CulturalMenuBench: Probing the Knowledge-Application Gap in Multimodal Culinary Reasoning

Bo Zeng[1], Linfeng Gao[1,2], Peiqin Lin[1], Yu Zhao[1]
Mingyan Zeng, Yu Tong, Xintong Wang[1,3*], Linlong Xu[1]
Longyue Wang[1], Weihua Luo[1], Qinggang Zhang[2], Jinsong Su[2]
[1]Alibaba Group [2]Xiamen University [3]University of Hamburg

Abstract

Multimodal language models achieve near-ceiling scores on food recognition benchmarks, yet it remains unclear whether this success reflects genuine cultural understanding or mere visual matching. To probe this distinction, we introduce CulturalMenuBench, a benchmark of 4,870 items in 10 languages across 18 regions; its 10 tasks pair final-dish and step-by-step cooking images with ingredients, procedural text, and regional labels, spanning basic recognition to process-grounded cultural attribution. Evaluating 12 models exposes a substantial knowledge-application gap: models exceeding 94% on standard multiple-choice tasks drop to at most 56% when attributing dishes to Chinese regional cuisines, despite an identical four-way format. Diagnostic analyses explain why: error patterns are consistent with random guessing, accuracy tracks visual distinctiveness rather than cultural structure, and models classify cuisines more accurately from dish names alone than from images (+7–18 points). The knowledge is thus present but cannot be activated through visual input. An ablation confirms these tasks genuinely require procedural evidence: removing sequential cooking images selectively degrades process-grounded tasks while others remain stable. Overall, CulturalMenuBench shows that near-perfect recognition can conceal an inability to apply cultural knowledge, motivating training that explicitly connects perception, procedure, and cultural context. Code and data are publicly available.[1]

1 Introduction

Do Large Language Models (LLMs) truly *understand* culinary culture, or do they merely *discriminate* surface patterns when prompted with explicit candidates? As LLMs are deployed in dietary recommendation, culinary tourism, and cross-cultural education, this distinction between knowledge *possession* and *application* becomes critical (Cao et al., 2024b; Fung et al., 2024; Li et al., 2024a). Genuine culinary understanding demands not only recognizing a dish from its image, but also inferring ingredient–technique associations from multi-step cooking processes and situating these observations within specific cultural and regional contexts (Liu et al., 2021; Cao et al., 2024a; Bugliarello et al., 2023; Qi et al., 2026; Bai et al., 2026).

Cuisine offers a fertile testbed for probing cultural competence, yet existing benchmarks (Romero et al., 2024; Vayani et al., 2025; Ma et al., 2023) remain circumscribed in both reasoning depth and cultural coverage (Tab. 1). FoodieQA (Li et al., 2024b) and WC-VQA (Winata et al., 2025) target direct visual recognition and question answering, omitting the cooking processes in which much of culinary understanding is encoded. IndiFoodVQA (Agarwal et al., 2024) advances toward compositional evaluation but is confined to English and does not capture intra-country regional distinctions. More critically, these shallow tasks engender a capability illusion: models readily surpass 90% on candidate matching, masking their true capacity for independent reasoning.

To address this, we present CulturalMenuBench, assembled through a scalable pipeline from web-mined menus. The benchmark offers three key advantages: 1) **rich process-level metadata**: beyond dish names and final images, it encompasses ingredients, step-by-step instructions, process images, and sub-regional tags; 2) **reasoning-intensive tasks**: 10 tasks spanning visual recognition, cross-modal matching, process-grounded reasoning, and cultural classification; and 3) **multi-granularity cultural coverage**: 18 regions across 10 countries in 10 languages, with provincial-level labels for Chinese cuisines enabling both cross-cultural and intra-cultural evaluation.

Systematic evaluation of 12 models uncovers a

*Correspondence to xintong.wang@uni-hamburg.de.
[1]https://github.com/BobTsang-NLP/CulturalMenuBench

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 40003–40020
October 24-29, 2026 ©2026 Association for Computational Linguistics

Benchmark	Reasoning	#T	#L	Food Info	Sub-region
FoodieQA	One-hop	3	1	dish name location flavor dish image	Yes
IndiFoodVQA	Multi-hop	1	1	dish name location raw material flavor dish image	No
WC-VQA	One-hop	2	30	dish name location dish image	No
Ours	Process-grounded	10	10	dish name location raw material practice dish image practice images	Yes

Table 1: **Statistics of representative culinary benchmarks.** #T: number of tasks, #L: number of languages, Food Info: food information, Sub-region: whether includes sub-region information.

knowledge-application gap: leading models attain 94–97% on standard four-way MC tasks, yet at most 56% when classifying sub-regional Chinese cuisines under the same four-way response interface. With the format held constant, this 40–50 point decline reflects a genuine failure of knowledge application rather than a response-format effect. Diagnostic error analysis exposes the mechanism: confusion patterns are consistent with random guessing, and per-cuisine accuracy strongly tracks visually distinctive cuisines (Sichuan 78% vs. Anhui 19%), indicating reliance on visual shortcuts rather than cultural reasoning.

Ablation corroborates that this finding is not an artifact of benchmark design: retaining only the first image per step group, thereby removing the remaining sequential context, selectively degrades process tasks by 5.7–11.7pp while non-process tasks remain stable ($\leq$1.0pp). This double dissociation establishes that the benchmark elicits authentic procedural reasoning, rendering the cultural reasoning failure all the more striking.

These findings expose a structural deficiency: the reasoning bridge from perception to cultural knowledge is not reliably activated, and this failure mode is consistent across architectures and scales. Our contributions are threefold:

- We introduce CulturalMenuBench, a multimodal benchmark pairing process-level culinary metadata (step-by-step images, ingredients, procedural text) with sub-regional labels across 10 languages and 18 regions.
- We devise a scalable pipeline yielding 4,870 items from web-mined menus, instantiating 10 tasks from visual recognition to process-grounded reasoning.
- We uncover a knowledge-application gap, including an approximately 40–50 point decline for leading matched systems under the same four-way response interface, and diagnose its mechanism: models rely on visual shortcuts rather than cultural reasoning, a failure consistent across 11 model/configuration conditions spanning 10 unique checkpoints.

2 CulturalMenuBench

To evaluate the process-grounded reasoning introduced above, CulturalMenuBench furnishes each of its 587 curated dishes with step-by-step cooking images, ingredient lists, and procedural text alongside the final dish image, which are necessary to trace technique sequences and intermediate states that distinguish regional traditions. The benchmark comprises 4,870 evaluation items spanning 10 languages and 18 regions. Construction proceeds in two stages (Fig. 1): menu collection (§2.1) and task construction (§2.2).

2.1 Menu Collection Pipeline

Pipeline Stage	Volume	Retention
1. Raw Data Crawling	239,686	100.0%
2. URL-based Dedup.	124,559	52.0%
3. Semantic-level Dedup.	93,492	75.1%
4. Multimodal Integrity Check	88,723	94.9%
5. **Final Curation**	**587**	**0.7%**

Table 2: **Data retention statistics across each stage.** The entire pipeline guarantees the completeness and consistency of core metadata (Dish Image + Practice Image + Textual Description).

To curate menus of requisite quality and diversity, we devise a dedicated collection pipeline. We source structured recipe data from MeishiChina[2], a large-scale Chinese culinary platform hosting user-contributed recipes spanning both domestic Chinese and internationally adopted cuisines (e.g., Japanese, Italian, Thai, Korean). This single-platform sourcing strategy is a deliberate design choice conferring two key advantages: (1) **structural homogeneity**: all recipes share a consistent schema of dish images, step-by-step instructions with photos, and ingredient lists, enabling uniform task construction across cuisines; and (2) **cross-cultural adaptation as an evaluation target**: the

[2] https://www.meishichina.com/

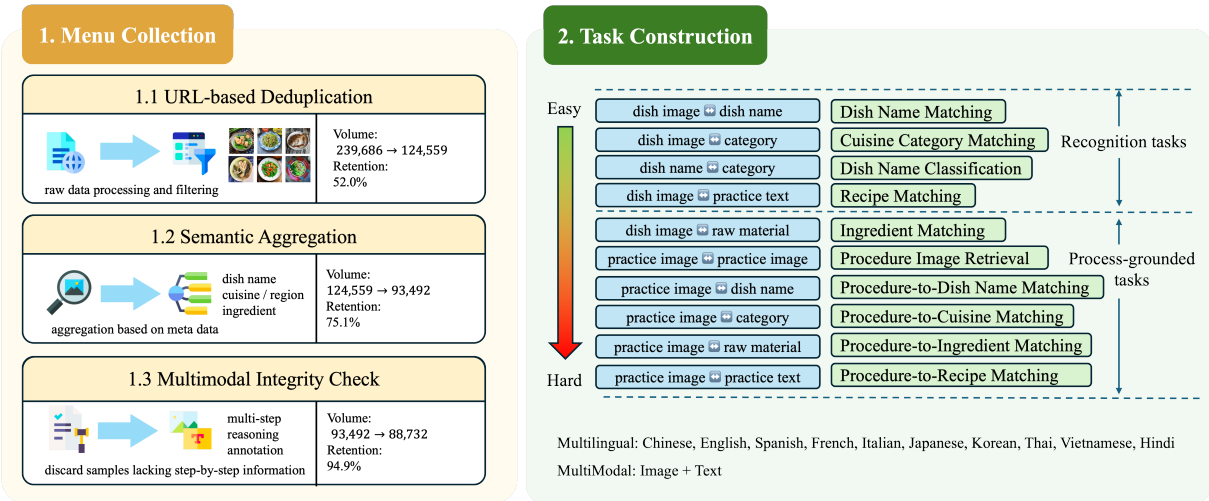


Figure 1: **Overview of CulturalMenuBench construction.** The pipeline comprises three-stage filtering (URL deduplication, semantic aggregation, integrity checks) from 240K raw records to 587 curated samples, followed by task construction spanning visual recognition to process-grounded cultural reasoning.

platform's international recipes reflect how dishes are adapted and reinterpreted across cultural boundaries, a phenomenon central to real-world culinary knowledge that models must handle. This unified-source design enables controlled comparison in which the primary variable is model capability, rather than confounding differences in data schema, annotation standards, or image quality across heterogeneous platforms. From the initial 239,686 records, we apply a multi-stage filtering regimen summarized in Tab. 2:

URL-based Deduplication We first eliminate redundant entries by deduplicating on the unique URL associated with each recipe. This stage discards mirrored or near-identical submissions, reducing the corpus to 124,559 distinct records.

Semantic Aggregation To address semantic redundancy, where different users submit recipes for the same dish under variant names or descriptions, we encode dish names using a multilingual sentence embedding model[3] and perform greedy deduplication: items are ranked by a composite quality score (number of practice images + text length), and each candidate is compared against all previously retained items via cosine similarity. A candidate is discarded if its similarity to any retained item exceeds a threshold of 0.9, unless a core-ingredient conflict is detected (e.g., "braised pork" vs. "braised fish"). This procedure yields 93,492 semantically unique dish entities.

Multimodal Integrity Check To support multimodal reasoning tasks that require aligned visual and textual evidence across intermediate reasoning steps, we enforce stringent completeness criteria: every retained sample must include both step-by-step instructional text and corresponding images. We implement this through a two-phase verification process. First, an automated filter discards records with missing or malformed metadata fields (as defined in Tab. 3). Second, human annotators manually inspect the remaining candidates to confirm the presence, alignment, and consistency of multimodal content. This procedure yields 88,723 fully multimodal candidates. We then perform a final curation step to ensure balanced sample distribution and strong category representativeness, with particular attention to the sub-regional diversity of Chinese cuisine, as shown in Tab. 11. For each category, we retain examples that satisfy three operationalized criteria: (1) **cultural typicality**: the dish is widely recognized as representative of its labeled cuisine (verified by the first author, who has domain expertise in Chinese culinary culture, and cross-checked against cuisine taxonomy references); (2) **multimodal consistency**: practice-step images visually correspond to the textual instructions without missing or misaligned steps; and (3) **instructional quality**: the recipe contains at least 5 practice steps with clear, non-repetitive photographs. The final procedure yields a curated benchmark of 587 high-quality samples.[4] The se-

[3]https://huggingface.co/machimachiop/paraphrase-multilingual-MiniLM-L2-v2

[4]An additional quality audit removed 25 items with incorrect cultural attribution (e.g., "Mexican bread", a Chinese

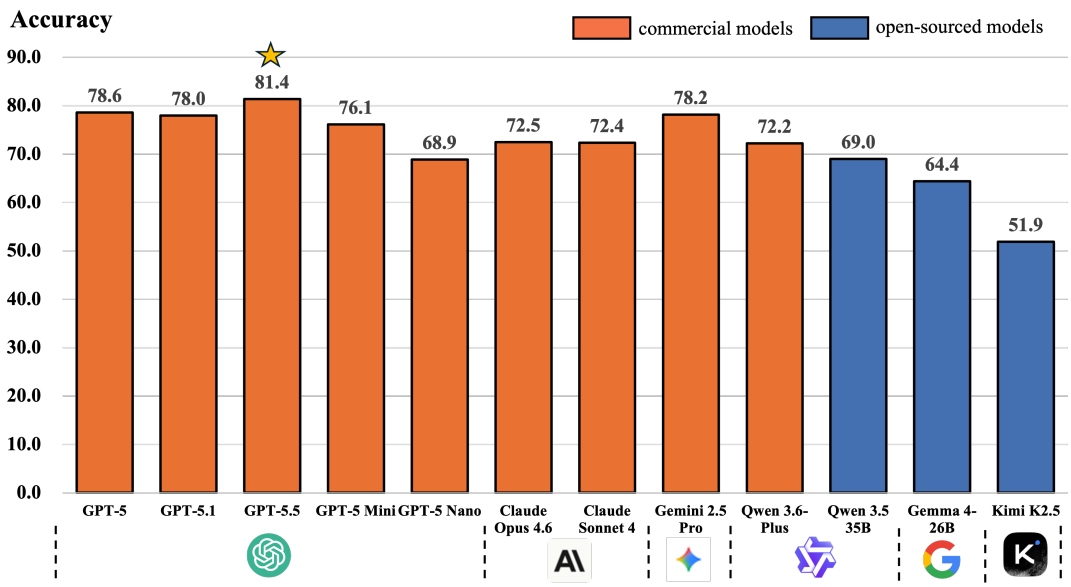


Figure 2: **Layer 2 (binary) performance of 12 LLMs.** The gap from near-ceiling MC scores (Tab. 5) quantifies the knowledge-application disconnect.

lection is filtered coverage under fixed, documented criteria rather than an unconstrained hand-picked set. Per-cuisine counts are not forced to be uniform, and thin cuisine classes limit the resolution of per-cuisine estimates.

Dataset Composition The 587 curated dishes span 200 Chinese and 387 non-Chinese entries across nine additional countries (Fig. 4, Appendix), totaling **7,814 images** (avg. 12.7 per sample). Chinese dishes constitute 34.1% of the benchmark and bear fine-grained provincial labels (Tab. 11), whereas non-Chinese dishes are labeled by country of origin. This asymmetry reflects two deliberate choices: (1) Chinese regional cuisines exhibit substantial inter-provincial diversity warranting fine-grained sub-regional study; and (2) the source platform naturally provides richer Chinese coverage. We report per-language performance throughout (§3) and confirm that Chinese accuracy is comparable to the overall mean, indicating no systematic inflation. The complete metadata schema is summarized in Tab. 3.

Cultural Localization For the non-Chinese portion, we restore native dish names and translate procedural instructions according to each dish's country label. Translation quality was assessed before final curation on 1,236 translated fields from 412 candidate dishes (three fields per dish). Three architecturally diverse models (Gemini-3.0-Pro (Google DeepMind, 2026), GPT-5-Mini (OpenAI, 2025a), and DeepSeek-V3 (DeepSeek-AI, 2024)) independently evaluated each field, with a majority vote (≥2/3) required for approval. This procedure attained a macro-average pass rate of **96.20%** (Tab. 10); translations that did not pass were manually corrected before release curation. Native speakers additionally audited Hindi, British English, and Vietnamese items against the Chinese sources; details are provided in Appendix I. Detailed error analysis is provided in Appendix I.

Overall, the pipeline's 0.7% retention rate (587 from 240K raw records) reflects intentionally aggressive filtering to prioritize multimodal completeness and cultural authenticity over scale.

Field	Modality	Description
`dish_name`	Text	Official name of the dish as it appears on the menu.
`category`	Text	Country of origin with subregion information when available.
`raw_material`	Text	Comprehensive list of primary ingredients of the dish.
`practice_text`	Text	Ordered, step-by-step instructions for cooking the dish.
`dish_image`	Image	High-resolution photograph of the finished dish.
`practice_image`	Image	Sequence of photos illustrating each step of the cooking process.

Table 3: **Metadata schema of CulturalMenuBench.**

2.2 Task Construction

Leveraging the rich metadata contained in the curated samples (Tab. 3), we devise 10 distinct evaluation tasks to probe LLMs' understanding and

bakery category name erroneously classified as Mexican cuisine) and corrected labels for 11 items.

reasoning capabilities across multiple dimensions of culinary culture. These tasks are organized into two tiers:

Recognition tasks (4 tasks) serve as controlled baselines that test single-step visual or textual matching, which pair dish images or names with categories or practice text. These can typically be solved through direct visual recognition or factual lookup without integrating sequential procedural evidence.

Process-grounded tasks (6 tasks) require reasoning over multi-step cooking procedures, ingredient–process alignment, or cross-modal procedural matching, which pair practice images with dish images, names, ingredients, or textual instructions. These demand that models track state changes across ordered visual frames and compose intermediate inferences.

Not all tasks in our benchmark require process-grounded reasoning; we retain recognition tasks as controlled baselines to isolate the specific difficulty introduced by procedural understanding. Details of all 10 tasks, including descriptions and prompts, are provided in Tab. 13 of §G. As an example of the most challenging tier, `practice_image` ↔ `practice_image` requires LLMs to first identify the dishes based on the practice images, then reason about their culinary origins, and finally verify whether the dishes belong to the same category.

To facilitate automated evaluation, we provide two complementary formats. The **binary format** frames each task as a yes/no verification question (random baseline: 50%), while the **multiple-choice (MC) format** presents four candidate answers per question (random baseline: 25%), with three carefully constructed distractors sampled from the same cuisine category or task domain to ensure plausibility. Both formats employ balanced data generation: for binary tasks, we produce equal numbers of positive and negative examples; for MC tasks, distractors are drawn from semantically proximate alternatives to preclude elimination by surface cues. All prompts are formulated in native languages to faithfully capture local cultural nuances.

For Chinese dishes, we generate 100 prompts per task, whereas for non-Chinese dishes, the number of prompts per task matches the total count of dishes in their respective category. Consequently, CulturalMenuBench contains a total of 4,870 samples, spanning 10 languages across 18 regions. We note that while the 10 tasks draw from a shared pool of 587 base dishes, each task is evaluated independently: no outputs or context are shared across tasks at inference time, and accuracy is computed per task; however, the shared dish pool should still be considered in joint analyses.

3 Experiments

3.1 Setup and Human Baseline

We evaluate CulturalMenuBench on 12 state-of-the-art LLMs (Fig. 2), encompassing both open-weight and proprietary systems. Exploiting the objective nature of our gold answers, we employ fully automated scoring. To contextualize model performance, five annotators with diverse cultural backgrounds independently answer a representative 350-sample subset.

Annotator	Background	Acc. (%)
Annotator 1	East Asian cuisine	87.14
Annotator 2	European cuisine	86.86
Annotator 3	South Asian cuisine	87.43
Annotator 4	Chinese regional cuisine	86.00
Annotator 5	Global cuisine	86.86
Average	–	**86.86**
Majority Vote	–	**86.86**

Table 4: **Human performance on a 350-sample subset of CulturalMenuBench.** Pairwise inter-annotator agreement reaches 0.946. The individual-score mean (86.858%) and majority-vote accuracy (304/350; 86.857%) independently round to 86.86%.

As shown in Tab. 4, annotators attain 86.00–87.43% accuracy (avg. 86.86%) with pairwise IAA of 0.946, confirming unambiguous task definitions and reliable gold labels. The ∼87% ceiling establishes that CulturalMenuBench is demanding yet well-defined; the best model (GPT-5.5, 81.4%) still trails humans by ∼6 points.

3.2 The Knowledge-Application Gap

How well do current LLMs *know* culinary culture? Under the MC format (4,085 items, random baseline 25%), the answer is: remarkably well. As shown in Tab. 5, leading proprietary models attain 94–97% accuracy, with GPT-5 reaching 96.8%, Claude Opus 4.6 scoring 96.2%, and even mid-tier systems like GPT-5 Mini achieving 94.7%. These scores vastly exceed chance and approach ceiling, spanning all 10 task types and 10 languages. By any conventional metric, this would suggest that current LLMs have internalized substantial culinary cultural knowledge.

Yet this apparent mastery conceals a striking fragility. When the *same* dish–attribute pairings are

tested via binary verification like "Is this dish image correctly paired with [X]?", accuracy plummets by 10–24 points across the board (Tab. 5). The most dramatic case is Claude Opus 4.6, which achieves 96.2% under MC yet only 72.5% on binary: a 24-point drop despite near-perfect knowledge confirmation. Even the strongest proprietary model (GPT-5.5: 81.4%) still trails humans (86.86%) by ∼6 points, while open-weight systems lag considerably further (Qwen3.5-35B: 69.0%, Gemma4-26B: 64.4%). This pattern persists across Chinese and non-Chinese subsets (GPT-5: 79.1% vs. 78.3%; see §D).

We term this **knowledge-application gap**: models that demonstrably *possess* cultural knowledge confirmed by near-perfect MC performance appear unable to *apply* it for independent judgment without candidate contrasts.

Is This Merely a Format Effect? A legitimate concern arises: MC and binary verification are structurally distinct tasks (4-way selection vs. yes/no judgment; baselines 25% vs. 50%), so some performance disparity is expected regardless of knowledge. Several observations motivate looking beyond a fixed format penalty. First, the gap varies 2.4× across models (from 9.8 for Qwen3.5-35B to 23.7 for Claude Opus 4.6). Second, some stronger binary performers show smaller gaps (e.g., GPT-5.5: 14.0 vs. Gemma4-26B: 24.0), although this pattern is not uniform. Third, within the GPT family, models with nearly matched MC scores (GPT-5: 96.8%, GPT-5.1: 96.6%) also show closely matched binary scores (78.6% vs. 78.0%). These cross-format patterns are suggestive rather than decisive.

However, the cross-format comparison alone cannot fully disentangle format difficulty from a genuine knowledge-application disconnect. A definitive test requires holding format constant while varying only the *type* of knowledge being probed. So we construct precisely such a test in §3.4, after first validating that the benchmark genuinely demands process reasoning (§3.3).

3.3 Validating Process-Grounded Reasoning

Before probing deeper into the knowledge-application gap, we must first establish that our benchmark measures what it claims. If models could solve process-grounded tasks through shallow pattern matching, the gap might simply reflect task difficulty rather than a reasoning deficit. Two analyses corroborate that this is not the case.

Model	Binary (%)	MC (%)	Gap
GPT-5.5	81.4	95.4	14.0
GPT-5.1	78.0	96.6	18.6
GPT-5	78.6	96.8	18.2
GPT-5 Nano	68.9	82.2	13.3
GPT-5 Mini	76.1	94.7	18.6
Claude Opus 4.6	72.5	96.2	23.7
Claude Sonnet 4	72.4	87.1	14.7
Gemini 2.5 Pro	78.2	94.6	16.4
Qwen3.6-Plus	72.2	95.1	22.9
Qwen3.5-35B†	69.0	78.8	9.8
Kimi-K2.5†	70.7	94.7	24.0
Gemma4-26B†	64.4	88.4	24.0
Random	50.0	25.0	–

Table 5: **Knowledge probe (MC) vs. application probe (Binary).** The gap quantifies the knowledge-application disconnect. †Open-source models.

Cross-Task Performance. The performance heatmap (Fig. 3, Appendix) reveals a pronounced hierarchy aligned with our two-tier taxonomy. Recognition tasks yield robust performance (e.g., "Dish Img ↔ Practice Text": GPT-5.1 93.0%), while more demanding relational and process-grounded tasks deteriorate markedly ("Dish Img ↔ Dish Img": 57.8%, "Practice Img ↔ Raw Material": 58.6%). This 35-point chasm between the easiest and hardest tasks shows that higher-order relational and process reasoning constitute a distinct and more demanding capability dimension.

Ablation: Removing Sequential Context. To verify that sequential visual context drives performance, we conduct an ablation on 500 stratified samples across three models (Qwen3.6-Plus, Claude 4.6, Kimi K2.5). We compare two conditions: (1) **Full**: all practice-step images; (2) **Single**: only the first image per step group, removing multi-step process information.

As shown in Tab. 6, all three models exhibit consistent accuracy drops when sequential visual context is reduced: overall accuracy decreases by 3.4–7.4 percentage points. Crucially, this effect is *selective*: practice-image tasks drop by 5.7–11.7pp, while non-practice tasks remain virtually unchanged (≤1.0pp). This double dissociation—sequential context matters precisely where process reasoning is required and nowhere else—confirms that our benchmark genuinely demands multi-step visual reasoning rather than shallow pattern matching.

The benchmark passes both tests: it measures what it claims, and it does so through genuine multi-

Model	Full (%)	Single (%)	Δ
Overall			
Claude 4.6	**96.4**	93.0	−3.4
Qwen3.6-Plus	95.4	88.0	−7.4
Kimi K2.5	92.4	88.4	−4.0
Practice-image tasks only			
Claude 4.6	97.3	91.7	−5.7
Qwen3.6-Plus	96.7	85.0	−11.7
Kimi K2.5	92.3	86.3	−6.0
Non-practice tasks only			
Claude 4.6	95.0	95.0	±0.0
Qwen3.6-Plus	93.5	92.5	−1.0
Kimi K2.5	92.5	91.5	−1.0

Table 6: **Ablation study**: Full (all practice-step images) vs. Single (one image per step group). Performance degradation confirms process-grounded reasoning dependence on sequential visual information.

step reasoning. We now turn to the harder question: *what happens when the task demands not merely procedural understanding, but cultural attribution?*

3.4 Diagnosing Cultural Reasoning Failure

To control response format, we use the same four-way MC interface as standard tasks: given a dish image, models select one of four displayed candidates from a 16-cuisine inventory. The target semantics, label space, and distractors differ; the substantive contrast is standard dish–attribute matching versus regional attribution, the latter requiring seasoning, ingredient, and technique knowledge beyond visual appearance.

We evaluate eight models on 378 four-way MC samples (Tab. 7), and all score only 38–56% under the same four-way response interface, compared with up to 97% on standard tasks. Consider GPT-5.1: 96.6% on standard MC (Tab. 5), yet 48.9% here, representing a 48 percentage point decline with the response format held fixed. The best-performing models, Kimi K2.6[5] and Claude 4.6, reach only 56.1%. This within-format comparison constitutes the cleanest evidence for the knowledge-application gap: models can recognize dishes but cannot reliably infer their cultural origins from visual input, even when given the same four-way response scaffolding. Human majority-vote accuracy reaches 69.0% (Tab. 12; IAA=0.634, reflecting inherent subjectivity), confirming the task is solvable, while the best-performing models still trail humans by 12.9 points.

[5]Kimi K2.6 is the updated release of K2.5 used in Tab. 5; we use the newer checkpoint for sub-region evaluation as it became available during our study.

Model	Dish Img (%)	Prac. Img (%)	Overall (%)
Kimi K2.6†	59.8	**52.4**	**56.1**
Claude 4.6	**60.3**	51.9	**56.1**
GPT-5.1	48.1	49.7	48.9
Qwen3.6-Plus	51.3	37.0	44.2
GPT-5 Mini	44.4	42.3	43.4
Qwen3.5-35B†	48.1	38.6	43.4
GPT-5 Nano	46.0	39.7	42.9
Gemma4-26B†	40.7	35.4	38.1
Random (1/4)	25.0	25.0	25.0

Table 7: **Layer 3: Sub-region classification.** The large gap from Layer 1 MC performance demonstrates weak visual cultural attribution despite strong standard MC performance. †Open-source models.

Why Do Models Fail? To augment statistical power for the diagnostic analysis below, we expand evaluation to 11 model/configuration conditions spanning 10 unique checkpoints (adding GPT-5, Qwen3.6-Flash, and a high-reasoning GPT-5 variant). Two hypotheses are tenable: (H1) models attempt cultural reasoning but lack fine-grained precision, or (H2) models bypass cultural reasoning entirely and rely on visual shortcuts. These hypotheses generate divergent predictions about error patterns, which we adjudicate below.

Confusion Pattern Analysis. We group the 16 cuisines into 5 culturally proximate families: Sichuan–Hunan–Guizhou (川湘贵), Cantonese–Fujian–Hakka–Taiwanese (粤闽客台), Shandong–Beijing–Northeastern–Henan (鲁北京东北豫), Zhejiang–Huaiyang–Anhui–Jiangxi (浙淮扬徽赣), and Northwestern (西北). Under H1, errors should cluster within families, as a model with partial knowledge might confuse Cantonese with Hakka, but not with Northeastern. Under H2, errors should disperse uniformly. The data support H2: the intra-family confusion ratio across all 11 model/configuration conditions ranges from 9.8% to 22.4%, closely matching the **random-guessing baseline of 18.1%** (Tab. 9). A one-sample binomial test against the 18.1% baseline yields $p > 0.05$ for 9 of 11 conditions; the two exceptions (Gemma4-26B and GPT-5 Mini) deviate *below* chance, showing even less cultural clustering than random, which reinforces rather than undermines H2. When models err on sub-region classification, their mistakes do not exhibit culturally informed structure; instead, they resemble guesses.

Visual Saliency Predicts Accuracy. If models rely on visual shortcuts rather than cultural reasoning, we would expect accuracy to track visual distinctiveness rather than cultural complexity. This is precisely what we observe. Sichuan cuisine (川菜),

characterized by red chili oil, Sichuan peppercorns, and an unmistakable color palette, achieves **78.4%** average accuracy across all 11 model/configuration conditions. Anhui dishes (徽菜), such as slow-braised ham and bamboo shoots, often lack distinctive visual markers; accuracy drops to **18.8%**, a 4.2× difference from Sichuan cuisine. Shandong cuisine (鲁菜) is historically prestigious but visually understated, averaging only 29.5%. This gradient is remarkably stable: the mean Spearman rank correlation of per-cuisine difficulty across all 55 condition pairs is 0.66, indicating that the models tend to find the same cuisines tractable or intractable. Models perform best when the answer is inscribed on the surface of the image.

Quantity, Not Quality. A final question: do stronger models at least *reason differently*, even if imperfectly? They do not. The intra-family confusion ratios of higher-performing models (Kimi K2.6: 21.7%, Claude 4.6: 16.9%) and lower-performing ones (Gemma4-26B: 9.8%, GPT-5 Mini: 12.6%) all oscillate around the 18.1% random baseline. Stronger models simply resolve more visually salient samples correctly. They exhibit superior *pattern discrimination*, not superior *cultural reasoning*. This finding carries a sobering implication: scaling alone appears insufficient to close the knowledge-application gap. The reasoning bridge from perception to cultural knowledge is not reliably activated by visual input.

Text-only Knowledge Probe. The diagnostics above demonstrate that models fail at visual cultural attribution, but is the cultural-geographic knowledge itself absent, or merely inaccessible through the visual pathway? To disentangle these possibilities, we remove images entirely and provide *only the dish name* (same 189 samples, identical 4-way options), stripping any cuisine-revealing keywords.[6] Strikingly, all three models perform *better* without images: Qwen3.6-Plus 68.8% (+17.5pp), GPT-5 69.8% (+11.6pp), Kimi K2.6 67.2% (+7.4pp; full results in Tab. 17, Appendix L). Models demonstrably *possess* cultural-geographic knowledge (67–70% from names alone) but cannot mobilize it through visual input. A complementary CoT experiment (Appendix L) supports the interpretation that the bottleneck is knowledge activation rather than reasoning depth: CoT yields only +1–2pp, whereas providing dish names alongside images improves by 10–14pp.

Synthesis. These diagnostics indicate that models harbor cultural-geographic knowledge in their text representations, but this knowledge becomes inaccessible when routed through visual input. What standard MC evaluations measure as "cultural knowledge" is in fact *visual-label discrimination*, not *cultural attribution*. A plausible explanation is that common multimodal objectives emphasize visual–text alignment without sufficient supervision for structured cultural grounding, making the gap difficult to detect under standard MC evaluation.

4 Related Work

Cultural Competence Benchmarks. Text-based benchmarks probe factual cultural recall (Yin et al., 2022; Wang et al., 2024; Myung et al., 2024) and cultural reasoning (Fabbri et al., 2025; Hasan et al., 2025; Arora et al., 2025). Multimodal evaluations, such as CVQA (Romero et al., 2024), CaMMT (Villa-Cueva et al., 2025), and broader suites (Yue et al., 2025; Vayani et al., 2025), assess cultural understanding at scale but remain limited to single-step retrieval or one-hop QA, affording no mechanism to distinguish knowledge *possession* from *application*. Methods for improving cultural awareness, like knowledge augmentation (Shi et al., 2024), regional alignment (Guo et al., 2025), targeted training (Feng et al., 2025), and red-teaming (Chiu et al., 2024), have advanced rapidly, yet verifying whether they yield genuine cultural reasoning requires benchmarks that probe beyond surface-level recognition.

Food-centered Cultural Evaluation. Food-centered tasks serve as a natural proxy for cultural competence (Zhou et al., 2025). FoodieQA (Li et al., 2024b) covers 14 Chinese regional cuisines but lacks procedural context; IndiFoodVQA (Agarwal et al., 2024) introduces multi-hop reasoning but is English-only; WC-VQA (Winata et al., 2025) achieves 30 languages yet offers only two task types with no process metadata. None incorporate step-by-step cooking images, the modality that encodes technique choices and preparation styles distinguishing regional cuisines. CulturalMenuBench addresses these lacunae by coupling process-level metadata with sub-regional labels and a three-layer diagnostic protocol that explicitly dis-

[6] 20/189 dish names contained explicit cuisine keywords. All were cleaned prior to evaluation.

entangles knowledge possession from application.

5 Conclusion

We present CulturalMenuBench, a multimodal benchmark (4,870 samples, 10 languages, 18 regions) evaluating process-grounded culinary reasoning and cultural knowledge. Our evaluation uncovers a **knowledge-application gap**: leading models attain 94–97% on standard MC yet reach only 38–56% on sub-region classification under the same four-way response interface, with diagnostics indicating reliance on visual shortcuts rather than cultural reasoning. Our results suggest that scaling alone is insufficient; future work must explicitly bridge perception and structured cultural knowledge.

Limitations

Food-specific scope. Cuisine, while a rich cultural marker, constitutes but one facet of cultural understanding. Our findings do not necessarily generalize to other domains (history, social customs, arts). Extending the three-layer diagnostic protocol to non-culinary cultural verticals is a natural extension.

Single-platform sourcing. All recipes originate from MeishiChina, a Chinese platform. Non-Chinese dishes therefore reflect cross-cultural adaptations rather than natively authored recipes. Future work should incorporate cuisine-native platforms (e.g., Cookpad for Japanese, Giallozafferano for Italian) to complement this perspective. A directional replication on 50 native Italian dishes from Giallozafferano is reported in Appendix B.

Translation quality. Multilingual text is LLM-translated and verified by three-model voting (96.2% macro-average pass rate in the pre-curation audit). However, culturally specific terminology may be transliterated rather than adapted, and LLM judges may share systematic blind spots. We therefore supplement the tribunal with targeted native-speaker validation (Appendix I), while broader native review remains future work.

Data reuse across tasks. All 10 tasks draw from the same 587-dish pool. Although each task is evaluated independently, users conducting joint analyses should be aware of potential memorization effects across task boundaries. Scaling to separate dish pools per task would strengthen independence guarantees.

Objective-only evaluation. Binary and MC formats enable automated scoring but exclude items requiring subjective interpretation or culturally contingent appropriateness, which are competencies central to authentic cross-cultural communication. Incorporating open-ended generation with human or LLM-as-judge evaluation would complement our diagnostic protocol.

Ethical Considerations

CulturalMenuBench is built from publicly accessible, user-contributed recipes on MeishiChina, collected solely for research evaluation; the released metadata contains no user names, account identifiers, or other personal information. Human evaluation was performed voluntarily by informed colleagues who are not authors of this paper, and only anonymized, aggregate results are reported. Risks of cultural misrepresentation arising from single-platform sourcing and translation are discussed in the Limitations section. AI assistants (Grammarly, ChatGPT, and Claude) were used only for grammar checking and rephrasing; all research ideas, experiments, and analyses are the authors' own. Code, data, and reproducibility materials are available.

Acknowledgments

We thank the anonymous reviewers and the area chair for their valuable feedback and constructive suggestions. This research was jointly supported by Alibaba Group and the Excellence Funds of Universität Hamburg.

References

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, St. Julian's, Malta, March. Association for Computational Linguistics.

Anthropic. 2025. System card: Claude Opus 4 and Claude Sonnet 4. Technical report, Anthropic.

Anthropic. 2026. Claude Opus 4.6 system card. Technical report, Anthropic.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 11772–11817, Vienna, Austria, July. Association for Computational Linguistics.

Songlin Bai, Xintong Wang, Linlin Yu, Bin Chen, Zhiang Xu, Yuyang Sheng, Changtong Zan, Xiaofeng Zhu, Yizhe Zhang, Jiru Li, and others. 2026. Industrybench: Probing the industrial knowledge boundaries of llms. *arXiv preprint arXiv:2605.10267*.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada, July. Association for Computational Linguistics.

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

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. 2024b. Exploring visual culture awareness in GPT-4V: A comprehensive probing. *arXiv preprint arXiv:2402.06015*.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalTeaming: AI-assisted interactive red-teaming for challenging LLMs' (lack of) multicultural knowledge. *arXiv preprint arXiv:2404.06664*.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria, July. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & LM benchmarking. *arXiv preprint arXiv:2402.09369*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Gemma Team. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.

Google DeepMind. 2026. Gemini 3 Pro model card. Technical report, Google DeepMind. Model released November 2025; model card last updated May 2026.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32866–32895, Suzhou, China, November. Association for Computational Linguistics.

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

Kimi Team. 2026. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, pages 611–626, Koblenz, Germany, October. Association for Computing Machinery.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems 37 (Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems)*, pages 84799–84838, Vancouver, Canada, December.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024b. FoodieQA: A multimodal dataset for fine-grained understanding of chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA, November. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2023. Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5674–5685, Ottawa, Ontario, Canada, October. Association for Computing Machinery.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 37 (Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems)*, pages 78104–78146, Vancouver, Canada, December.

OpenAI. 2025a. GPT-5 system card. Technical report, OpenAI.

OpenAI. 2025b. GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum. Technical report, OpenAI.

OpenAI. 2026. GPT-5.5 system card. Technical report, OpenAI.

Haonan Qi, Jin Cao, Yongqi Zhang, Xintong Wang, Weidong Tang, Bin Chen, Chengfu Huo, Haojun Pan, Hengyu You, Jing Li, and others. 2026. Industrybench-mipu: Benchmarking multi-image attribute value extraction for industrial products. *arXiv preprint arXiv:2606.14383*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems 37 (Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems)*, pages 11479–11505, Vancouver, Canada, December.

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

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Hafizi Amirudin, Muhammad Ridzuan, Daniya Najiha Abdul Kareem, Ketan Pravin More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Febian Farestam, Muztoba Rabbani, Sanoojan Ballah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Augusto Zacarias Xavier, Amit Bhatkal, Hawau Olamide Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Shahbaz Khan. 2025. All languages matter: Evaluating LMMs on culturally diverse 100 languages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19565–19575, Nashville, Tennessee, USA, June.

Emilio Villa-Cueva, Sholpan Bolatzhanova, Diana Turmakhan, Kareem Elzeky, Henok Biadglign Ademtew, Alham Fikri Aji, Vladimir Araujo, Israel Abebe Azime, Jinheon Baek, Frederico Belcavello, Fermin Cristobal, Jan Christian Blaise Cruz, Mary Dabre, Raj Dabre, Toqeer Ehsan, Naome A. Etori, Fauzan Farooqui, Jiahui Geng, Guido Ivetta, Thanmay Jayakumar, Soyeong Jeong, Zheng Wei Lim,

Aishik Mandal, Sofía Martinelli, Mihail Minkov Mihaylov, Daniil Orel, Aniket Pramanick, Sukannya Purkayastha, Israfel Salazar, Haiyue Song, Tiago Timponi Torrent, Debela Desalegn Yadeta, Injy Hamed, Atnafu Lambebo Tonja, and Thamar Solorio. 2025. CaMMT: Benchmarking culturally aware multimodal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22423–22441, Suzhou, China, November. Association for Computational Linguistics.

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

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Christabelle Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico, USA, April. Association for Computational Linguistics.

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, December. Association for Computational Linguistics.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*, Singapore, April. Official proceedings do not assign page numbers.

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

A Scope of the Diagnostic Evidence

Our confusion and saliency analyses are grounded in Chinese regional cuisines, which is the only subset with fine-grained sub-regional labels. However, three observations suggest the diagnosed mechanism generalizes beyond this specific cultural context. First, the failure is architecture-agnostic: it manifests across 11 model/configuration conditions spanning 10 unique checkpoints, multiple model families, and both open-source and commercial systems. Second, the distinction between visual discrimination and cultural attribution is not specific to Chinese food; it applies equally to domains where surface appearance diverges from categorical membership. Third, our Layer 1–2 results show that the MC-to-binary gap is consistent across all 10 languages (12–26pp; Appendix D), indicating that the knowledge-application disconnect is cross-cultural even where we lack sub-regional labels to perform the full diagnostic. We view the Chinese sub-region analysis as a detailed case study of a general phenomenon. A directional pilot on 50 native Italian dishes reproduces the recognition–region gap (Appendix B), but broader validation requires scaling verified sub-regional labels, including Italian regional and Indian state-level cuisines. Future expansion can combine native platforms, LLM-assisted pre-filtering with human adjudication, and audited community contributions.

B Native-Data Replication beyond Chinese Cuisine

To test whether the knowledge-application gap is specific to cross-cultural adaptation, translation, or the source platform, we collect 50 native Italian dishes from Giallozafferano[7], using region labels from its "ricette regionali" taxonomy. We rerun dish recognition and region attribution as four-way MC tasks (25% random baseline) on four models.

[7] https://www.giallozafferano.com/

Model	Recognition (%)	Region (%)	Gap (pp)
GPT-5.1	96.0	**80.0**	16.0
Kimi K2.6†	96.0	68.0	28.0
Qwen3.6-Plus	92.0	56.0	36.0
Qwen3.5-35B†	**98.0**	60.0	38.0

Table 8: Native Italian replication ($N = 50$, four-way MC). The recognition–region gap appears across all four models. †Open-source models.

The gap reproduces in direction for every model (16–38pp), with recognition at or above 92%. Because this pilot covers only 50 dishes, it provides directional cross-platform evidence rather than a comprehensive estimate of Italian regional-cuisine performance.

C Confusion Diagnosis

Model	Acc. (%)	Errors	Intra-fam. (%)	Δ vs. rand.
GPT-5‡	57.1	162	17.9	−0.2
Claude 4.6	56.1	166	16.9	−1.2
Kimi K2.6†	56.1	166	21.7	+3.6
GPT-5	53.2	177	14.7	−3.4
GPT-5.1	48.9	193	18.7	+0.6
Qwen3.6-Flash	46.6	202	16.3	−1.8
Qwen3.6-Plus	44.2	211	20.9	+2.8
GPT-5 Mini	43.4	214	12.6	−5.5
Qwen3.5-35B†	43.4	214	22.4	+4.3
GPT-5 Nano	42.9	216	15.7	−2.4
Gemma4-26B†	38.1	234	9.8	−8.3
Random baseline	25.0	—	18.1	0.0

Table 9: Diagnosing the cultural reasoning gap across 11 model/configuration conditions (10 unique checkpoints). **Intra-fam.**: percentage of errors where the predicted cuisine belongs to the same cultural family as the ground truth. **Δ vs. rand.**: deviation from the 18.1% random baseline. No condition shows meaningful above-chance cultural clustering. †Open-source. ‡High reasoning effort.

D Additional Analyses

Cross-Lingual Consistency. Figure 3 (a) shows performance variation across languages, where high-resource European languages lead (French 83.4%, Italian 84.8%) while low-resource Asian languages trail (Korean 73.2%, Thai 73.2%). Crucially, the MC-to-binary gap persists across all ten languages (range: 12–26pp), confirming that the knowledge-application disconnect is not an artifact of any particular linguistic or cultural subset.

Response Bias. Ground-truth labels are approximately uniform across A/B/C/D. Most models preserve this balance (χ^2 test, $p > 0.05$), though GPT-5 Nano shows a significant preference for option A (29.1%, $p < 0.0001$). In binary tasks, both annotators and models exhibit a rejection bias (58–65% "no" responses despite balanced ground truth), indicating genuine task difficulty rather than annotation artifacts.

Disjoint-Subset and Bootstrap Analysis. We split the 177 dishes shared by the Layer 1 and Layer 3 MC tasks into five disjoint random subsets and recompute the within-format gap for the six models evaluated on both layers. The gap remains positive in every subset for every model (no subset gap below 0.30), with a standard deviation across subsets of at most 0.098. Across 1,000 bootstrap samples, the proportion with a positive gap is 1.000 for all six models, showing that the aggregate gap is not carried by a small set of dishes.

Reasoning Chain Verification. On Layer 2 process-reasoning tasks (where models perform well), do they actually engage in structured multi-step reasoning, or succeed through shortcuts? We collect reasoning chains from Claude 4.6 and Qwen3.6-Plus on 30 practice-image samples. Both exhibit structured patterns: visual observation (83–100%), process-level reasoning (87–93%), and ingredient–procedure cross-referencing (50–93%). This confirms that the benchmark elicits genuine multi-step reasoning on process tasks, making the Layer 3 cultural reasoning failure all the more striking: models *can* reason over procedures, but cannot bridge from procedural understanding to cultural attribution.

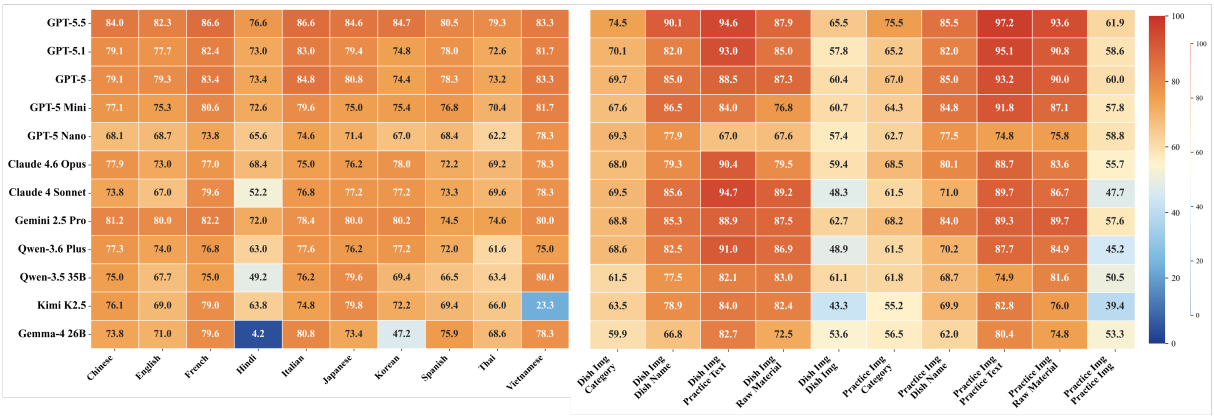
E Supplementary Tables

Language	Pass Rate (%)	Passed/Total
English / Spanish / French	94.4 / 96.1 / 96.7	449/468
Hindi / Italian / Thai	94.0 / 95.3 / 93.3	424/450
Japanese / Korean / Viet.	98.7 / 97.3 / 100.0	312/318
Average	**96.20**	**-**

Table 10: Pre-curation translation audit on 412 candidate non-Chinese dishes (1,236 fields) via an LLM voting tribunal (Gemini-3.0-Pro, GPT-5-Mini, DeepSeek-V3). A majority vote (≥2/3) is required for approval; 96.20% is the macro-average across languages.

Regional Cuisine	Count	Regional Cuisine	Count
Lu Cuisine (Shandong)	32	Guizhou Cuisine	6
Yue Cuisine (Cantonese)	28	Hakka Cuisine	5
Chuan Cuisine (Sichuan)	21	Taiwanese Cuisine	4
Xiang Cuisine (Hunan)	19	Yu Cuisine (Henan)	4
Zhe Cuisine (Zhejiang)	15	Gan Cuisine (Jiangxi)	4
Beijing Cuisine	11	Yunnan Cuisine	3
Northeastern Cuisine	10	Hong Kong Cuisine	2
Northwestern Cuisine	8	Su Cuisine (Jiangsu)	2
Min Cuisine (Fujian)	8	Jin Cuisine (Shanxi)	2
Huaiyang Cuisine	7	Xinjiang Cuisine	1
Hui Cuisine (Anhui)	7	E Cuisine (Hubei)	1

Table 11: Distribution of dishes by regional Chinese cuisine in CulturalMenuBench.



(a) Performance Heatmap Across Languages (b) Performance Heatmap Across Tasks

Figure 3: Performance heatmap of 12 LLMs on CulturalMenuBench across different languages and tasks.

Annotator	Background	Acc. (%)
Annotator 1	Sichuan/Hunan cuisine	65
Annotator 2	Cantonese/Fujian cuisine	55
Annotator 3	Shandong/Northeastern cuisine	63
Annotator 4	Zhejiang/Huaiyang cuisine	60
Annotator 5	Generalist food critic	63
Average	–	**61.2**
Majority Vote	–	**69.0**

Table 12: Human performance on sub-region classification (100 samples, 16 Chinese regional cuisines, 4-way MC). Pairwise inter-annotator agreement: 0.634.

F Sample Distribution

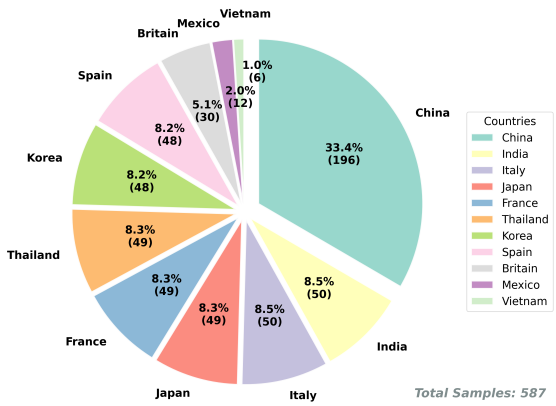


Figure 4: Distribution of the curated menu samples across 10 countries.

G Details of Tasks

Details of the 10 tasks in CulturalMenuBench are listed in Tab. 13. For each task, two metadata fields were selected for matching, and we ensured that each task included image metadata to create complex multimodal reasoning tasks. To ensure objective evaluation, we constrained the model to make binary judgments (yes or no) in the prompts for each task. Additionally, for cuisines from different countries, we used the corresponding national language in the prompts (not shown in the table), which effectively evaluates the model's multilingual understanding capabilities.

H Multiple-Choice Format Details

To complement the binary evaluation, we construct a multiple-choice (MC) variant of CulturalMenuBench comprising 4,085 items across 14 directed relations. For each item, three plausible distractors are sampled from the same cuisine category or metadata domain as the correct answer, ensuring that surface-level elimination strategies are ineffective. The MC format reduces the random baseline from 50% (binary) to 25%, providing a more discriminating evaluation of model capabilities. The per-task MC accuracy for complete models is reported in Tab. 14.

Task	GPT-5 Mini	Claude Son. 4	GPT-5 Nano
DI ↔ DN	92	86	88
DI ↔ PI	95	82	75
DI ↔ PT	97	92	86
DI ↔ RM	94	86	85
DN ↔ DI	93	86	84
DN ↔ PI	90	82	78
PI ↔ DI	97	82	54
PI ↔ DN	92	84	82
PI ↔ PT	99	94	88
PI ↔ RM	96	88	91
PT ↔ DI	96	92	84
PT ↔ PI	98	93	81
RM ↔ DI	94	85	81
RM ↔ PI	96	83	78
Overall	94.7	87.1	82.2

Table 14: Per-task MC accuracy (%). DI = Dish Image, DN = Dish Name, PI = Practice Image, PT = Practice Text, RM = Raw Material. Practice-image tasks (PI ↔ DI) show the largest gap between models, with GPT-5 Nano dropping to 54%.

Task: `dish_image` ↔ `dish_name` Description: Given a dish image and a candidate dish name, decide whether the name correctly identifies the food shown. Prompt: "Is this image *[dish_name]*? *[dish_image]*"
Task: `dish_image` ↔ `category` Description: Given a dish image and a cuisine label, judge whether the dish belongs to that cuisine. Prompt: "Is this image *[category]* cuisine? *[dish_image]*"
Task: `dish_image` ↔ `raw_material` Description: Given a dish image and a list of main ingredients, verify whether the ingredients match the pictured dish. Prompt: "Is the main ingredient of this dish *[raw_material]*? *[dish_image]*"
Task: `dish_image` ↔ `practice_text` Description: Given a dish image and a cooking-method description, decide whether the method matches the dish. Prompt: "The cooking method for this dish is: *[practice_text]*. Is this description correct? *[dish_image]*"
Task: `practice_image` ↔ `dish_name` Description: Given several cooking-process images and a dish name, determine whether the images depict the preparation of that dish. Prompt: "Do these cooking process images show the preparation of *[dish_name]*? *[practice_image]*"
Task: `practice_image` ↔ `category` Description: Given several cooking-process images and a cuisine label, judge whether the technique belongs to that cuisine. Prompt: "Do these cooking process images show the preparation of *[category]* cuisine? *[practice_image]*"
Task: `practice_image` ↔ `raw_material` Description: Given several cooking-process images and a list of main ingredients, verify whether the ingredients fit the depicted dish. Prompt: "Do these cooking process images show a dish whose main ingredient is *[raw_material]*? *[practice_image]*"
Task: `practice_image` ↔ `practice_text` Description: Given several cooking-process images and a cooking-method description, decide whether they correspond. Prompt: "The cooking method shown in these images is: *[practice_text]*. Is this description correct? *[practice_image]*"
Task: `dish_image` ↔ `dish_image` Description: Given two dish images, determine whether they belong to the same cuisine category. Prompt: "Do the following two dish images belong to the same cuisine category? `dish_image1` `dish_image2`"
Task: `practice_image` ↔ `practice_image` Description: Given two sets of cooking-process images, assess whether they represent dishes from the same cuisine category. Prompt: "Do the following two groups of preparation step images belong to the same cuisine category? `practice_image1` `practice_image2`"

Table 13: Task categories and their corresponding prompt templates. Each task is expressed in three consecutive lines: task name, description, and prompt. All tasks require a binary (yes/no) answer expressed in the target language.

I Translation Quality Details

The LLM voting tribunal exhibits meaningful inter-judge disagreement: only 84.0% of items receive unanimous 3/3 votes, with the remaining 16% showing split decisions, confirming that judges are not rubber-stamping translations. Error analysis of the rejected items reveals that information omission (46%) is the dominant failure mode (e.g., omitting "Korean-style" from a dish name), followed by terminology errors (6%) and unnaturalness (6%). Failure rates vary across both field and cuisine. Among fields, dish names are the most difficult to translate, with a 90.0% pass rate, compared with 98.5% for procedures and 99.3% for ingredient lists. Across cuisines, Indian and British cuisines have lower pass rates for dish names, at 80% and 83%, respectively, suggesting greater cross-cultural variation in naming conventions.

Native-Speaker Validation. Native speakers who read Chinese audited 18 items each in Hindi, British English, and Vietnamese against the Chinese sources, covering dish names, ingredients, and cooking steps. The annotations were pre-drafted with LLM assistance and then reviewed and signed off by the native speakers. The audit found no outright mistranslations and 2–3 under-specification cases per language, such as a specific dish receiving a more generic name. It also identified a duplicated source string and residual Chinese punctuation. These cases are disambiguated or corrected in the release.

J Error Analysis by Failure Type

We analyze errors from a high-performing model (GPT-5.1, 3.4% error rate) and a lower-performing full-coverage model (GPT-5 Nano, 17.8% error rate), classifying them into three categories.

Visual Confusion accounts for 49.6% of GPT-5.1 errors but only 20.7% of GPT-5 Nano errors. These errors arise from visually similar dishes (e.g., confusing two curry-based soups or pasta variants). Indian cuisine is disproportionately affected due to its curry-heavy visual homogeneity, accounting for 24.7% of shared hard cases despite being only 11.8% of the data.

Process Reasoning Failure is the dominant error mode for GPT-5 Nano (51.4% of its errors) but only 22.6% for GPT-5.1. The hardest task, `practice_image`→`dish_image`, yields a 46.0% error rate for Nano vs. 1.0% for GPT-5.1, which suggests that reasoning about cooking procedures scales steeply with model capacity.

Ingredient Recognition Failure constitutes ∼28% of errors for both models, though the absolute count in Nano is 6.3× higher, indicating that ingredient identification difficulty scales proportionally with overall capability.

Language and Cuisine Effects. Vietnamese (57 samples) has the highest error rate for both models: 13.7% for GPT-5.1 and 36.8% for Nano. Together, Vietnamese and Hindi account for 28.7% of GPT-5.1 errors despite representing only 13.4% of the data, revealing compound difficulty from low-resource languages paired with visually homogeneous cuisines. Of GPT-5.1's 115 errors, 85 (74%) are shared with Nano, indicating that residual errors cluster on genuinely ambiguous items.

K Experimental Setup Details

Model Configurations. Table 15 summarizes the 12 models evaluated in Layers 1–2 and the 11 model/configuration conditions (10 unique checkpoints) used for Layer 3 sub-region diagnosis. Model provenance is documented by the corresponding system cards and technical reports for GPT-5/5.1/5.5 (OpenAI, 2025a,b, 2026), Claude 4/4.6 (Anthropic, 2025, 2026), Gemini 2.5 (Gemini Team, 2025), Kimi K2.5 (Kimi Team, 2026), and Gemma 4 (Gemma Team, 2026). All API-based models are accessed via official endpoints; the exact provider documentation is available for Kimi K2.6[8] and Qwen3.6-Plus/Flash.[9] The Qwen3.5-35B checkpoint follows its official model card.[10] Open-weight models are served locally via vLLM v0.19.1 (Kwon et al., 2023) with tensor parallelism across 2 GPUs.

Model	Identifier	Access
Proprietary (API)		
GPT-5.5	gpt-5.5-2025-11-20	OpenAI
GPT-5	gpt-5-2025-08-07	OpenAI
GPT-5.1	gpt-5.1-2025-11-13	OpenAI
GPT-5 Mini	gpt-5-mini-2025-08-07	OpenAI
GPT-5 Nano	gpt-5-nano-2025-08-07	OpenAI
Gemini 2.5 Pro	gemini-2.5-pro	Google AI
Claude Opus 4.6	claude-opus-4-6	Anthropic
Claude Sonnet 4	claude-sonnet-4	Anthropic
Kimi K2.5/K2.6	kimi-k2.5 / kimi-k2.6	Moonshot AI
Qwen3.6-Plus	qwen3.6-plus-2026-04-02	Alibaba
Qwen3.6-Flash	qwen3.6-flash-2026-04-02	Alibaba
Open-weight (local inference)		
Qwen3.5-35B	Qwen/Qwen3.5-35B-A3B	vLLM, TP=2
Gemma4-26B	gemma-4-26B-A4B-it	vLLM, TP=2

Table 15: Model configurations. All OpenAI models use the GlobalStandard variant. Kimi K2.6 is used for sub-region evaluation (Layer 3); K2.5 for Layers 1–2. Open-weight models are served with BF16 precision and max sequence length of 8,192 tokens.

Inference Parameters. For Layer 1–2 evaluation, all models use greedy decoding (temperature=0) with max output tokens of 64 (binary and MC tasks require only short outputs). For the sub-region diagnosis (Layer 3), GPT-family models use `reasoning_effort=low` to reduce latency; Claude Opus 4.6 uses extended thinking with a budget of 1,024 tokens. The text-only probe uses temperature=0 with max 64 tokens.

Prompt Templates. We provide the exact prompts used for each experimental condition:

- **Standard sub-region MC** (image input):

```
[dish image]
这张菜品图片属于以下哪个菜系?
A. 贵州菜  B. 浙菜  C. 客家菜  D. 川菜
请仅输出A、B、C、D选项的字母，不要输出任何其他内容。
```

- **Text-only probe** (no image, dish name input):

```
根据菜名判断，"{dish_name}" 属于以下哪个菜系?
A. ...  B. ...  C. ...  D. ...
请仅输出A、B、C、D选项的字母，不要输出任何其他内容。
```

  Dish names are cleaned to remove cuisine-revealing keywords (e.g., "四川小吃红油抄手"→"红油抄手"). 20/189 names required cleaning.

- **CoT prompting** (image + reasoning instruction):

[8] Official API documentation: Moonshot AI platform.
[9] Official model pages: Qwen3.6-Plus and Qwen3.6-Flash.
[10] Official model card: Qwen3.5-35B-A3B.

```
[dish image]
请仔细观察这张菜品图片，按以下步骤推理:
1. 首先识别这道菜是什么
2. 分析其食材、调味料和烹饪手法的特征
3. 根据这些特征判断它最可能属于哪个菜系
A. ...   B. ...   C. ...   D. ...
请先给出你的推理过程，最后在新的一行输出答案字母。
```

- **Text+Image combined** (dish name + image):

```
[dish image]
这道菜叫做「{dish_name}」。结合菜名和图片，判断它属于以下哪个菜系?
A. ...   B. ...   C. ...   D. ...
请仅输出A、B、C、D选项的字母，不要输出任何其他内容。
```

Distractor Construction. For all 4-way MC tasks (Layers 1–3), distractors are sampled from the same cultural family when possible. For sub-region classification, each item's three distractors are drawn from the pool of 16 regional cuisines, with at least one from the same cultural family (e.g., for a Sichuan dish: one Hunan/Guizhou option plus two from other families). This ensures that surface-level elimination is insufficient.

Evaluation Protocol. Answer extraction uses regex matching for the first (standard tasks) or last (CoT) occurrence of [A-D] in model output. Samples with empty responses or unparseable outputs are retried up to 3–5 times with exponential backoff. The concurrency is set to 4–10 parallel requests per model depending on API rate limits.

L CoT and Text-Anchor Ablation

CoT Prompting. On the same 177 dish-image samples, we instruct models to reason step-by-step before answering: (1) identify the dish, (2) analyze ingredients, seasonings, and cooking techniques, (3) determine the most likely regional cuisine. All models employ their strongest available reasoning configuration (GPT-5: built-in reasoning tokens; Qwen3.6-Plus: extended thinking enabled; Gemma4-26B: prompt-level CoT).

Text+Image Combined. We provide the dish name alongside the image (e.g., "Based on the name '宫保鸡丁' and image, which cuisine does it belong to?"), testing whether textual anchoring can bridge the visual-to-cultural reasoning gap.

Model	Baseline	+CoT	+Name
GPT-5	57.6	59.3 (+1.7)	**67.8** (+10.2)
Qwen3.6-Plus	58.3	59.3 (+1.0)	**67.8** (+9.5)
Gemma4-26B	43.5	45.2 (+1.7)	**57.1** (+13.6)

Table 16: CoT prompting vs. text-anchor ablation on sub-region classification. CoT yields marginal improvement, whereas providing the dish name boosts accuracy by 10–14pp, supporting the interpretation that the bottleneck is knowledge activation rather than reasoning depth. Notably, even the strongest text-anchored result (67.8%) only *matches* text-only performance rather than exceeding it, suggesting that visual input contributes little additive cultural signal beyond what the dish name alone provides.

Model	Image (%)	Text-only (%)	Δ
Qwen3.6-Plus	51.3	**68.8**	+17.5
Kimi K2.6	59.8	**67.2**	+7.4
GPT-5	58.2	**69.8**	+11.6

Table 17: Text-only vs. image-based sub-region classification (same 189 samples, same 4-way MC options). Removing images *improves* accuracy by 7–18pp, showing that cultural-geographic knowledge exists but is not reliably activated through visual input.

M Case Study Details

The two tasks in Fig. 5 illustrate the multi-step reasoning required by CulturalMenuBench:

Task 1: Semantic Disambiguation. The model must disambiguate "English Cheese Tea Treats", where "cheese tea" evokes East Asian beverage culture while "treats" suggests solid desserts. The model decomposes the query ("English" → origin; "Cheese Tea" → beverage; "Treats" → dessert confusion), notes Chinese watermarks contradicting the Anglophone framing, and aligns against visual evidence (berries, cream, ramekins → crumble/dessert) to conclude "No."

Task 2: Visual Abstraction. Both images depict Hasselback potatoes with characteristic fan-cut geometry, but differ in plating and garnish. The model must abstract away surface differences and focus on core identity markers: cut pattern, primary ingredient, and cooking method. It then recognizes both as Western/European-inspired preparations, demonstrating categorical equivalence despite stylistic variation.

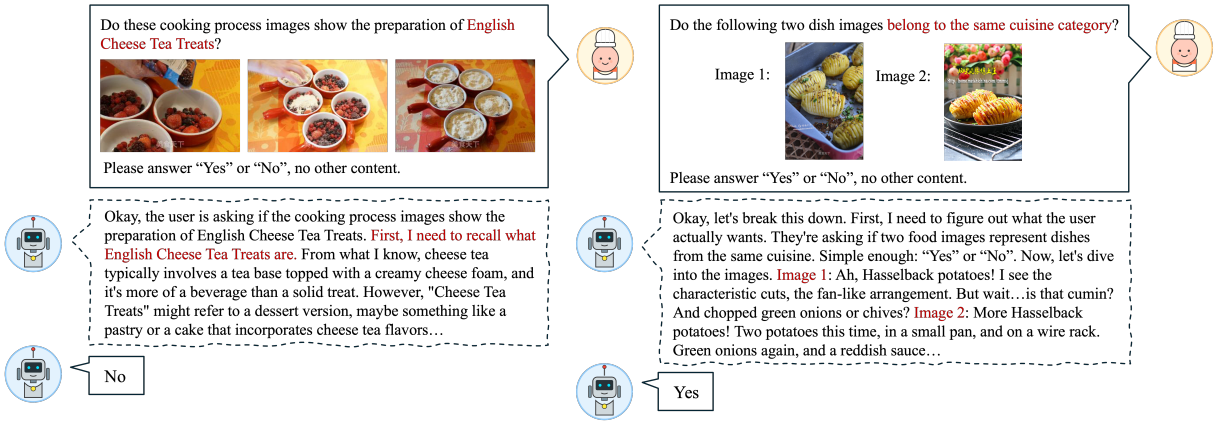


Figure 5: Two representative examples from CulturalMenuBench. Task 1 requires cross-cultural semantic disambiguation; Task 2 requires abstracting away surface-level plating differences to recognize categorical equivalence. Both demand multi-step reasoning integrating visual, procedural, and cultural knowledge.