities, we craft prompts in six languages written by native speakers, including English (*en*), Chinese (*zh*), Arabic (*ar*), Korean (*ko*), Russian (*ru*), and Hebrew (*he*). 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. These prompt illustrations and language details are depicted in Appendix B.

Code-switching. To simulate real-world scenarios where mixed-language expression often occurs, we implement code-switching prompts, varying the main language (ML) and subject language (SL). Specifically, we define the code-switching prompt setting as $\mathrm{P(ML, SL)}$, where $\mathrm{ML}, \mathrm{SL} \in \{en, zh, ar, ko, ru, he\}$, and $\mathrm{ML} \neq \mathrm{SL}$. Our objective is to ensure that the language of the predicted object remains consistent with the main language (ML). For instance, the prompt "勃艮第牛肉 (bó gěn dì niú ròu) is a dish made with [Y]" follows the $\mathrm{P}(en, zh)$ setting, with the expected output for [Y] in English.

4.3 Probing Metric

Despite considerable prior work on 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 verbalization of ingredients is not unique (e.g., in Chinese, both 盐 (yán) and 食盐 (shí yán) can denote *salt*), and (4) ingredients can be flexible (e.g., in the Italian dish *frico*, listed with *cheese* as an ingredient in Wikidata, either *mozzarella* or *feta* are actually valid). (5) The reference ingredients label in Wikidata is incomplete.

Given these constraints, we introduce two automatic metrics: the absolute-match metric, Mean Average Precision (mAP), and the fuzzy-match metric, Mean Word Similarity (mWS), along with a Manual Evaluation Score (MES) that incorporates both LLM-simulated and real human assessments for implementation. All mAP, mWS, and MES metrics are in the range $[0, 1]$.

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 reference 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 reference ingredients set, and 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:

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

where n refers to the size of the candidate object set, and rel@k is 1 if the object at rank k is relevant to food item i, otherwise 0. 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 dish 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 reference objects. First, we define the similarity score $S(i, g)$ for each ingredient g within each dish i. Only the predicted objects in the top-l rankings of the model prediction that are most similar to g

contribute to the evaluation score, where $l = |\text{ing}_i|$. $S(i,g)$ is defined as follows:

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

where w_g and w_p are the `Fasttext` (Bojanowski et al., 2017; Joulin et al., 2016) vectors for the ingredient g and the predicted object p, respectively, and $\cos(\cdot)$ 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:

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

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

Manual Evaluation Score. MES considers only the top-l predicted ingredients for evaluation in each dish, similar to the mWS metric, where l is the number of reference ingredients. A predicted ingredient is considered correct if it satisfies any of the following criteria: (1) *Direct Match*: The ingredient is exactly in the Wikidata reference; (2) *Substitutability*: The ingredient can replace one in the reference; (3) *Missing Traditional Ingredient*: The ingredient is traditionally or commonly used in the dish, but not listed in the reference. Given a manual evaluator $\mathrm{M}(\cdot)$, the MES is defined as follows:

$$\text{S}_i = \frac{1}{l} \sum_{p \in \text{topl}_i} \mathrm{M}(p, \text{ing}_i) \quad (9)$$

$$\text{MES} = \frac{1}{|G|} \sum_{i \in \text{G}} \text{S}_i \quad (10)$$

where $\mathrm{M}(\cdot)$ is 1 if the top-l predicted ingredient p is evaluated as meeting the criteria; otherwise, 0.

5 Experiments: Existing Cultural Biases

5.1 Experimental setup

Baselines We explore encoder-only LLMs, including BERT (Devlin et al., 2019) and mBERT, encoder-decoder LLMs, such as T5 (Raffel et al., 2020) and mT5 (Xue et al., 2020), as well as decoder-only LLMs like Qwen2 (Yang et al., 2024), Llama2 (Touvron et al., 2023), and Llama3 (AI@Meta, 2024). Of these, BERT and T5 are English monolingual LLMs, while the remaining five are multilingual LLMs. [4] Additionally, we employ a 'dumb' baseline, in which the model is assumed to consistently predict the top-10 most common ingredients for each dish.

Cultural groups We identify the top 30 countries with the most dishes across six languages, take their intersection to ensure sufficient data, consider geographical diversity, and ultimately narrow our focus to 14 cultural groups, as shown in Table 1.

Metrics We evaluate all LLMs using two automated metrics. Due to cost constraints, we conduct manual evaluation only on the LLM that achieve the highest probing performance in the automated metrics. Specifically, for the evaluator $\mathrm{M}(\cdot)$ used to compute MES, we utilize GPT-4o (OpenAI et al., 2024)[5] as a simulated evaluator and recruit real human evaluators from Prolific.[6] For each country group, we hire 3 evaluators familiar with the related cuisine, with a pay rate of £9/hour, considered mid-level on the platform. Details on the GPT-4o evaluation prompts and the human evaluation platform are provided in the Appendix C.

5.2 Automatic evaluation results

The probing results based on English prompt on the filtered dataset FMLAMA-*en* are illustrated in Table 1. The average Pearson correlation between mAP and mWS across all LLMs is 0.72, indicating a strong positive relationship and demonstrating that the experimental results are consistent across both automatic metrics. Overall, without fine-tuning, decoder-only LLMs exhibit significantly better cultural knowledge recall in such challenging tasks compared to encoder-only and encoder-decoder LLMs, highlighting their superior capacity as knowledge bases (Petroni et al., 2019). Besides, results across all models (Avg. Column) indicate that the groups from the U.S. and India, whose official languages include English, perform the best generally. They consistently rank in the top 3 in both automatic evaluation metrics. This pattern is especially pronounced in the results for encoder-only LLMs and encoder-decoder LLMs. Apart from the Iran group[7], these two cultural groups also

[4] Monolingual LLM configurations, as well as results for the five languages on their respective filtered sub-datasets, are provided in the Appendix D.1.

[5] Version: gpt-4o-2024-08-06

[6] `https://www.prolific.co/`

[7] See detailed analysis in Appendix E.

Origin	Count	Dumb Base	Encoder-only LLMs			Encoder-Decoder LLMs		Decoder-only LLMs			Avg.
			Bb	Bl	mB	T5	mT5	Qwen2	Llama2	Llama3	
Italy	215 (13.9%)	18.14	8.12±1.14	7.52±0.43	9.31±1.33	4.54±0.56	6.73±1.34	26.82±4.36	19.37±3.68	29.62±3.30	14.00
U.S.	285 (18.4%)	11.10	**19.98±2.00**	**18.83±1.95**	**19.91±5.88**	**10.24±2.62**	**16.50±4.86**	32.21±2.05	21.64±5.75	36.58±3.58	21.99
Turkey	98 (6.3%)	12.90	11.80±1.03	9.84±3.81	14.98±3.89	6.10±1.57	10.25±2.91	25.22±1.46	22.96±4.04	30.01±2.30	16.39
Japan	186 (12.0%)	9.35	11.10±1.19	9.14±1.81	11.52±1.89	5.45±1.90	6.38±1.00	23.15±2.01	20.24±3.10	25.47±1.73	14.06
France	175 (11.3%)	16.50	9.33±1.24	7.51±0.90	8.80±2.34	3.86±1.62	4.57±1.11	29.56±2.38	21.95±4.83	30.01±2.19	14.45
U.K.	83 (5.4%)	**18.67**	12.47±1.77	13.72±2.20	14.20±4.01	6.16±2.89	9.88±3.06	29.06±2.47	21.82±3.43	32.15±4.06	17.43
Mexico	57 (3.7%)	9.40	8.25±1.04	10.35±2.88	8.87±2.18	3.41±0.49	5.96±2.36	25.14±1.58	22.03±5.65	28.97±3.54	14.12
India	132 (8.5%)	11.57	18.42±1.35	16.94±3.10	17.51±1.75	9.18±2.98	7.80±1.15	35.82±1.79	33.88±5.00	37.62±1.89	**22.15**
Germany	57 (3.7%)	14.02	8.54±1.77	7.84±1.01	9.00±2.39	4.38±1.48	6.68±1.56	29.05±2.63	20.42±5.56	28.86±3.78	14.35
China	97 (6.3%)	8.60	15.06±2.45	14.94±1.37	15.24±5.31	7.80±1.97	12.39±2.62	28.96±3.15	21.02±4.78	30.39±2.70	18.23
Iran	21 (1.4%)	12.60	9.16±1.75	8.19±1.51	13.56±3.16	9.00±2.04	6.21±0.15	**40.26±8.19**	**35.36±9.81**	**45.72±3.65**	20.93
Greece	21 (1.4%)	15.07	4.09±1.10	3.48±1.45	4.85±1.37	3.81±1.09	1.32±0.15	31.68±8.85	15.88±2.54	31.60±3.17	12.09
Spain	95 (6.1%)	16.05	9.18±0.93	7.18±1.04	7.42±1.51	3.98±1.37	3.33±0.46	25.21±5.07	17.00±4.23	25.36±1.68	12.33
Russia	27 (1.7%)	10.64	4.72±0.98	7.10±2.81	6.44±1.97	1.41±0.60	2.21±0.33	14.83±4.80	10.47±0.98	11.77±0.59	7.37
ALL	1549 (100.0%)	13.29	12.56±1.06	11.58±1.29	12.96±2.84	6.32±1.42	8.68±2.03	28.56±2.27	21.89±4.06	30.96±2.09	16.69
Coefficient of Variation (CV)		24.15	40.87	40.76	36.89	43.29	54.29	20.39	28.27	23.84	-

(a) Performance results evaluated using **mAP** (%). CV indicates the extent of cultural bias in each model's performance.

Origin	Encoder-only LLMs			Encoder-Decoder LLMs		Decoder-only LLMs			Avg.
	Bb	Bl	mB	T5	mT5	Qwen2	Llama2	Llama3	
Italy	0.3813±0.02	0.3624±0.01	0.3573±0.02	0.3129±0.02	0.3195±0.01	0.4930±0.02	0.4510±0.02	0.5057±0.01	0.3979
U.S.	**0.4461±0.02**	**0.4358±0.02**	**0.4219±0.06**	0.3516±0.02	**0.4076±0.03**	0.5147±0.02	0.4656±0.04	0.5411±0.02	**0.4481**
Turkey	0.4130±0.02	0.3886±0.03	0.3954±0.04	0.3387±0.03	0.3419±0.02	0.4689±0.01	0.4612±0.03	0.5022±0.01	0.4137
Japan	0.3778±0.02	0.3501±0.03	0.3446±0.03	0.3059±0.03	0.2932±0.01	0.4416±0.01	0.4304±0.02	0.4436±0.01	0.3734
France	0.3750±0.02	0.3660±0.02	0.3497±0.04	0.3279±0.02	0.3367±0.01	0.5201±0.02	0.4697±0.03	0.5099±0.01	0.4069
U.K.	0.3708±0.02	0.3775±0.02	0.3583±0.04	0.3031±0.02	0.3320±0.03	0.4532±0.03	0.4291±0.01	0.4620±0.04	0.3857
Mexico	0.3509±0.01	0.3654±0.03	0.3255±0.03	0.3287±0.01	0.3262±0.02	0.4547±0.02	0.4762±0.04	0.4863±0.03	0.3892
India	0.4402±0.01	0.4193±0.04	0.4135±0.02	0.3380±0.04	0.3141±0.01	0.5106±0.00	0.5140±0.03	0.5190±0.02	0.4336
Germany	0.3699±0.03	0.3492±0.03	0.3698±0.03	0.2987±0.02	0.3358±0.01	0.4987±0.02	0.4758±0.05	0.4983±0.01	0.3995
China	0.3676±0.02	0.3532±0.02	0.3317±0.05	0.3368±0.03	0.3670±0.03	0.4521±0.04	0.4285±0.02	0.4671±0.02	0.3880
Iran	0.3950±0.03	0.3747±0.03	0.3972±0.03	0.3384±0.04	0.3303±0.02	**0.5510±0.06**	**0.5498±0.07**	**0.6086±0.03**	0.4431
Greece	0.3854±0.03	0.3707±0.03	0.3461±0.02	**0.3569±0.04**	0.3257±0.01	0.5415±0.08	0.4459±0.02	0.5527±0.02	0.4156
Spain	0.3761±0.01	0.3386±0.02	0.3290±0.02	0.3215±0.02	0.2923±0.01	0.4630±0.03	0.4247±0.03	0.4641±0.01	0.3762
Russia	0.3686±0.01	0.3778±0.03	0.3286±0.01	0.3386±0.01	0.2912±0.02	0.4272±0.04	0.3969±0.02	0.4053±0.00	0.3668
ALL	0.3959±0.01	0.3803±0.02	0.3699±0.03	0.3276±0.02	0.3380±0.01	0.4864±0.01	0.4573±0.03	0.4981±0.01	0.4067
Corr.	0.35	0.59	0.38	0.57	0.83	0.86	0.71	0.92	-

(b) Performance results evaluated using **mWS** with `Fasttext`. Corr. represents the correlation between the FastText and BERT base embedding evaluation results.

Table 1: **Automatic Evaluation**: Probing performance comparison with English prompts and FMLAMA-*en* sub-dataset. "B/mB" respectively represent abbreviations for BERT and mBERT, and "b/l" stands for base/large. **Bold** and underline indicate the best and second-best performing cultural groups, respectively, for each model (within each column). The average `Pearson correlation` between mAP and mWS across LLMs is 0.72.

	Bb	Bl	mB	T5
mAP	0.60	0.49	0.48	0.56
mWS	0.48	0.38	0.36	0.41
	mT5	**Qwen2**	**Llama2**	**Llama3**
mAP	0.38	0.05	0.04	0.13
mWS	-0.06	0.05	-0.04	0.02

Table 2: Correlation between Country Dish Counts and Probing Performance.

demonstrate the best performance in decoder-only LLMs. Especially, for the 'dumb' baseline, the U.S. probing performance is not the best, suggesting that the data distribution does not influence the LLM's cultural judgment capabilities. This further strengthens the persuasiveness of our experiment.

Additionally, (1) to assess the variation in model performance across different cultural groups, we calculate the Coefficient of Variation based on the absolute-match mAP metric. This serves as an indicator of the degree of cultural bias in each model's performance. Our results show that decoder-only LLMs exhibit less cultural bias compared to other model types. Among these, the newer version, Llama3, demonstrates greater consistency in performance than Llama2. Similarly, the non-English-dominant Qwen2 LLM also shows smaller variations in cultural bias. (2) To explore the impact of different embedding methods in mWS, we use BERT Base with mean pooling to evaluate probing performance under the mWS metric. Results

Origin	GPT-4o	Human	Origin	GPT-4o	Human	Origin	GPT-4o	Human
Italy	52.40	55.70	**U.K.**	64.12	**72.33**	**Iran**	53.39	51.70
U.S.	**70.06**	<u>70.26</u>	**Mexico**	47.05	51.95	**Greece**	47.68	51.23
Turkey	55.19	53.59	**India**	<u>64.22</u>	63.72	**Spain**	43.02	53.16
Japan	50.76	56.07	**Germany**	51.71	59.02	**Russia**	31.42	44.86
France	52.03	58.28	**China**	61.52	68.96	**All**	56.58	-

Table 3: **Manual Evaluation:** Llama3 probing results evaluated using **MES**. The human evaluation score is the average MES from three evaluators per country group. Since evaluators vary by country group, the overall result (ALL) is shown as "-".

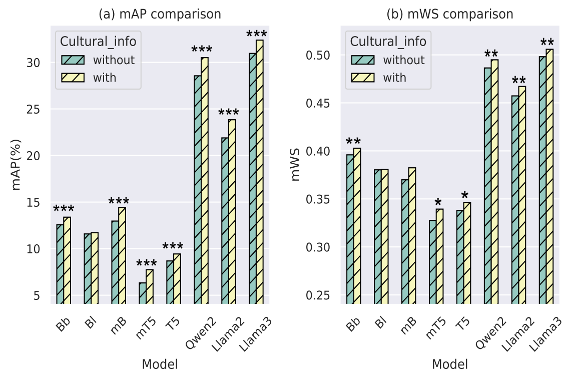


Figure 2: Comparative probing results on FMLAMA-*en*: incorporating cultural information about the origin of dishes into English prompts can enhance the probing of cultural knowledge. Significance levels are indicated by $*$ for $p < 0.1$, $**$ for $p < 0.05$, and $***$ for $p < 0.01$.

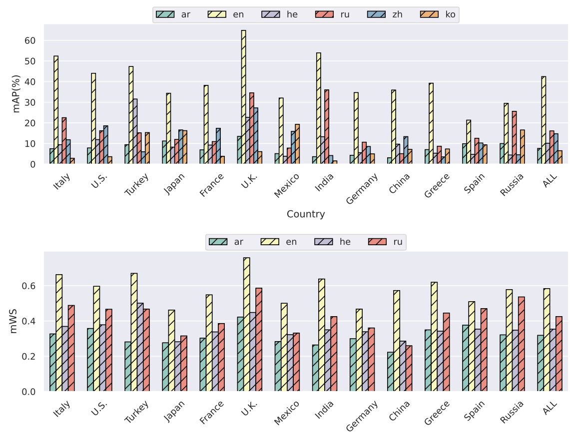


Figure 3: Comparison of probing results on Llama3 with prompts in different languages, showing that the English prompt exhibits the best performance.

(see Table 8) show that the U.S. cultural group achieves the highest performance, while decoder-only LLMs exhibit significant improvement, consistent with findings using FastText. We calculate the correlation between mWS evaluations using BERT Base and FastText for each LLM. The results reveal a strong correlation for decoder-only LLMs, as shown in Table 1. (3) To explore the correlation between training data size and cultural probing performance, we use country dish counts as a proxy for Wikidata's cultural data size and calculate their correlation with probing performance. The correlation results are show as Table 2. Based on the probing performance evaluated by the two automated metrics, LLMs like Bb, B1, and T5 consistently exhibit little correlations with cultural data size. In contrast, decoder-only LLMs show almost no correlations, indicating limited sensitivity to cultural data size as represented by dish counts. Therefore, in more advanced LLMs, cultural competency isn't directly tied to the training sample size of specific cultural groups.

5.3 Manual evaluation results

Table 1 shows that Llama3 achieve the best probing performance, so we conduct manual evaluation on its probing results. Table 3 displays the results of the GPT-4o simulator and human evaluations, showing a high Pearson correlation of 0.88. Both evaluation methods consistently show that the highest MES scores come from English-speaking cultural groups, including the U.S., U.K., and India, with the U.S. performing especially well in both manual evaluations. This aligns closely with the automatic evaluation results, reinforcing the argument that LLMs tend to exhibit cultural bias toward these groups.

Upon reviewing the evaluation results for the GPT-4o simulator, we find that it treats synonymous expressions as *Direct Matches*, e.g., *cooked rice* vs. *gohan* (the Japanese term) and *cauliflower* vs. *Brassica oleracea var. botrytis* (the botanical term). Additionally, reference labels from Wikidata often exhibit inconsistencies, with terms like "shrimp or prawn" and "fish as food" marked as ingredients. Furthermore, GPT-4o flags pre-made

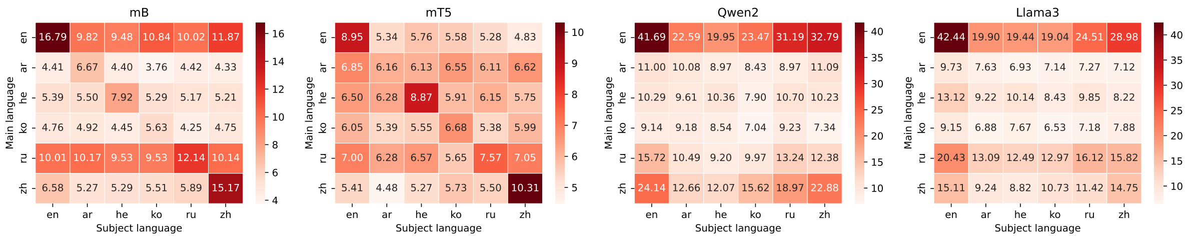


Figure 4: Probing results using code-switching prompts on Qwen2 and LLaMA3.

ingredients as *Substitutable*; for example, *flour* in Garganelli with pre-made *pasta*. For human evaluators, evaluation results may vary due to individual subjective biases. The average Inter-Rater Reliability (IRR) across country groups, measured by Fleiss' kappa, is 0.57,[8] indicating moderate agreement. While this indicates acceptable reliability for subjective tasks, it also highlights that evaluators from the same cultural background may still show variation in their judgments. Although GPT-4o can partially simulate human evaluations, its hallucination issues (Ji et al., 2023) raise concerns about the reliability of its assessments. These observations underscore the inherent challenges of using automatic metrics for culturally-related tasks.

6 Analysis: Designing Better Prompts

We examine the impact of different prompt settings on cross-cultural knowledge exploration, focusing on cultural background references, multilingual prompts, and code-switching scenarios.

Cultural background analysis. Using the English prompts from Figure 9 as a basis, we incorporate information about the dish country of origin into the probing prompts for each specific dish (as described in §4.2). Figure 2 presents the comparative probing results on FMLAMA-*en* across different LLMs, considering the inclusion of cultural information mentions in the prompts. We find that 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.

Multilingual prompt analysis. We utilize prompts in six different languages to conduct knowledge probing on the multilingual LLM, Llama3.[9] To ensure a fair comparison of probing results across different language prompts, we filter a sub-dataset where all food instances have both subject and ingredient labels available in all the languages involved. Figure 3 displays probing results on the filtered sub-dataset with prompts in different languages.[10] The English prompt performs significantly better than prompts in other languages across all cultural groups, probably due to more training data.

Code-switching analysis. Figure 4 presents the probing results using code-switching prompts, for each pair out of six languages, applied to the two best-performing decoder-only LLMs, Qwen2 and Llama3. When English is the main language and the dish subject is presented in another language, probing performance significantly declines but still outperforms probing conducted solely in other languages. This suggests that the instruction language is important, not just the language of the dish name. Furthermore, when the main language of the prompt is non-English and the subject language is English, performance does not significantly decline and even shows a slight improvement. This further emphasizes the dominant role of English in multilingual LLMs. The detailed probing results of each cultural group and other multilingual LLMs are in shown Appendix D.3.

7 Error Analysis

A manual inspection of a data sample[11] followed by an automated evaluation on the complete data reveals recurring mistake patterns: (1) **Coarse-grained confusions** (foreign ingredients, e.g., *coffee* in a Chinese dish), where the main issue lies in the model's lack of cultural attribute awareness for

[8]IRR and the average evaluation time for each country group are shown in Appendix C.

[9]The probing results comparison on other multilingual LLMs are shown in Appendix D.2.

[10]Some Chinese and Korean objects are missing Fasttext vectors. Therefore, mWS cannot be calculated for them.

[11]Conducted by one of the authors on 500 predictions by Llama3, Llama2, Bert-base-uncased and T5-base.

Country	Top 5 Most Falsely Predictions (Llama3)
Italy	pasta, bread, meat, egg, pork
Japan	rice, fish, soy, tofu, chicken
France	butter, cream, meat, caramel, bread
Russia	meat, beef, cabbage, pork, beet

(a) Coarse-grained confusions

Model	Top 5 Most Frequent Predictions
BERT	dried meat, ham, beef, pork, fish
T5	vinaigrette, scrambled eggs, buckwheat flour, ricotta, sultana
Llama2	rice, meat, bread, white meat, egg
Llama3	rice, meat, sugar, chicken, bread

(b) Fine-grained confusions

Table 4: Error Analysis: Coarse and fine-grained confusions in ingredient predictions by various cultural groups and models.

both the dish and its ingredients. (2) **Fine-grained confusions** (wrong local ingredients, e.g., *Chinese wine* in *Mapo tofu*), and (3) **Inconsistent confusions**, where the model predicts different incorrect ingredients across prompts for the same dish.

Table 4 illustrates the described error patterns. In the results for Llama3 in English, 47% of the incorrect predictions for Italian dishes are the ingredient *pasta*, 45% of the incorrect predictions for French dishes are *butter*, and 40% of the incorrect predictions for Japanese dishes are *tofu*. These are examples of fine-grained confusions. Additionally, *rice*, *wheat* and *flour* account for 11% of the Llama3's total predictions, regardless of the dish's origin. This is a coarse confusion, where the model often predicts the same foreign ingredient.

These errors reveal a lack of cultural knowledge about the dishes and suggest different types of guessing. While coarse and inconsistent confusions seem to reflect more general guesses, possibly influenced by the frequency of certain ingredients in the training data, fine-grained confusions indicate that the model has some knowledge of the dish's origin and is attempting to guess a local ingredient accordingly (i.e. overfitting to the context). Specifically, for fine-grained errors, even when the cultural attributes of candidate ingredients align with the dish, LLMs still make incorrect predictions. While many studies have shown that LLMs can function as knowledge bases (Nguyen et al., 2024; Pan et al., 2024) and memorize factual information (Fierro et al., 2024a,b), the familiarity with specific entities is crucial in shaping their predictive expectations (Du et al., 2024).

8 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.

9 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. Furthermore, 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.

ACKNOWLEDGMENTS

This research is supported by the project of Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002), Shenzhen Science and Technology Program (Grant No. ZDSYS20230626091302006), Key Project of Shenzhen Higher Education Stability Support Program (Grant No. 2024SC0009).

References

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

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.

AI@Meta. 2024. Llama 3 model card.

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, 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.

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.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. 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, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2024a. How do multilingual models remember? investigating multilingual factual recall mechanisms. *arXiv preprint arXiv:2410.14387*.

Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. 2024b. MuLan: A study of fact mutability in language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 762–771, Mexico City, Mexico. Association for Computational Linguistics.

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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.

Anjishnu Mukherjee, Aylin Caliskan, Ziwei Zhu, and Antonios Anastasopoulos. 2024. Global gallery: The fine art of painting culture portraits through multilingual instruction tuning. 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 6398–6415, Mexico City, Mexico. Association for Computational Linguistics.

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.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1774–1784.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

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, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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.

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.

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

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*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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 details

A.1 Examples

Figure 5 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, $la = \text{LANG}\,(na) = \text{LANG}\,(pa) = \text{LANG}\,(ma)$, which indicates the descriptive language for all attributes remains consistent. For a dish with the same url, there may be several instances with different languages. Table 5 contrasts our dataset with prior cultural knowledge collections.

A.2 Statistics

There are a total of 33,600 dishes in FmLAMA. We count the number of dishes corresponding to each language and country in the dataset, as shown in Figures 6 and 7. Figure 6a shows the top 20 languages ranked by the number of dishes, with English (en) having the highest count at 2804, followed by Spanish (es), French (fr), and Japanese (ja). Figure 6b displays the distribution of dish counts across different intervals for all languages, indicating that most languages have fewer than 200 dishes. Similarly, Figure 7a shows the top 20 countries ranked by the number of dishes, with Italy having the highest count at 2975, followed by France, America, and Japan. Figure 7b shows the distribution of dish counts across various intervals for all countries, indicating that only 8 countries have more than 200 dishes. Figure 8 shows the dish statistics by ingredient count. The majority of dishes contain only one ingredient, totaling 18,546. Dishes with two ingredients are the second most common, with a count of 6,426. The counts decrease as the number of ingredients increases, with dishes containing eleven ingredients being the least common, at 32.

B Probing Templates

B.1 Prompt illustration

Probing templates in six involved languages are shown in Figure 9, in which [X] represents the subject (dish) and [Y] indicates the object (ingredient).

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 5: Examples of FmLAMA.

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	33600

Table 5: Comparison of cultural knowledge datasets. Size is in number of instances.

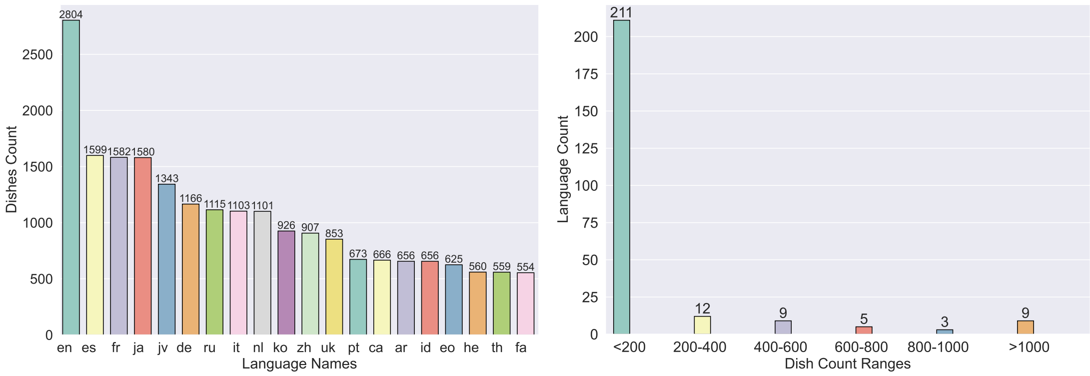
B.2 Language comparison

B.2.1 Geographic differences

In this section, we discuss the representative countries where each selected language is spoken and their geographic distribution: (1) *English*: Mainly spoken in U.S., Canada, Australia, Guyana, etc., spanning North and South America, Europe, Australia, and Africa. (2) *Chinese*: Mainly spoken in China, Singapore, etc., covering huge area of Asia. (3) *Arabic*: Mainly spoken in Saudi Arabia, Egypt, United Arab Emirates, etc., covering the Arabian Peninsula, North Africa, and some sub-Saharan African countries. (4) *Korean*: Mainly spoken in South Korea, North Korea. (5) *Russian*: Mainly spoken in Russia, Belarus, Kazakhstan, etc., covering Eastern Europe, Central Asia, and Northern Asia. (6) *Hebrew*: Mainly spoken in Israel, with Hebrew-speaking communities worldwide. These languages represent cultural diversity, being spoken on different continents by groups with rich and distinct cultural backgrounds, demonstrating the broad geographic and cultural diversity represented by these languages.

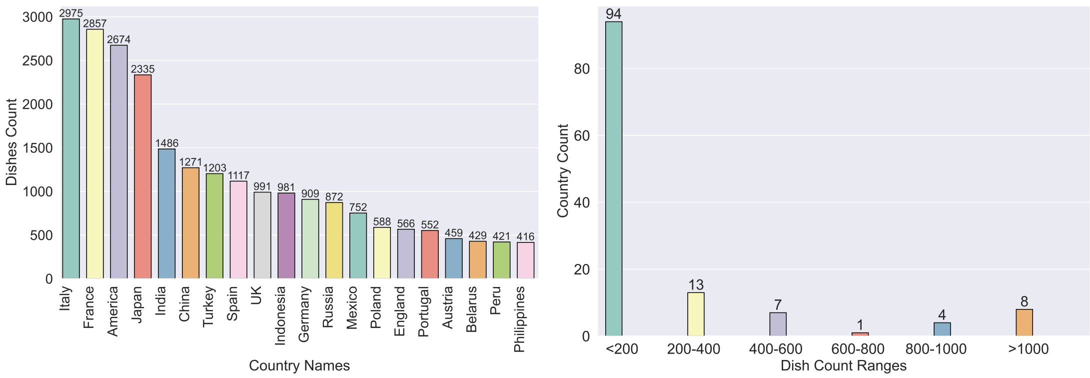
B.2.2 Grammatical differences

In addition to geographical differences in usage, these six languages also display the following grammatical and reading distinctions: (1) *Word Order*: English and Chinese follow an Subject-Verb-Object word order, Arabic typically uses Verb-Subject-Object, Korean follows Subject-Object-Verb, while Russian and Hebrew are relatively flexible but generally adhere to Subject-Verb-Object. (2) *Tense*: English, Arabic, Russian, and Hebrew possess complex tense systems. Chinese lacks a strict tense structure, whereas Korean expresses tense through verb conjugation. (3) *Gender*: Arabic, Russian, and Hebrew have grammatical gender, while English distinguishes gender only in pronouns. Chinese and Korean do not have grammatical gender. (4) *Articles*: English, Arabic, and Hebrew have an article system, whereas Chinese, Korean, and Russian do not use articles. (5) *Reading Direction*: English, modern Chinese, modern Korean, and Russian are read from left to right, while Arabic and Hebrew are read from right to left. Traditional Chinese and traditional Korean are read vertically from top to bottom, with columns arranged from right to left. (6) *Case System*: Ara-



(a) The Top20 languages ranked by the dishes count (b) Languages distributed across various dish count ranges

Figure 6: Dish statistics by languages



(a) The Top20 countries ranked by the dishes count (b) Countries distributed across various dish count ranges

Figure 7: Dish statistics by countries

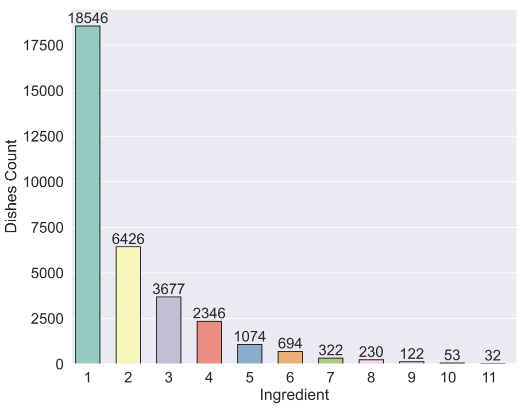


Figure 8: Dish statistics by ingredient count

bic and Russian have complex case systems, while Korean employs particles to indicate grammatical relations. English, Chinese, and Hebrew do not use case inflections.

C Manual Evaluation Setup

C.1 GPT-4o simulator

The evaluation prompts for the GPT-4o simulator are shown in Table 6.

C.2 Human evaluator

Figure 10 shows the human evaluation website. The Inter-Rater Reliability (IRR) and the average evaluation time for the three evaluators from each country group are provided in Table 7.

D Supplementary Experiments and Analyses

D.1 Probing results in each FmLAMA-*la*

In addition to the multilingual LLMs discussed in §5, configure probing for monolingual, encoder-only LLMs in 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`,

La.	Prompt	La.	Prompt
en	[X] is a dish made with [Y]. [X] is a type of food that includes [Y]. The dish [X] is made with [Y]. The dish [X] uses [Y] as an ingredient. The dish [X] is made from [Y].	ar	[1] هو طبق مصنوع مع [2]. [1] هي نوع من الأطعمة التي تحتوي على [2]. طبق [1] مصنوع مع [2]. الطبق [1] يستخدم [2] كمكون. الطبق [1] مصنوع من [2].
ko	[X]는 [Y]를 곁들여 만든 요리예요. [X]는 [Y]가 포함된 음식이에요. 요리 [X]는 [Y]를 곁들여 만들어져요. 요리 [X]는 [Y]를 재료로 사용해요. 요리 [X]는 [Y]로 만들어져요.	zh	[X]是一道菜肴，主要使用的原料是[Y]。 [X]是一种食物，其中包含[Y]。 [X]这道菜是需要用[Y]制作。 [X]这道菜的配料包括[Y]。 [X]这道菜的主要成分是[Y]。
ru	[X] это блюдо которое делают с [Y]. [X] это еда содержащая [Y]. Блюдо [X] готовится с [Y]. В блюде [X] в качестве ингредиента используется [Y]. Блюдо [X] делают из [Y].	he	[X] הוא מאכל שמכינים עם [Y]. [X] הוא מאכל שמכיל את [Y]. את המאכל [X] מכינים עם [Y]. במאכל [X] משתמשים במרכיב [Y]. מאכל [X] עשוי מ [Y].

Figure 9: Probing templates in six involved languages, with [X] representing the subject and [Y] indicating the object that can be substituted.

`klue/bert-base` (Park et al., 2021) and `bert-base-chinese`.

The probing results with prompts in the other five languages (Arabic, Chinese, Hebrew, Korean, and Russian) on the corresponding filtered sub-datasets (FMLAMA-*ar*, FMLAMA-*zh*, FMLAMA-*he*, FMLAMA-*ko*, and FMLAMA-*ru*) are depicted in Table 9, 10, 11, 12, and 13, 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.

D.2 Language analysis on other multilingual LLMs

Figures 11, 12, 13, and 14 present a comparison of probing results for the multilingual LLMs—mBERT, mT5, Qwen2, and Llama2—using prompts in various languages on the corresponding filtered sub-datasets. Decoder-only LLMs, such as Qwen2, Llama2, and Llama3, while representing the cutting edge of current LLM technology, exhibit greater performance variation under multilingual prompt settings.

D.3 Code-switching probing results in each cultural group

Figure 15- 28 presents a detailed comparison of the probing results for each cultural group and other multilingual LLMs.

E Cultural bias *vs* Dish popularity

Table 1 shows that some LLMs perform better in probing tasks for low-resource language groups. Specifically, the decoder-only LLM performs well in the Iran group (both mAP and mWS metric), while Greece shows decent results with both T5 and decoder-only LLMs (only mWS metric).

To understand the strong performance of the Iran group across different metrics, we analyzed the results for the 21 dishes. Table 14 shows 17 involved dishes where the predicted ingredients ranked in the top 10 (Llama3). Since the top-ranked predicted ingredients in the table are not all commonly used, we can rule out the possibility that the strong performance of the Iran group is due to the LLM's common ingredient prediction errors. We hypothesize two possible reasons for this performance: (1) LLM Cultural Ability: Decoder-only LLMs may be more familiar with the Iran cultural group. (2) Data Bias: The dishes in the Iran group are likely well-known and representative of Iranian cuisine, so the LLM might show stronger memory performance for familiar entities (Du et al., 2024).

Instructions

You are a professional evaluator. I will provide you with a dish, its reference ingredient label sourced from Wikidata, and a list of predicted ingredients. Your task is to evaluate whether each predicted ingredient is appropriate for the respective dish.
 	
### Evaluation Criteria: 	
For each predicted ingredient, mark it as **Correct**, **Maybe**, or **Incorrect** using the following guidelines:
 	
• **Correct**: The ingredient meets one of these:
– **Direct Match**: It is explicitly listed in the reference ingredient label from Wikidata.
– **Substitutability**: It can replace a specific ingredient in the reference label during cooking.
– **Missing Traditional Ingredient**: It is traditionally or commonly used in the dish, but not listed in the reference label.
• **Maybe**: The ingredient could be used in some variations of the dish, but its use is uncommon or ambiguous.
• **Incorrect**: The ingredient is rarely or never used in the dish.
 	
### Output Format: 	
Please directly return the evaluation as a Python dictionary with JSON format where the key is the predicted ingredient, and the value is a tuple consisting of:
1. The evaluation ("Correct", "Maybe", "Incorrect")
2. The reason ("Direct Match", "Substitutability", "Missing Traditional Ingredient", or a brief explanation)
 	
### Example: 	
`Input:`
`Dish: Spaghetti Bolognese`
`Reference ingredient label: spaghetti, ground beef, onion, garlic, Carrot`
`Predicted ingredients: spaghetti, basil, chicken, tomato, beef`
 	
`Output`
`{`
`    'ground beef': ('Correct', 'Substitutability'),`
`    'meat': ('Maybe', 'Ambiguous'),`
`    'beef': ('Correct', 'Substitutability'),`
`    'pasta': ('Incorrect', 'Pasta is not part of the sauce itself'),`
`    'turkey meat': ('Maybe', 'It can be used in some variations'),`
`}`

Table 6: GPT-4o simulator: Evaluation Instructions

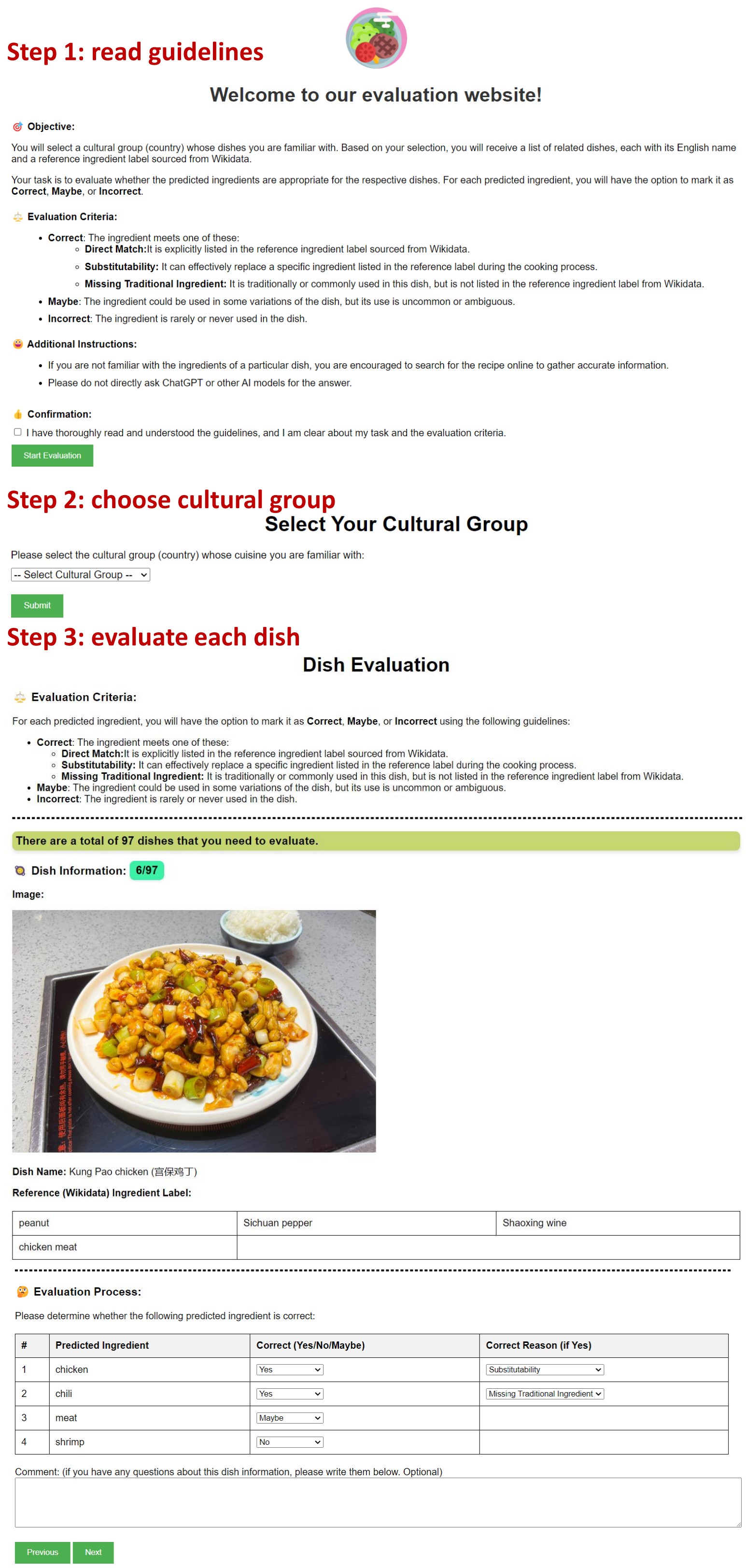


Figure 10: Human evaluation website

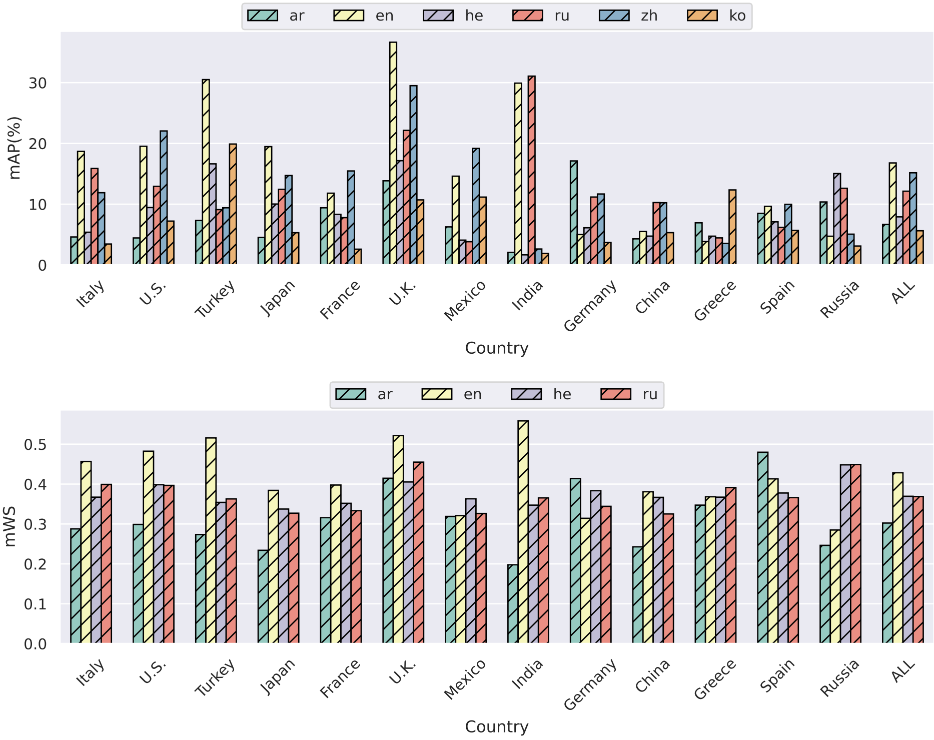


Figure 11: Comparison of probing results on mBERT with prompts in different languages,

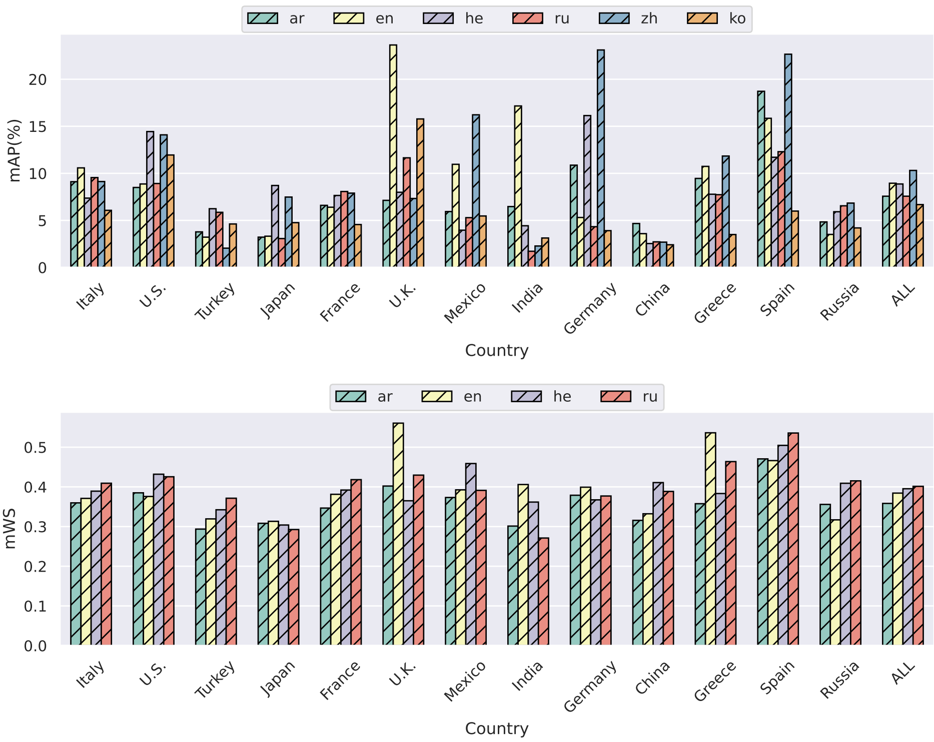


Figure 12: Comparison of probing results on mT5 with prompts in different languages,

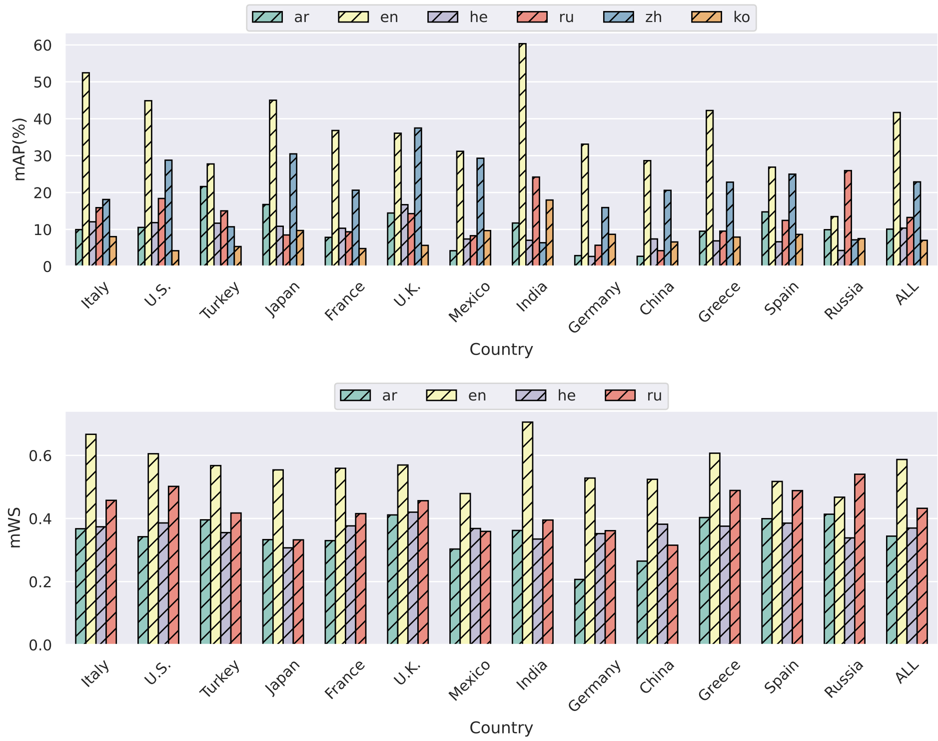


Figure 13: Comparison of probing results on Qwen2 with prompts in different languages,

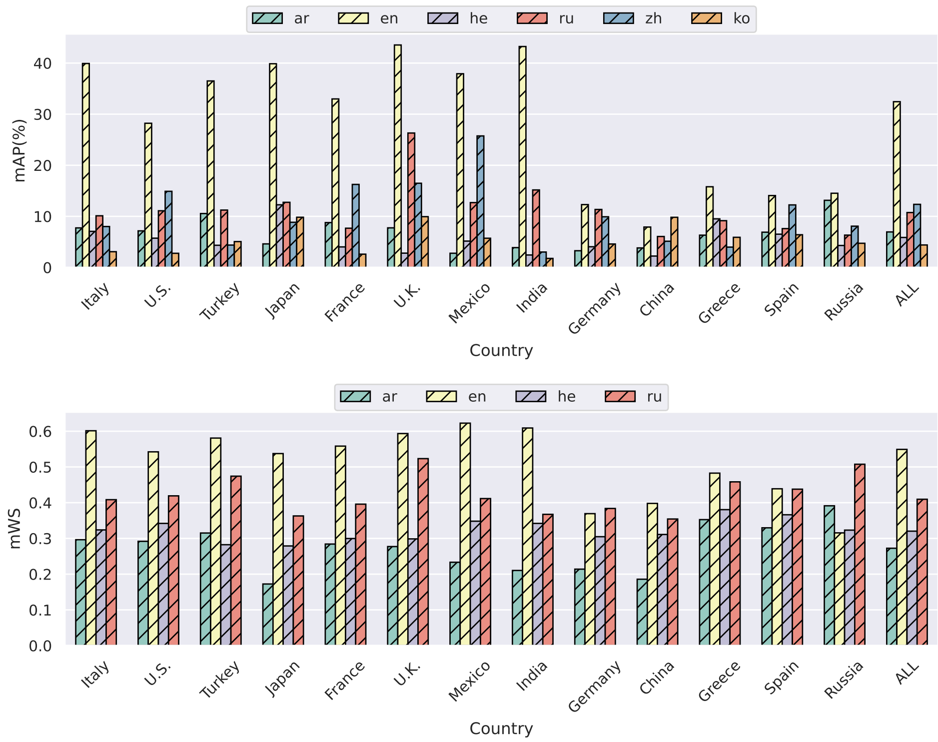


Figure 14: Comparison of probing results on Llama2 with prompts in different languages,

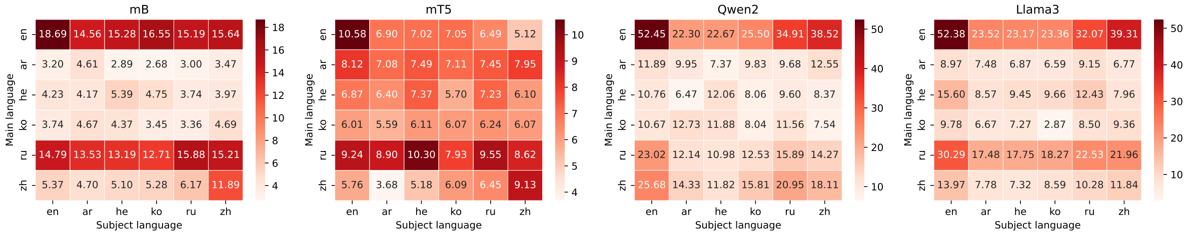


Figure 15: **Italy**: Probing results of code-switching prompts.

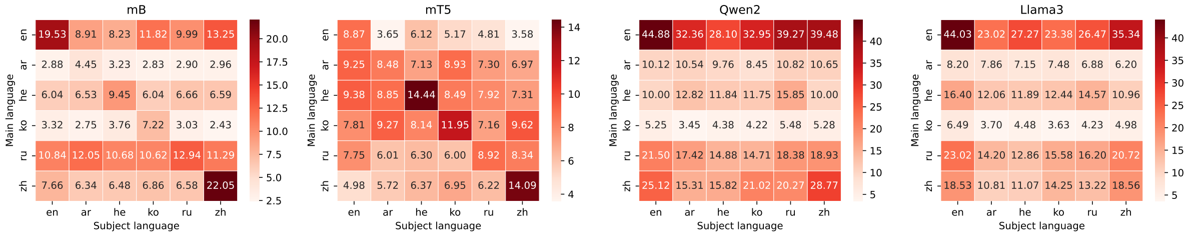


Figure 16: **U.S.**: Probing results of code-switching prompts.

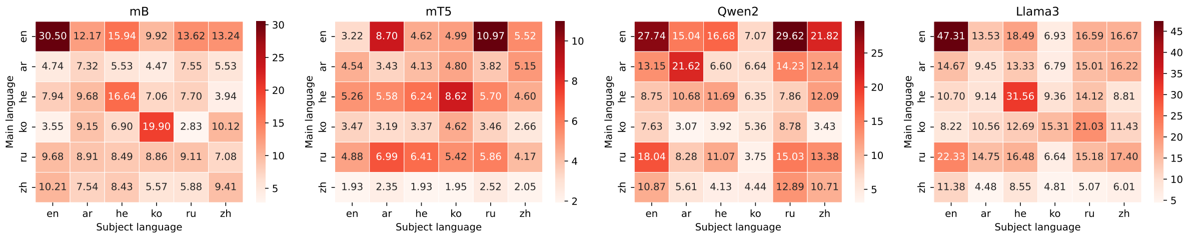


Figure 17: **Turkey**: Probing results of code-switching prompts.

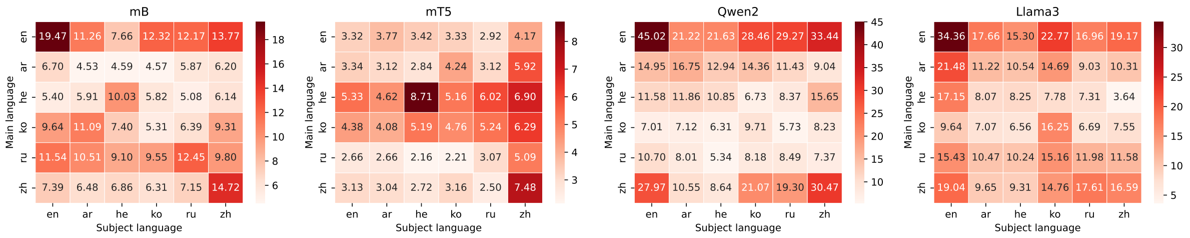


Figure 18: **Japan**: Probing results of code-switching prompts.

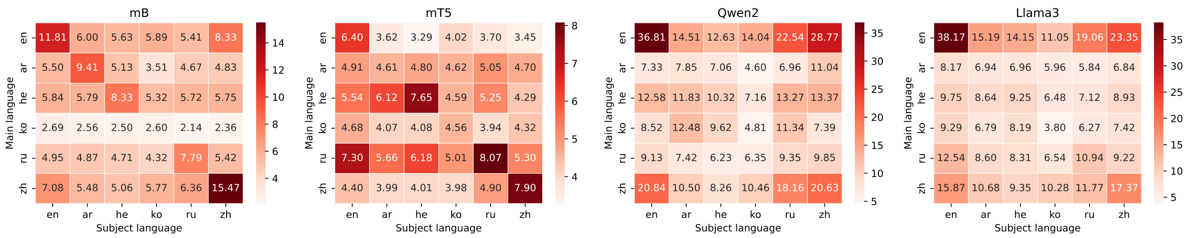


Figure 19: **France**: Probing results of code-switching prompts.

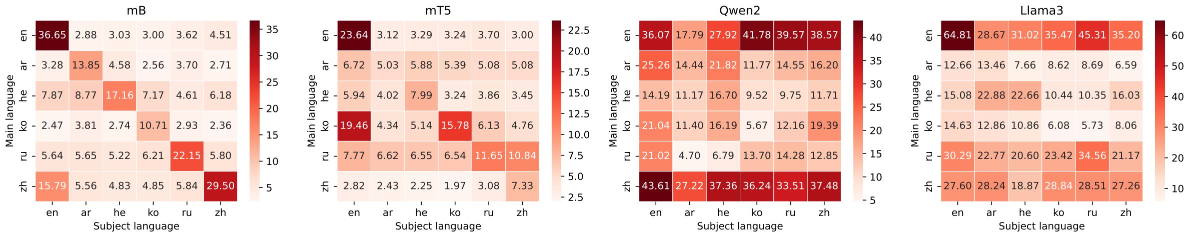


Figure 20: **U.K.**: Probing results of code-switching prompts.

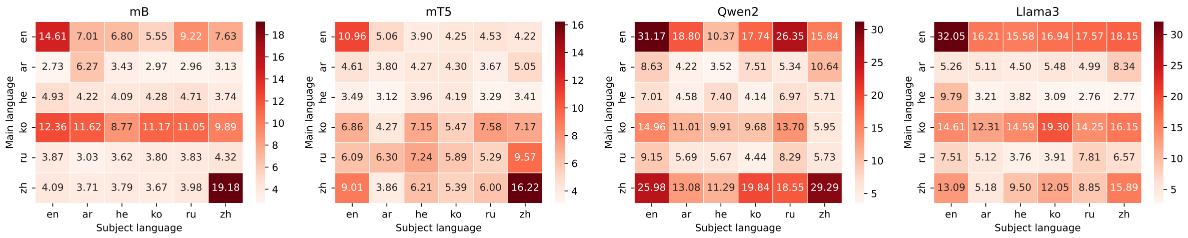


Figure 21: **Mexico**: Probing results of code-switching prompts.

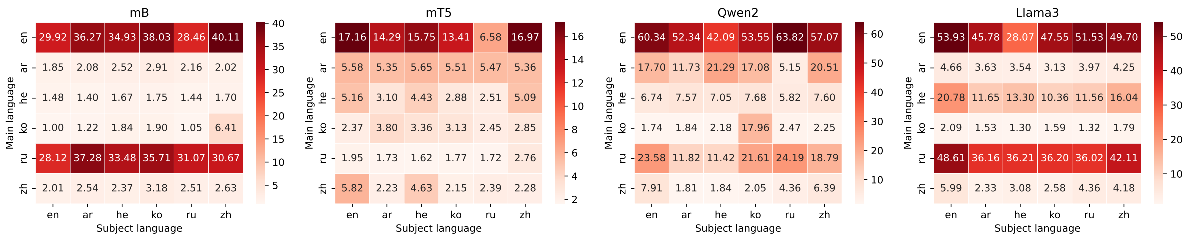


Figure 22: **India**: Probing results of code-switching prompts.

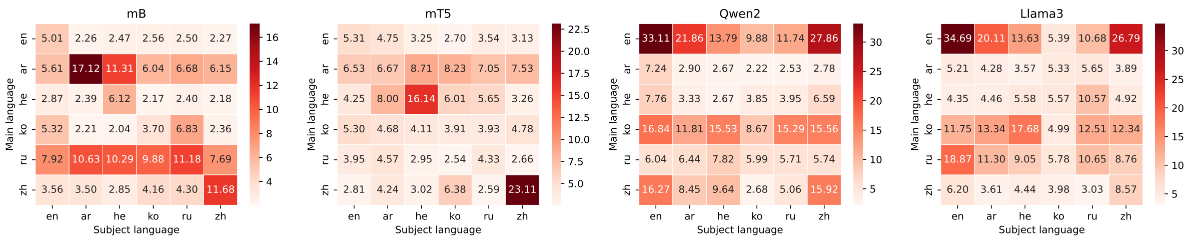


Figure 23: **Germany**: Probing results of code-switching prompts.

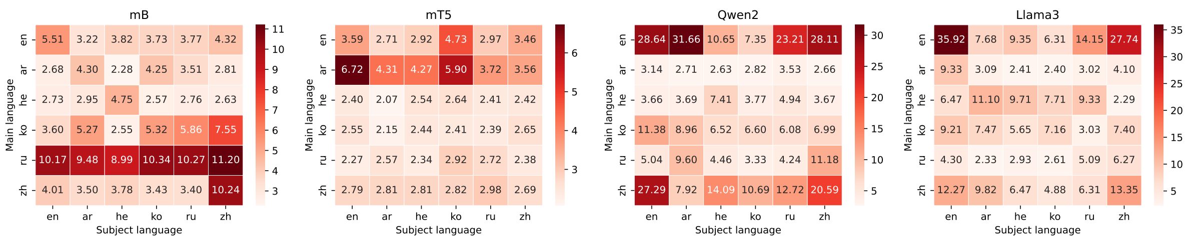


Figure 24: **China**: Probing results of code-switching prompts.

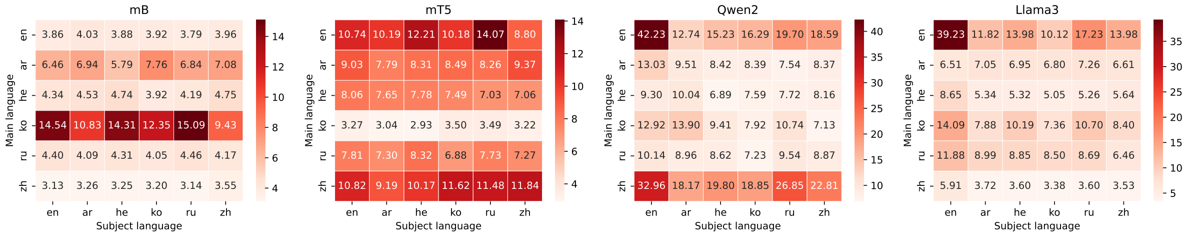


Figure 25: **Greece**: Probing results of code-switching prompts.

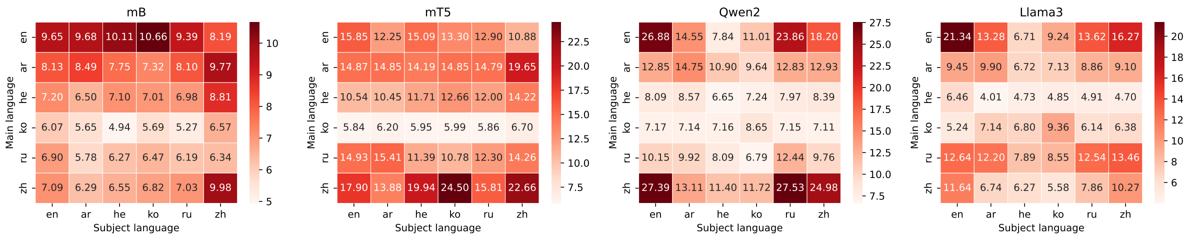


Figure 26: **Spain**: Probing results of code-switching prompts.

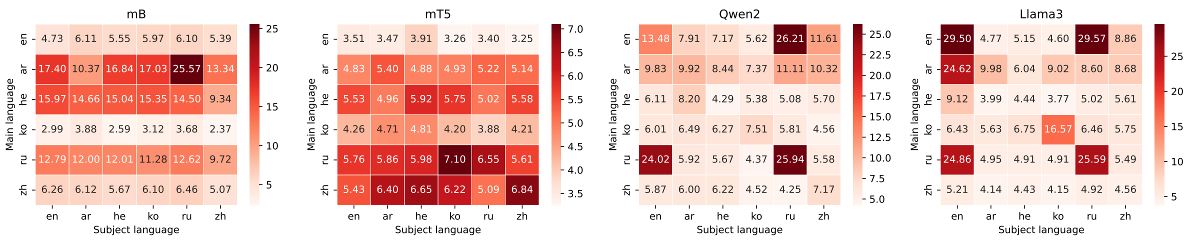


Figure 27: **Russia**: Probing results of code-switching prompts.

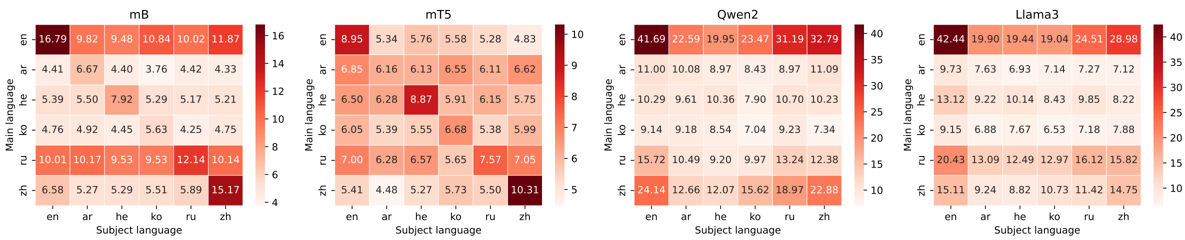


Figure 28: **ALL**: Probing results of code-switching prompts.

Group	Dishes	IRR	Time(m)
Italy	215	0.60	74.93
U.S.	285	0.48	104.25
Turkey	98	0.78	32.05
Japan	186	0.64	49.25
France	175	0.63	71.83
U.K.	83	0.46	30.33
Mexico	57	0.44	22.25
India	132	0.58	40.25
Germany	57	0.69	20.03
China	97	0.59	48.50
Iran	21	0.61	13,33
Greece	21	0.35	11.30
Spain	95	0.57	41.50
Russia	27	0.51	26.50
Mean	111	0.57	-

Table 7: Human evaluator: IRR and average evaluation time for the three evaluators from each country group.

Origin	Encoder-only LLMs			Encoder-Decoder LLMs		Decoder-only LLMs			Avg.
	Bb	Bl	mB	mT5	T5	Qwen2	Llama2	Llama3	
Italy	0.7393±0.01	0.7425±0.01	0.7438±0.00	0.6865±0.02	0.6579±0.01	0.7749±0.01	0.7658±0.01	0.7869±0.01	0.7372
U.S.	**0.7748±0.01**	**0.7763±0.01**	**0.7815±0.01**	0.7071±0.02	**0.7178±0.03**	0.7968±0.01	0.7839±0.02	0.8107±0.01	0.7686
Turkey	0.7357±0.01	0.7461±0.01	0.7545±0.01	0.7015±0.02	0.6703±0.01	0.7681±0.01	0.7571±0.01	0.7830±0.01	0.7395
Japan	0.7113±0.00	0.7144±0.00	0.7218±0.01	0.6788±0.02	0.6442±0.01	0.7446±0.00	0.7446±0.01	0.7516±0.00	0.7139
France	0.7398±0.01	0.7520±0.02	0.7513±0.01	0.6755±0.02	0.6642±0.01	0.7941±0.01	0.7796±0.01	0.7846±0.00	0.7426
U.K.	0.7460±0.01	0.7500±0.01	0.7550±0.00	0.6697±0.02	0.6745±0.02	0.7678±0.01	0.7701±0.01	0.7790±0.01	0.7390
Mexico	0.7273±0.01	0.7313±0.01	0.7291±0.01	0.7171±0.01	0.6734±0.01	0.7605±0.01	0.7752±0.02	0.7852±0.02	0.7374
India	0.7477±0.01	0.7523±0.00	0.7568±0.00	0.6998±0.01	0.6619±0.01	0.7867±0.00	**0.7944±0.01**	0.7935±0.01	0.7491
Germany	0.7348±0.01	0.7305±0.03	0.7493±0.00	0.6755±0.01	0.6535±0.01	0.7835±0.01	0.7769±0.02	0.7849±0.01	0.7361
China	0.7273±0.01	0.7319±0.01	0.7317±0.01	0.7024±0.01	0.7054±0.01	0.7698±0.01	0.7593±0.01	0.7793±0.01	0.7384
Iran	0.7064±0.01	0.7013±0.01	0.7310±0.01	0.7042±0.02	0.6686±0.01	0.7905±0.03	0.7914±0.04	0.8026±0.01	0.7370
Greece	0.7488±0.01	0.7618±0.02	0.7566±0.02	0.7190±0.03	0.6369±0.00	**0.8137±0.04**	0.7739±0.01	**0.8247±0.01**	0.7544
Spain	0.7402±0.00	0.7465±0.01	0.7472±0.00	0.6883±0.01	0.6391±0.01	0.7748±0.01	0.7663±0.01	0.7765±0.01	0.7349
Russia	0.7467±0.01	0.7571±0.01	0.7573±0.00	**0.7216±0.02**	0.6710±0.01	0.7716±0.01	0.7680±0.01	0.7721±0.00	0.7457
ALL	0.7419±0.00	0.7466±0.01	0.7507±0.00	0.6924±0.01	0.6723±0.01	0.7778±0.01	0.7711±0.01	0.7859±0.00	0.7423
Corr.	0.35	0.59	0.38	0.83	0.57	0.86	0.71	0.96	-

Table 8: Probing performance comparison with **English** prompts and **FMLAMA-*ar*** sub-dataset. The results are evaluated using mWS with **BERT base**. *Corr.* denotes the correlation of mWS evaluations using BERT Base and FastText for each LLM, which reveal a strong correlation on decoder-only LLMs,

Origin	Count	Bb-ar	mB	mT5	Qwen2	Llama2	Llama3	Avg.
Italy	76 (19.2%)	2.56±0.69	3.07±1.02	5.12±1.20	5.13±3.42	3.30±1.13	4.05±0.81	4.03
U.S.	54 (13.7%)	5.29±1.18	7.22±3.91	5.68±1.53	9.41±4.12	3.93±1.00	5.95±2.38	5.09
Turkey	52 (13.2%)	4.07±0.88	6.66±3.09	3.30±0.53	9.20±5.71	3.33±1.04	**9.54±3.48**	5.47
Japan	44 (11.11%)	2.69±1.50	2.56±0.50	2.13±0.78	9.91±3.73	3.75±1.12	5.43±1.62	5.20
France	37 (9.4%)	7.51±2.43	6.50±4.55	4.51±1.21	7.16±3.23	5.71±0.82	4.63±1.77	4.70
U.K.	26 (6.6%)	5.80±2.15	9.17±6.85	6.72±1.67	11.77±6.35	5.72±4.86	8.62±6.52	5.75
Mexico	20 (4.6%)	3.52±1.57	7.37±1.01	3.43±0.61	3.60±0.55	2.38±0.48	3.72±1.68	2.82
India	18 (3.2%)	4.39±1.09	8.86±5.19	3.08±1.17	10.82±9.44	3.04±1.56	3.03±1.38	7.29
Germany	13 (3.3%)	7.39±2.52	**10.65±4.95**	5.87±2.04	9.53±4.72	2.75±2.12	7.67±3.29	4.69
China	13 (3.3%)	2.07±0.53	4.86±4.36	2.43±0.93	6.79±2.17	2.75±0.70	6.54±2.51	4.58
Iran	11 (2.8%)	**9.30±4.85**	8.19±5.41	7.38±2.56	11.23±7.82	**6.31±2.85**	6.67±5.02	5.90
Greece	11 (2.8%)	3.12±0.98	5.21±3.26	4.25±1.23	9.25±4.78	3.13±1.09	4.54±3.95	4.63
Spain	10 (2.5%)	5.84±1.24	4.13±2.32	**8.01±1.35**	**12.74±3.84**	3.41±1.18	6.90±4.08	7.34
Russia	10 (2.5%)	3.62±0.80	3.13±0.79	2.48±0.26	4.20±0.98	4.88±2.78	4.76±0.94	4.34
ALL	395 (100.0%)	4.41±0.35	5.80±2.56	4.48±0.60	8.24±3.65	3.85±1.00	5.85±1.47	4.96

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

Origin	Bb-ar	mB	mT5	Qwen2	Llama2	Llama3	Avg.
Italy	0.3280±0.01	0.2259±0.02	0.3547±0.05	0.3213±0.06	0.2252±0.05	0.2802±0.05	0.2892
U.S.	0.3355±0.02	0.2746±0.05	0.3581±0.03	0.3261±0.05	0.2419±0.03	0.3194±0.02	0.3093
Turkey	0.3141±0.01	0.2434±0.04	0.3154±0.03	0.3111±0.06	0.2210±0.03	0.3164±0.06	0.2869
Japan	0.2876±0.02	0.2411±0.04	0.3150±0.05	0.3030±0.05	0.1834±0.04	0.2553±0.04	0.2642
France	0.3599±0.02	0.2984±0.04	0.3462±0.03	0.3129±0.04	0.2608±0.02	0.2844±0.01	0.3104
U.K.	0.3450±0.02	0.3118±0.08	0.3651±0.04	0.3450±0.06	0.2406±0.07	0.3435±0.05	0.3252
Mexico	0.3098±0.02	0.2891±0.04	0.3335±0.07	0.2687±0.03	0.2163±0.02	0.2665±0.02	0.2807
India	0.2850±0.03	0.2474±0.07	0.3051±0.03	0.3355±0.09	0.2217±0.05	0.2455±0.03	0.2734
Germany	0.3300±0.02	0.2859±0.07	0.3530±0.04	0.2772±0.06	0.2046±0.04	0.3172±0.05	0.2946
China	0.2938±0.02	0.2836±0.06	0.3255±0.07	0.2889±0.03	0.1907±0.04	0.2774±0.04	0.2767
Iran	0.3453±0.04	0.2644±0.06	0.3634±0.03	0.3488±0.04	0.2405±0.06	0.3257±0.02	0.3147
Greece	0.3241±0.02	0.2559±0.04	0.3502±0.05	0.3358±0.06	0.2325±0.03	0.3052±0.04	0.3006
Spain	**0.3659±0.02**	**0.3213±0.03**	**0.4071±0.06**	**0.3626±0.06**	0.2869±0.05	**0.3576±0.05**	0.3502
Russia	0.3323±0.02	0.2772±0.02	0.3347±0.02	0.3451±0.02	**0.2964±0.03**	0.3145±0.00	0.3167
ALL	0.3243±0.01	0.2625±0.03	0.3421±0.04	0.3176±0.05	0.2282±0.04	0.2957±0.03	0.2951

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

Table 9: Probing performance comparison with **Arabic** prompts and **FMLAMA-*ar*** sub-dataset.

Origin	Count	Bb-zh	mB	mT5	Qwen2	Llama2	Llama3	Avg.
Italy	49 (8.7%)	17.18±1.67	13.07±4.92	10.55±0.84	24.54±2.87	9.32±4.48	13.58±4.08	14.71
U.S.	101 (17.9%)	19.32±1.47	13.79±3.56	12.37±1.09	26.50±2.94	12.91±3.61	13.17±6.45	16.34
Turkey	12 (2.1%)	8.92±3.86	6.30±2.12	5.22±0.18	18.16±3.49	8.18±3.58	11.00±6.39	9.63
Japan	114 (20.2%)	15.79±2.32	11.18±4.41	7.85±0.60	17.67±2.73	11.15±2.66	11.78±4.54	12.57
France	70 (12.4%)	19.45±4.13	16.04±5.31	9.14±0.98	20.96±1.45	13.36±3.62	14.71±5.89	15.61
U.K.	38 (6.7%)	**27.88±2.65**	**19.55±7.36**	10.64±0.68	28.43±4.16	**18.02±8.33**	**21.46±6.75**	21.00
Mexico	19 (3.4%)	21.12±0.86	13.40±3.51	12.05±0.79	16.25±4.70	10.60±1.32	8.21±3.83	13.61
India	24 (4.3%)	5.47±1.65	5.02±2.68	1.13±0.17	12.37±3.79	4.57±2.20	6.58±5.13	5.86
Germany	16 (2.8%)	10.31±2.53	12.46±3.54	6.81±1.75	10.41±2.12	4.08±2.38	7.62±4.63	8.62
China	87 (15.4%)	22.55±4.86	18.32±6.75	7.88±0.91	**33.48±6.15**	13.83±3.72	19.96±9.42	19.34
Iran	7 (1.2%)	7.72±1.44	8.19±4.58	3.54±0.20	8.39±2.08	3.40±0.38	8.82±3.50	6.68
Greece	7 (1.2%)	4.11±0.68	2.88±1.08	5.43±0.43	14.81±3.05	1.40±0.37	3.46±1.23	5.35
Spain	12 (2.1%)	13.00±1.83	10.21±1.58	**12.72±1.17**	21.02±0.93	5.53±3.77	12.41±6.05	12.48
Russia	8 (1.4%)	5.58±0.93	3.98±0.76	2.29±0.38	5.89±1.09	4.06±2.84	5.14±1.07	4.49
ALL	564 (100%)	17.84±2.37	13.56±4.36	8.96±0.44	22.71±2.37	11.46±3.32	13.76±5.56	14.71

Table 10: Probing performance comparison with **Chinese** prompts and **FMLAMA-*zh*** sub-dataset.

Origin	Count	mB	mT5	Qwen2	Llama2	Llama3	Average
Italy	59 (18.6%)	4.18±2.54	4.30±1.49	7.78±2.37	3.97±1.47	6.29±3.14	5.30
U.S.	67 (21.1%)	8.62±7.22	8.13±2.02	9.09±4.52	3.62±0.81	10.17±4.44	7.93
Turkey	12 (3.9%)	8.99±8.65	6.33±4.01	9.21±5.92	2.93±1.18	10.68±6.33	7.63
Japan	19 (6.0%)	6.69±5.23	5.99±3.64	6.60±4.19	6.92±3.85	4.39±3.12	6.12
France	65 (20.5%)	5.43±3.01	4.54±0.92	6.40±2.74	2.72±0.54	5.50±1.47	4.92
U.K.	23 (7.3%)	**9.43±6.79**	6.95±2.26	9.79±5.06	3.97±2.49	10.08±5.66	8.04
Mexico	11 (3.5%)	2.22±0.75	2.26±0.74	4.33±2.44	2.87±1.99	2.14±2.01	2.76
India	18 (5.7%)	8.12±3.72	3.31±1.07	2.80±2.77	3.26±1.95	4.73±2.46	4.44
Germany	11 (3.5%)	7.57±5.98	**14.84±6.23**	8.59±3.46	5.92±1.83	9.38±5.42	9.26
China	9 (2.8%)	3.61±3.34	2.12±0.78	3.75±3.62	1.64±1.29	4.05±1.62	3.03
Iran	6 (1.9%)	8.58±5.37	5.05±2.90	**19.40±9.90**	2.95±0.57	**20.63±10.96**	11.32
Greece	6 (1.9%)	2.76±0.60	4.31±0.49	12.96±9.55	**7.67±2.85**	3.34±2.19	6.21
Spain	7 (2.2%)	4.58±1.81	7.02±1.87	3.86±1.66	4.04±1.62	2.82±1.26	4.46
Russia	4 (1.3%)	4.31±3.65	3.35±0.07	2.10±0.28	2.56±0.46	2.56±0.44	2.98
ALL	317 (100%)	6.41±4.30	5.77±1.59	7.57±2.66	3.74±0.85	7.17±2.63	6.13

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

Origin	mB	mT5	Qwen2	Llama2	Llama3	Average
Italy	0.3102±0.03	0.3320±0.03	0.3197±0.03	0.2973±0.04	0.3265±0.05	0.3171
U.S.	0.3435±0.06	0.3626±0.03	0.3524±0.03	0.3042±0.03	0.3517±0.04	0.3429
Turkey	0.2976±0.06	0.3225±0.05	0.3110±0.04	0.2621±0.05	0.3296±0.10	0.3046
Japan	0.3134±0.05	0.3167±0.04	0.3114±0.05	0.2839±0.03	0.2899±0.02	0.3031
France	0.3082±0.03	0.3509±0.04	0.3334±0.02	0.2838±0.04	0.3134±0.03	0.3179
U.K.	0.3400±0.06	0.3418±0.04	0.3414±0.04	0.2930±0.02	0.3167±0.06	0.3266
Mexico	0.2959±0.05	{0.3829±0.06}	0.3420±0.06	0.2847±0.06	0.2865±0.04	0.3184
India	{0.3614±0.05}	0.3437±0.06	0.3354±0.01	0.3133±0.04	0.3192±0.05	0.3346
Germany	{0.3595±0.04}	0.3763±0.02	{0.3756±0.03}	{0.3339±0.05}	{0.3586±0.06}	0.3608
China	0.2863±0.03	0.3131±0.06	0.2714±0.06	0.2683±0.03	0.2432±0.04	0.2765
Iran	0.3435±0.05	0.3521±0.07	{0.4569±0.07}	0.2902±0.02	{0.4416±0.09}	0.3769
Greece	0.3014±0.02	0.3423±0.02	0.3722±0.08	0.3332±0.03	0.2951±0.04	0.3288
Spain	0.3493±0.03	{0.4229±0.07}	0.3547±0.02	{0.3635±0.06}	0.3260±0.03	0.3633
Russia	0.3276±0.03	0.3748±0.06	0.3319±0.04	0.3308±0.01	0.3179±0.04	0.3366
ALL	0.3237±0.04	0.3484±0.03	0.3369±0.02	0.2969±0.03	0.3248±0.04	0.3261

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

Table 11: Probing performance comparison with **Hebrew** prompts and **FMLAMA-*he*** sub-dataset.

Origin	Count	Bb-ky	Bb-kl	mB-u	mT5	Qwen2	Llama2	Llama3	Avg.
Italy	88 (16.7%)	6.90±1.57	7.75±2.34	3.85±2.54	6.02±0.64	3.50±1.10	1.19±0.22	2.23±0.75	4.49
U.S.	64 (12.2%)	11.37±2.66	10.87±3.86	**6.32±5.20**	**10.15±0.34**	2.46±0.69	1.68±0.12	2.95±0.93	6.54
Turkey	19 (3.6%)	2.75±2.47	4.13±2.62	4.07±3.17	1.82±0.44	3.86±1.00	2.16±0.12	**7.47±1.45**	3.75
Japan	101 (19.2%)	7.34±2.12	7.74±1.86	1.93±1.24	5.25±0.99	6.56±5.34	2.19±0.22	2.62±1.15	4.80
France	67 (12.7%)	2.33±0.43	3.09±0.44	3.15±0.17	2.79±0.30	3.10±0.68	1.40±0.09	2.41±1.27	2.61
U.K.	27 (5.1%)	**14.07±2.83**	**15.89±4.43**	6.00±4.42	7.55±1.34	3.21±1.02	1.79±0.31	6.89±5.70	7.91
Mexico	13 (2.5%)	5.18±3.69	8.37±2.85	4.89±3.29	2.31±0.40	7.12±3.13	2.75±1.99	4.18±1.61	4.97
India	32 (6.1%)	7.98±2.29	5.74±1.32	3.28±2.09	2.96±0.61	**7.69±6.42**	2.01±0.32	2.36±0.67	4.57
Germany	15 (2.9%)	2.50±0.94	5.08±1.25	2.42±1.68	3.20±1.18	2.63±0.95	1.33±0.48	1.91±0.21	2.72
China	54 (10.3%)	9.02±2.76	10.12±3.04	2.36±2.22	4.05±1.50	4.43±1.74	2.16±0.48	3.06±1.48	5.03
Iran	8 (1.5%)	7.99±5.23	3.76±1.41	1.62±0.67	1.52±0.24	1.50±0.14	2.87±2.49	2.12±1.83	3.05
Greece	9 (1.7%)	2.32±0.73	6.76±3.73	2.74±0.76	5.76±4.22	4.59±1.31	**3.67±2.94**	3.35±1.87	4.17
Spain	21 (4.0%)	4.05±2.25	5.90±2.70	1.79±1.24	5.82±1.33	3.99±1.38	2.50±0.97	3.85±0.99	3.99
Russia	8 (1.5%)	3.70±0.99	4.22±1.83	2.68±1.12	2.12±0.05	3.01±1.41	1.72±0.20	3.07±1.46	2.93
ALL	526 (100%)	7.05±1.47	7.68±2.00	3.45±2.10	5.19±0.42	4.32±1.76	1.86±0.18	3.07±0.95	4.66

Table 12: Probing performance comparison on **mAP** (%) with **Korean** prompts and **FMLAMA-*ko*** sub-dataset.

Origin	Count	Bb-ru	mB	mT5	Qwen2	Llama2	Llama3	Avg.
Italy	101 (15.9%)	4.29±2.04	6.57±3.51	4.52±1.27	7.99±2.61	4.30±2.32	10.61±2.60	6.38
U.S.	79 (12.5%)	9.76±4.95	9.99±5.25	5.51±0.71	12.01±3.53	9.04±4.33	11.97±3.75	9.71
Turkey	28 (4.4%)	5.75±2.12	9.46±1.68	3.58±1.61	11.08±8.01	10.49±2.90	16.37±2.97	9.46
Japan	68 (10.7%)	7.89±4.76	7.13±2.34	1.76±0.56	6.10±2.74	8.50±5.79	14.11±0.74	7.58
France	93 (14.7%)	7.08±3.04	6.83±2.85	5.03±1.86	7.03±2.48	8.85±4.96	10.55±3.43	7.56
U.K.	33 (5.2%)	8.42±5.82	10.03±5.01	5.54±3.74	11.89±6.16	**13.17±8.89**	16.82±8.31	10.98
Mexico	22 (3.5%)	4.08±0.89	2.01±0.82	2.08±0.95	4.70±1.40	5.64±3.25	4.21±1.34	3.79
India	21 (3.3%)	**11.46±9.78**	**19.02±2.71**	4.27±3.11	16.80±6.73	12.95±5.70	22.93±2.60	14.57
Germany	45 (7.1%)	7.53±5.26	7.95±3.59	4.39±2.03	9.41±2.54	11.74±4.92	11.57±4.32	8.77
China	30 (4.7%)	9.67±4.64	11.86±3.63	3.28±2.19	10.44±4.98	12.63±6.07	12.82±4.43	10.12
Iran	13 (2.1%)	10.86±6.84	11.06±5.99	**15.51±4.53**	**22.88±6.61**	10.77±8.30	**23.88±6.87**	15.83
Greece	18 (2.8%)	3.45±2.42	1.74±0.43	3.38±1.24	8.56±2.57	4.91±2.85	4.77±1.28	4.47
Spain	38 (6.0%)	3.35±1.51	6.51±2.61	3.37±1.37	5.27±2.52	3.88±2.07	5.61±1.84	4.67
Russia	45 (7.1%)	4.79±2.05	3.75±0.77	2.80±0.92	5.21±2.94	5.31±2.18	5.47±2.27	4.55
ALL	634 (100%)	6.85±3.14	7.76±1.83	4.28±1.11	8.84±2.67	8.19±4.04	11.52±2.82	7.91

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

Origin	Bb-ru	mB	mT5	Qwen2	Llama2	Llama3	Avg.
Italy	0.3319±0.04	0.3217±0.06	0.3633±0.03	0.4073±0.04	0.3774±0.02	0.3779±0.05	0.3633
U.S.	**0.3796±0.04**	0.3556±0.05	0.3692±0.02	0.4393±0.03	0.3938±0.05	0.4068±0.05	0.3907
Turkey	0.3199±0.06	0.3317±0.04	0.3300±0.04	0.3966±0.06	0.4012±0.02	0.4178±0.04	0.3662
Japan	0.2955±0.03	0.2796±0.04	0.2740±0.02	0.3085±0.04	0.3110±0.05	0.3319±0.01	0.3001
France	0.3438±0.05	0.3132±0.03	0.3652±0.04	0.3875±0.02	0.3954±0.04	0.3752±0.04	0.3634
U.K.	0.3647±0.08	0.3442±0.05	0.3771±0.04	0.4327±0.04	0.4149±0.08	0.4179±0.09	0.3919
Mexico	0.3150±0.01	0.3081±0.04	0.3604±0.04	0.3688±0.03	0.3909±0.02	0.3225±0.04	0.3443
India	0.3475±0.10	**0.3808±0.05**	0.3062±0.03	0.4375±0.06	0.3665±0.07	0.4236±0.04	0.3770
Germany	0.3500±0.08	0.3330±0.05	0.3402±0.04	0.3968±0.03	0.4053±0.04	0.3743±0.07	0.3666
China	0.3318±0.04	0.3348±0.05	0.3276±0.03	0.3741±0.07	0.3819±0.06	0.3520±0.05	0.3504
Iran	0.3252±0.06	0.3358±0.03	**0.4221±0.06**	**0.4916±0.06**	**0.4418±0.04**	**0.4818±0.09**	0.4164
Greece	0.3360±0.05	0.3015±0.04	0.3573±0.03	0.4045±0.01	0.4012±0.03	0.3365±0.05	0.3562
Spain	0.3240±0.03	0.3283±0.04	0.3625±0.04	0.3880±0.04	0.3638±0.03	0.3313±0.04	0.3497
Russia	0.3583±0.03	0.3199±0.04	0.3899±0.03	0.3732±0.04	0.3907±0.01	0.3591±0.03	0.3652
ALL	0.3395±0.04	0.3247±0.04	0.3515±0.03	0.3940±0.03	0.3824±0.03	0.3750±0.04	0.3612

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

Table 13: Probing performance comparison with **Russian** prompts and **FMLAMA-*ru*** sub-dataset.

ID	Subject (dish)	Involved Object (ingredients)	Rank
0	pomegranate soup	pomegranate	0
1	Albaloo polo	rice	1
2	Kofta	minced meat	0
3	Sabzi polo	herb	5
4	falooda	milk, vermicelli	1, 0
5	ghormeh sabzi	parsley, herb	0, 4
6	Bastani	milk, sugar, egg	8, 7, 6
7	zeytoon parvardeh	olive	8
8	rock candy	sugar	0
9	Faloodeh	vermicelli	0
10	muhallebi	milk, rice flour	2, 1
11	rogan josh	lamb meat	1
12	ash-e doogh	yogurt	0
13	Mahyawa	fish	4
14	sholezard	rice	1
15	Eggplant Caviar	eggplant	0
16	nân-e panjere	flour	6

Table 14: Top 17 dishes from the Iran group with ingredients ranked in the top 10 by Llama3, excluding common prediction errors.