WORLDCUISINES: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines

Genta Indra Winata[*♠,1,2], **Frederikus Hudi**[*♠,2,3], **Patrick Amadeus Irawan**[*♠,2,4], **David Anugraha**[*♠,5], **Rifki Afina Putri**[*♠,2,6], **Yutong Wang**[♠,7], **Adam Nohejl**[♠,3], **Ubaidillah Ariq Prathama**[♠,4], **Nedjma Ousidhoum**[♠,8], **Afifa Amriani**[9], **Anar Rzayev**[6], **Anirban Das**[1], **Ashmari Pramodya**[3], **Aulia Adila**[7], **Bryan Wilie**[10], **Candy Olivia Mawalim**[7], **Ching Lam Cheng**[11], **Daud Abolade**[12,13], **Emmanuele Chersoni**[14], **Enrico Santus**[9], **Fariz Ikhwantri**[9], **Garry Kuwanto**[15], **Hanyang Zhao**[16], **Haryo Akbarianto Wibowo**[17], **Holy Lovenia**[2], **Jan Christian Blaise Cruz**[2,17], **Jan Wira Gotama Putra**[9], **Junho Myung**[6], **Lucky Susanto**[18], **Maria Angelica Riera Machin**[3], **Marina Zhukova**[19], **Michael Anugraha**[9], **Muhammad Farid Adilazuarda**[2,17], **Natasha Santosa**[20], **Peerat Limkonchotiwat**[2,21], **Raj Dabre**[22], **Rio Alexander Audino**[4], **Samuel Cahyawijaya**[2,23], **Shi-Xiong Zhang**[1], **Stephanie Yulia Salim**[7], **Yi Zhou**[8], **Yinxuan Gui**[11], **David Ifeoluwa Adelani**[♣,12,24,25,26], **En-Shiun Annie Lee**[♣,5,27], **Shogo Okada**[♣,7], **Ayu Purwarianti**[♣,2,4], **Alham Fikri Aji**[♣,2,17,18], **Taro Watanabe**[♣,3], **Derry Tanti Wijaya**[♣,15,18], **Alice Oh**[♣,6], **Chong-Wah Ngo**[♣,11],

[1]Capital One [2]SEACrowd [3]NAIST [4]ITB [5]UofT [6]KAIST [7]JAIST [8]Cardiff University
[9]Independent [10]HKUST [11]SMU [12]Masakhane [13]University of Lagos [14]HK PolyU
[15]Boston University [16]Columbia University [17]MBZUAI [18]Monash University [19]UCSB
[20]Tokyo Tech [21]AI Singapore [22]NICT [23]Cohere [24]McGill [25]MILA
[26]Canada CIFAR AI Chair [27]Ontario Tech
♠Main Authors ♣Senior Authors

Abstract

Vision Language Models (VLMs) often struggle with culture-specific knowledge, particularly in languages other than English and in underrepresented cultural contexts. To evaluate their understanding of such knowledge, we introduce WORLDCUISINES, a massive-scale benchmark for multilingual and multicultural, visually grounded language understanding. This benchmark includes a visual question answering (VQA) dataset with text-image pairs across 30 languages and dialects, spanning 9 language families and featuring over *1 million data points*, making it the largest multicultural VQA benchmark to date. It includes tasks for identifying dish names and their origins. We provide evaluation datasets in two sizes (12k and 60k instances) alongside a training dataset (1 million instances). Our findings show that while VLMs perform better with correct location context, they struggle with adversarial contexts and predicting specific regional cuisines and languages. To support future research, we release a knowledge base with annotated food entries and images along with the VQA data.

Figure 1: Images of stuffed pasta and dumplings from our dataset showcase a similar culinary concept across different cultures: wrapping meat, dairy (such as cheese), or vegetables in dough. These dishes can be prepared in various ways, including pan-frying, deep-frying, steaming, or boiling.

1 Introduction

Food is an essential medium for the exchange of regional cultures, serving to connect diverse peoples and traditions (Wahlqvist, 2007). Analyzing various culinary practices provides valuable insights into the cultural values, historical narratives, and social customs of the communities that produce and consume these foods (Holtzman, 2006). Furthermore, food plays a significant role in shaping lan-

[*] These authors contributed equally. This is an open-source project, and the work was done outside of their affiliations. Contacts: genta.winata@capitalone.com and frederikus.hudi.fe7@naist.ac.jp.

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 3242–3264

	# VQA	# Lang./Dialect†	# Countries	# Food Entries	# Images	Parallel Data	License
FoodieQA (Li et al., 2024b)	659	2	1	60	389	×	CC BY-NC-ND 4.0
World Wide Dishes (Magomere et al., 2024)	765	131	63	765	301	×	CC-BY 4.0
xGQA (Pfeiffer et al., 2022)	12,578	8	8	N/A	398	✓	CC-BY 4.0
MaXM‡ (Changpinyo et al., 2023)	2,142	7	7	N/A	335	×	Custom
EVJVQA (Nguyen et al., 2023)	33,790	3	1	N/A	4,909	×	N/A
CulturalVQA (Nayak et al., 2024)	2,378	1	11	N/A	2,328	×	N/A
SEA-VQA (Urailertprasert et al., 2024)	1,999	1	8	N/A	515	×	Custom
CVQA (Romero et al., 2024)	9,000	26	28	1,834	4,560	✓	Various
IndiFoodVQA (Agarwal et al., 2024)	16,716	1	1	255	414	×	N/A
WC-VQA	**1,152,000**	30	189	2,414	6,045	✓	CC BY-SA 4.0

Table 1: Data statistics for WC-VQA compared to existing VQA datasets. The data samples are sourced from their respective publications. ‡The reported numbers are based on their human-annotated test set. †This entry includes the language variations we collected for all languages.

guage, which serves as a proxy for cultural knowledge (Freedman, 2021). Food choices often reflect intricate community histories, societal transformations, and both individual and collective identities, thereby creating a rich tapestry of cultural expression (Almerico, 2014). The relationship between culture and food is dynamic; both evolve in tandem over time, resulting in the emergence of new dishes that are influenced by historical culinary traditions (Anderson, 2014).

As a result, similar food concepts can be found across different countries, reflecting a shared human culinary heritage. Researchers use food as a proxy to model and analyze cultural dynamics, helping to quantify cultural differences across regions (Adilazuarda et al., 2024). Many cultures have developed their own versions of "stuffed pasta" or "dumplings", each with unique ingredients and preparation methods, often known by different names (Gallani, 2015), as illustrated in Figure 1. Small details like how the dumpling is shaped can signal the cultural background. Conversely, some dishes share the same name but have different meanings; for example, "jelly" in the U.S. refers to a fruit spread, while in the U.K. and parts of Asia, it refers to a gelatinous dessert (Poppe, 1992; Abe, 2013). This culinary diversity presents a challenge for Vision Language Models (VLMs), which must accurately recognize and differentiate food items based on cultural context for applications like food recognition. These models navigate the complexities of names, ingredients, and preparation styles that vary widely across regions. VLMs have shown effectiveness in text captioning (Liu et al., 2024b,c) and have been adapted to support multiple languages (Geigle et al., 2023; Shin et al., 2024).

However, there is limited research on evaluating the multicultural capabilities of VLMs, particularly in terms of multilinguality. The study by Romero et al. (2024) introduce visual question answering (VQA) from a multicultural perspective, but it mainly focuses on knowledge and situational context at a specific moment, which does not fully assess the ability of VLMs to reason and differentiate between cultures within a single question. Moreover, another study on food VQA is limited to Chinese culture and does not explore the broader spectrum of global cultures (Li et al., 2024b). An earlier investigation into cultural bias in language models also found that cultural knowledge is lacking (Naous et al., 2023). Therefore, further research is necessary to address these limitations and enhance our understanding of VLMs' multicultural and multilingual capabilities.

To facilitate a comprehensive analysis of multilingual and multicultural research, we develop resources for evaluating VLMs. Table 1 summarizes how our work compares to previous studies. Our benchmark stands out for its cultural diversity, offering more VQA datasets and broader language and dialect coverage. Our major contributions can be summarized in three-fold:

- We present WORLDCUISINES, the first massive scale benchmark consisting of 1 million high-quality multilingual and multicultural text-image pairs annotated by native speakers in their local languages. We publicly release our resources, i.e., datasets, [1] code,[2] and leaderboard[3] to advance future research in this rapidly evolving field.

- We evaluate open-source and commercial

[1]We release WC-VQA at https://huggingface.co/datasets/worldcuisines/vqa and WC-KB consisting food, location, cuisine, and prompt templates at https://huggingface.co/worldcuisines.

[2]We release our code at https://github.com/worldcuisines/worldcuisines.

[3]We release our leaderboard at https://huggingface.co/spaces/worldcuisines/worldcuisines.

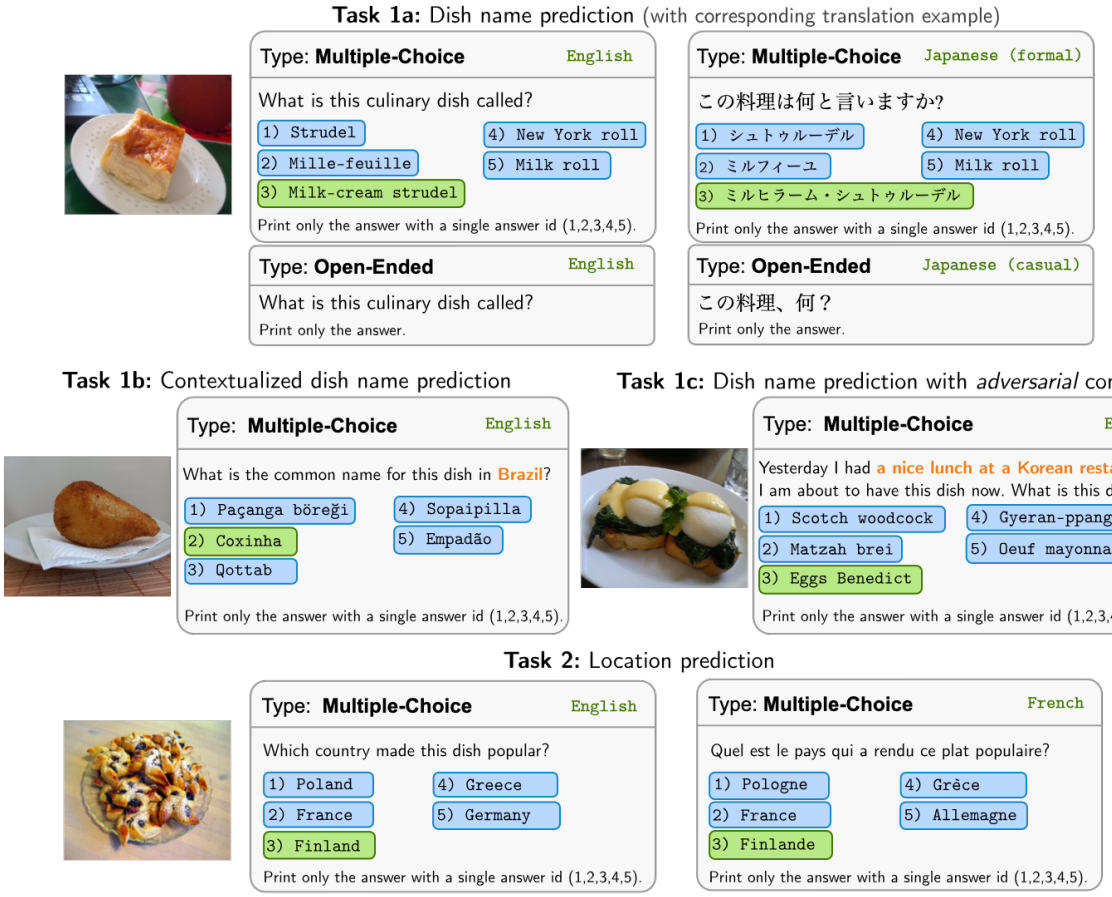


Figure 2: WC-VQA in WORLDCUISINES comprises two primary tasks: (1) predicting dish names and (2) predicting regional cuisines. Task 1 is further divided into three subtasks: (a) no-context, (b) contextualized, and (c) adversarial. We also include two answer types: multiple-choice question (MCQ) and open-ended question (OEQ).

VLMs for cultural awareness through two VQA tasks: predicting dish names from images and context, and identifying their geographical origin. We also assess the impact of context, including adversarial scenarios.

- We create multilingual templates for queries and context (such as the questions in QA pairs) while preserving language varieties, including dialects and registers. This is achieved by creating translations that incorporate different inflections, articles, and contractions. Our goal is to ensure naturalness in each translation and to use appropriate inflections for place names.

2 WORLDCUISINES

We propose WORLDCUISINES, an open-source benchmark designed to evaluate the cultural relevance and understanding of VLMs. Figure 2 displays VQA examples in English, alongside selected parallel translations in Japanese and French.

2.1 Overview

We develop both a VQA dataset (WC-VQA) and a curated KB for world cuisines (WC-KB). The WC-VQA dataset is constructed using WC-KB, which serves as the primary data source. We design two tasks as follows:

- **Task 1:** Dish Name Prediction. This task involves predicting the name of a dish based on its image, a question, and contextual information. It comprises three subtasks, each with distinct query types: *(a) no-context* question, *(b) contextualized* question, and *(c) adversarial contextualized* question.

- **Task 2:** Location prediction. The task is to predict location where the food is commonly consumed and originated given the dish image, question, and a context.

WC-KB. A KB encompassing 2,414 dishes worldwide includes 6,045 images and metadata,

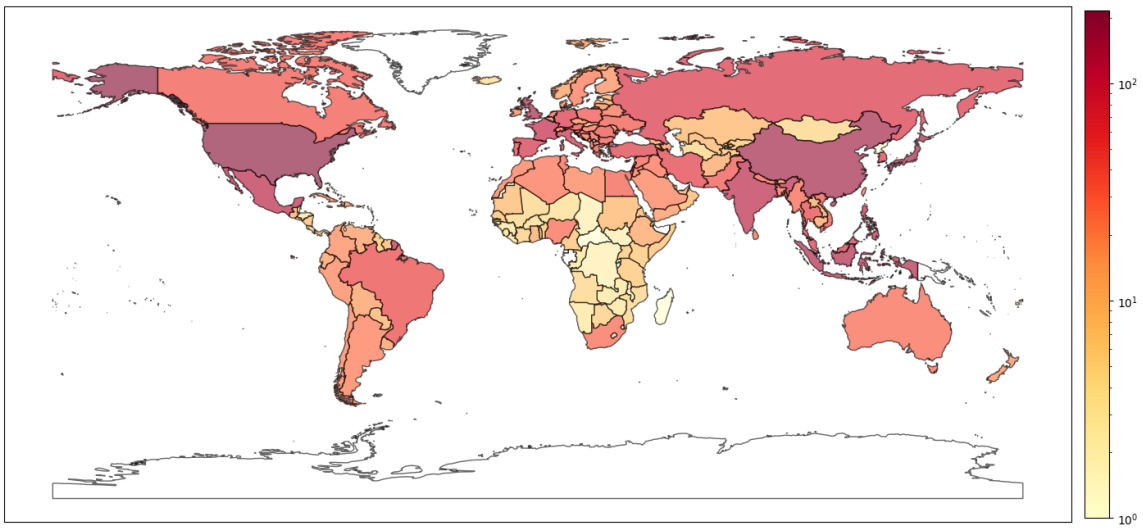


Figure 3: WORLDCUISINES distribution of food entries by country in the World Map. The food entries are distributed across 189 countries, with the highest concentration found in Asia, Europe, and North America. There are also some entries from the continents of Africa, Oceania, and Central and South America.

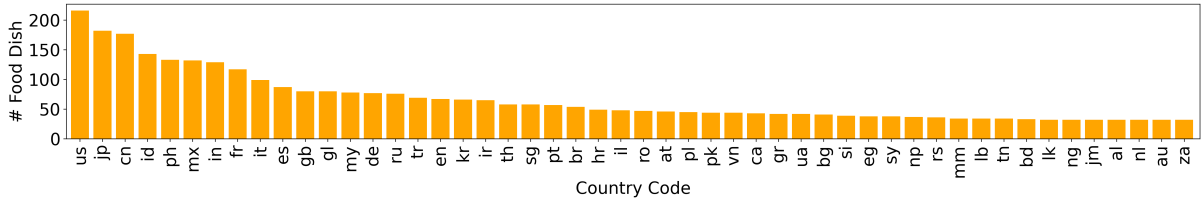


Figure 4: Countries by number of assigned dishes, showing the top 50 countries.

covering both coarse-grained (e.g., stew) and fine-grained categories (e.g., beef stew), locations, and regional cuisines. It also features multilingual translations of 90 crowd-sourced prompt templates and 401 parallel data entries (i.e., multilingual information) for location and regional cuisine information.

WC-VQA. A multilingual parallel VQA dataset with 1 million samples encompassing over 30 languages and dialects, including various varieties and registers, such as formal and casual styles, with high-quality human annotations. The VQA is designed to evaluate models' ability to understand cultural food names and their origins.

2.2 WC-KB Construction

Our data sources are gathered from Wikipedia[4] and Wikimedia Commons[5] to ensure they can be easily redistributed under an accepted open-source license. The data construction process involves four key steps: (1) dish selection, (2) metadata annotation, (3) quality assurance, and (4) data compilation. Figure 3 provides statistics on the regions covered in our dataset, with detailed information available in Table 9 in the Appendix. Figure 4 shows the distribution of dish frequencies, highlighting the top 50 countries with the most dishes.

2.2.1 Dish Selection

We compile a comprehensive list of dish names sourced from Wikipedia. We manually review pages that feature lists of dishes to determine whether each dish is a specialty unique to a specific culture, as we aim to focus on dishes that have distinct cultural significance. We exclude generic categories, such as ice cream, which lacks a specific cultural association. We ensure that each dish on our list has its own dedicated Wikipedia page. If a dish does not have a Wikipedia page, it is also excluded from our compilation. This meticulous approach ensures that our dataset is both culturally relevant and well-documented.

[4]Wikipedia web pages can be accessed at https://wikipedia.org.

[5]Wikimedia Commons web pages can be accessed at https://commons.wikimedia.org.

Data Split	Task 1 (Dish Name) (a) no-context # VQA	(a) no-context # Images	(b) contextualized # VQA	(b) contextualized # Images	(c) adversarial # VQA	(c) adversarial # Images	Task 2 (Location) # VQA	Task 2 (Location) # Images	Total # VQA
Train (1M)	270,300	3,383	267,930	3,555	271,770	3,589	270,000	3,361	1,080,000
Test Small (12k)	3,000	100	3,000	100	3,000	100	3,000	100	12,000
Test Large (60k)	15,000	500	15,000	500	15,000	499	15,000	499	60,000

Table 2: Dataset statistics for WC-VQA tasks for train, test small, and test large data splits. Total #VQA represents the total number of VQA from Task 1 and Task 2.

2.2.2 Metadata Annotation

Given a dish name and its corresponding Wikipedia page link, we then ask annotators to manually compile metadata based on the provided information. This metadata includes:

- **Visual Representation**: Images sources from Wikimedia Commons are included, along with their license information.
- **Categorization**: Dishes are classified into both coarse-grained (e.g., rice, bread) and fine-grained (e.g., fried rice, flatbread) categories.
- **Description**: Annotators provide a description of each dish based on the content from its Wikipedia page, avoiding the use of the dish's name, origin, or any distinctive keywords that uniquely identify the dish.
- **Cuisine**: The dish's origin cuisine and any cuisines with which it is strongly associated.
- **Geographic Distribution**: This includes the dish's associated countries, area (city or region), and broader continental region.

The metadata description, along with the example, is further elaborated in the Appendix Table 4.

2.2.3 Quality Assurance

Before starting the quality assurance process, we first identify common issues that arise during the annotation and develop automated rules to detect easily identifiable annotation errors, such as incorrect string formatting. Annotators are then asked to correct these errors. To further ensure data quality and validity, we conduct several rounds of quality assurance. Initially, we focus on image quality by removing instances where images are blurry, dark, or contain distracting elements such as people or other dishes. We also verify image licenses by cross-referencing them with information on Wikimedia Commons. Next, we refine the dish categorization and descriptions, ensuring consistency in category assignments and maintaining descriptions free from "information breaches" (e.g., excluding regional details from the description). We standardize cuisine names and eliminate any redundancies. Finally, we meticulously review all country and area information to ensure its accuracy. This comprehensive approach guarantees the integrity and reliability of our dataset.

2.2.4 Data Compilation

In this phase, we verify the overall quality check done by annotators, and identify any potential inconsistencies that are missed during the quality assurance. Then, we compile the dataset by collecting the metadata into a single file.

2.3 VQA Generation

In this phase, we generate VQA data by sampling from WC-KB. An entry of VQA data comprises visual image, question text, and answer text. This process involves four stages: (1) conducting a similarity search for dish names, (2) constructing questions and contexts, (3) translating these elements into multiple languages, and (4) generating the VQA triplets.

2.3.1 Dish Names Similarity Search

To identify similar dishes in our dataset, we follow the approach from Winata et al. (2024) to employ a multilingual model E5_{LARGE} Instruct (Wang et al., 2024) for computing text embedding. Formally, given a dish x with name x_{name} and text description x_{desc}, we use a multilingual model θ to compute the embedding vector $v_x = \theta(\{x_{\text{name}}; x_{\text{desc}}\})$, then apply cosine similarity to compute a score $s = \text{similarity}(v_i, v_j)$ between dish i and dish j. For each dish, we consider the top-k most similar dishes to generate distractors in the multiple choice question.

2.3.2 Question and Context Construction

Dish name prediction (Task 1) is divided into three question variations depending on the context: (1a)

Model	Task 1 (Dish Name) (a) no-context		(b) contextualized		(c) adversarial		Task 2 (Location)		Average	
	MCQ	OEQ	MCQ	OEQ	MCQ	OEQ	MCQ	OEQ	MCQ	OEQ
Open-Source										
Llava1.6 Vicuna 7B	34.57	1.59	43.48	4.03	34.84	1.41	32.24	9.29	36.28	4.08
Llava1.6 Vicuna 13B	40.17	2.79	48.17	5.85	39.05	2.57	37.79	10.16	41.30	5.34
Qwen2 VL Instruct 2B	41.65	7.98	42.29	8.13	39.69	6.74	47.85	14.55	42.87	9.35
Qwen2 VL Instruct 7B	61.48	6.76	67.85	10.36	53.52	6.12	55.90	21.03	59.69	11.07
Qwen2 VL Instruct 72B	74.19	12.67	80.79	21.31	62.43	8.37	61.90	27.27	69.83	17.40
Llama 3.2 Instruct 11B	59.93	18.75	64.12	22.96	53.17	13.39	57.93	31.58	58.79	21.67
Llama 3.2 Instruct 90B	77.69	16.93	82.92	23.60	63.96	10.87	67.87	31.31	73.11	20.68
Molmo-E 1B	18.81	0.01	24.22	0.23	19.55	0.01	18.97	1.54	20.39	0.45
Molmo-D 7B	46.01	2.89	55.95	3.66	41.61	2.31	33.35	11.45	44.23	5.08
Molmo-O 7B	39.96	5.15	44.93	6.03	38.41	3.51	29.81	10.07	38.28	6.19
Pangea 7B‡	52.35	1.52	63.07	2.73	49.17	1.57	48.71	20.15	53.33	6.49
Aria 25B	58.61	4.99	69.29	9.17	52.82	3.39	42.82	16.20	55.89	8.44
Phi-3.5 Vision 4B	43.37	2.91	48.71	4.23	40.87	2.07	35.01	9.22	41.99	4.61
Pixtral 12B	56.65	1.22	70.69	2.94	52.12	1.09	46.67	14.43	56.53	4.92
NVLM-D 72B	69.82	4.71	78.93	10.29	52.12	2.89	51.97	16.68	63.21	8.64
Proprietary										
GPT-4o	**88.45**	**21.88**	**91.57**	**37.51**	**82.29**	**14.79**	66.52	**37.13**	**82.21**	**27.83**
GPT-4o Mini	72.80	10.28	81.65	20.87	57.76	5.72	52.37	25.79	66.14	15.66
Gemini 1.5 Flash	77.05	12.81	80.97	15.16	69.13	6.46	**71.53**	30.03	74.67	16.12

Table 3: Accuracy (%) results of WC-VQA for Test Large (60k). MCQ and OEQ indicate multiple-choice question and open-ended question, respectively. Best and second-best are **bolded** and underlined, respectively. ‡We employ an optimized prompt provided by the authors (see Subsection E.1 in the Appendix for further details).

no-context question, where we simply ask for the name of the dish without any provided context; (1b) *contextualized question* where we provide additional information related to cuisine or location; and (1c) *adversarial contextualized question* which are similar to the contextualized questions but may include misleading location information to assess the model's robustness to irrelevant details.

For example, consider *coxinha* from **Brazil**, shown in Figure 2 (1b). A query with additional context here would be: "`What is the common name for this dish in` **`Brazil`**`?`" Here, the origin of *coxinha*, **Brazil**, serves as the context. In contrast, adversarial context involves providing misleading or irrelevant information in terms of location or type of cuisine to assess the model's robustness to such distractions. For instance, in the case of eggs benedict shown in Figure 2 (1c), an adversarial context would be: "`Yesterday I had a nice lunch at a` **`Korean`** `restaurant. I am about to have this dish now. What is this dish called?`" In this scenario, the model should ignore the irrelevant detail ("`nice lunch at a Korean restaurant`") and focus solely on the image and the question.

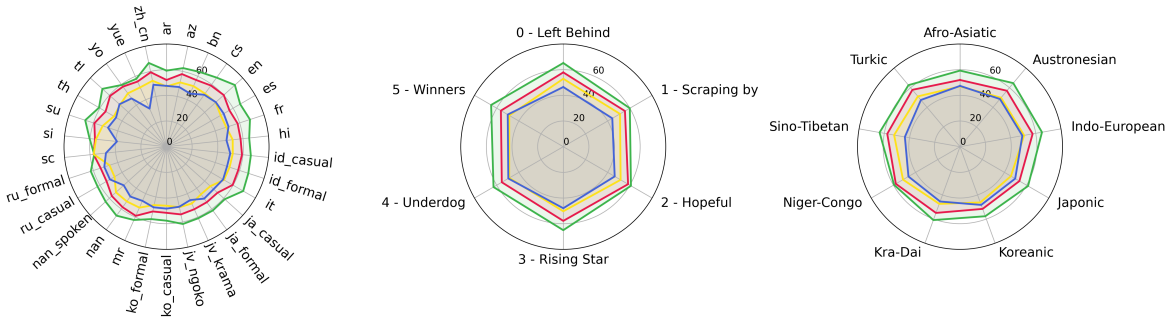
Only basic question without any provided context is available for regional cuisine prediction (Task 2). The data statistics for each task are presented in Table 2.

2.3.3 Multiple Language Translation

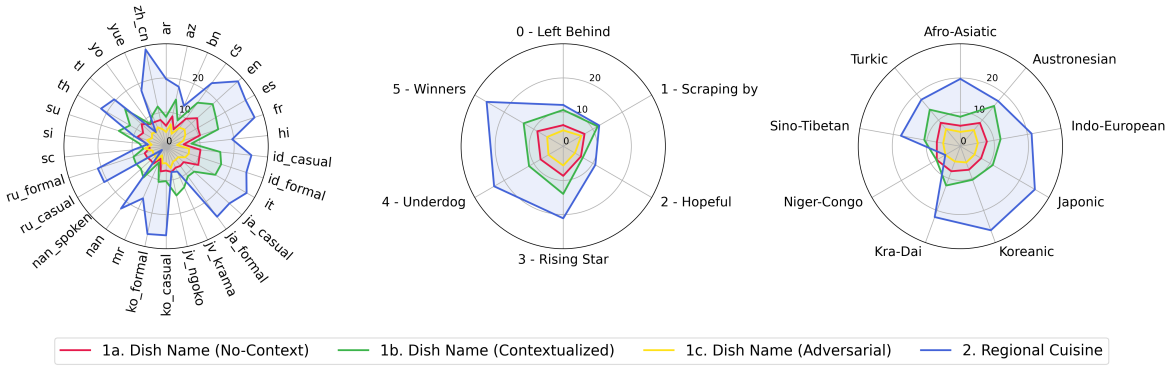
Question and Context. All questions and contexts are initially collected in English, which are then carefully translated by native human speakers into 30 language varieties: 23 different languages with 7 languages having two different varieties each. We instructed the translators to prioritize the naturalness, and then followed by the diversity of translations when the duplication occurs.

Food Name Alias. Using Wikipedia pages as our primary source, we can verify if the English page has translations available in other languages. This enables us to extract dish names in multiple languages and compile them as translations for each dish. We utilize both the Wikipedia page titles in various languages and the alias text found on the English page. These translations are especially valuable for multilingual prompt translation, as they allow us to use the dish's native name instead of its English equivalent, enhancing cultural relevance and accuracy. We use the English name as default when the translation is unavailable.

Locations and Cuisines. As there are more than 400 unique locations, including countries, cities,



(a) Multiple-choice question (MCQ).



(b) Open-ended question (OEQ).

Figure 5: Accuracy (%) categorized by language **(left)**, language vitality **(center)**, and language family **(right)**. We classify the language vitality by following the classification proposed by Joshi et al. (2020).

and areas, we first translate the English locations into other languages by using GPT-4o, followed by proofreading each translation by the native speakers. The string values for the regional cuisines, i.e., the adjective form of the location in English, are translated in the same manner as location.

Morphological Inflections. Indo-European languages, such as Czech or Spanish, are rich in inflectional morphology which involves word modification to express different grammatical categories, such as number, gender, or case. For example, the equivalents of the phrases "in Japan" and "from Japan" in Czech are "v Japonsk*u*" and "z Japonsk*a*", respectively. We provide a framework for the human translators to use the inflections in the prompt template to prioritize the naturalness while keeping the inflections as few as possible.

2.3.4 Generating VQA Triplets

To ensure no overlap in train and test subsets, we split the dishes and the multilingual-questions into two subsets each, to ensure no dish or multilingual questions leakage between train and test. For every subset, we apply random sampling to get a pair of dish and its multilingual-questions. We use the dish entry in our WorldCuisines KB dataset to pick the image and the location to be injected to the context, if any. The answer candidates for multiple-choice were picked by utilizing similarity search (Section 2.3.1). We repeat this process until we reach the desired number of training or test samples, or until all possible dish and question combinations are used, discarding any duplicates.

3 Experiments

3.1 Experimental Setup

Metrics. We use accuracy as the primary metric to evaluate predictions. For Task 2 (open-ended), we employ BERTScore (Zhang et al., 2019) with XLM-R Large (Conneau and Lample, 2019) as a secondary metric to determine if the model-generated content includes food names similar to those in the gold labels. For open-ended questions, we compute the accuracy of each test sample against multiple references, including translations of the dish in different languages. This approach allows us to accommodate predictions that may not be in the expected language.

Models. We evaluate our benchmark on various available VLMs, including 15 open-source models and 3 proprietary models. During the inference

of the open-source model, we use 16-bit floating point and employ greedy decoding. We access the proprietary models through API. The complete list of the models is available in Table 3.

4 Results and Discussion

4.1 Overall Results

The results for WC-VQA are summarized in Table 3. The multiple-choice question (MCQ) results without any context exhibit significant variability, ranging from 30% to 80%, indicating considerable differences in model performance. This variability indicates that predicting MCQs remains a challenging task for many models. Notably, proprietary models, particularly GPT-4o, demonstrate exceptional performance, outperforming all other models. In the open-ended question (OEQ) setting, the task proves even more difficult than the MCQ, with models achieving a maximum accuracy of less than 20% for dish name predictions and slightly higher for location predictions when no context is provided. However, incorporating context enhances performance across all settings, highlighting that context effectively guides the models in making better predictions. Interestingly, when the adversarial context is introduced, it misleads the models, leading to incorrect predictions and adding further complexity to the task. Among the models evaluated, Llama 3.2 Instruct significantly outperforms other open-source model families, while Qwen2 performs relatively better than Llava 1.6 and Molmo, despite having smaller model sizes.

4.2 The Role of Context

For dish name prediction (Task 1), incorporating more relevant context significantly enhances performance across all language families. However, when adversarial context is introduced, performance drops significantly. The adversarial context included in the prompt significantly affects the prediction. Instead of relying solely on the image input, the model often sways and makes predictions based on incorrect location or cuisine information, even when the context is unrelated to the query. This observation is particularly intriguing, as it signifies that such prompts can shift the model's attention and influence its generation process.

4.3 Results by Language

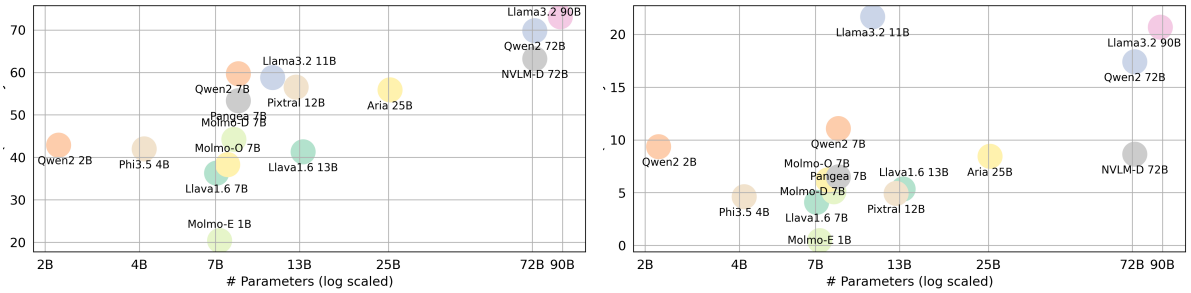
In Task 1 with OEQ setting (Figure 5b), some languages with non-Latin scripts, such as Arabic, Korean, Japanese, and Marathi, tend to perform poorly, with the exception of Chinese. For Task 2 with OEQ setting, most models struggle with Sino-Tibetan languages (i.e., Chinese, Cantonese, and Hokkien) and Niger-Congo languages (i.e., Yoruba). In contrast, the models demonstrate relatively strong performance with Japonic, Koreanic, Kra-Dai (i.e., Thai), and Turkic (i.e., Azerbaijani) languages. We also observe that answering OEQs in underrepresented languages remains particularly challenging for the models, as shown by the relatively lower results for the "left behind", "scraping by", and "hopeful" languages. Interestingly, lower performance in the OEQ does not necessarily translate to the lower performance in the MCQ setting (Figure 5a) where the performance gap between language categories is less pronounced. The gap between OEQ and MCQ, especially for underrepresented languages, suggests that the bottleneck might lie in the factors beyond cultural understanding, such as text generation capabilities.

4.4 Scaling Law

It is evident that large models perform better than smaller ones, showing the scaling law still exists in this experiment, as shown in Figure 6. It is very interesting to see the same trend across different model families (e.g., Llava, Qwen, and even GPT-4o series). However, it is pretty clear for open-source models, Llama 3.2 Instruct has the lead for overall performance, which may be due the coverage of multilingual data used in its training, although it is still unclear since there is no evidence or supporting information that can back up the finding. Regardless, NVLM-D model does not perform as good as their base model Qwen2 VL Instruct in our benchmark. One reason could be the NVLM model is highly tuned for English, but not in languages other than English.

5 Related Work

Cultural VQA. Several prior studies have focused on developing culturally relevant VQA benchmarks, including FM-IQA (Gao et al., 2015), MCVQA (Gupta et al., 2020), xGQA (Pfeiffer et al., 2022), MaXM (Changpinyo et al., 2023), MTVQA (Tang et al., 2024), MABL (Kabra et al., 2023), MAPS (Liu et al., 2024a), and MaRVL (Liu et al., 2021). Additionally, CVQA (Romero et al., 2024) and CulturalVQA (Nayak et al., 2024) provide VQA datasets that cover various regions and



(a) MCQ Accuracy vs. Parameters.

(b) OEQ Accuracy vs. Parameters.

Figure 6: Scaling matters for MCQ (6a) and OEQ (6b).

diverse topics, including food, with CVQA also offering questions in multiple languages alongside English translations. SEA-VQA (Urailertprasert et al., 2024) specifically benchmarks the South East Asian region. In contrast, FoodieQA (Li et al., 2024b) and World Wide Dishes (Magomere et al., 2024) are benchmark focusing on food. Our work is similarly motivated by using food as a cultural proxy, but it distinguishes itself with a significantly larger dataset and broader coverage of languages and cultures.

Multi-modal LLMs. Recent advancements in VLMs have led to the emergence of multi-modal LLMs that can process both images and text. LLaVA (Liu et al., 2024c) exemplifies this approach by utilizing Vicuna (Zheng et al., 2023) as an image encoder, thereby enhancing visual understanding. This architecture has set a precedent for other VLMs, including Qwen2-VL (Bai et al., 2023), Llama 3.2 (Dubey et al., 2024), Pixtral (Agrawal et al., 2024), Phi-3.5 Vision (Abdin et al., 2024), Molmo (Deitke et al., 2024), Aria (Li et al., 2024a), Pangea (Yue et al., 2024), and NVLM (Dai et al., 2024), each leveraging their respective large language models for multi-modal tasks. In a specialized application, FoodLMM (Yin et al., 2023) focuses specifically on the food domain, training on publicly available food datasets and conversational data generated by GPT-4 (Achiam et al., 2023). Our work evaluates the capabilities of these models within the food domain, offering insights into their performance and potential applications in culinary-related tasks across multicultural settings.

6 Conclusion

We introduce WORLDCUISINES, an open-source, large-scale benchmark designed for multilingual and multicultural, visually grounded language understanding. It comprises over 1 million data points across 30 languages and dialects. Our findings reveal that this benchmark remains challenging for VLMs, particularly with dishes from specific regions and in low-resource languages. This provides insight into how well models understand regional cuisines. To enhance usability, we offer a dedicated evaluation split with two datasets of varying sizes. Our evaluation shows that while VLMs perform better with the correct context, they struggle with adversarial contexts intended to mislead them. Additionally, we are releasing a comprehensive knowledge base, VQA dataset, code, and leaderboard as open-source resources to support future research.

Acknowledgements

We extend our gratitude to everyone who has supported our project, especially the numerous annotators who provided meticulous and comprehensive annotations and conducted thorough quality checks. Special thanks to Francesca Porcu for her assistance with the Sardinian language and to Shintaro Ozaki for his help with Japanese. We are also deeply appreciative of Nayeon Lee and Wenliang Dai for their insightful discussions and for integrating NVLM into our benchmark. Additionally, we thank Xiang Yue and Yueqi Song for their help in integrating Pangea into our benchmark.

Limitations

In this paper, we limit our investigation to avoid exhaustively evaluating all possible models due to resource constraints. Our primary focus is on developing a benchmark that facilitates exploration for future research. We also provide a training data split for reference, allowing other researchers to utilize it to enhance their VLMs and evaluate their

models against our test sets. Currently, we include 30 different languages and dialects, establishing one of the largest and most diverse benchmarks for comprehensive multilingual VQA. We aim to extend this benchmark to encompass additional languages in the future, making it more inclusive and representative of a broader range of linguistic diversity.

It is important to note that our food entries are currently sourced from English Wikipedia. Although we aim to include as many diverse dishes as possible, we acknowledge that this approach limits the coverage of some regions. This is due to language affects commonsense and its specific knowledge (Sakai et al., 2024), which in turns suggesting insufficiency of sourcing only English Wikipedia. Nevertheless, our dataset serves as a valuable starting point. In future work, we plan to incorporate entries from non-English Wikipedia pages to improve regional representation and cultural diversity. For evaluation purposes, we include accuracy metrics for overall model performance and BERTScore for more detailed analysis. However, we recognize that evaluating VQA model performance on multicultural data remains an open challenge. Appropriate evaluation metrics are needed to effectively model the diversity of cultural contexts and linguistic variations. Addressing this issue will be a key focus of our future research efforts.

Ethical Considerations

Our research focuses on evaluating VLMs within the context of multilingual and multicultural VQA, a field that holds significant implications for diverse multilingual communities. We are committed to conducting our data collection and evaluations with the highest standards of transparency and fairness. To achieve this, we have adopted a crowd-sourcing approach for the annotation process, inviting volunteers to contribute and become co-authors if they provide significant contributions. We follow the guidelines from ACL for authorship eligibility as shown in https://www.aclweb.org/adminwiki/index.php/Authorship_Changes_Policy_for_ACL_Conference_Papers. In line with our commitment to openness and collaboration, we will release our dataset under an open-source license, CC-BY-SA 4.0.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Marié Abe. 2013. Tokyo, japan. In *The Ethnomusicologists' Cookbook*, pages 40–45. Routledge.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Pulkit Agarwal, Settaluri Sravanthi, and Pushpak Bhattacharyya. 2024. Indifoodvqa: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1158–1176.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Gina M Almerico. 2014. Food and identity: Food studies, cultural, and personal identity. *Journal of International Business and Cultural Studies*, 8:1.

Eugene Newton Anderson. 2014. *Everyone eats: Understanding food and culture*. NYU Press.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. Maxm: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Freedman. 2021. *Why Food Matters*. Yale University Press.

Barbara Gallani. 2015. *Dumplings: A global history*. Reaktion Books.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 900–913.

Jon D Holtzman. 2006. Food and memory. *Annu. Rev. Anthropol.*, 35(1):361–378.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. 2024a. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, et al. 2024b. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *arXiv preprint arXiv:2406.11030*.

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

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Foutse Yuehgoh, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Elizaveta Semenova, et al. 2024. You are what you eat? feeding foundation models a regionally diverse food dataset of world wide dishes. *arXiv preprint arXiv:2406.09496*.

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

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

Ngan Luu-Thuy Nguyen, Nghia Hieu Nguyen, Duong TD Vo, Khanh Quoc Tran, and Kiet Van Nguyen. 2023. Vlsp2022-evjvqa challenge: Multilingual visual question answering. *arXiv preprint arXiv:2302.11752*.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

J Poppe. 1992. Gelatin. In *Thickening and gelling agents for food*, pages 98–123. Springer.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214. Association for Computational Linguistics.

DongJae Shin, HyeonSeok Lim, Inho Won, ChangSu Choi, Minjun Kim, SeungWoo Song, HanGyeol Yoo, SangMin Kim, and KyungTae Lim. 2024. X-llava: Optimizing bilingual large vision-language alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2463–2473.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. Sea-vqa: Southeast asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185.

Mark L Wahlqvist. 2007. Regional food culture and development. *Asia Pacific journal of clinical nutrition*, 16:2.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. 2024. Miners: Multilingual language models as semantic retrievers. *arXiv preprint arXiv:2406.07424*.

Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2023. Foodlmm: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Data Statement

A.1 Executive Summary

WORLDCUISINES is a vision-language benchmark comprised of two resources: (1) WC-VQA, a multilingual parallel question answering dataset covering 30 languages and dialects where each dish image is accompanied by questions and context constructed through human translation; and (2) WC-KB, a knowledge base containing images and metadata associated with the dishes.

A.2 Curation Rationale

The goal of WORLDCUISINES is to evaluate the cultural understanding of vision-language models (VLMs) within the food domain. To achieve this, we develop WC-VQA and WC-KB. Dish names and their information are collected from English Wikipedia, and the images are selected from Wikimedia Commons to ensure a permissive license, with an emphasis on representing a wide range of food categories and geographic origins (or where the dish is popular). This selection strategy aims to provide insights into the VLMs' ability to generalize across diverse culinary and cultural contexts.

A.3 Language Variety

WORLDCUISINES covers 30 languages and dialects spoken across diverse countries and regions. The complete list of languages and dialects is shown in Table 5. An example of the multilingual prompt is shown in Table 6.

A.4 Annotator Demographic

Over 30 annotators are involved in building WORLDCUISINES, specifically in translating the query and context for the WC-VQA dataset. Most annotators are native speakers of the target languages or dialects included in our data; some are L2 speakers with more than 10 years of study in their respective languages. The detailed demographics for each language are elaborated below.

A.4.1 Austronesian

Indonesian Two native Indonesian speakers are involved as translators. One is in the 26–35 age

Attribute	Value	Description	Example
Name	String	Name of the dish.	"Dorayaki"
Alias	List<Dict>	Name alias, i.e. the name in the original language.	[{"どら焼き": "Japanese"}]
Coarse-grained categories	List<String>	Coarse-level categories.	["Pancake", "Dessert"]
Fine-grained categories	List<String>	Fine-level categories.	["Wagashi Pancake"]
Cuisines	String	Name of cuisine.	"Japanese"
Associated Cuisines	String	Associated cuisines to the dish.	"Japanese"
Area	String	Specific region where the dish is originated	"Ueno"
Countries	String	Specific region where the dish is originated	"Japan"
Region[1..5]	String	Specific continent where the dish is originated	"Eastern Asia"
Text Description	String	Short description of the dish, including the ingredients used to prepare the dish or the cooking method.	"The dish consists of two small pancake-like patties made from castella wrapped around a filling of sweet bean paste."
Image[1..8] URL	String	Image link to Wikimedia Commons.	". . . /commons/9/9c/Dorayaki_001_(3).jpg" ()
Image[1..8] License	String	License of the image	"CC BY-SA 3.0"

Table 4: WC-KB attributes in WORLDCUISINES.

Language Name	Language Vitality†	Resource Classification‡	Linguistic Register	Additional Notes
Austronesian				
Indonesian	Institutional	3 - Rising Star	Formal Casual	
Tagalog	Institutional	3 - Rising Star		
Sundanese	Stable	1 - Scraping by	Loma	Common speech form
Javanese	Institutional	1 - Scraping by	Krama Ngoko	Central-Java dialect, polite form Central-Java dialect, casual form
Japonic				
Japanese	Institutional	5 - Winners	Formal Casual	Polite form or teinei-go Daily conversation
Sino-Tibetan				
Chinese	Institutional	5 - Winners		Standard Mandarin
Cantonese	Institutional	1 - Scraping by		
Hokkien	Institutional	0 - Left Behind	Written Spoken	Medan dialect Medan dialect
Koreanic				
Korean	Institutional	4 - Underdog	Formal Casual	
Kra-Dai				
Thai	Institutional	3 - Rising Star		
Indo-European				
English	Institutional	5 - Winners		
Spanish	Institutional	5 - Winners		Latin-American dialect
French	Institutional	5 - Winners		
Russian	Institutional	4 - Underdog	Formal Casual	
Czech	Institutional	4 - Underdog		
Italian	Institutional	4 - Underdog		
Hindi	Institutional	4 - Underdog		
Bengali	Institutional	3 - Rising Star		
Marathi	Institutional	2 - Hopeful		
Sardinian	*Endangered*	1 - Scraping by		Logudorese (src)
Sinhala	Institutional	0 - Left Behind	Formal	Spoken form
Afro-Asiatic				
Arabic (MSA)	Institutional	5 - Winners		
Niger-Congo				
Yoruba	Institutional	2 - Hopeful		
Turkic				
Azerbaijani	Institutional	1 - Scraping by		North Variety (azj)

Table 5: The details of languages used in the prompt generation for our VQA dataset. †Taken from Ethnologue (Campbell and Grondona, 2008). ‡Based on Joshi et al. (2020).

range, and the other is in the 16–25 age range.

Tagalog One native Tagalog speaker in the 16–25 age range is involved as a translator.

Sundanese Two L2 Sundanese speakers contribute to the translation. One, in the 16–25 age range with 15 years of experience with the Sun-

Language	Question Prompt		Answer	
	Multi-choice question (MCQ)	Open-ended question (OEQ)	ID	Text
English	Yesterday I had a nice lunch at a Japanese restaurant. I am about to have this dish now. What is this dish called? 1. Hangtown fry 2. Zucchini slice 3. Chawanmushi 4. Rolex 5. Egg foo young Print only the answer with a single answer id (1,2,3,4,5).	Yesterday I had a nice lunch at a Japanese restaurant. I am about to have this dish now. What is this dish called? Print only the answer.	5	Egg foo young
French	Hier, j'ai pris un bon déjeuner dans un restaurant japonais. Je suis sur le point de manger ce plat maintenant. Comment appelle-t-on ce plat ? 1. Hangtown fry 2. Zucchini slice 3. Chawanmushi 4. Rolex 5. Fu yung hai Print only the answer with a single answer id (1,2,3,4,5).	Hier, j'ai pris un bon déjeuner dans un restaurant japonais. Je suis sur le point de manger ce plat maintenant. Comment appelle-t-on ce plat ? Print only the answer.	5	Fu yung hai
Indonesian (Formal)	Kemarin, saya menyantap makan siang yg nikmat di restoran Jepang. Sekarang saya akan menyantap hidangan ini. Disebut apakah hidangan ini? 1. Hangtown fry 2. Zucchini slice 3. Chawanmushi 4. Rolex 5. Puyunghai Print only the answer with a single answer id (1,2,3,4,5).	Kemarin, saya menyantap makan siang yg nikmat di restoran Jepang. Sekarang saya akan menyantap hidangan ini. Disebut apakah hidangan ini? Print only the answer.	5	Puyunghai
Indonesian (Casual)	Kemarin aku makan siang enak di restoran Jepang. Sekarang mau makan makanan ini. Makanan ini disebut apa? 1. Hangtown fry 2. Zucchini slice 3. Chawanmushi 4. Rolex 5. Puyunghai Print only the answer with a single answer id (1,2,3,4,5).	Kemarin aku makan siang enak di restoran Jepang. Sekarang mau makan makanan ini. Makanan ini disebut apa? Print only the answer.	5	Puyunghai
Japanese (Formal)	昨日、私は日本料理店で美味しい昼食を食べました。 今まさにこの料理を食べようとしています。 この料理の名前は何ですか? 1. Hangtown fry 2. Zucchini slice 3. 茶碗蒸し 4. Rolex 5. 芙蓉蛋 Print only the answer with a single answer id (1,2,3,4,5).	昨日、私は日本料理店で美味しい昼食を食べました。 今まさにこの料理を食べようとしています。 この料理の名前は何ですか? Print only the answer.	5	芙蓉蛋
Japanese (Casual)	昨日日本料理のお店で美味しいランチを食べたんだけど、 今まさに食べてるこの料理の名前は何？ 1. Hangtown fry 2. Zucchini slice 3. 茶碗蒸し 4. Rolex 5. 芙蓉蛋 Print only the answer with a single answer id (1,2,3,4,5).	昨日日本料理のお店で美味しいランチを食べたんだけど、 今まさに食べてるこの料理の名前は何？ Print only the answer.	5	芙蓉蛋
Javanese (Krama)	Kaping wingi kula nedha nikmat ing restoran Jepang. Kula kepengin nedha menika malih sakmenika. Naminipun nopo dhaharan menika? 1. Hangtown fry 2. Zucchini slice 3. Chawanmushi 4. Rolex 5. Endhog foo young Print only the answer with a single answer id (1,2,3,4,5).	Kaping wingi kula nedha nikmat ing restoran Jepang. Kula kepengin nedha menika malih sakmenika. Naminipun nopo dhaharan menika? Print only the answer.	5	Endhog foo young
Javanese (Ngoko)	Wingi aku mangan enak ndek restoran Jepang. Aku pengen mangan neh saiki. Opo jenenge panganan iki? 1. Hangtown fry 2. Zucchini slice 3. Chawanmushi 4. Rolex 5. Endhog foo young Print only the answer with a single answer id (1,2,3,4,5).	Wingi aku mangan enak ndek restoran Jepang. Aku pengen mangan neh saiki. Opo jenenge panganan iki? Print only the answer.	5	Endhog foo young

Table 6: Multilingual prompt example of Task 1 (c) adversarial in 8 language variants (out of 30). The visual image given is an image of Egg foo young, a Chinese cuisine. The "qa_id" of this example is 1806.

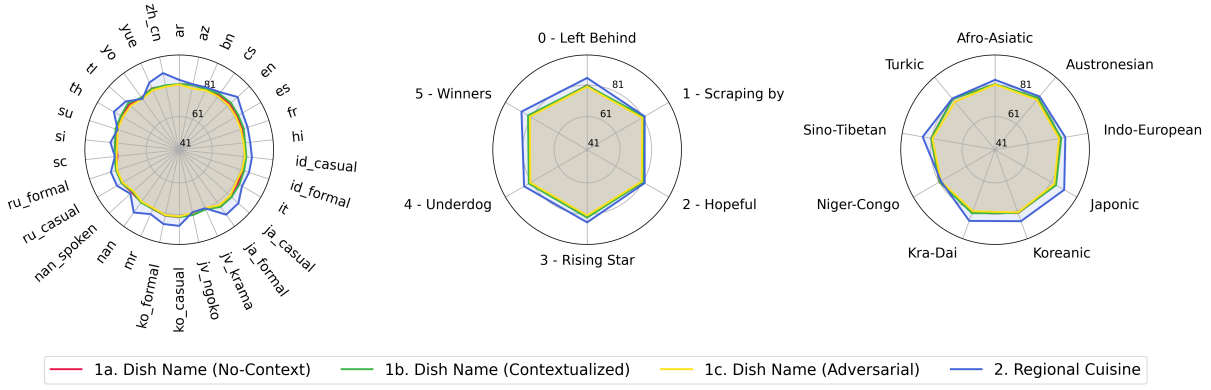


Figure 7: BERTScore (%) categorized by language **(left)**, language vitality **(center)**, and language family **(right)**. We classify the language vitality by following the classification from Joshi et al. (2020).

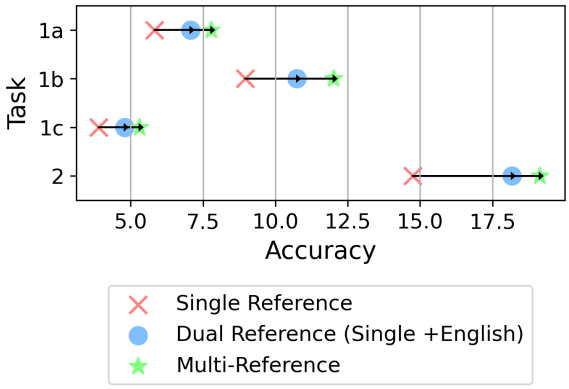


Figure 8: Model performance evaluated with different references on open-ended question.

danese language, assists with translation. The other, in the 26–35 age range with 25 years of experience with the language, primarily serves as the proofreader.

Javanese One native Javanese speaker with Central Java dialect in the 16–25 age range translates for both registers of the language (Krama and Ngoko).

A.4.2 Japonic

Japanese Three L2 Japanese speakers with over 10 years of language study contribute to the Japanese translation. Two are in the 26–35 age range, and one is in the 36–45 age range. A native Japanese speaker then proofreads the translated sentences. Additionally, one native Japanese speaker from Western Japan in the 16–25 age range gives input for the casual form.

A.4.3 Sino-Tibetan

Chinese One native Chinese speaker in the 16–25 age range is involved as a translator.

Cantonese Two native Cantonese speakers are involved as translators. One is in 36–45 age range, and the other is in the 16–25 age range.

Hokkien Two native Hokkien speakers in the Medan dialect translate for both written and spoken versions of the language. Both are in the 26–35 age range.

A.4.4 Koreanic

Korean One native Korean speaker in the 16–25 age range translates the formal and casual versions of the language.

A.4.5 Kra-Dai

Thai One native Thai speaker in the 26–35 age range is involved as a translator.

A.4.6 Indo-European

English Query and context in English are constructed. All are L2 English speakers with over 20 years of study and have lived in the English speaking countries. Four of the annotators are in the 26–35 age range, and one is in 36–45 age range. Two native English speakers skimmed through the prompt templates.

Spanish One native Spanish speaker in the 26–35 age range translates the Latin-American versions or dialects of the language.

French One native French speaker and one L2 speaker are involved as translators. The native speaker is in the 26–35 age range, and the L2 speaker is in the 36–45 age range.

Russian One native Russian speaker in the 26–35 age range is involved as translators. One L2 speaker in 36–45 proofreads the template for inflection.

Czech One native Czech speaker in the 36–45 age range is involved as a translator.

Italian Two native Italian speakers, both in the 36–45 age range, are involved as translators.

Hindi One native Hindi speaker in the 26–35 age range is involved as a translator.

Bengali One native Bengali speaker in the 26–35 age range is involved as a translator.

Marathi One native Marathi speaker in the 26–35 age range is involved as a translator.

Sardinian One native Logudorese Sardinian speaker in the 36–45 age range is involved as a translator.

Sinhala One native Sinhala speaker in the 26–35 age range is involved as a translator.

A.4.7 Afro-Asiatic

Arabic (MSA) One native Arabic speaker in the 26–35 age range is involved in the Modern Standard Arabic (MSA) translation.

A.4.8 Niger-Congo

Yoruba One native Yoruba speaker in the 16–25 age range is involved as a translator.

A.4.9 Turkic

Azerbaijani One native Azerbaijani speaker in the 16–25 age range is involved as a translator.

B Open-Source Collaborative Effort

The WORLDCUISINES data collection and benchmark construction is a fully open-source project. We invite contributions from researchers, practitioners, and grassroots communities, such as local NLP communities, who are interested in participating. Contributions can include data collection, annotation, quality checks, and evaluation. To ensure high-quality data, we engage native speakers of local languages in the annotation process with strict quality control (QC). The contributors who provide substantial contribution are invited to have co-authorship on this paper. We follow the guidelines from ACL for authorship eligibility.[6] Our goal is to develop a resource and benchmark that will have a meaningful impact on future research. To achieve this, we are dedicated to expanding language coverage and ensuring that contributions are as *inclusive and diverse* as possible.

[6]The ACL guidelines can be found at https://www.aclweb.org/adminwiki/index.php/Authorship_Changes_Policy_for_ACL_Conference_Papers.

C Detailed Dataset Construction

C.1 Dataset Compilation

Our dataset, comprising 2,414 dishes and 6,084 images, was meticulously compiled and verified manually. Key metadata includes dish name, alias, coarse- and fine-grained categories, cuisines, regions, descriptions, images, and their licenses. The compilation process followed these steps:

- We listed dish names for annotation.
- Annotators filled metadata fields and selected up to 8 licensed images per dish, guided by instruction documentation with examples for consistent accuracy.
- Post-annotation, annotators formed subgroups to verify specific metadata categories, ensuring detailed and consistent data across fields such as categories, cuisines, regions, descriptions, and images.

C.2 Negative Sampling

Recall that from our annotations, we have detailed metadata for all 2,414 dishes, including the dish name, coarse-grained categories, fine-grained categories, countries, and text descriptions. The negative answers were sampled using the following procedure:

(1) We used a multilingual model, specifically E5-LARGE Instruct, to compute the text embeddings. Each embedding was generated by concatenating the dish name with its corresponding text description.

(2) To identify negative samples, we computed the cosine similarity between the embeddings of the target dish and those of all other dishes in the dataset. The top-K most similar dishes were selected under three different conditions:

- Same Fine-grained Category: Select top-K dishes from the same fine-grained category as the target dish.
- Same Coarse-grained Category: Select top-K dishes from the same coarse-grained category but potentially different fine-grained categories.

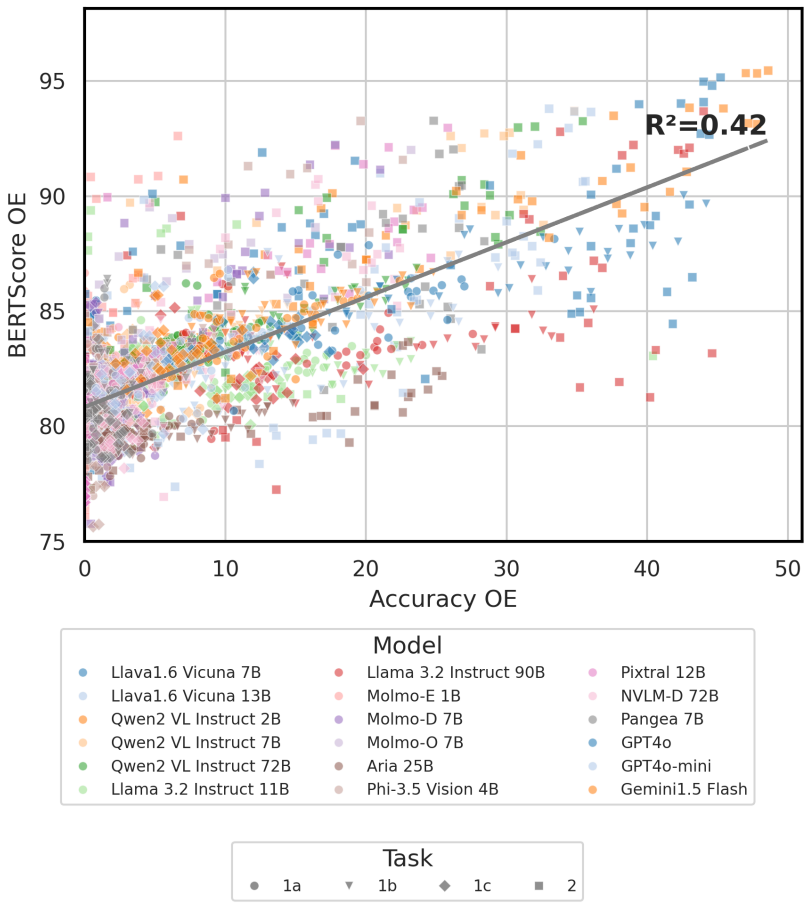


Figure 9: Regression Analysis for BERTScore OE vs. Accuracy OE.

- No Category Restriction: Select top-K dishes from the entire dataset without any restriction on categories.

Here, we used K=15, resulting in 45 candidate dishes in total.

(3) Each MCQ consists of five options: one correct answer and four negative answers. The negative answers were chosen as follows:

- Two Difficult Options: The first two negative answers were selected from dishes in the same fine-grained category. These are intended to be more challenging for the model to distinguish.
- One Medium Option: The third negative answer was selected from dishes in the same coarse-grained category.
- One Easy Option: The fourth negative answer was selected from dishes without any category restriction, making it likely to be easier to identify as incorrect.

This approach ensures a balanced difficulty among the negative options, with two difficult, one medium, and one easy negative answer.

(4) *(Optional)* Specifically for task 2, where the question involves identifying the correct location (country) of a dish, we followed a slightly modified approach:

- From the previously retrieved 4 negative options, we identified the countries associated with each dish.
- We then excluded the countries that are valid locations for the correct dish. The

remaining countries were used to create the negative options for the location-based question.

D More Results

D.1 Primary Metric: Accuracy (%)

Table 7 presents the comprehensive results of WC-VQA for both Test Small and Test Large. Additionally, we examine the performance gap between different references used in the evaluation, with the results displayed in Figure 8.

D.2 Secondary Metric: BERTScore

As a secondary metric, we employ BERTScore using XLM-R Large as the base model. Table 8 presents the comprehensive results of WC-VQA for both Test Small and Test Large. Figure 7 illustrates the model's performance categorized by language, language vitality, and language family.

Robustness and Error Analysis. Figure 9 illustrates the correlation between BERTScore and accuracy in the open-ended setting through regression analysis. The R-squared value is 0.41, indicating a low correlation between BERTScore and accuracy. Despite this, BERTScore remains a useful metric for assessing whether the model's predictions have semantic similarity to the gold labels, even if they are not exact matches.

E Evaluation

E.1 Prompt Sensitivity

We use the same prompts for all models, with the exception of the Pangea 7B model (Yue et al., 2024). This model is particularly sensitive and lacks robustness in handling diverse prompt instructions, often struggling to follow instructions accurately, especially in multiple-choice questions (MCQs), unless a specific template is applied. In contrast, models like Llama 3.2 Instruct and Qwen2 VL Instruct are more adaptable to varied instructions. After consulting with the authors, we adopted the prompt "Answer with the option letter from the given choices directly." for MCQ queries when using the Pangea 7B model.

Model (Accuracy %)	Task 1 (Dish Name) (a) no-context		(b) contextualized		(c) adversarial		Task 2 (Location)		Average	
	MCQ	OEQ	MCQ	OEQ	MCQ	OEQ	MCQ	OEQ	MCQ	OEQ
Test Small (12k)										
Open-Source										
Llava1.6 Vicuna 7B	33.63	0.87	43.13	2.83	28.67	0.60	27.77	7.93	33.30	3.06
Llava1.6 Vicuna 13B	40.87	1.00	50.30	4.17	38.37	1.60	31.07	8.63	40.15	3.85
Qwen2 VL Instruct 2B	40.97	3.33	44.40	4.60	47.07	3.43	48.37	12.50	45.20	5.96
Qwen2 VL Instruct 7B	63.83	4.07	67.20	8.57	57.00	3.90	56.80	21.23	61.21	9.44
Qwen2 VL Instruct 72B	76.13	10.40	81.63	17.43	67.23	6.27	56.73	26.07	70.43	15.04
Llama 3.2 Instruct 11B	57.93	14.37	65.57	19.20	56.27	9.50	46.60	27.23	56.59	17.58
Llama 3.2 Instruct 90B	77.33	14.27	83.43	22.30	71.23	9.00	64.70	29.73	74.17	18.82
Molmo-E 1B	21.87	0.00	24.53	0.13	20.23	0.00	19.60	1.27	21.56	0.35
Molmo-D 7B	50.67	1.00	57.00	2.23	48.67	1.73	36.73	11.70	48.27	4.16
Molmo-O 7B	46.03	2.13	43.27	4.37	41.60	2.10	26.83	9.03	39.43	4.41
Pangea 7B	45.33	0.43	59.40	1.33	22.17	0.63	34.10	17.90	40.25	5.07
Pangea 7B‡	54.87	0.43	65.77	1.33	55.00	0.63	48.47	17.90	56.03	5.07
Aria 25B	65.77	2.67	71.43	6.47	57.13	1.80	39.60	15.70	58.48	6.66
Phi-3.5 Vision 4B	49.27	1.90	53.03	3.03	42.90	1.33	31.23	8.43	44.11	3.67
Pixtral 12B	57.57	0.60	72.33	1.83	55.40	0.57	44.73	12.83	57.51	3.96
NVLM-D 72B	75.50	3.13	78.20	7.37	54.67	1.37	54.13	17.40	65.62	7.32
Proprietary										
GPT-4o	**88.40**	**16.60**	**90.43**	**35.47**	**82.23**	**12.60**	63.60	**35.53**	**81.17**	**25.05**
GPT-4o Mini	75.33	7.30	83.00	17.67	64.83	3.53	52.87	26.90	69.01	13.85
Gemini 1.5 Flash	78.17	16.30	82.07	23.53	71.33	7.33	**66.00**	32.30	74.39	19.86
Test Large (60k)										
Open-Source										
Llava1.6 Vicuna 7B	34.57	1.59	43.48	4.03	34.84	1.41	32.24	9.29	36.28	4.08
Llava1.6 Vicuna 13B	40.17	2.79	48.17	5.85	39.05	2.57	37.79	10.16	41.30	5.34
Qwen2 VL Instruct 2B	41.65	7.98	42.29	8.13	39.69	6.74	47.85	14.55	42.87	9.35
Qwen2 VL Instruct 7B	61.48	6.76	67.85	10.36	53.52	6.12	55.90	21.03	59.69	11.07
Qwen2 VL Instruct 72B	74.19	12.67	80.79	21.31	62.43	8.37	61.90	27.27	69.83	17.40
Llama 3.2 Instruct 11B	59.93	18.75	64.12	22.96	53.17	13.39	57.93	31.58	58.79	21.67
Llama 3.2 Instruct 90B	77.69	16.93	82.92	23.60	63.96	10.87	67.87	31.31	73.11	20.68
Molmo-E 1B	18.81	0.01	24.22	0.23	19.55	0.01	18.97	1.54	20.39	0.45
Molmo-D 7B	46.01	2.89	55.95	3.66	41.61	2.31	33.35	11.45	44.23	5.08
Molmo-O 7B	39.96	5.15	44.93	6.03	38.41	3.51	29.81	10.07	38.28	6.19
Pangea 7B	41.38	1.52	57.95	2.73	21.77	1.57	37.15	20.15	39.56	6.49
Pangea 7B‡	52.35	1.52	63.07	2.73	49.17	1.57	48.71	20.15	53.33	6.49
Aria 25B	58.61	4.99	69.29	9.17	52.82	3.39	42.82	16.20	55.89	8.44
Phi-3.5 Vision 4B	43.37	2.91	48.71	4.23	40.87	2.07	35.01	9.22	41.99	4.61
Pixtral 12B	56.65	1.22	70.69	2.94	52.12	1.09	46.67	14.43	56.53	4.92
NVLM-D 72B	69.82	4.71	78.93	10.29	52.12	2.89	51.97	16.68	63.21	8.64
Proprietary										
GPT-4o	**88.45**	**21.88**	**91.57**	**37.51**	**82.29**	**14.79**	66.52	**37.13**	**82.21**	**27.83**
GPT-4o Mini	72.80	10.28	81.65	20.87	57.76	5.72	52.37	25.79	66.14	15.66
Gemini 1.5 Flash	77.05	12.81	80.97	15.16	69.13	6.46	**71.53**	30.03	74.67	16.12

Table 7: Accuracy (%) results of WC-VQA. MCQ and OEQ indicate multiple-choice question and open-ended question, respectively. Best and second-best are **bolded** and underlined, respectively. ‡We employ an optimized prompt provided by the authors (see Subsection E.1 in the Appendix for further details).

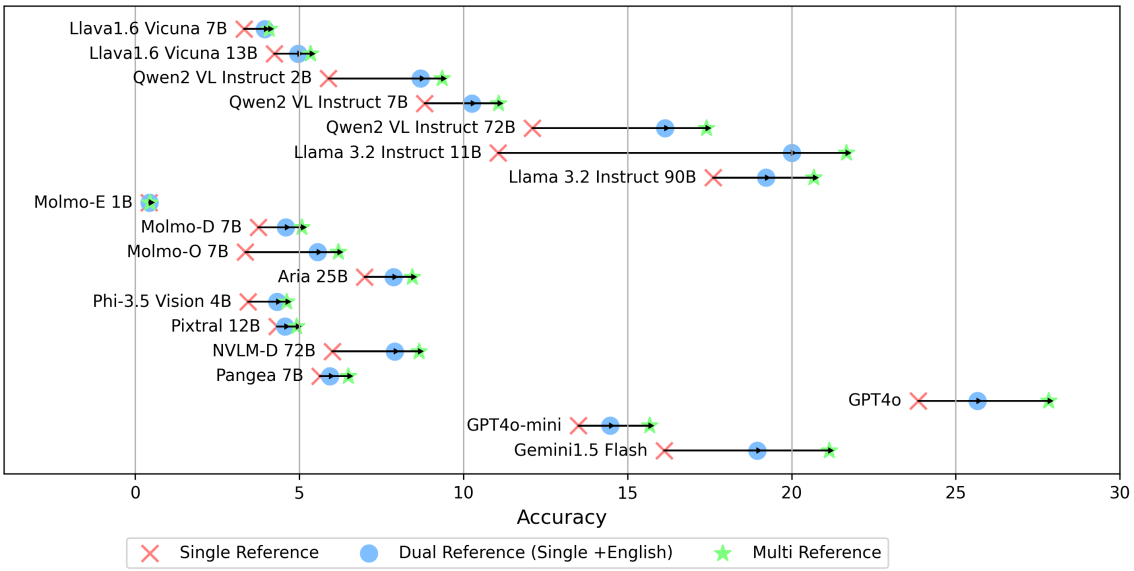


Figure 10: Model performance with different references on open-ended question.

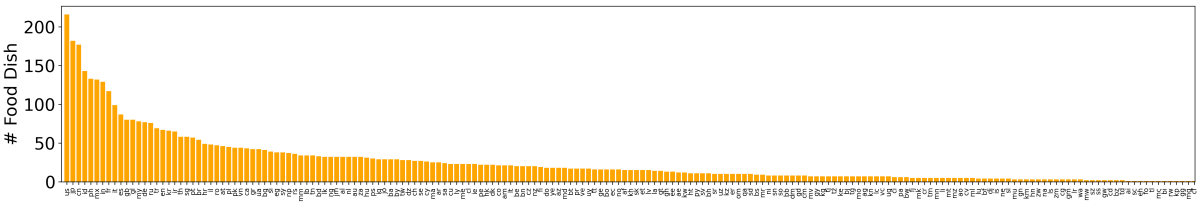


Figure 11: Dish frequency by country showing 189 countries.

Model (BERTScore)	Task 1 (Dish Name) (a) no-context	(b) contextualized	(c) adversarial	Task 2 (Location)	Average
Test Small (12k)					
Open-Source					
Llava1.6 Vicuna 7B	81.49	82.13	81.56	85.45	82.66
Llava1.6 Vicuna 13B	80.50	80.65	80.14	81.77	80.77
Qwen2 VL Instruct 2B	82.48	82.75	82.34	84.29	82.97
Qwen2 VL Instruct 7B	82.65	83.13	82.10	87.22	83.78
Qwen2 VL Instruct 72B	83.78	84.63	83.06	87.10	84.64
Llama 3.2 Instruct 11B	82.45	82.93	81.64	82.59	82.40
Llama 3.2 Instruct 90B	82.82	83.44	81.98	85.70	83.48
Molmo-E 1B	81.17	81.12	81.24	83.58	81.78
Molmo-D 7B	81.26	81.65	80.55	84.87	82.08
Molmo-O 7B	82.14	82.24	81.44	84.38	82.55
Pangea 7B	81.29	81.78	80.19	86.31	82.39
Aria 25B	79.85	80.26	79.86	80.53	80.12
Phi-3.5 Vision 4B	80.82	79.66	76.77	83.25	80.12
Pixtral 12B	78.84	79.12	78.90	86.40	80.81
NVLM-D 72B	81.39	82.05	79.98	85.64	82.27
Proprietary					
GPT-4o	**84.86**	**86.92**	**83.89**	<u>88.98</u>	**86.16**
GPT-4o Mini	83.10	83.91	82.16	87.34	84.13
Gemini 1.5 Flash	<u>84.68</u>	<u>85.09</u>	<u>83.11</u>	**89.15**	<u>85.51</u>
Test Large (60k)					
Open-Source					
Llava1.6 Vicuna 7B	81.63	82.10	81.58	85.81	82.78
Llava1.6 Vicuna 13B	80.65	80.70	80.12	81.86	80.83
Qwen2 VL Instruct 2B	82.95	83.10	82.81	84.51	83.34
Qwen2 VL Instruct 7B	82.92	83.42	82.30	87.39	84.01
Qwen2 VL Instruct 72B	83.72	<u>85.10</u>	83.11	87.42	84.84
Llama 3.2 Instruct 11B	82.54	82.79	81.64	82.88	82.46
Llama 3.2 Instruct 90B	83.05	83.51	81.95	85.85	83.59
Molmo-E 1B	81.17	81.10	81.13	83.87	81.82
Molmo-D 7B	81.39	81.63	80.73	85.10	82.21
Molmo-O 7B	82.27	82.21	81.52	84.63	82.66
Pangea 7B	81.40	81.91	80.23	86.79	82.58
Aria 25B	79.89	80.20	79.83	80.63	80.14
Phi-3.5 Vision 4B	80.98	79.55	77.61	83.31	80.36
Pixtral 12B	79.00	79.33	78.98	86.75	81.02
NVLM-D 72B	81.54	82.17	80.05	85.67	82.36
Proprietary					
GPT-4o	**85.04**	**86.93**	**83.92**	<u>89.06</u>	**86.24**
GPT-4o Mini	83.19	84.05	82.38	87.30	84.23
Gemini 1.5 Flash	<u>84.47</u>	84.97	<u>83.14</u>	**89.43**	<u>85.50</u>

Table 8: BERTScore results of WC-VQA. Only the results from open-ended (OEQ) are used. Best and second-best are **bolded** and <u>underlined</u>, respectively.

Continents/Regions	# Countries	# Food Entries	% in Our Data
Global*	N/A	**96**	**3.98%**
Africa	**52**	**190**	**7.87%**
Eastern Africa	18	40	1.7%
Middle Africa	6	17	0.7%
Northern Africa	7	67	2.8%
Southern Africa	5	33	1.4%
Western Africa	16	60	2.5%
America	**37**	**472**	**19.55%**
Caribbean	15	60	2.5%
Central America	8	134	5.6%
Northern America	2	230	9.5%
South America	12	109	4.5%
Europe	**47**	**808**	**33.47%**
Eastern Europe	10	164	6.8%
Northern Europe	15	237	9.8%
Southern Europe	13	300	12.4%
Western Europe	9	233	9.7%
Asia	**53**	**1,052**	**43.58%**
Central Asia	5	10	0.4%
Eastern Asia	9	420	17.4%
South Eastern Asia	12	362	15.0%
Southern Asia	9	200	8.3%
Western Asia	18	155	6.4%
Oceania	**3**	**37**	**1.53%**
Australia & New Zealand	2	33	1.4%
Melanesia	1	4	0.2%
Micronesia	-	-	-
Polynesia	-	-	-

Table 9: Geographical distribution of WC-KB, corresponds to Figure 3. Note that there are food entries linked to multiple regions, with some linked to multiple continents. *__Global__ denotes entries with more than five regions.

Country	Count	%	Country	Count	%	Country	Count	%
United States	216	9.47	Argentina	25	1.10	Grenada	8	0.35
Japan	182	7.98	Saudi Arabia	24	1.05	Cameroon	8	0.35
China	177	7.76	North Macedonia	24	1.05	Somalia	8	0.35
Indonesia	143	6.27	Cuba	23	1.01	Antigua and Barbuda	7	0.31
Philippines	133	5.83	Libya	23	1.01	Maldives	7	0.31
Mexico	132	5.78	Montenegro	23	1.01	Kyrgyzstan	7	0.31
India	129	5.65	Chile	23	1.01	Tajikistan	7	0.31
France	117	5.12	Ireland	23	1.01	Togo	7	0.31
Italy	99	4.34	Peru	22	0.96	Uganda	7	0.31
Spain	87	3.81	Hong Kong	22	0.96	Benin	7	0.31
United Kingdom	80	3.51	Denmark	22	0.96	Macau	7	0.31
Global	80	3.51	Colombia	21	0.92	Guyana	7	0.31
Germany	77	3.37	Armenia	21	0.92	Saint Kitts and Nevis	7	0.31
Russia	76	3.33	Lithuania	21	0.92	Saint Lucia	7	0.31
Turkey	69	3.02	Belgium	20	0.88	Saint Vincent and the Grenadines	7	0.31
Korea	66	2.89	Brunei Darussalam	20	0.88	Fiji	5	0.22
Iran	65	2.85	Czech Republic	20	0.88	Mongolia	5	0.22
Thailand	58	2.54	New Zealand	20	0.88	Liechtenstein	5	0.22
Singapore	58	2.54	Finland	19	0.83	Macedonia	5	0.22
Portugal	57	2.50	Dominican Republic	18	0.79	Malta	5	0.22
Brazil	54	2.37	Yemen	18	0.79	Mozambique	5	0.22
Israel	48	2.10	Azerbaijan	18	0.79	Angola	5	0.22
Romania	47	2.06	Moldova	18	0.79	Cabo Verde	5	0.22
Austria	46	2.02	Bhutan	17	0.75	Turkmenistan	5	0.22
Poland	45	1.97	Puerto Rico	17	0.75	Costa Rica	5	0.22
Pakistan	44	1.93	Venezuela	17	0.75	Burkina Faso	4	0.18
Vietnam	44	1.93	Uruguay	17	0.75	Luxembourg	4	0.18
Canada	43	1.89	Bolivia	16	0.70	Djibouti	4	0.18
Greece	42	1.84	Trinidad and Tobago	16	0.70	Iceland	4	0.18
Ukraine	42	1.84	Georgia	16	0.70	Sierra Leone	4	0.18
Bulgaria	41	1.80	Norway	16	0.70	Niger	4	0.18
Slovenia	39	1.71	Cambodia	15	0.66	Mauritius	3	0.13
Egypt	38	1.67	Afghanistan	15	0.66	Guinea	3	0.13
Syria	38	1.67	Slovakia	15	0.66	Zimbabwe	3	0.13
Nepal	37	1.62	Ethiopia	15	0.66	Namibia	3	0.13
Serbia	36	1.58	Latvia	15	0.66	Lesotho	3	0.13
Myanmar	34	1.49	Laos	14	0.61	Zambia	3	0.13
Lebanon	34	1.49	Guatemala	14	0.61	Congo	3	0.13
Tunisia	34	1.49	Ghana	13	0.57	Gambia	3	0.13
Bangladesh	33	1.45	United Arab Emirates	12	0.53	Liberia	3	0.13
Malaysia	32	1.40	Kuwait	12	0.53	Comoros	3	0.13
Sri Lanka	32	1.40	Paraguay	11	0.48	South Korea	3	0.13
Nigeria	32	1.40	El Salvador	11	0.48	Wales	3	0.13
Jamaica	32	1.40	Bahrain	11	0.48	Honduras	3	0.13
Netherlands	32	1.40	Haiti	11	0.48	Anguilla	1	0.04
Albania	32	1.40	Uzbekistan	10	0.44	Western Sahara	1	0.04
South Africa	32	1.40	Kazakhstan	10	0.44	Faroe Islands	1	0.04
Australia	32	1.40	Eritrea	10	0.44	Seychelles	1	0.04
Hungary	31	1.36	Oman	10	0.44	Burundi	1	0.04
Palestine	30	1.32	Qatar	10	0.44	Rwanda	1	0.04
Iraq	29	1.27	Sudan	10	0.44	North Korea	1	0.04
Jordan	29	1.27	Suriname	10	0.44	Timor-Leste	1	0.04
Bosnia and Herzegovina	29	1.27	Mauritania	9	0.39	Guernsey	1	0.04
Taiwan	28	1.23	Bahamas	9	0.39	Madagascar	1	0.04
Algeria	28	1.23	Nicaragua	8	0.35	Central African Republic	1	0.04
Switzerland	27	1.18	Senegal	8	0.35	Monaco	1	0.04
Cyprus	26	1.14	Barbados	8	0.35			
Morocco	25	1.10	Dominica	8	0.35			

Table 10: Distribution of food entries by country.