

MMAC: A Multilingual, Multimodal Alignment Framework for Cultural Grounding Evaluation

Weihua Zheng^{1,2}, Zhengyuan Liu¹, Tanmoy Chakraborty³, Weiwen Xu⁴, Xiaoxue Gao¹, Bryan Chen Zhengyu Tan², Bowei Zou¹, Chang Liu⁶, Yujia Hu², Xing Xie⁵, Xiaoyuan Yi⁵, Jing Yao⁵, Chaojun Wang⁴, Long Li⁴, Rui Liu², Huiyao Liu⁷, Koji Inoue⁸, Ryuichi Sumida⁸, Tatsuya Kawahara⁸, Fan Xu⁹, Lingyu Ye⁹, Wei Tian⁹, Dongjun Kim¹⁰, Jimin Jung¹⁰, Jaehyung Seo¹², Nadya Yuki Wangsajaya¹¹, Pham Minh Duc¹, Ojasva Saxena³, Palash Nandi³, Xiyao Tao¹, Wiwik Karlina¹, Tuan Luong¹, Keertana Arun Vasani¹¹, Roy Ka-Wei Lee^{2*}, Nancy F. Chen^{1*}

¹Agency for Science, Technology and Research, Singapore ²Singapore University of Technology and Design

³Indian Institute of Technology Delhi ⁴Alibaba DAMO Academy ⁵Microsoft Research Asia

⁶Shanghai University of Finance and Economics ⁷Inner Mongolia University ⁸Kyoto University

⁹Jiangxi Normal University ¹⁰Korea University ¹¹Nanyang Technological University ¹²Konkuk University

Abstract

The global deployment of Large Language Models (LLMs) underscores the urgent need to evaluate their cultural alignment. However, assessing genuine “cultural awareness” across modalities (text, vision, speech) and languages remains a significant challenge. To comprehensively investigate this domain, we propose a Multilingual, Multimodal Alignment framework for Cultural grounding evaluation (MMAC). This systematic framework encompasses a tri-modally aligned cultural benchmark creation pipeline and a five-dimensional evaluation protocol to assess cross-country awareness disparities, evaluate cross-lingual and cross-modal consistency, and verify cultural knowledge generalization and grounding validity. Given the prevailing Western cultural bias in current models, we focus on 8 Asian countries as our dataset foundation to more acutely reveal potential cultural deficiencies in LLMs. Our dataset, MMAC-bench, features 27,000 human-curated questions across 10 languages. Crucially, it is the first dataset aligned at the input level across text, image, and speech, enabling direct cross-modal transfer tests. Each question consists of multiple-choice options accompanied by open-ended generated explanations, where 79% require multi-step reasoning grounded in cultural context, moving beyond simple memorization. We probe the causes of modal divergence, offering insights into fostering culturally robust MLLMs.

Code: <https://github.com/ZWH-ASTAR/MMAC>

1 Introduction

As multimodal, multilingual models proliferate (Chen et al., 2024; Bai et al., 2023; Jiang et al., 2023; OpenAI et al., 2024; Touvron et al., 2023; Hao et al., 2025), a fundamental question arises: Does a model possess a coherent underlying under-

standing of a culture, or does it merely exhibit fragmented knowledge depending on the input modality and language? Existing culture-centric datasets (Myung et al., 2025; Wang et al., 2024; Zhou et al., 2025; Li et al., 2025a) frequently suffer from two key limitations: (i) insufficient alignment of instances across modalities, (ii) inadequate representation of low-resource Asian languages. Furthermore, evaluation processes are easily hacked through memorization or elimination in multiple choice questions (MCQs), which bypass the genuine reasoning capabilities (Wang et al., 2025; Hartmann et al., 2023; Liu et al., 2025; Yong et al., 2025). As a result, we still lack a principled way to separate actual cultural competence from artifacts.

To address this research gap, we introduce MMAC, an explainable evaluation framework grounded in a new dataset, MMAC-bench. Under the framework, we conducted a five-axis evaluation: (i) cultural awareness disparity, defined as the differences exhibited by models across national cultures, languages, and modalities. (ii) cross-modal consistency in cultural awareness, defined as the extent to which a model gives stable answers to semantically equivalent inputs when the representation (text, image+question, or spoken question); (iii) cross-lingual consistency in cultural awareness, defined as the extent to which a model gives stable answers to semantically equivalent inputs when the language changes; (iv) cultural awareness grounding, defined as whether correct answers rely on appropriate cultural signals rather than exploitable shortcuts; and (v) cultural awareness generalization, defined as whether a model that has access to the relevant cultural knowledge can perform the required reasoning within those cultural contexts (Balepur et al., 2024; Molfese et al., 2025; Zheng et al., 2023; Zhang et al., 2026a; Kreutzer et al., 2025). Motivated by prior findings that current LLMs exhibit systematic biases favoring Western contexts over non-mainstream Asian cultures (Chiu

*Corresponding authors.

et al., 2025; Romero et al., 2024; Vayani et al., 2025; Weihua et al., 2026; Myung et al., 2025; Ng et al., 2025), this work centers on evaluating model performance across 8 specific Asian nations. Our dataset aligns tri-modal items (textual question, image+question, and Text-to-Speech (TTS)-spoken question) with identical semantics and provides parallel local-language and English versions authored by native experts across 8 countries and 10 languages to make a comprehensive evaluation.

Using MMAC-bench, we evaluate 14 multilingual and multimodal LLMs (e.g., GPT-4o, Qwen, Llama). We find that accuracy drops markedly in low-resource Asian languages compared to English; cross-modal consistency lags text-only performance, indicating incomplete transfer from language to vision and speech; and grounding controls reduce a non-trivial fraction of apparent “wins”, revealing shortcut use. We also analyze multi-step, culture-specific reasoning errors and where visual or linguistic cues fail to connect. We summarize our contributions as follows:

- **Aligned tri-modal, multilingual benchmark.** We release MMAC-bench with 27k multimodal questions authored by in-country experts across 8 countries and 10 languages.
- **Five-axis evaluation protocol.** We formalize cultural awareness *consistency* (cross-modal/cross-lingual) and *grounding* with negative controls and ablations, plus generalization tests under held-out themes/countries.
- **Extensive baselines and analyses.** We report zero-shot baselines for 14 model families (multilingual LLMs and VLMs), including common-support subsets, and diagnose failure modes by modality, language, and reasoning step count.

2 Related Work

Cultural knowledge in text. Recent benchmarks assess culture-specific knowledge via MCQs (Kim et al., 2024; Wang et al., 2024; Susanto et al., 2025; Myung et al., 2025; Chiu et al., 2025; Cho et al., 2025), consistently showing performance gaps favoring English/high-resource settings and sensitivity to formatting. However, most lack aligned multimodal counterparts to test cross-modal cultural understanding.

Cultural perception in images (VQA). Multilingual VQA datasets use community-sourced images and questions (Romero et al., 2024; Nayak et al., 2024; Vayani et al., 2025), revealing vision-

language gaps and language sensitivity. Yet they typically lack text/speech-parallel versions of identical items, making it hard to isolate whether failures stem from cultural knowledge, visual grounding, or language handling.

Speech and accent robustness. Speech datasets reveal substantial accent-related biases (Eisenstein et al., 2023; Roh et al., 2025; Chen et al., 2026), and perturbations can drastically change outcomes. However, they rarely evaluate cultural knowledge directly or analyze trimodal alignment and consistency.

Consistency and grounding. MCQ performance may reflect shortcuts rather than grounded reasoning (Balepur et al., 2024; Molfese et al., 2025; Zheng et al., 2023; Aakanksha et al., 2024; Kirk et al., 2024). Benchmarks seldom include negative controls or report cross-lingual/cross-modal consistency which is essential for distinguishing cultural competence from artifacts.

How is MMAC-bench different from others?

Existing datasets (Myung et al., 2025; Wang et al., 2024; Kim et al., 2024) evaluate cultural understanding within a single modality or language at a time, without tightly *aligning* instances across modalities and languages, and no built-in *grounding controls*. MMAC-bench addresses these gaps by: (i) providing semantically aligned tri-modal items (text, image+text, speech+text) in parallel local-language and English versions; (ii) adopting *cross-modal* and *cross-lingual* consistency as prime metrics; and (iii) integrating targeted ablations and negative controls to test whether answers rely on the intended cultural signal rather than shortcuts. This design enables clearer attribution of failure modes: knowledge vs. language vs. modality, and more reliable measurement of cultural awareness in multimodal, multilingual models. We summarize key differences among representative datasets in Tab. 1.

3 Benchmark Construction

MMAC-bench was collaboratively constructed by research teams from eight countries: China, Singapore, Japan, South Korea, Mongolia, Vietnam, Indonesia, and India. The pipeline comprised five stages: §3.1 annotator selection, §3.2 selection of representative cultural themes and languages, §3.3 collection of text and image materials, §3.4 question authoring and annotation by country sub-teams, and §3.5 human review and revision for

Benchmark	MLA	MDA	RI	# of Ctries	# of Langs	Modality	Multi-step reasoning	Question forms	Total samples	Total domains
SeaEval	✗	✗	✗	4	1		✓	Diverse	415	-
CLiCk	✗	✗	✗	1	2		-	Diverse	1,995	11
BLEnD	✓	✗	✗	16	13		-	Fixed	52,557	6
Culturalbench	✗	✗	✗	45	1		-	Fixed	1,696	17
CVQA	✓	✗	✗	30	31		-	Fixed	10,000	10
CulturalbenchVQA	✗	✗	✗	11	1		-	Diverse	2,378	5
ALM-bench	✓	✗	✗	73	100		-	Diverse	22,763	19
Md3	✗	✗	✗	3	1		-	Diverse	3,689	2
MULTI-AUDIOJAIL	✓	✗	✗	6	6		-	Diverse	102,720	-
Ours	✓	✓	✓	8	10		✓	Diverse	27,000	9

Table 1: Comparison of existing culture-related benchmark datasets with MMAC-bench. and denote the text, image and speech modalities, respectively. “MLA”, “MDA” and “RI” denote “multilingual alignment”, “multimodal alignment” and “Response interpretability”, respectively.

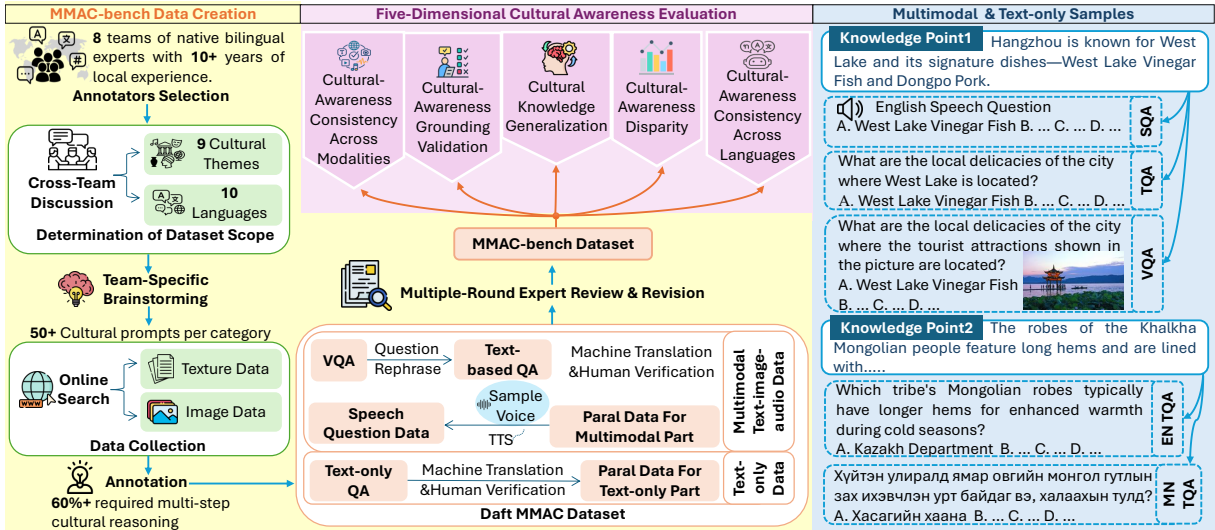


Figure 1: An overview of the MMAC evaluation framework: data creation pipeline (yellow), evaluation dimensions (pink), and representative dataset samples (blue).

quality and cultural representativeness. For the definition of cultural themes, we followed the framework proposed by Adilazuarda et al. (2024).

3.1 Annotator Selection

An in-country expert team curated each national subset. All team members were native speakers of the local language and proficient in English. Annotators had lived in the respective country for more than ten years, ensuring deep familiarity with local cultural contexts. Detailed annotator information is in Appendix A.1. Before annotation began, we held project-wide briefings to explain the scope and requirements. We also distributed a detailed English annotation guideline (see Appendix A.3).

3.2 Cultural Themes and Languages

Through collaborative discussions, we finalized 9 cultural themes, *Daily life habits/Culture*, *Food/Cuisine*, *Transportation*, *Buildings*, *History*, *Geographical location and climate*, *Education*,

Fashion/Clothing and Language/Race. Each national subset includes the country’s official language(s) and English. For India, we selected Hindi as the representative language due to its large speaker base among the 22 official languages. For Singapore, we included all four official languages: English, Chinese, Malay, and Tamil. In total, MMAC-bench covers ten languages; full details are available in Appendix A.2.

3.3 Text and Image Data Collection

For both the text-only and multimodal tracks, teams generated at least 56 keywords or short phrases per category as *cultural prompts*. If a category could not supply enough prompts, the shortfall was filled using prompts from other categories. These prompts were designed to capture both diversity and geographic breadth. Using these prompts, team members retrieved relevant texts and images from the web and extracted short passages to serve as the basis for question authoring. When source

content was ambiguous, we cross-checked with multiple references to ensure authenticity. All images were obtained from Creative Commons (CC)-licensed resources (see details in Appendix A.2).

3.4 Question Creation and Annotation

Teams curated multiple-choice QA data with open-ended explanations based on collected materials; over 60% require multi-step cultural reasoning, defined as sequential derivation or synthesis rather than single-fact recall (detailed examples in Appendix A.2). To ensure stylistic variety, rigid templates were avoided. To preserve stylistic diversity, question templates were not fixed. Each national subset comprises a *multimodal* and a *text-only* component. All items were authored in official local languages and translated into English with manual verification, preserving culture-specific terms via phonetic transliteration or local renderings when no standard equivalents exist.

Multimodal Component. Annotators created VQA items where the correct answer requires visual understanding. Each VQA question was rephrased into a semantically-equivalent text-only MCQ. The answer options and the correct answer were kept unchanged. We provided both the original and rephrased items in English and the local language. We also generated speech inputs by converting the text to audio using high-quality TTS systems, supplemented with a subset of human-recorded utterances. (Appendix A.5 provides TTS toolkit and speech data building details). For Spoken QA, we considered two configurations: (i) converting only the question stem to speech while keeping textual options, and (ii) converting both the stem and the options to speech. To preserve comparability with VQA under controlled variables and to reduce ambiguity introduced by fully spoken options, our main experiments adopt the “spoken stem + textual options” configuration across five evaluation dimensions. Results for the fully spoken setting (spoken stem and options) on the test set are reported in Appendix A.16 for reference. To reflect accent effects, we produced English speech in both accent-neutral and locally accented versions.

Text-only Component. This component contains questions that were not suitable for pairing with an image or are inherently text-based. We applied the same requirement that at least 60% of questions involve multi-step reasoning. All questions were created in multiple languages.

For each question in multimodal and text-only

components, annotators identified the requisite knowledge points, the minimal information needed for the correct answer, typically summarized in a few sentences (shown in Fig. 1). These knowledge points are used for model’s cultural awareness grounding validation (Sec. 4.3).

3.5 Human Review and Revision

After each country team completed a draft, in-country linguists conducted quality reviews. The review covered: ambiguity in wording, accuracy of English translations, clarity and fluency of the speech data, completeness of knowledge points, and appropriateness of answer options. Details of the review procedures and consistency checks are provided in Sec. A.4. Teams revised their subsets based on this feedback, yielding the final high-quality release.

4 Experiments

We evaluate existing LLMs under the MMAC framework in a unified zero-shot setting, using a shared prompt template that matches the question language unless stated otherwise (see Appendix A.8 for prompts and experimental details). We report results along five dimensions and, for each, analyze the factors that drive performance. Following established protocols from prior benchmarks (Romero et al., 2024; Vayani et al., 2025; Myung et al., 2025; Wang et al., 2024), we select one representative model from each model family to ensure systematic comparison. Our evaluation includes three closed-source models: GPT-4o (OpenAI et al., 2024), Claude Sonnet 4, and Gemini 2.5 Flash (Comanici et al., 2025), and eleven open-source multilingual or multimodal models, including the Qwen (Bai et al., 2025; Yang et al., 2025), LLaMA (Touvron et al., 2023), and GLM (GLM et al., 2024; Team et al., 2025c) families. Models and tasks are detailed in Appendix A.7. For models without multilingual support, we report English-only scores for comparability. For speech evaluation, we include only models that accept *speech tokens* directly; models that require intermediate automatic speech recognition (ASR) are excluded.

4.1 Main Results

The MMAC framework offers 2 evaluation modes: MCQ-only and MCQ+Explanation (open-ended generation). The former facilitates model evaluation and reflects performance trends across

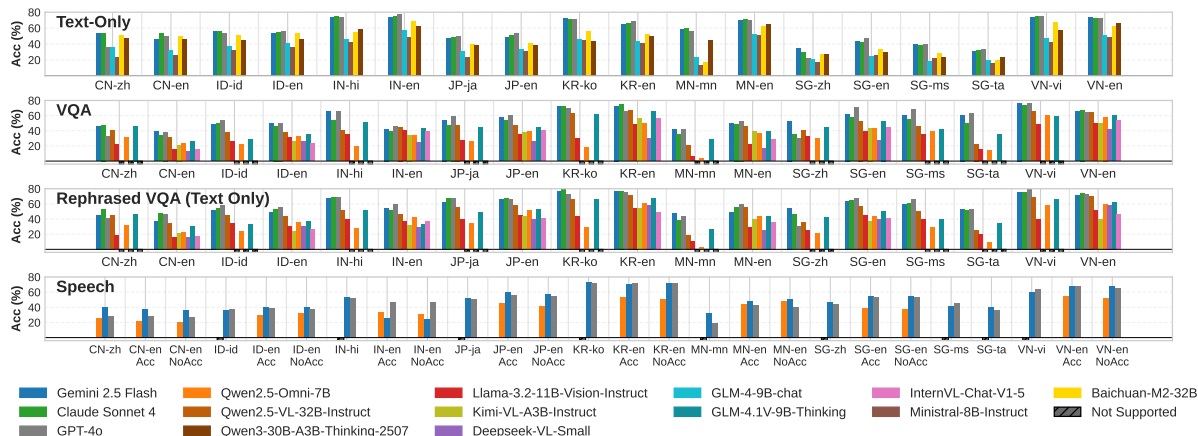


Figure 2: **Performance of LLMs on MMAC-bench across modalities.** Exact values are provided in Tables 19, 20, 21, 22. Each country-modality pair has 500 multilingual questions. Y-axis: Accuracy (%) = correct matches / 500. X-axis {Country}–{Language}: cultural dataset and presentation language.

different cultures, while the latter reveals the groundedness of the model’s cultural awareness. Fig. 2 summarizes MCQ accuracy on MMAC-bench. Nearly all state-of-the-art models score below 80% and most below 50%, exposing the limitations of existing LLMs in grasping diverse cultural nuances.

4.2 Cultural Awareness Disparity

Across countries (language factor controlled).

To isolate country effects, we fix the modality and compare each model across countries. For each country, we report the model’s highest score over its available languages (Fig. 2) to avoid conflating cultural competence with language proficiency. Models show higher awareness for Korean culture on average (63.98% across 4 modalities), plausibly due to global diffusion and richer data availability (Jang et al., 2024; Dal Yong, 2018). Remarkably, Vietnamese culture (62.96%) is on par with Korean, likely reflecting Vietnam’s high social-media penetration (79.8% of the population) and thus large volumes of user-generated content (DataReportal et al., 2025). In contrast, China and India exhibit larger gaps, consistent with multilayered cultural forms and greater regional heterogeneity. Mongolia trails further, consistent with low-resource settings and sparser training corpora. **Across languages (within culture).** English prompts often outperform low-resource languages, reflecting the breadth of English corpora and limited cross-lingual transfer (Hu et al., 2025; Weihua et al., 2025). This advantage diminishes or reverses for medium/high-resource local languages (e.g., Chinese, Japanese), where culture-specific

terms and proper names are well represented locally but rare in English corpora, hurting retrieval and grounding. For example, “乌护 (Wuhu),” a lineage among Uyghur ancestral groups, lacks a standard English equivalent; transliteration is rare and ambiguous in English data. Thus, when the model is competent in the relevant local language, using that language can yield better cultural grounding than English.

Across modalities (holding language fixed).

We observe a consistent ordering: *text-only* > *VQA* > *spoken QA*. Data availability follows the same order (text ≫ image–text ≫ raw speech). Speech adds uncertainty (noise, homophony), and many architectures encode modalities separately and then fuse downstream, introducing alignment/compression losses that widen gaps. Interestingly, in Speech, Qwen and Gemini outperform their standard English baselines in 6 and 5 country-specific cultural settings, respectively (Fig. 2). Accents appear to serve as a prior cue for specific cultures, enhancing the models’ accuracy on corresponding tasks. We attribute this to the co-occurrence of accents and their related cultural content within the data (see Appendix A.18 for detailed analysis).

From the foregoing analysis, it is evident that LLMs in Asian cultural contexts also display cultural and modality biases shaped by data distributions; furthermore, given the limited effectiveness of cross-lingual cultural knowledge transfer, English cannot be assumed to perform reliably better on culture-related tasks. In contrast, in the speech modality, accents, often treated as noise, paradoxically serve as effective cultural cues that activate relevant context and improve performance. Further

analysis on cultural category impacts and culture-language relationships appears in Section A.13.

4.3 Cultural Awareness Grounding Validation

MCQs are convenient but can be solved via shortcuts (e.g., option elimination) rather than grounded knowledge (Myung et al., 2025; Romero et al., 2024; Wang et al., 2024, 2025; Hu et al., 2026). To mitigate this, we adopt two measures: (i) **Retained background knowledge**: Each item accompanied by its *knowledge points* (supporting evidence). (ii) **Explainable responses**: Models must provide a textual rationale for their choice.

We employ a majority-voting LLM-as-Judge mechanism to verify whether the generated explanation accurately reflects the supporting knowledge points. To reduce variability across judges, we use English-only questions and require explanations in English (see Appendix A.11 for judge details, human consistency checks, and the multimodel consistency evaluation). Sec. A.12 details the advantages and risks of this setup. We introduce the *Rationale Unfaithfulness Rate* (RUR) to quantify spurious success, defined as the proportion of correct answers accompanied by explanations that contradict or omit the required knowledge.

As shown in Fig. 3, even proprietary models (Claude, GPT, Gemini) exhibit an RUR between 5% and 20%. The discrepancy is more severe in open-source models: Llama-3.2-11B-Vision-Instruct and Qwen3-30B-A3B-Thinking have the highest RUR, with Llama’s rate particularly elevated on non-English inputs. Qwen3-30B-A3B-Thinking often produces long explanations that contain hallucinations and sometimes derives the correct option from premises that contradict the ground truth, indicating reliance on generic heuristics rather than culturally grounded reasoning. These findings underscore that MCQs alone are insufficient metrics for cultural understanding. Consequently, we advocate for an *MCQ+Explanation* evaluation protocol, where a model is credited only if both the option and the rationale are correct. By removing spurious successes, this method provides a more faithful estimate of cultural awareness. Full results for these evaluations are presented in the Appendix, Tabs. 23, 24, 25, 26.

4.4 Cultural-Awareness Consistency Across Languages

Cross-lingual cultural-awareness consistency measures model output consistency across seman-

tically equivalent multilingual prompts, regardless of accuracy (Eq. 1) (Wang et al., 2024).

$$\text{Cons.}_s = \frac{1}{N} \sum_{i=1}^N \frac{1}{\binom{m_i}{s}} \sum_{\substack{S \subseteq L_i \\ |S|=s}} \mathbf{1}(|\{a_i^\ell : \ell \in S\}| = 1), \quad (1)$$

where L_i denotes the set of languages available for question i , and $m_i = |L_i|$ is its cardinality; $S \subseteq L_i$ with $|S| = s$ denotes any size- s language subset; a_i^ℓ is the model’s answer to question i when prompted in language ℓ ; $\mathbf{1}(\cdot)$ is the indicator function; and $|\{a_i^\ell : \ell \in S\}| = 1$ asserts that all answers within S are identical. If $m_i < s$, there are no valid subsets and item i contributes zero.

Fig. 4a shows that multilingual LLMs lack cross-lingual consistency on cultural tasks in text and VQA, particularly for language pairs with significant resource gaps. For Mongolian culture, the disparity between Mongolian and English yields only 65.2% consistency for Claude, while all open-source models remain below 50%. By contrast, Korean culture shows higher consistency, plausibly reflecting the global diffusion of contemporary Korean media and the resulting multilingual exposure to related knowledge (Jang et al., 2024; Dal Yong, 2018). Consistency also declines sharply as the number of evaluated languages increases. For Singapore-related items, when Chinese, English, Tamil, and Malay are assessed jointly, the maximum consistency does not exceed 45% (Gemini on VQA), despite relatively high pairwise values of 60.60% (EN-TA), 64.20% (EN-MS), and 55.20% (EN-ZH). In some culturally challenging cases, visual context can partially bridge languages: for Indian culture, GLM-4.1 achieves 44.20% cross-lingual consistency on Hindi VQA, which is 13.2 points higher than its rephrased text-only counterpart, although both remain low.

Consistency relies on resources and cultural exposure, worsening with resource asymmetry but improving with cultural prominence. It shows nonlinear decay across languages, where pairwise agreement implies no global coherence. Visual cues can narrow gaps in certain settings, but are insufficient to overcome the structural limitations of low-resource languages.

4.5 Cultural-Awareness Consistency Across Modalities

Cross-modal cultural-awareness consistency measures output invariance across modalities for semantically equivalent queries. Mirroring cross-lingual trends, Fig. 4b shows that model-averaged

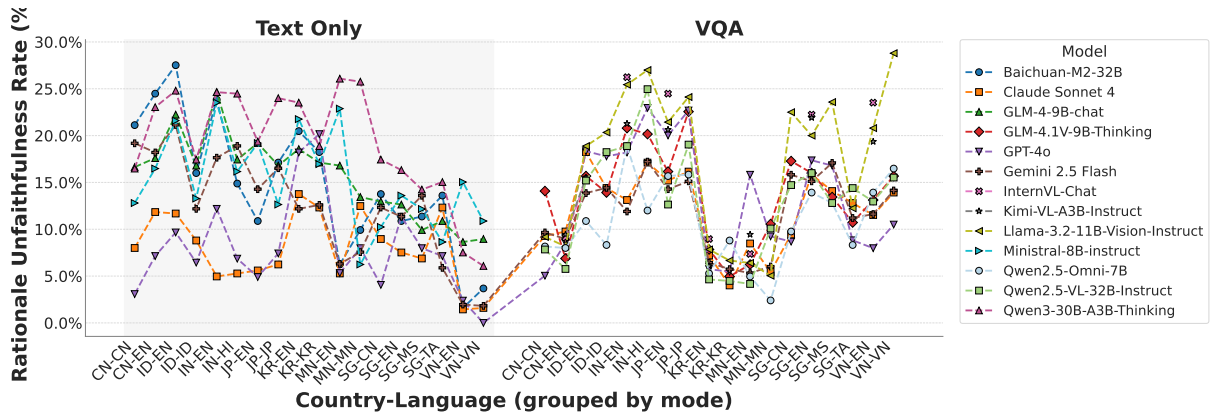


Figure 3: **Rationale Unfaithfulness Rates (RUR) of LLMs across text-only and VQA.** Similar trends are observed for Rephrase VQA and Spoken QA; detailed results are provided in Appendix A.17.

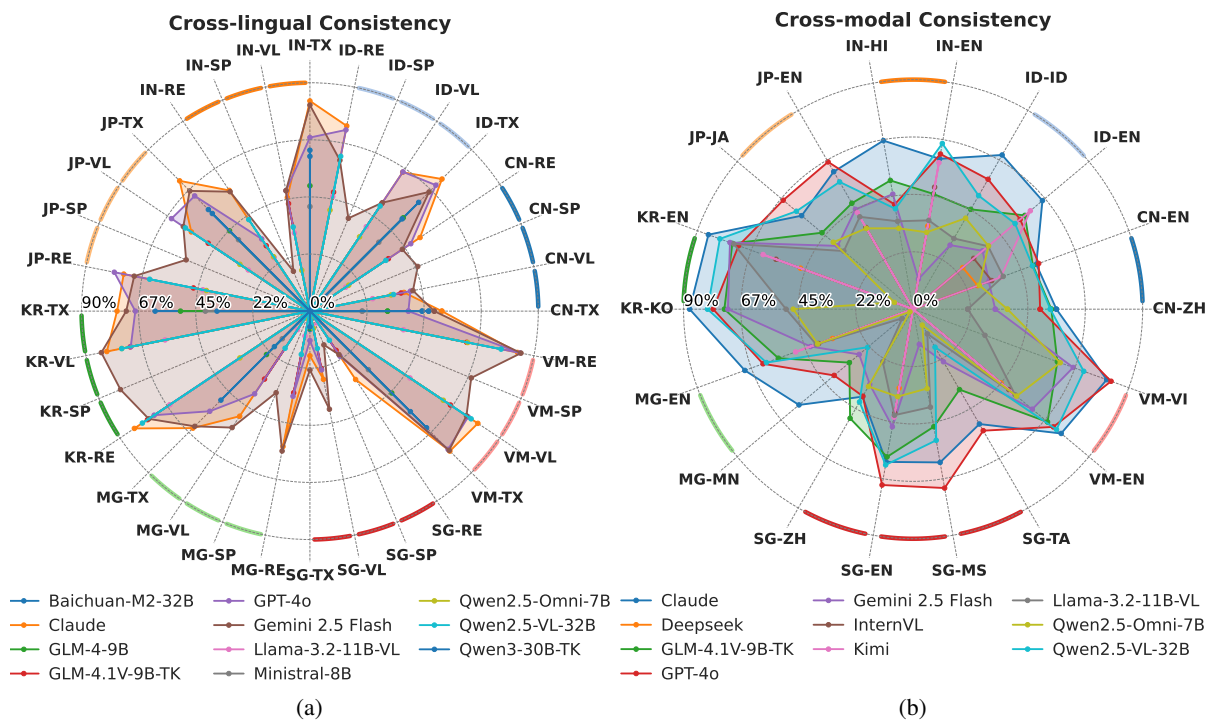


Figure 4: (a) **Cross-lingual consistency with fixed country and modality** and (b) **cross-modal consistency with fixed language and country.** TX/VL/RE/SP represent text QA, visual QA, rephrase QA, and speech QA.

cross-modal consistency across eight Asian countries remains below 67%, suggesting asymmetric cultural transfer. Under low-resource language settings, almost all models struggle to maintain stable cross-modal answers. Within the same national context, medium- to high-resource local languages typically yield higher cross-modal consistency than English. To examine the observed ordering *text-only* > *VQA* > *spoken QA* (see Sec. 4.2), we conduct a detailed error analysis with Qwen2.5-VL-32B-Instruct. Because spoken QA adds additional complexities (noise, accents, intonation), our analysis in this section focuses on VQA versus text-only

performance; we leave a fuller study of speech to future work.

We specifically isolate instances where the model succeeded with pure text input but failed in the VQA context. Apart from the most common errors arising from a lack of understanding of culture-related image contexts, our analysis reveals two additional predominant categories of errors:

The pitfall of prompt-guided selective attention. Models often over-rely on explicit prompt objects, whereas cultural VQA demands the interpretation of culture-specific visual cues. To assess whether model focus aligns with these cues, we

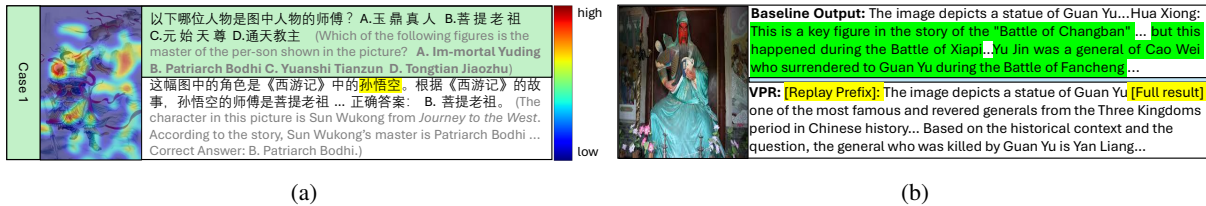


Figure 5: (a) **Attention heatmap visualization over image regions during incorrect model answers.** Color scale from blue (low) to red (high) indicates increasing model attention. (b) **Comparison of baseline vs. “VPR” outputs, with hallucinations highlighted in green,** when prompted with “*In official history, which general listed below was killed by the person pictured?*” The full results of (a) and (b) are presented in Fig. 19 and 11.

extract visual evidence using answer-conditioned multi-layer Grad-CAM (Selvaraju et al., 2019), interpolating the resulting heatmaps and mapping them back to the original images for visualization (see Appendix A.9 for implementation details).

This phenomenon is clearly revealed by the heatmaps in Fig. 5a, showing how the model focuses on subjects explicitly mentioned in the prompt when answering questions. However, this selective attention can unfortunately lead the model to overlook other critically important local details within the image, consequently resulting in erroneous inferences. In Case 1 (Fig. 5a), when the model is presented with the question “Which of the following figures is the master of the person shown in the picture?” in Chinese, its attention is predominantly drawn to the figure on the left side of the image, who is explicitly referenced in the prompt. However, the model overlooks the Howling Celestial Dog in the upper-right corner of the image, a crucial clue for identifying Erlang Shen. Additional attention-visualization examples and analyses are given in Appendix A.19.

Image tokens contribute to reasoning hallucinations. We find that visual content increases reasoning hallucinations in models compared to text-only QA, despite accurate image recognition capabilities. In Fig. 5b, while the baseline model correctly identifies “Guan Yu”, it still produces multiple reasoning hallucinations (highlighted in green). However, text-only Rephrase VQA queries show no such hallucinations (Appendix A.10), indicating that reasoning errors likely originate from image tokens and suggest modality-specific bias in multimodal reasoning. To test this hypothesis, we propose Vision-ablated Prefix Replay (VPR), which generates image descriptions then removes visual conditions while maintaining fixed prefixes for subsequent reasoning (details in Appendix A.10). In Fig. 5b, VPR conditions the model’s reasoning on the generated caption ‘The image depicts a statue of

Guan Yu’ while discarding visual tokens, eliminating hallucinations and producing the correct answer. Furthermore, we select 200 questions where VQA answers contained hallucinations but corresponding text-only queries were correct. VPR eliminates hallucinations and produced correct answers for 77 of these cases (38.5%), supporting our hypothesis. Multimodal models demonstrate significant inconsistency in cultural awareness across modalities, indicating flawed cultural knowledge transfer. In VQA, this deficiency stems from two core issues: “selective attention pitfall” where models over-focus on text-prompted subjects while missing key visual cues, and visual token-induced “reasoning hallucinations”.

4.6 Cultural Knowledge Generalization

Prior work (Balepur et al., 2024; Molfese et al., 2025; Zheng et al., 2023) suggests that scaling increases LLMs’ factual memory but not genuine logical generalization. To distinguish whether cultural multi-step reasoning errors stem from knowledge gaps or generalization deficits, we conduct a deconstruction study. We decompose each question into atomic sub-questions testing single knowledge points, which models answer first (see Tab. 10 for an example). We then evaluate the original question under two conditions: (i) with in-context “sub-question → model answer” pairs, and (ii) from scratch. If models solve all subquestions but fail the original question from scratch, it demonstrates that they possess the underlying knowledge but struggle to integrate it, highlighting generalization deficits. If they err on sub-questions, failure likely reflects missing culture-specific knowledge.

We analyze the failure modes of Claude Sonnet 4, GLM-4, Mistral, and Qwen3 on multi-step reasoning tasks across English, Chinese, and Indonesian. By employing a majority-voting LLM-as-Judge to validate decomposed sub-questions, we identify distinct failure patterns driven by resource

Model	En			Zh			Id		
	SC	IF	SF	SC	IF	SF	SC	IF	SF
Claude	35.3	2.6	62.1	41.1	3.5	55.4	20.2	9.2	70.6
Qwen	18.5	10.7	70.8	35.5	2.6	61.9	13.4	3.7	82.9
Mistral	18.9	1.4	80.7	25.2	1.0	73.8	4.4	1.1	94.5
GLM	30.9	2.0	67.1	40.5	2.8	56.7	6.5	5.3	88.2

Table 2: Cultural knowledge generalization results (%). **SC**: correct sub-questions and final answer; **IF**: correct sub-questions but incorrect synthesis; **SF**: errors in initial reasoning.

availability and architecture (see Tabs. 2 and 12).

The “Lost in Integration” Phenomenon: We observe a significant misalignment between factual retrieval and logical synthesis. Specifically, models often answer sub-questions correctly yet fail the final synthesis. This “integration failur” is notable in Qwen’s English context (10.7%) and Claude’s Indonesian context (9.2%), significantly higher than the $\sim 2\%$ baseline in other models. This indicates that as task complexity increases, logical integration collapses faster than fact retention.

Divergent Failure Mechanisms in Cultural Generalization: Cultural generalization remains difficult across all metrics. In high-resource scenarios, closed-source models lack generalization despite broad knowledge, while open-source models struggle with both knowledge and transferability gaps. In low-resource scenarios, the disparity widens: closed-source models show sharp declines in generalization, while open-source failures are overwhelmingly driven by knowledge deficits ($>88.5\%$), underscoring severe data scarcity.

5 Conclusion

We introduced MMAC, a framework centered on MMAC-bench, a high-quality, human-curated benchmark characterized by its symmetrical trimodal structure across 8 Asian countries and 10 languages. Using MMAC, we conducted evaluations using a five-dimensional protocol that assesses disparity, consistency, generalization, and grounding validity. Results show persistent data-driven cultural bias, uneven cross-lingual transfer, and fragile multimodal reasoning (selective visual attention and image-induced hallucinations). Meanwhile, accented speech can act as a useful cultural cue. Models also struggle to integrate known facts into multi-step reasoning, indicating a generalization bottleneck. We argue for consistency- and grounding-aware evaluation, alongside methods that strengthen cross-modal alignment and cover-

age in low-resource languages. MMAC provides data, protocols, and baselines to track progress toward culturally reliable multimodal LLMs.

Limitations

This work proposes MMAC, a multi-faceted framework designed to evaluate the cultural awareness of Large Language Models (LLMs). Evaluated against an initial benchmark of eight Asian countries, the framework features a scalable data creation pipeline capable of adapting to any cultural context. Currently, the assessment is conducted through Multiple-Choice Questions (MCQs) and free-form explanation generation. While this study currently focuses on a subset of Asian cultures, future research will aim to extend this coverage to Western regions and additional linguistic groups, while simultaneously incorporating more diverse evaluation modalities to capture a more holistic view of cultural intelligence.

Acknowledgments

This research project is supported by the National Research Foundation, Singapore, under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

This research project is supported by the National Research Foundation, Singapore, under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

This research is supported by A*STAR under its Japan-Singapore Joint Call: Japan Science and Technology Agency (JST) and A*STAR 2024 (R24I6IR136).

This project is supported by Microsoft Research via the Agentic AI Research and Innovation (AARI) Initiative.

Tanmoy Chakraborty acknowledges the financial support of Anusandhan National Research Foundation (ANRF/ARG/2025/001505/QSS).

Fan Xu acknowledges the Natural Science Fund project in Jiangxi province under Grant 20224ACB202010.

Ethics Statement

Bias and Representation: Our benchmark is designed to *reveal* culture-, modal- and language-dependent performance gaps of LLMs in Asian contexts, including biases shaped by imbalanced training data and uneven cross-lingual/cross-modal transfer. At the same time, MMAC may still reflect residual biases in web-sourced materials (e.g., overrepresentation of mainstream narratives) or biases introduced by question framing; some culturally grounded items could also inadvertently reinforce stereotypes if not carefully curated. To mitigate these risks, we focus on verifiable cultural facts, require evidence-backed “knowledge points” for each item, and apply human review to identify and remove problematic or misleading content.

Data Provenance and Annotation: MMAC is collaboratively constructed by in-country expert teams. Annotators are native speakers of the local language, proficient in English, and have lived in the corresponding country for more than ten years; all annotators received project-wide briefings and followed detailed annotation guidelines. For data sourcing, teams used culturally diverse prompts to retrieve texts and images from the web and cross-checked ambiguous claims with multiple references to ensure authenticity. All images are drawn from Creative Commons (CC)-licensed resources permitted for research use. Questions are authored in local languages and translated into English with manual verification, preserving key terms via transliteration or local renderings where necessary. Speech inputs are generated with high-quality TTS, including both accent-neutral and locally accented English to study accent effects. We additionally filter hate speech and misinformation and anonymize privacy-related information, including privacy-preserving image processing when applicable.

Intended Benefit: We aim to provide a transparent benchmark and evaluation protocol that helps diagnose cultural bias, fragile cross-lingual transfer, and cross-modal misalignment, rather than merely ranking models. By pairing multiple-choice questions with evidence-like “knowledge points” and requiring explainable rationales that are checked against these points, the benchmark encourages culturally grounded reasoning and discourages shortcut solving. We will release the dataset and supporting resources for research use to facilitate reproducibility, bias analysis, and future methods that

broaden high-quality coverage for lower-resource languages and improve cross-modal cultural understanding.

References

- Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreuzer, Marzieh Fadaee, and Sara Hooker. 2024. [The multilingual alignment prism: Aligning global and local preferences to reduce harm](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. [Artifacts or abduction: How do LLMs answer multiple-choice questions without the question?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.
- Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2025. [Long chain-of-thought reasoning across languages](#). *Preprint*, arXiv:2508.14828.
- Xi Chen, Wei Xue, and Yike Guo. 2026. [Actormind: Emulating human actor reasoning for speech role-playing](#). *Preprint*, arXiv:2604.11103.
- Xi Chen and Min Zeng. 2025. [Prototype conditioned generative replay for continual learning in NLP](#). 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 12754–12770, Albuquerque, New Mexico. Association for Computational Linguistics.

- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. **Cultural-bench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming**. *Preprint*, arXiv:2410.02677.
- Seungho Cho, Changgeon Ko, Eui Jun Hwang, Junmyeong Lee, Huije Lee, and Jong C. Park. 2025. **Language over content: Tracing cultural understanding in multilingual large language models**. *Preprint*, arXiv:2510.16565.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Coqui.ai. 2025. Tts: A deep learning toolkit for text-to-speech. <https://github.com/coqui-ai/TTS>. Accessed: 2025-09-24.
- Jin Dal Yong. 2018. An analysis of the korean wave as transnational popular culture: North american youth engage through social media as tv becomes obsolete. *International Journal of Communication*, 12:404–422.
- DataReportal, Kepios, We Are Social, and Meltwater. 2025. **Digital 2025: Vietnam**. Accessed 21 Sep 2025.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.
- Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. **Md3: The multi-dialect dataset of dialogues**. *Preprint*, arXiv:2305.11355.
- Xiaoxue Gao, Huayun Zhang, and Nancy Chen. 2025. Multigen: Child-friendly multilingual speech generator with llms. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction*, pages 199–204.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. **Chatglm: A family of large language models from glm-130b to glm-4 all tools**. *Preprint*, arXiv:2406.12793.
- Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein, and Mrinmaya Sachan. 2025. **Multilingual performance biases of large language models in education**. *Preprint*, arXiv:2504.17720.
- Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, and 1 others. 2025. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*.
- Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. **Sok: Memorization in general-purpose large language models**. *Preprint*, arXiv:2310.18362.
- Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. **Large language models are cross-lingual knowledge-free reasoners**. *Preprint*, arXiv:2406.16655.
- Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. 2026. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications*.
- Moonkyoung Jang, Dokyung Kim, and Hyunmi Baek. 2024. More than just a fan: the influence of k-pop fandom on the popularity of k-drama on a global ott platform. *Applied Economics Letters*, 31(2):152–157.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. **Mistral 7b**. *Preprint*, arXiv:2310.06825.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. **CLiCK: A benchmark dataset of cultural and linguistic intelligence in Korean**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning*, pages 5530–5540. PMLR.

- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. [The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models](#). *Preprint*, arXiv:2404.16019.
- Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. 2025. [Déjà vu: Multilingual llm evaluation through the lens of machine translation evaluation](#). *Preprint*, arXiv:2504.11829.
- Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, and 1 others. 2025a. [Chemvlm: Exploring the power of multimodal large language models in chemistry area](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.
- Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025b. [Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 736–763.
- Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. 2026. [Re-track: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23373–23381.
- Kailin Liang, Bin Liu, Yifan Hu, Rui Liu, Feilong Bao, and Guanglai Gao. 2022. [Mnnts2: An open-source multi-speaker mongolian text-to-speech synthesis dataset](#). In *National Conference on Man-Machine Speech Communication*, pages 318–329. Springer.
- Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. [SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6119–6136, Albuquerque, New Mexico. Association for Computational Linguistics.
- Francesco Maria Molfese, Luca Moroni, Luca Giofrè, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. [Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering](#). *arXiv preprint arXiv:2503.14996*.
- 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, and 3 others. 2025. [Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages](#). *Preprint*, arXiv:2406.09948.
- 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](#). *Preprint*, arXiv:2407.10920.
- Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulgadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. [Sea-lion: Southeast asian languages in one network](#). *Preprint*, arXiv:2504.05747.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. 2023. [Facial geometric detail recovery via implicit representation](#). In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*.
- Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. 2025. [Multilingual and multi-accent jailbreaking of audio llms](#). *Preprint*, arXiv:2504.01094.
- 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, and 57 others. 2024. [Cvqa: Culturally-diverse multilingual visual question answering benchmark](#). *Preprint*, arXiv:2406.05967.
- Ashwin Sankar, Srija Anand, Praveen Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. [Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts](#). *Advances in Neural Information Processing Systems*, 37:68161–68182.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual llms think in english?](#) *Preprint*, arXiv:2502.15603.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,

- and Dhruv Batra. 2019. [Grad-cam: Visual explanations from deep networks via gradient-based localization](#). *International Journal of Computer Vision*, 128(2):336–359.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners](#). *Preprint*, arXiv:2210.03057.
- Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengaranjan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. [Sea-helm: South-east asian holistic evaluation of language models](#). *Preprint*, arXiv:2502.14301.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 76 others. 2025a. [Kimi-vl technical report](#). *Preprint*, arXiv:2504.07491.
- M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, and et al. Yijie Zhou. 2025b. [Baichuan-m2: Scaling medical capability with large verifier system](#). *Preprint*, arXiv:2509.02208.
- V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025c. [Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning](#). *Preprint*, arXiv:2507.01006.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadgign Ademteu, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. [All languages matter: Evaluating llms on culturally diverse 100 languages](#). *Preprint*, arXiv:2411.16508.
- 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.
- Xinyi Wang, Antonis Antoniadis, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. [Generalization v.s. memorization: Tracing language models’ capabilities back to pretraining data](#). *Preprint*, arXiv:2407.14985.
- Zheng Weihua, Xin Huang, Zhengyuan Liu, Tarun Kumar Vangani, Bowei Zou, Xiyan Tao, Yuhao Wu, AiTi Aw, Nancy F. Chen, and Roy Ka-Wei Lee. 2026. [Adamcot: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33863–33871.
- Zheng Weihua, Roy Ka-Wei Lee, Zhengyuan Liu, Wu Kui, AiTi Aw, and Bowei Zou. 2025. [CCL-XCoT: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1768–1788, Suzhou, China. Association for Computational Linguistics.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#). *Preprint*, arXiv:2412.10302.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. [MMLU-ProX: A multilingual benchmark for advanced large language model evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Zheng-Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen H. Bach, and Julia Kreutzer. 2025. [The state of multilingual llm safety research: From measuring the language gap to mitigating it](#). *Preprint*, arXiv:2505.24119.

Mingyu Zhang, Zixu Li, Zhiwei Chen, Zhiheng Fu, Xiaowei Zhu, Jiajia Nie, Yinwei Wei, and Yupeng Hu. 2026a. [Hint: Composed image retrieval with dual-path compositional contextualized network](#). *Preprint*, arXiv:2603.26341.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025a. [CM-align: Consistency-based multilingual alignment for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25689–25702. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025b. [Less, but better: Efficient multilingual expansion for LLMs via layer-wise mixture-of-experts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17948–17963, Vienna, Austria. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025c. [Multilingual knowledge editing with language-agnostic factual neurons](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2026b. [Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning](#). *Preprint*, arXiv:2510.07300.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025d. [Cross-modal information flow in multimodal large language models](#). *Preprint*, arXiv:2411.18620.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

Wenrui Zhou, Mohamed Hendy, Shu Yang, Qingsong Yang, Zikun Guo, Yuyu Luo, Lijie Hu, and Di Wang. 2025. [Flattery in motion: Benchmarking and analyzing sycophancy in video-llms](#). *Preprint*, arXiv:2506.07180.

A Appendix

Dataset Construction & Quality Assurance

A.1 Annotator Demographic

Our annotation team comprises members from eight different countries. All team members are native speakers of their local languages and proficient in English, with professional backgrounds in natural language processing or speech processing.

Each annotator has lived in the respective country for more than ten years. Table 3 presents the annotators’ details and professional backgrounds; to protect privacy, we replace personal names with numeric identifiers within each country. After each team completes the first round of annotations, in-country linguistic experts conduct a data review. Table 4 lists the language experts’ information; likewise, we anonymize personal names.

A.2 Data Statistics

Under the MMAC framework, the dataset covers 8 countries and 10 languages, with each country’s split presented in both English and its local language, totaling 27,000 questions. Table 5 lists the countries included in our dataset and their corresponding local language(s). Over 79% of all items are multi-step cultural reasoning questions. We define a multi-step cultural reasoning item as one whose solution requires sequential derivation and/or synthesis from at least two independent knowledge components, rather than mere recall or paraphrase of a single cultural fact. Table 6 presents an example multi-step question with its analysis. The proportion of multi-step items by country and modality is shown in Figure 7. The dataset spans nine categories—*Daily Life/Culture, Food/Cuisine, Transportation, Buildings, History, Geographical Location & Climate, Education, Fashion/Clothing, and Language/Ethnicity*—with per-country category distributions summarized in Figure 6.

The following outlines the potential contents included in each category.

- **Daily Life/Culture:** Covers everyday etiquette, customs, and values—greeting practices, daily routines and social decorum, family and community interactions, as well as festivals and folk practices. It may also include consumption and leisure preferences, common life scenarios, and behavioral norms.
- **Food/Cuisine:** Regional cuisines, representative dishes and ingredients, cooking techniques, dietary taboos and table manners, utensils, and dining settings. Also includes festive foods, street snacks, and regional taste differences.
- **Transportation:** Transportation across historical periods and the evolution of vehicles, along with regional differences in modes of transport. May also cover landmark transit systems and commuting culture.

Table 3: Data Annotator Demographics and Skills

Country	ID	Gender	Age	Education	English Prof.	Local Lang. Prof.	Professional Background
China	1	Male	28	Ph.D.	Proficient	Proficient	NLP
	2	Male	25	Master Deg.	Proficient	Proficient	NLP
	3	Male	30	Ph.D.	Proficient	Proficient	NLP
	4	Male	25	Master Deg.	Familiar	Fluent	NLP
	5	Male	24	Master Deg.	Familiar	Proficient	NLP
Japan	1	Male	23	Master Deg.	Proficient	Proficient	SP
	2	Male	35	Ph.D.	Fluent	Proficient	SP
Mongolia	1	Male	26	Ph.D.	Proficient	Proficient	SP
	2	Male	25	Master Deg.	Proficient	Proficient	SP
Korea	1	Male	31	Ph.D.	Proficient	Proficient	NLP
	2	Male	26	Master Deg.	Proficient	Proficient	NLP
	3	Female	25	Master Deg.	Proficient	Proficient	NLP
India	1	Male	25	Bachelor’s Deg.	Proficient	Proficient	NLP
	2	Male	31	Ph.D.	Proficient	Proficient	NLP
Vietnam	1	Male	21	Bachelor’s Deg.	Proficient	Proficient	SP
	2	Male	20	Bachelor’s Deg.	Proficient	Proficient	SP
	3	Male	21	Bachelor’s Deg.	Proficient	Proficient	SP
	4	Male	21	Bachelor’s Deg.	Proficient	Proficient	SP
	5	Male	29	Master Deg.	Proficient	Proficient	SP
Indonesia	1	Female	21	Bachelor’s Deg.	Proficient	Proficient	NLP
	2	Female	26	Ph.D.	Proficient	Proficient	NLP
Singapore	1	Female	26	Ph.D.	Proficient	Proficient	NLP
	2	Female	18	Bachelor’s Deg.	Proficient	Proficient	NLP
	3	Male	18	Bachelor’s Deg.	Proficient	Proficient	NLP
	4	Male	21	Bachelor’s Deg.	Proficient	Proficient	NLP
	5	Female	28	Ph.D.	Proficient	Proficient	SP

Note: NLP: Natural Language Processing; SP: Speech Processing.

Table 4: Data Reviewer Demographics and Skills

Country	ID	Gender	Age	Education	English Prof.	Local Lang. Prof.
China	1	Female	30	Master Deg.	Proficient	Proficient
Japan	1	Female	32	Master Deg.	Proficient	Proficient
Mongolia	1	Male	25	Master Deg.	Proficient	Proficient
Korea	1	Male	54	Bachelor’s Deg.	Proficient	Proficient
India	1	Male	28	Master Deg.	Proficient	Proficient
Vietnam	1	Male	27	Master Deg.	Proficient	Proficient
Indonesia	1	Female	28	Bachelor’s Deg.	Proficient	Proficient
Singapore	1	Female	26	Bachelor’s Deg.	Proficient	Proficient
	2	Female	22	Bachelor’s Deg.	Proficient	Proficient
	3	Male	30	Ph.D.	Proficient	Proficient
	4	Male	26	Ph.D.	Proficient	Proficient

Table 5: Question Distribution by Country and Modality

Country	Languages	Questions per Language	Questions per Modality (Text-Only/VQA/Speech)
China	Chinese, English	500	500/500/500
Singapore	English, Chinese, Malay, Tamil, English	500	500/500/500
Japan	Japanese, English	500	500/500/500
Korea	Korean, English	500	500/500/500
India	Hindi, English	500	500/500/500
Indonesia	Indonesian, English	500	500/500/500
Vietnam	Vietnamese, English	500	500/500/500
Mongolia	Mongolian, English	500	500/500/500

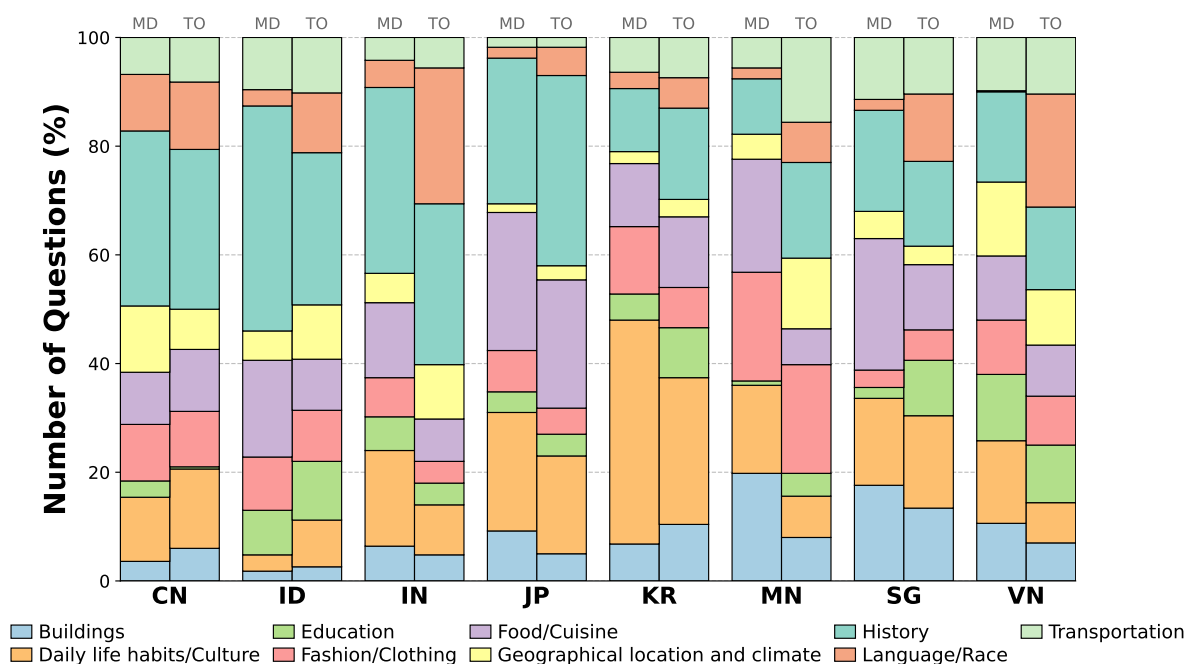


Figure 6: Distribution of question categories across countries and modalities

- **Buildings:** Traditional and modern architectural styles; religious and public buildings; housing forms and materials; city skylines and the preservation of historic districts. Can also address symbolic meanings in architecture and region-specific structural features.
- **History:** Major historical periods and events, notable figures and heritage sites, and how historical memory shapes contemporary society and culture. May also include colonial/independence histories and cultural change driven by migration and war.
- **Geographical Location & Climate:** Landforms and terrain, climate zones and seasonal variation, natural resources, and ecosystems. Extends to lifestyle, clothing, and dietary adaptations shaped by geography and climate.
- **Education:** Renowned national works of literature, art, or music, as well as the structure of the education system and pathways to advancement, including stories surrounding prestigious institutions.
- **Fashion/Clothing:** Traditional attire and its ceremonial contexts; modern dress styles and aesthetic trends; accessories and color preferences; occupational/school uniforms and seasonal clothing. May also discuss cultural symbolism embedded in garments.
- **Language/Ethnicity:** Code-switching (where applicable); official and commonly used lan-

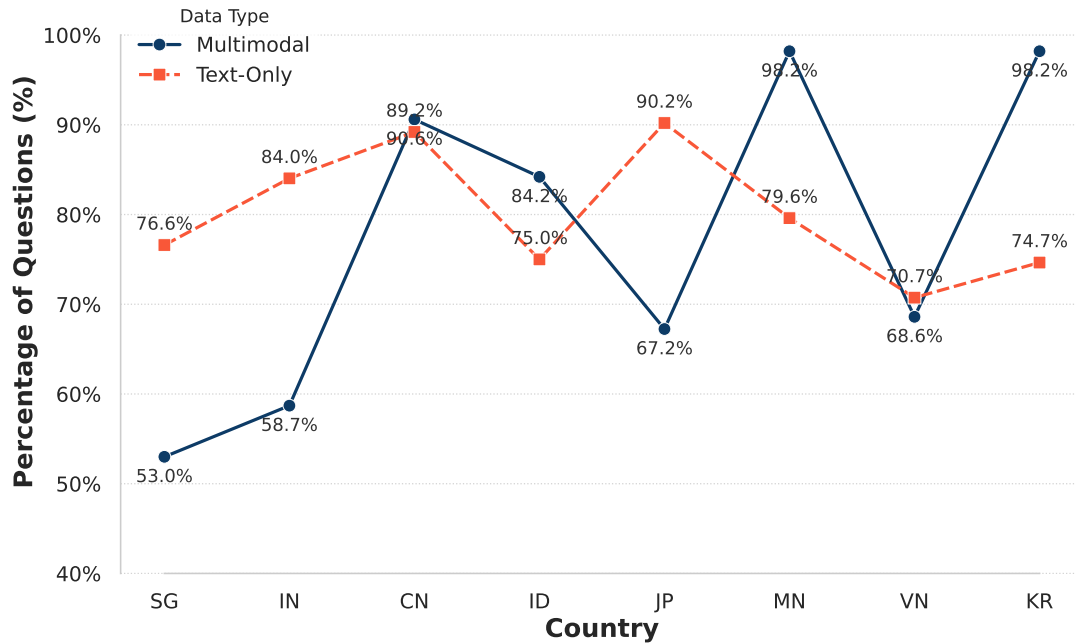


Figure 7: Multi-step reasoning question proportions across countries

guages; dialects and accent features; writing systems and naming conventions; multi-ethnic compositions and cultural practices. Also includes politeness strategies in language and norms of cross-group communication.

Figure 8 and Figure 9 show some samples of our dataset.

A.3 Annotation Guideline and Statistics of Translation Issues

Annotation Guideline. The content of the guideline distributed to annotators is shown in Figure 10. To minimize heuristic cues arising from non-cultural knowledge, we add a consistency constraint on distractors: they must belong to the same category as the correct option and closely resemble it in observable attributes and semantic representation. We also encourage each team to uncover cultural elements unique to their own country, rather than focusing only on widely known aspects. For the Language category, if code-switching is prevalent in the annotators’ country, we strongly encourage including such language-assessment examples in the Text-Only portion of the dataset. All content involving racism or hate speech is prohibited from inclusion in our dataset.

Statistics of Translation Issues. We quantified the primary challenges encountered when translating from local languages into English. The distribution of errors is as follows:

- **Terminology translation errors** – 23%
- **Culture-specific terms with no direct translation** – 61%
- **Literal mistranslation of culture-related vocabulary** – 16%

To address terminology translation errors, we correct them by cross-referencing dictionaries and consulting expert linguists. For terms lacking a direct equivalent, we employ transliteration. Regarding literal mistranslations of culture-related vocabulary, we first paraphrase the original expression to capture its equivalent meaning before translating.

A.4 Data Quality Assurance and Annotation Consistency

Multi-Stage Verification Pipeline. To ensure the high fidelity of our dataset, each generated instance underwent a multi-stage verification pipeline. Following initial creation, every question was subjected to cross-verification by two independent annotators within the same data sub-group. Instances with unanimous agreement were approved directly; in cases of discrepancy, a linguistic expert native to the respective country adjudicated the conflict to reach a majority-based consensus.

Annotators evaluated each instance based on the following rigorous criteria:

- **Linguistic Clarity:** Verification that the

Table 6: An example of multi-step reasoning question

Question	An ancient tower became famous due to a poem by the Tang Dynasty poet Cui Hao. In which period was it proposed to be constructed with iron materials? (A) 16th year of Guangxu, Qing Dynasty (B) 1st year of Jiaqing, Qing Dynasty (C) 7th year of Tongzhi, Qing Dynasty (D) 8th year of the Republic of China
Reasoning Decomposition	<p>Step 1: Identify the tower. The poem is Cui Hao’s “Yellow Crane Tower,” so the tower is <i>Yellow Crane Tower</i>.</p> <p>Step 2: Recall historical events. In the 10th year of Guangxu (1884) the tower was destroyed by fire. In the 16th year of Guangxu (1890), Zhang Zhidong (Governor-General of Hubei and Hunan) first proposed rebuilding the tower using iron materials.</p> <p>Step 3: Match with the options. A (1890): Correct, matches the historical fact. B (1796): Incorrect, too early. C (1868): Incorrect, before the fire. D (1919): Incorrect, after the Qing Dynasty and not the first proposal.</p> <p>Final Answer: A. 16th year of Guangxu, Qing Dynasty.</p>

prompt is devoid of ambiguity or syntactic obscurities that might impede a definitive response.

- **Option Plausibility and Factuality:** Ensuring that the distractor set contains at least one viable answer and that the ground truth is anchored in verifiable, widely accepted facts.
- **Content Safety:** A dual-filter for hate speech, derogatory descriptions, or culturally insensitive content.
- **Visual Grounding (VQA-specific):** Confirming the modality dependency, ensuring the question cannot be resolved without the associated visual context.
- **Paraphrase Semantic Invariance:** Validating that rephrased questions maintain functional equivalence with the original VQA prompts and remain compatible with the pre-defined options.
- **Privacy Preservation:** Screening for the inclusion of any Personally Identifiable Information (PII).
- **Acoustic Fidelity (Speech-specific):** Assessing the generated audio for intelligibility, clarity, and adherence to standard phonetic norms.
- **Knowledge Self-Sufficiency:** Determining whether the provided knowledge metadata contains sufficient information to deduce the correct answer without external leakage.

Human Verification Protocols. To maintain the integrity of our evaluation, we implemented the following “blind” testing protocols:

- **Ablation-based VQA Validation:** To verify image dependency, annotators attempted to answer VQA questions with the image withheld. Questions that remained answerable were flagged as “visually deficient” and discarded or revised.
- **Paraphrase Robustness:** Rephrased questions were audited to ensure they retained the necessary visual descriptors required to bridge the question and the image content.
- **Cross-Regional Knowledge Validation:** To mitigate annotator bias, specifically the risk of annotators inadvertently relying on their own internal knowledge to fill gaps in the data, we introduced a cross-regional check. In addition to internal team reviews, an evaluator from a different national background was assigned to verify the “Knowledge Completeness” for each country-specific instance. This ensured that the provided knowledge points were truly self-contained and sufficient for an outsider to reach the correct conclusion.

The intra-group review consistency, as measured by Cohen’s kappa coefficient, is reported in Table 7.

A.5 Details of the TTS Tool and Procedure for Building the Speech Data

Before generating the English audio, we standardized the input text through normalization of numbers and symbols, handling of abbreviations and special terms, and sentence segmentation.

To ensure high-quality speech synthesis, we curated a suite of state-of-the-art TTS tools tailored to the linguistic characteristics of each language. As demonstrated by the performance metrics in Table 8, the various speech generation tools deliver

Knowledge Point	Source Question	English Question
The Mogao Caves, also known as the Thousand Buddha Grottoes, feature the "Transformation Tableau of the Medicine Buddha Sutra" in Cave 220, which depicts scenes from the High Tang period, and the "Mural of Zhang Yichao's Army in Procession" in Cave 156, which showcases scenes from the Late Tang period.	某石窟有千佛洞之称，其在唐代有以下哪些壁画作品？ A. 《药师经变画》 B. 《炽盛光佛图》 C. 《张议潮统军出行图》 D. 《五台山图》	Which mural paintings from the Tang Dynasty are found in a certain grotto known as the Thousand Buddha Caves? A. Transformation Tableau of the Medicine Buddha Sutra B. Painting of Tejaprabhā Buddha C. Mural of Zhang Yichao's Army in Procession D. Mount Wutai
Gyeongbokgung Palace was built with Bukaksan as its main mountain, strategically placing its buildings on a spacious site with Gwanghwamun as its main gate, opening onto a wide boulevard that formed the center of Hanyang, the capital of the Joseon Dynasty. Gyeongbokgung was destroyed during the Imjin War in 1592, the 25th year of King Seonjo's reign.	백악산을 주산으로 넓은 지형에 건물을 배치하고 정문 앞으로 넓은 육조거리가 펼쳐진 한양의 중심이 1592년 소실된 계기는 무엇인가? A. 6.25 전쟁 B. 화재 발생 C. 임진왜란 D. 을미사변	What was the cause of the destruction in 1592 of the central area of Hanyang, where buildings were laid out across a wide terrain with Bugaksan as the main mountain and a broad Yukjo Street stretching out in front of the main gate? A. Korean War B. Fire outbreak C. Imjin War D. Eulmi Incident
Knowledge Point	Source Question	English Question
Char Kway Teow: This stir-fried flat noodle dish often contains prawns. It is typically served with lime. Fried Hokkien Mee: a stir-fried dish of yellow noodles and rice vermicelli cooked in a rich stock made from pork bones and prawn heads. It is almost always served with a wedge of calamansi lime on the side.	பின்வரும் பொதுவான சிங்கப்பூர் நூடுல்ஸ் உணவுகளில் எது அடிக்கடி இறால் மற்றும் எலுமிச்சை சேர்க்கப்பட்டிருக்கும்? A. சார்க்வே டியோ B. லெக்சா C. வறுத்த ஹொக்கியன் மீ D. கிரே.பிஷ் நூடுல்ஸ்	Which of the following common Singaporean noodle dishes often include prawns and lime? A. Char kway teow B. Leksa C. Fried Hokkien Mee D. Crayfish Noodles
The word "Gamcha" is Bengali/Assamese word which comes from two very simple and commonly used Bengali/Assamese words, (গা) ga which means "Body", nd (মুছা) mucha which means "wipe". It is often used as traditional full sized handkerchief. And it looks like a towel and acts as one too in case of necessity, it acts as a mask which means dust and pollution stays away from you, it also acts as sun protection so that you don't get much tanned or skin burn.	रवि दक्षिण एशिया में उष्णकटिबंधीय जलवायु के लिए अनुकूलित पारंपरिक परिधानों का अध्ययन कर रहे हैं। उन्होंने पश्चिम बंगाल में ग्रामीण पुरुषों के एक समूह को पसीना पोंछने, धूप में सिर ढकने और यहां तक कि अस्थायी मास्क के रूप में हल्के सूती कपड़े का उपयोग करते हुए देखा। वह संभवतः किस पारंपरिक वस्तु का उल्लेख कर रहे हैं? A. कर्ता B. धोती C. गमछा D. शेरवानी	Ravi is studying traditional garments in South Asia adapted for tropical climates. He notices a group of rural men in West Bengal using a light cotton cloth for wiping sweat, covering their heads in the sun, and even as a makeshift mask. Which traditional item is he likely referring to? A. Kurta B. Dhoti C. Gamcha D. Sherwani

Figure 8: Text-Only Question Examples

high-fidelity and natural-sounding speech, meeting the quality requirements of real-world applications: low error rates (e.g., a WER of 3.71 for English and a CER of 2.10 for Tamil) and high naturalness, with Mean Opinion Scores (MOS) consistently exceeding 4.30 for Mongolian, Vietnamese, and Indonesian.

For English, we employed CosyVoice (Du et al., 2025), which supports voice cloning from sampled speakers and can preserve the timbre and accent of a reference voice. We collected representative recordings from native speakers across eight countries to capture diverse accents for English speech synthesis. For standard English, we adopted CosyVoice's built-in default English voice

(female). CosyVoice was also used to generate Chinese, Japanese, and Korean audio. For Tamil and Malay, we used the TTS system proposed in (Gao et al., 2025), while Mongolian speech was synthesized using the tool described in (Liang et al., 2022). Vietnamese audio was generated using an in-house high-quality TTS system developed by our speech processing team. In addition, the Coqui-ai TTS toolkit (Coqui.ai, 2025) was used for Indonesian and Hindi. Table 9 summarizes the models used for speech generation across different languages.

Each generated audio sample was individually verified. When errors occurred, such as inappropriate pauses, missing segments, or mispronun-



Knowledge Point	Image	Question
Multicolor woodblock printing enabled full-color ukiyo-e and mass popularity; in Edo, portraitists like Sharaku and Toyokuni thrived, and yakusha-e and bijin-ga became bestsellers. And most yakusha-e are a subset of kabuki-e.		Question: この画像のような、絵をなんというでしょう？ A.美人絵 B.墨摺絵 C.歌舞伎絵 D.役者絵 Rephrased Question: 浮世絵で歌舞伎役者や当時の人気俳優の姿を描いた絵を何というでしょう？
Tugu Proklamasi (the Proclamation Monument) is a monument commemorating the Proclamation of Independence of the Republic of Indonesia. It stands within the Proclamation Park complex on Jalan Proklamasi in Central Jakarta. The house where the proclamation was read was demolished in the 1960s.		Di lokasi pada gambar dulu ada sebuah rumah; kapan rumah itu dirobohkan? A. Tepat setelah hari kemerdekaan Indonesia B. Tahun 1960-an C. Tahun 1970-an D. Awal tahun 2000 Rephrased Question: Kapan rumah yang terletak di tugu proklamasi tersebut dibongkar?

Figure 9: VQA/Rephrase VQA Question Examples

ciations. We first adjusted the input text and re-synthesized the audio, as TTS systems are often highly sensitive to textual variations. If repeated corrections still failed, we resorted to manual re-recording. Unlike in other language tasks, our requirement here was not fluency or naturalness, but rather clear articulation of the questions and answer options.

All audio files were standardized in WAV format with a 16 kHz sampling rate. And our speech generation uses two input types.

1. Only the question stem from the *Rephrase VQA (text-only)* item;
2. The entire *Rephrase VQA (text-only)* item, including the question and its answer options.

A.6 Sub-question Decomposition

Table 10 illustrates examples of deconstructing multi-step reasoning questions into single-fact verification subquestions. We enumerate all single-fact subquestions embedded in each multi-step item to assess the model’s generalization.

Experimental Setup & Evaluation Methodology

A.7 Model Selection

Open-source multilingual text-only LLMs. We evaluate Qwen3-30B-A3B-Thinking-2507 (Yang

et al., 2025), Baichuan-M2-32B (Team et al., 2025b), GLM-4-9B-Chat (GLM et al., 2024), and Ministral-8B-Instruct (Jiang et al., 2023) for their multilingual capabilities.

Open-source multimodal LLMs. We evaluate nine vision–language (and omni) models on image and text: Qwen2.5-VL-32B-Instruct (Bai et al., 2025), Llama-3.2-11B-Vision-Instruct (Touvron et al., 2023), Kimi-VL-A3B-Instruct (Team et al., 2025a), DeepSeek-VL-Small (Wu et al., 2024), GLM-4.1V-9B-Thinking (Team et al., 2025c), InternVL-Chat-V1-5 (Chen et al., 2024), and Qwen2.5-Omni-7B (Xu et al., 2025). Kimi-VL-A3B-Instruct, DeepSeek-VL-Small, and InternVL-Chat-V1-5 are evaluated in English only; Qwen2.5-VL-32B-Instruct, Llama-3.2-11B-Vision-Instruct, and GLM-4.1V-9B-Thinking are evaluated in multiple languages. Qwen2.5-Omni-7B is evaluated across image, text, and speech in a multilingual setting. We summarize the information of the LLMs evaluated in our experiments in Table 11.

A.8 Prompt Templates and Experimental Settings

Prompt templates for evaluation tasks across modalities. Table 13 presents the English prompts used in our evaluations across different modalities. When the query switches to another language, the corresponding translated version of the prompt will be used to ensure input-language

Table 7: Inter-annotator Agreement and Statistics by Dimension

Dimension (Country & Modality)	n_{11}	n_{22}	n_{12}	n_{21}	P_o	P_e	Kappa (κ)
China (Text-Only)	460	25	8	7	0.9700	0.8784	0.7532
China (VQA)	455	32	6	7	0.9740	0.8573	0.8178
China (Rephrase VQA)	440	45	10	5	0.9700	0.8122	0.8404
Japan (Text-Only)	448	42	4	6	0.9800	0.8291	0.8830
Japan (VQA)	462	30	4	4	0.9840	0.8732	0.8738
Japan (Rephrase VQA)	480	15	2	3	0.9900	0.9322	0.8524
South Korea (Text-Only)	420	65	10	5	0.9700	0.7506	0.8797
South Korea (VQA)	415	72	8	5	0.9740	0.7379	0.9008
South Korea (Rephrase VQA)	450	38	6	6	0.9760	0.8396	0.8504
Singapore (Text-Only)	462	30	4	4	0.9840	0.8738	0.8732
Singapore (VQA)	490	6	2	2	0.9920	0.9685	0.7459
Singapore (Rephrase VQA)	485	12	1	2	0.9940	0.9477	0.8854
Indonesia (Text-Only)	435	52	8	5	0.9740	0.7917	0.8752
Indonesia (VQA)	442	48	5	5	0.9800	0.8090	0.8953
Indonesia (Rephrase VQA)	410	75	10	5	0.9700	0.7256	0.8906
Mongolia (Text-Only)	405	82	7	6	0.9740	0.7095	0.9105
Mongolia (VQA)	425	60	8	7	0.9700	0.7672	0.8711
Mongolia (Rephrase VQA)	430	58	6	6	0.9760	0.7760	0.8929
India (Text-Only)	458	32	5	5	0.9800	0.8636	0.8533
India (VQA)	465	28	3	4	0.9860	0.8824	0.8809
India (Rephrase VQA)	418	72	6	4	0.9800	0.7435	0.9220
Vietnam (Text-Only)	422	70	5	3	0.9840	0.7507	0.9358
Vietnam (VQA)	445	45	6	4	0.9800	0.8196	0.8891
Vietnam (Rephrase VQA)	452	42	3	3	0.9880	0.8360	0.9268

Table 8: Performance Metrics Across Different Languages

Language	Performance		
	WER	CER	MOS
English	3.71	–	–
Chinese	–	3.01	–
Japanese	–	5.27	–
Korean	–	4.01	–
Mongolian	–	–	4.58
Vietnamese	–	–	4.36
Hindi	–	9.40	–
Tamil	–	2.10	–
Malay	–	2.30	–
Indonesian	–	–	4.43

consistency. Table 14 presents the prompts used to invoke closed-source model APIs for translation, answer-consistency evaluation, and answer extraction.

Experiments setting. All evaluations in this work are conducted in the zero-shot setting, using single-turn inference for each model on an NVIDIA H100 80G. For image inputs that exceed a model’s maximum allowable resolution, we proportionally downscale the image until it is under 1 megapixel before testing. We decode with greedy search (no sampling; `do_sample=false`, `num_beams=1`), so

temperature/top- p /top- k are not used; the maximum output length is set to 2048 tokens to ensure reproducibility. GPT-4o and Gemini 2.5 Flash are accessed via OpenRouter API platform, and Claude via the Anthropic API. For all closed-source models, we set temperature=0 to minimize randomness and improve reproducibility. Our speech inputs are no longer than 30 seconds and sampled at 16 kHz—well below Gemini 2.5 Flash’s maximum speech-input duration and Qwen2.5-Omni-7B’s maximum input token limit. So we do not perform any input-length processing.

A.9 Answer-conditioned multi-layer Grad-CAM

To assess whether model focus aligns with these cues, we extract visual evidence using answer-conditioned multi-layer Grad-CAM (Selvaraju et al., 2019), interpolating the resulting heatmaps and mapping them back to the original images for visualization.

Specifically, given an image v , a textual prompt x and an answer $y_{a:b} = (y_a, \dots, y_b)$ autoregressively produced by a model with parameters θ , we define an answer-conditioned objective on the log-likelihood of the answer tokens as Eq. 2. Here

Table 9: Models used for speech data generation

Language	Model
English with accent, Korean, Japanese, Chinese	CosyVoice2-0.5B (Du et al., 2025)
English without accent	CosyVoice-300M (Du et al., 2024)
Vietnamese	Internally developed TTS systems
Tamil, Malay	MultiGen (Gao et al., 2025)
Mongolian	MnTTS2 (Liang et al., 2022)
Indonesian	Indonesian TTS (Kim et al., 2021)
Hindi	AI4Bharat Indic-TTS (Sankar et al., 2024)

Table 10: Sub-question decomposition of multi-step reasoning questions.

Multi-Step Question	Sub-Question	Answer for Sub-Question
Which of the following buildings belongs to the same ethnic characteristic architecture as the stilted building? A. Tulou B. Moxiaolou C. Diaofang D. Yaodong	What ethnic group is associated with the stilted building?	The Tujia ethnic group.
	What ethnic group is associated with the Tulou?	Hakka
	What ethnic group is associated with the Moxiaolou?	The Tujia ethnic group.

we use the token-sum objective. For memory efficiency, only the visual tower is set to require gradients. We denote the forward activation of the ℓ -th visual block as $A(\ell) \in \mathbb{R}^{C \times H \times W}$, the gradient of the block is $\mathbf{G}^{(\ell)}$. A per-layer Grad-CAM is built via channelwise inner product followed by ReLU (Eq. 3).

$$J_{\text{sum}} = - \sum_{t=a}^b \log p_{\theta}(y_t | v, x, y_{<t}) \quad (2)$$

$$\mathbf{M}^{(\ell)} = \text{ReLU} \left(\sum_{c=1}^C \mathbf{G}_c^{(\ell)} \odot \mathbf{A}_c^{(\ell)} \right) \in \mathbb{R}^{H \times W} \quad (3)$$

After lightweight smoothing and normalization of $\mathbf{M}^{(\ell)}$, the resulting $\hat{\mathbf{M}}^{(\ell)}$ then undergoes cross-layer aggregation and bilinear interpolation upsampling on the k blocks at the end of the visual tower, as defined by Eqs. 4 and 5, to the original image resolution. For more robust results, we specifically choose the last 3 blocks and do cross-layer aggregation by averaging them.

$$\mathbf{M} = \text{Agg}_{\ell \in \mathcal{L}} (\hat{\mathbf{M}}^{(\ell)}) \quad (4)$$

$$\tilde{\mathbf{M}} = \text{Bilinear}(\mathbf{M}, H_{\text{img}}, W_{\text{img}}) \in [0, 1]^{H_{\text{img}} \times W_{\text{img}}} \quad (5)$$

A.10 Vision-ablated Prefix Replay

We found that visual content increases reasoning hallucinations in models compared to text-only QA, despite accurate image recognition capabilities. To validate our hypothesis, we propose a ‘‘Vision-ablated Prefix Replay’’ (VPR) method. This method enables a model with parameters θ to first describe the image contents based on image \mathbf{x}^{img} and text prompt \mathbf{x}^{text} . After this initial description, we structurally ablate the visual condition and fix the prefix $\hat{\mathbf{S}}_{1:n}$ for subsequent reasoning generation.

This evaluates the marginal contribution of visual conditions to reasoning. Specifically, assuming the model completes image description within the first n tokens, we remove visual conditions starting from the $(n + 1)$ -th token and use only the text prompt and generated tokens as prefix. The joint probability distribution of the subsequent sequence $\mathbf{S}_{>n} = (s_{n+1}, \dots, s_T)$ can be expressed as:

$$\begin{aligned} p_{\theta}(\mathbf{S}_{>n} | \mathbf{x}_{\text{text}}, \emptyset, \hat{\mathbf{S}}_{1:n}) \\ = \prod_{t=n+1}^T p_{\theta}(s_t | \mathbf{x}_{\text{text}}, \emptyset, \hat{\mathbf{S}}_{1:n}, \mathbf{s}_{n+1:t-1}). \end{aligned} \quad (6)$$

Previous work has explored related ideas. For example, Zhang et al. (2025d) attempt to suppress visual leakage by blocking the attention paths to image-token positions during decoding. However, this cannot fully eliminate the influence of visual content: in a causal language model, the information of earlier image tokens is encoded into subsequent question tokens, so residual visual information remains even when attention to image tokens is blocked. In contrast, our method first elicits a textual description of the image, then removes the visual input and recomputes the representations of the prefix tokens, thereby purging visual information and enabling a more precise assessment of the image content’s contribution to the model’s reasoning process.

For the question in Figure 5b, the completed responses under different generation modes are provided in Table 15.

Table 11: Overview of selected LLMs

Model	Type	Language Coverage	Modalities
Qwen2.5-VL-32B-Instruct	Vision–Language (VL)	Multilingual	Image + Text
Llama-3.2-11B-Vision-Instruct	Vision–Language (VL)	Multilingual	Image + Text
Kimi-VL-A3B-Instruct	Vision–Language (VL)	English only	Image + Text
DeepSeek-VL2-Small	Vision–Language (VL)	English only	Image + Text
GLM-4.1V-9B-Thinking	Vision–Language (VL)	Multilingual	Image + Text
InternVL-Chat-V1-5	Vision–Language (VL)	English only	Image + Text
Qwen2.5-Omni-7B	Omni-modal	Multilingual	Image + Text + Speech
Qwen3-30B-A3B-Thinking-2507	Text-only LLM	Multilingual	Text
Baichuan-M2-32B	Text-only LLM	Multilingual	Text
GLM-4-9B-Chat	Text-only LLM	Multilingual	Text
Ministral 8B-Instruct	Text-only LLM	Multilingual	Text
GPT-4o	Vision–Language (VL)	Multilingual	Image + Text
Claude Sonnet 4	Vision–Language (VL)	Multilingual	Image + Text
Gemini 2.5 Flash	Omni-modal	Multilingual	Image + Text + Speech

Table 12: Number of Incorrect Cases Across Different Models and Languages.

Model	Number of Incorrect Cases		
	English	Chinese	Indonesian
GLM-4-9B-chat	337	320	312
Mistral-8B-Instruct	370	384	341
Claude Sonnet 4	232	231	222
Qwen3-30B-A3B	271	265	276

A.11 Consistency Analysis and Hyperparameters Setting for LLM-as-Judge

Majority-voting LLM-as-Judge To mitigate potential bias from any single evaluator, we employ a majority-voting LLM-as-a-Judge framework to assess the consistency between model-generated explanations and their corresponding knowledge points. We selected three frontier models, GPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4, to serve as our independent judges. For each instance, the knowledge point and generated explanation are fed into the models, which are then tasked with providing a direct binary consistency verdict. The final judgment is determined by the majority consensus among the three models. The specific prompts are provided in Table 14. Before adopting the LLM-as-judge paradigm, we conducted a small-scale study to assess human–model agreement and inter-model agreement. This majority-voting scheme yields near-perfect alignment with human annotations (Cohen’s $\kappa = 0.913$), supporting its reliability for mitigating single-judge bias. In addition, we further investigated the accuracy of using the LLM-as-Judge method in English versus non-English

languages.

Human–model agreement. We sampled 100 items from the dataset, each comprising a multi-step reasoning question and its decomposed sub-questions. Three annotators independently judged whether the model’s answer was semantically consistent with the gold answer for each (binary: *yes/no*). For each sub-question, the human judgment was determined by majority vote. We then queried **Claude Sonnet 4**, **GPT-4o**, and **Gemini 2.5 Flash** via API to obtain their judgments on the same items. An item was counted as *consistent* for a model only if the model’s judgments for all sub-questions and the final question matched the human judgments. Results showed human–model agreement of **98%** for GPT-4o, **98%** for Claude, and **96%** for Gemini.

Inter-model agreement on English Data. Given the binary nature of the task and the observed human–model agreement rates (98%, 98%, 96%), the conservative lower-bound on inter-model agreement is **96%**.

Inter-model agreement on Tamil and Malay Data. We further examined inter-model agreement in Malay and Tamil. Specifically, the agree-

Table 13: English prompts for different modals.

Modal	Prompt Template
Text-Only/Rephrase	Please answer the following culture-related question.\n{question}\n{options}\nThis is a multiple-choice question. Please first return all possible option letters, then explain your choice in English.
VQA	Based on the image, please answer the following culture-related question.\n{question}\n{options}\nThis is a multiple-choice question. Please first return all possible option letters, then explain your choice in English.
Speech question & text options	This is a culture-related question.\n Based on the question mentioned in this audio, please choose the correct answers from the following provided options. {options}\nThis is a multiple-choice question. Please first return all possible option letters, then explain your choice in English.
Speech question & options	This is a culture-related question. Based on the question and options mentioned in this audio, please choose the correct options. This is a multiple-choice question. Please first return all possible option letters, then explain your choice in English.

Table 14: Prompts Used with Closed-Source APIs for Translation, Answer Extraction, and Answer–Reference Consistency Checking

Task	Prompt
Translation	Translate the following sentence into English. \n{Input sentence}\nThis is a multiple-choice question in the cultural domain of {Country}. Pay particular attention to the terms in the input and use their official translations; if no official translation exists, you may use transliteration. Ensure accuracy, faithfulness, and fluency. Return only the translation; do not include any additional hints or analysis.
LLM-as-Judge	You are an expert evaluator. Your task is to determine if the 'Model Answer' correctly and completely incorporates the information from the 'Knowledge Point'. Knowledge Point:\n{knowledge_point}\nModel Answer\n{model_answer}\n\n. If the Model Answer correctly and completely includes the information from the Knowledge Point, or if the model's response aligns with the Knowledge Point's content and viewpoint. And the Model Answer doesn't contain any factual error, Answer 'yes'. Otherwise Answer 'no'. Please only return 'yes' or 'no'.
Answer Extraction	This is a model's response to a multiple-choice question. First, understand the text, then extract the model's chosen options, returning only the option letters (e.g., A, B, C). Do not include the option content. Output the result in a format like [A, B]. If the response does not contain any final choice, return NA.

Table 15: Model outputs across different modes for the same question.

Mode	Answer
Rephrase VQA baseline	<p>Answer: B. Yan Liang. Explanation: In historical records and the classic Chinese novel <i>Romance of the Three Kingdoms</i>, Guan Yu is known for his legendary bravery and martial prowess. One of his most famous feats occurred during the Battle of Hulao Gate (虎牢关之战) in 194 CE, which took place during the late Eastern Han dynasty. Here’s the context: Hua Xiong: During the Battle of Hulao Gate, Hua Xiong was a general serving under Dong Zhuo, a powerful warlord who controlled the Han imperial court. Hua Xiong challenged the coalition forces led by Yuan Shu, Yuan Shao, and others. Guan Yu volunteered to fight Hua Xiong and defeated him in battle. This event is well-known as one of Guan Yu’s early displays of martial skill. Yan Liang and Wen Chou: After defeating Hua Xiong, Guan Yu continued to distinguish himself in battle. He later encountered two other generals, Yan Liang and Wen Chou, who were also serving under Dong Zhuo. In the same battle, Guan Yu fought and killed both Yan Liang and Wen Chou. This event further solidified his reputation as a formidable warrior. Yu Jin: Yu Jin was a general who served under Cao Cao, a rival of Guan Yu’s lord, Liu Bei. While Yu Jin and Guan Yu did face each other in battle, Yu Jin was not killed by Guan Yu. Instead, Yu Jin surrendered to Guan Yu during the Battle of Fancheng (樊城之战) in 219 CE, after being trapped by water flooding caused by Guan Yu. Why the Correct Answer is B (Yan Liang): Yan Liang was explicitly mentioned in historical records and the <i>Romance of the Three Kingdoms</i> as one of the generals killed by Guan Yu during the Battle of Hulao Gate. The other options do not align with the historical context: Hua Xiong was defeated but not killed by Guan Yu in the same battle. Wen Chou was also killed by Guan Yu, but the question asks for a single correct answer, and Yan Liang is more commonly associated with this specific event. Yu Jin was not killed by Guan Yu; he surrendered instead. Thus, the correct answer is B. Yan Liang.</p>
VQA baseline	<p>The image depicts a statue of Guan Yu, one of the most famous and revered generals from the Three Kingdoms period in Chinese history. Guan Yu is known for his loyalty, bravery, and righteousness, and he is often associated with the Shu Han kingdom. Key Points to Consider: Guan Yu’s Achievements: Guan Yu is celebrated for his military prowess and is known for several significant victories during the Three Kingdoms era. Hua Xiong (华雄): This is a key figure in the story of the “Battle of Changban” (or “Battle of Changping”). According to historical records and the novel <i>Romance of the Three Kingdoms</i>, Guan Yu famously defeated Hua Xiong in battle. Hua Xiong was a general serving under Dong Zhuo, and his defeat marked a turning point in the coalition against Dong Zhuo. Analysis of the Options: A. Hua Xiong: This is correct. Hua Xiong was indeed killed by Guan Yu during the Battle of Changban. This event is well-documented in both historical records and the <i>Romance of the Three Kingdoms</i>. B. Yan Liang (颜良): While Guan Yu did kill Yan Liang, this occurred later in his career, during the Battle of Baima (White Horse). However, the question specifically asks about the person in the picture, and the most iconic victory associated with Guan Yu is his defeat of Hua Xiong. C. Wen Chou (文丑): Wen Chou was another general defeated by Guan Yu, but this happened during the battle of Xiapi, which came after the Battle of Changban. D. Yu Jin (于禁): Yu Jin was a general of Cao Wei who surrendered to Guan Yu during the Battle of Fancheng. He was not killed by Guan Yu. Final Answer: The person in the picture (Guan Yu) is most famously known for killing Hua Xiong. Therefore, the correct answer is: A.</p>
Vision-ablated Prefix Replay	<p>The image depicts a statue of Guan Yu, one of the most famous and revered generals from the Three Kingdoms period in Chinese history. Guan Yu is known for his loyalty, bravery, and martial prowess. Analysis of the question: We need to determine which general was killed by Guan Yu based on historical records. Hua Xiong (华雄): Hua Xiong was a general serving under Dong Zhuo during the late Eastern Han dynasty. He was killed by Guan Yu in the Battle of Hulao Gate, as described in the <i>Romance of the Three Kingdoms</i>. This event is one of the most iconic moments in Guan Yu’s early career. Yan Liang (颜良): Yan Liang was a general serving under Yuan Shao. He was killed by Guan Yu during the Battle of Baima (白马之战) in 200 CE. This battle was part of the larger conflict between Cao Cao and Yuan Shao. Wen Chou (文丑): Wen Chou was another general serving under Yuan Shao. He was also killed by Guan Yu shortly after the Battle of Baima, during the same campaign against Yuan Shao. Yu Jin (于禁): Yu Jin was a general serving under Cao Cao. He was not killed by Guan Yu but rather surrendered to Guan Yu during the Battle of Fancheng in 219 CE when Guan Yu’s forces surrounded him. Yu Jin later returned to Cao Cao’s service after Guan Yu’s defeat. Correct Answer: Based on the historical context and the question, the general who was killed by Guan Yu is Yan Liang. Final Answer: B.</p>

ment rate is 88% for Malay (disagreement: 12%) and 72% for Tamil (disagreement: 28%). These results reinforce the observation that, in mid- to low-resource languages, agreement is systematically weaker than in high-resource languages such as English. This trend holds not only for Human–Model agreement, but also for agreement across models, with the reduction being particularly severe in low-resource settings.

Motivated by this finding, we further explored whether prompting design could improve agreement in mid- to low-resource languages. Table 16 summarizes the three prompt variants considered in our experiments. We then evaluated their impact on both Human–Model agreement and inter-model agreement. The corresponding results are reported in Table 17.

As shown in Table 17, alternative prompt-

ing strategies lead to improvements in both Human–Model agreement and inter-model agreement for LLM-as-a-Judge in low- and mid-resource languages. For instance, three-model agreement rises to 89% for Malay and 77% for Tamil. Nevertheless, these gains are limited in magnitude, and the resulting agreement levels remain below the reliability threshold needed to support trustworthy evaluation.

Human–model agreement rates under different language settings. Prior work has extensively evaluated the performance of different LLMs on the same tasks across different languages, revealing significant gaps between English and low-resource language settings (Gupta et al., 2025; Xuan et al., 2025; Zhang et al., 2025a,c). Xuan et al. (2025) utilize the MMLU-ProX benchmark to demonstrate that leading models, including GPT-4o and Gemini, suffer a massive accuracy drop, up to 30%, when transitioning from English to low-resource languages. Furthermore, in the educational domain, (Gupta et al., 2025) report that GPT-4o and Gemini exhibit significant performance biases, struggling with complex pedagogical tasks such as feedback generation in languages like Telugu and Farsi compared to their English performance. We have also conducted experiments on using LLMs as evaluators in non-English scenarios. We compared the accuracy of using LLMs to evaluate outputs in English, Malay, and Tamil, with this evaluation performed on 100 semantically identical questions across different languages. We used human judgment as ground truth to calculate each model’s Human-Model consistency across different languages. As shown in Table 18, model evaluation accuracy in non-English scenarios is significantly lower than in English scenarios, especially in low-resource languages. Therefore, our evaluation experiments for model’s explanation are currently conducted only in English settings to ensure accurate reflection of model capabilities.

A.12 Analysis of English-only Rationales

In our Cultural Awareness Grounding Validation experiments, we set the model’s explanation language to English and analyzed the advantages and potential risks of this configuration.

Advantages of Standardizing English as the Explanation Language.

- **Mitigating Systematic Bias in LLM-as-a-Judge Frameworks:** Given that leading proprietary LLMs consistently demonstrate superior

evaluative capabilities in English, particularly when contrasted with their performance in low-resource linguistic settings, adopting English as the unified language for explanations serves to minimize stochastic noise and systematic bias inherent in automated cross-lingual evaluations. In Sec. A.11, we elucidate the performance of the LLM-as-a-judge paradigm across both English and non-English contexts by synthesizing prior literature with our own empirical findings.

- **Overcoming Expressive Bottlenecks in Low-Resource Contexts:** Prevailing research (Schut et al., 2025) posits that due to the English-centric nature of large-scale pre-training corpora, the internal latent representations of many multilingual LLMs operate primarily within an English-dominant semantic space, with cross-lingual alignment occurring predominantly in the final output layers. Furthermore, empirical evidence (Barua et al., 2025; Chen and Zeng, 2025; Li et al., 2025b; Shi et al., 2022; Li et al., 2026; Zhang et al., 2025b, 2026b) indicates that model performance during reasoning tasks (e.g., Chain-of-Thought) in English often matches or exceeds that in non-English languages. Consequently, utilizing English as the medium for explanations provides a more high-fidelity window into the model’s conceptual understanding, effectively bypassing the expressive constraints and performance degradation often associated with low-resource languages.

Potential Risks and Limitations. We acknowledge that in specific domain-sensitive or culturally nuanced contexts, a model’s reasoning may be intrinsically tied to the idiosyncratic lexicon of a particular language. In such instances, enforcing English-only explanations may induce *semantic drift*, potentially leading to mischaracterizations by the LLM-judge and subsequent evaluative discrepancies.

In light of these considerations, we argue that English explanations constitute a robust “latent-to-explicit projection” of a model’s internal reasoning. Although this approach carries a marginal risk of semantic attenuation, it remains the most rigorous and comparable evaluation protocol within current benchmarking constraints. Future iterations of this work will involve expanded human-in-the-loop evaluations to extend the RUR framework to broader multilingual contexts.

Table 16: Alternative prompt variants introduced for evaluating answer consistency in Malay and Tamil.

Type of prompt	Prompt
Fine-grained logical verification	You are a rigorous Fact-Checking Auditor. Your task is to perform a granular comparison between the Knowledge Point and the Model Answer.\n\n Knowledge Point: {knowledge_point}\n Model Answer: {model_answer}\n\n Evaluation Protocol:\n 1. Deconstruction: Identify every individual fact, constraint, and relationship in the Knowledge Point.\n 2. Verification: Confirm if each identified unit is accurately reflected in the Model Answer.\n 3. Error Detection: Check for any added information in the Model Answer that contradicts or misinterprets the Knowledge Point.\n\n If the Model Answer is 100% consistent with the Knowledge Point without any omissions or factual errors, answer 'yes'. Otherwise, answer 'no'. \n Return only 'yes' or 'no'.
Mismatch Point Detection	You are an Adversarial Evaluator. Your goal is to find any discrepancy, contradiction, or factual hallucination in the Model Answer when compared against the Knowledge Point.\n\n Reference (Knowledge Point): {knowledge_point}\n Candidate (Model Answer): {model_answer}\n\n Rejection Criteria:\n - Does the Model Answer omit a critical qualifier (e.g., "always", "sometimes", "only")?\n - Does the Model Answer introduce external facts not supported by the Knowledge Point?\n - Is there any logical deviation from the source text?\n\n If you find even one factual discrepancy or significant omission, answer 'no'. If the Model Answer is a perfect and error-free representation of the Knowledge Point, answer 'yes'. \n Return only 'yes' or 'no'.
Semantic comparison	You are a Semantic Alignment Specialist. You must determine if the Model Answer and the Knowledge Point are semantically equivalent and factually synchronized.\n\n Knowledge Point: {knowledge_point}\n Model Answer: {model_answer}\n\n Judgment Logic:\n - The Model Answer must capture the full essence and all specific data points of the Knowledge Point.\n - Paraphrasing is acceptable as long as the truth value remains unchanged.\n - Any "hallucination" (info not present in the Knowledge Point) or "distortion" (info changed from the Knowledge Point) results in a failure.\n\n Answer 'yes' if the Model Answer is factually perfect and complete according to the Knowledge Point. Answer 'no' otherwise.\n Return only 'yes' or 'no'.

Table 17: Inter-model agreement in Malay and Tamil across prompt types.

Prompt type	Language	GPT-4o	Claude Sonnet 4	Gemini 2.5 Flash	Inter-Model
Fine-grained logical verification	Malay	90%	86%	93%	89%
	Tamil	72%	83%	78%	68%
Mismatch Point Detection	Malay	87%	85%	88%	84%
	Tamil	69%	73%	76%	67%
Semantic comparison	Malay	85%	88%	90%	86%
	Tamil	80%	76%	82%	77%

Table 18: Human–model agreement rates for different models used as LLM-as-judge under different language settings.

Language	GPT-4o	Claude Sonnet 4	Gemini 2.5 Flash
English	98%	98%	96%
Malay	90%	88%	92%
Tamil	76%	80%	72%

Additional Results & Detailed Data

A.13 Performance Variation Across Data Categories

To investigate how different cultural categories affect model performance, we evaluate models by category under each modality, taking the average accuracy across all languages within a modality as the category’s accuracy. As shown in Figures 12, 13, and 14, under the Text-Only and VQA modalities, models consistently struggle more with

the “Fashion/Clothing” and “Transportation” categories compared to others, while performing strongly on “Daily life habits/Culture” and “Education.” In the speech modality, models show slightly weaker performance on “Food/Cuisine” and “History,” but still maintain strong results on “Daily life habits/Culture” and “Education”.

These patterns likely reflect both data and task asymmetries across categories. Daily life habits/culture and education rely more on high-frequency, broadly documented facts and relatively

coarse-grained reasoning, making them easier for models trained on abundant, well-aligned multilingual data. By contrast, fashion/clothing and transportation encode highly time-sensitive and region-specific concepts (e.g., changing trends, local garment names, route and line names), while food/cuisine and history also depend on rare, culturally bound proper nouns; in the speech modality, models must directly map variable acoustic realizations of these low-frequency terms to the correct concepts, further depressing performance in these categories.

We further explore the relationship between cultural types and languages based on Figure 16. When Mongolian is employed as the input language, the model exhibits markedly superior performance in the education category relative to other categories. Conversely, when Hindi, Vietnamese, and Tamil are utilized as input languages, the model demonstrates suboptimal performance in the Fashion/Clothing, Geographical location and climate, and education categories, respectively.

We posit that the underlying cause of this phenomenon may be attributed to the fact that for low-resource languages such as Mongolian, casual text from the internet (e.g., social media, forums) is relatively scarce. High-quality corpora for such languages predominantly originate from government documents, digitized textbooks, academic papers, or Wikipedia. This compels the model to “absorb” a substantial volume of formal, educational texts during the training phase. Consequently, the model exhibits a form of “overfitting” with exceptionally high performance when processing the education category. In contrast, Hindi internet data may be dominated by news, politics, religion, or literature. Contemporary fashion and clothing-related content on the Indian internet likely exists primarily in English (or Hinglish, a Hindi-English code-mixed variety). If the model is trained exclusively on pure Hindi, it consequently lacks the vocabulary and contextual framework necessary to describe “fashion”. Similarly, if Vietnamese training corpora lack specific technical descriptions of geography and climate, the model will exhibit domain-specific knowledge gaps. Tamil, being widely used across diverse regions including Sri Lanka, India, and Singapore, encounters substantial educational disparities across these countries, and such regional conflicts may impede the model’s ability to adequately fit educational domain knowledge.

A.14 Performance of LLMs on MMAC-bench across modalities

The exact data corresponding to the bar chart in Section 4.2, Figure 2 are presented in Tables 19, 20, 21, 22.

A.15 Joint MCQ Answering and Explanation Performance of LLMs on MMAC-bench Across Modalities

The model performance based on both the MCQ choices and the model-generated explanations are provided in Tables 23, 24, 25, 26.

A.16 Results for fully spoken question and answering

We considered two configurations when constructing the TTS-Spoken QA dataset: (i) converting only the question stem to speech while keeping the answer options as text, and (ii) converting both the stem and the options to speech. To preserve comparability with VQA under controlled variables and to minimize ambiguity introduced by fully spoken options, our main experiments adopt the “spoken stem + textual options” setting across five evaluation dimensions. Results for the fully spoken setting (spoken stem and spoken options) on the test set are provided in Figure 17 and Table 27 for reference. We find that converting both the question and options to speech leads to a significant performance drop compared with the “spoken question + textual options” configuration, indicating that spoken options introduce greater uncertainty than the spoken question itself. This warrants further investigation in future work.

A.17 Rationale Unfaithfulness Rates Across Rephrase VQA and Spoken QA

The results for LLMs’ Rationale Unfaithfulness Rates (RUR) across Rephrase VQA and Speech are shown in Fig. 18. We observe patterns consistent with the Text-Only and VQA modalities: closed-source models generally have lower RURs than open-source models, though they still fall within the 5%–20% range. Among open-source models, Llama shows a markedly higher RUR on non-Spanish languages than on Spanish, which we attribute to linguistic bias stemming from the disproportionately large share of Spanish in Llama’s training data relative to other languages.

Table 19: Text-only modality performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500. “-” means “not support”. Results are presented as Mean $_{SE}$ [95% CI], where the bolded figure is the primary metric, followed by its standard error in subscript and the confidence interval in brackets. The better-performing result among different languages within the same country is **bolded**.

Model	CN-zh		CN-en		ID-id		ID-en	
Gemini 2.5 Flash	53.6 _{2.23}	[49.2, 57.9]	45.4 _{2.23}	[41.1, 49.8]	56.2 _{2.22}	[51.8, 60.5]	53.2 _{2.23}	[48.8, 57.5]
Claude Sonnet 4	53.8 _{2.23}	[49.4, 58.1]	53.6 _{2.23}	[49.2, 57.9]	55.6 _{2.22}	[51.2, 59.9]	54.8 _{2.23}	[50.4, 59.1]
GPT-4o	36.0 _{2.15}	[31.9, 40.3]	50.2 _{2.24}	[45.8, 54.6]	53.6 _{2.23}	[49.2, 57.9]	55.8 _{2.22}	[51.4, 60.1]
GLM-4-9B	36.0 _{2.15}	[31.9, 40.3]	32.6 _{2.10}	[28.6, 36.8]	37.6 _{2.17}	[33.4, 41.9]	49.4 _{2.24}	[45.0, 53.7]
Mistral-8B	23.2 _{1.89}	[19.7, 27.1]	26.0 _{1.96}	[22.3, 30.0]	31.8 _{2.08}	[27.8, 36.0]	35.6 _{2.14}	[31.5, 39.8]
Baichuan-M2	51.2 _{2.24}	[46.8, 55.5]	49.8 _{2.24}	[45.4, 54.1]	50.6 _{2.24}	[46.2, 54.9]	53.2 _{2.23}	[48.8, 57.5]
Qwen3-30B	47.0 _{2.23}	[42.6, 51.3]	45.8 _{2.23}	[41.4, 50.1]	44.8 _{2.22}	[40.5, 49.1]	46.2 _{2.23}	[41.8, 50.5]

Model	IN-hi		IN-en		JP-ja		JP-en	
Gemini 2.5 Flash	73.0 _{1.99}	[68.9, 76.7]	73.6 _{1.97}	[69.6, 77.3]	47.4 _{2.23}	[43.1, 51.8]	48.2 _{2.23}	[43.9, 52.6]
Claude Sonnet 4	74.4 _{1.95}	[70.4, 78.0]	74.2 _{1.96}	[70.2, 77.8]	48.8 _{2.24}	[44.4, 53.2]	50.6 _{2.24}	[46.2, 55.0]
GPT-4o	73.2 _{1.98}	[69.2, 76.9]	77.2 _{1.88}	[73.3, 80.7]	49.8 _{2.24}	[45.4, 54.2]	53.2 _{2.23}	[48.8, 57.5]
GLM-4-9B	45.6 _{2.23}	[41.2, 49.9]	57.0 _{2.21}	[52.6, 61.2]	30.2 _{2.05}	[26.3, 34.3]	33.6 _{2.11}	[29.6, 37.8]
Mistral-8B	42.2 _{2.21}	[37.9, 46.5]	49.0 _{2.24}	[44.6, 53.3]	23.6 _{1.90}	[20.0, 27.5]	30.2 _{2.05}	[26.3, 34.3]
Baichuan-M2	54.6 _{2.23}	[50.2, 58.9]	69.0 _{2.07}	[64.8, 72.9]	39.6 _{2.19}	[35.4, 43.9]	40.8 _{2.20}	[36.5, 45.1]
Qwen3-30B	58.6 _{2.20}	[54.2, 62.8]	62.2 _{2.17}	[57.8, 66.3]	38.4 _{2.18}	[34.2, 42.7]	38.0 _{2.17}	[33.8, 42.3]

Model	KR-ko		KR-en		MN-mn		MN-en	
Gemini 2.5 Flash	72.2 _{2.00}	[68.1, 76.0]	65.2 _{2.13}	[60.9, 69.2]	58.2 _{2.21}	[53.8, 62.4]	70.0 _{2.05}	[65.8, 73.9]
Claude Sonnet 4	71.4 _{2.02}	[67.3, 75.2]	66.0 _{2.12}	[61.7, 70.0]	59.4 _{2.20}	[55.0, 63.6]	71.0 _{2.03}	[66.9, 74.8]
GPT-4o	71.4 _{2.02}	[67.3, 75.2]	68.8 _{2.07}	[64.6, 72.7]	55.4 _{2.22}	[51.0, 59.7]	69.4 _{2.06}	[65.2, 73.3]
GLM-4-9B	45.4 _{2.23}	[41.0, 49.7]	42.8 _{2.21}	[38.5, 47.2]	23.0 _{1.88}	[19.5, 26.9]	52.6 _{2.23}	[48.2, 56.9]
Mistral-8B	44.8 _{2.22}	[40.5, 49.1]	40.8 _{2.20}	[36.6, 45.2]	13.0 _{1.50}	[10.3, 16.2]	50.8 _{2.24}	[46.4, 55.2]
Baichuan-M2	56.4 _{2.22}	[52.0, 60.6]	51.6 _{2.23}	[47.2, 56.0]	16.8 _{1.67}	[13.8, 20.3]	61.6 _{2.18}	[57.3, 65.8]
Qwen3-30B	43.4 _{2.22}	[39.1, 47.7]	49.8 _{2.24}	[45.4, 54.2]	44.6 _{2.22}	[40.3, 49.0]	65.0 _{2.13}	[60.7, 69.1]

Model	SG-zh		SG-en		SG-ms		SG-ta	
Gemini 2.5 Flash	34.0 _{2.12}	[30.0, 38.3]	43.4 _{2.22}	[39.1, 47.8]	39.0 _{2.18}	[34.8, 43.3]	31.4 _{2.08}	[27.5, 35.6]
Claude Sonnet 4	29.8 _{2.05}	[26.0, 34.0]	41.8 _{2.21}	[37.6, 46.2]	37.8 _{2.17}	[33.7, 42.1]	31.6 _{2.08}	[27.7, 35.8]
GPT-4o	22.4 _{1.86}	[19.0, 26.3]	47.6 _{2.23}	[43.3, 52.0]	39.4 _{2.19}	[35.2, 43.8]	33.2 _{2.11}	[29.2, 37.4]
GLM-4-9B	20.2 _{1.80}	[16.9, 23.9]	24.8 _{1.93}	[21.2, 28.8]	18.4 _{1.73}	[15.3, 22.0]	19.4 _{1.77}	[16.2, 23.1]
Mistral-8B	17.2 _{1.69}	[14.2, 20.8]	25.2 _{1.94}	[21.6, 29.2]	22.4 _{1.86}	[19.0, 26.3]	15.4 _{1.61}	[12.5, 18.8]
Baichuan-M2	26.6 _{1.98}	[22.9, 30.6]	33.6 _{2.11}	[29.6, 37.9]	28.8 _{2.03}	[25.0, 32.9]	19.4 _{1.77}	[16.2, 23.1]
Qwen3-30B	26.4 _{1.97}	[22.7, 30.4]	29.2 _{2.03}	[25.4, 33.3]	23.6 _{1.90}	[20.1, 27.5]	22.8 _{1.88}	[19.3, 26.7]

Model	VN-vi		VN-en	
Gemini 2.5 Flash	73.6 _{1.97}	[69.6, 77.3]	73.0 _{1.99}	[68.9, 76.7]
Claude Sonnet 4	74.4 _{1.95}	[70.4, 78.0]	72.2 _{2.00}	[68.1, 76.0]
GPT-4o	74.8 _{1.94}	[70.8, 78.4]	71.8 _{2.01}	[67.7, 75.6]
GLM-4-9B	47.6 _{2.23}	[43.3, 52.0]	51.0 _{2.24}	[46.6, 55.4]
Mistral-8B	41.8 _{2.21}	[37.6, 46.2]	48.6 _{2.24}	[44.3, 53.0]
Baichuan-M2	67.6 _{2.09}	[63.4, 71.6]	62.8 _{2.16}	[58.5, 66.9]
Qwen3-30B	57.8 _{2.21}	[53.4, 62.1]	65.8 _{2.12}	[61.5, 69.8]

Qualitative Analysis & Case Studies

A.18 Analysis of Speech as a Cultural Prior

Compared to images and text, speech input introduces greater uncertainty through environmental noise, homophony, and accents—with accents closely tied to cultural context. Our research reveals that accents function beyond mere noise. Testing synthetic speech in standard English versus multiple national accents, we found Qwen and Gemini outperformed their standard English baselines in 6 and 5 country-specific cultural settings

(Figure 2), respectively. Notably, Qwen achieved 2.8% and 3.6% accuracy gains for Indonesian and Japanese accents (Table 22). We attribute this to systematic co-occurrence of accented English with country-specific entities and contexts in training corpora, enabling accents to serve as cultural and lexical priors during inference. Our findings demonstrate that accents can function as valuable cultural cues rather than simply noise sources for model exploitation.

Table 20: VQA modality performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500. “-” means “not support”. Results are presented as Mean $_{SE}$ [95% CI], where the bolded figure is the primary metric, followed by its standard error in subscript and the confidence interval in brackets. The better-performing result among different languages within the same country is **bolded**.

Model	CN-zh		CN-en		ID-id		ID-en	
Gemini 2.5 Flash	45.6 _{2.23}	[41.3, 50.0]	38.8 _{2.18}	[34.6, 43.1]	48.6 _{2.24}	[44.3, 53.0]	49.8 _{2.24}	[45.4, 54.2]
Claude Sonnet 4	47.6 _{2.23}	[43.3, 52.0]	33.6 _{2.11}	[29.6, 37.9]	49.8 _{2.24}	[45.4, 54.2]	46.4 _{2.23}	[42.1, 50.8]
GPT-4o	32.8 _{2.10}	[28.8, 37.0]	38.4 _{2.18}	[34.2, 42.7]	53.6 _{2.23}	[49.2, 57.9]	50.2 _{2.24}	[45.8, 54.6]
Qwen2.5-VL-32B	40.2 _{2.19}	[36.0, 44.6]	30.8 _{2.06}	[26.9, 35.0]	37.6 _{2.17}	[33.5, 41.9]	37.8 _{2.17}	[33.7, 42.1]
Llama-3.2-11B	21.6 _{1.84}	[18.2, 25.4]	15.2 _{1.61}	[12.3, 18.6]	26.2 _{1.97}	[22.5, 30.2]	31.2 _{2.07}	[27.3, 35.4]
Kimi-VL-A3B	-	-	21.2 _{1.83}	[17.8, 25.0]	-	-	25.4 _{1.95}	[21.8, 29.4]
Qwen2.5-Omni-7B	30.6 _{2.06}	[26.7, 34.8]	23.6 _{1.90}	[20.1, 27.5]	21.8 _{1.85}	[18.4, 25.6]	32.4 _{2.09}	[28.5, 36.6]
Deepseek-VL	-	-	12.2 _{1.46}	[9.6, 15.4]	-	-	25.4 _{1.95}	[21.8, 29.4]
GLM-4.1V-9B	46.4 _{2.23}	[42.1, 50.8]	26.2 _{1.97}	[22.5, 30.2]	28.6 _{2.02}	[24.8, 32.7]	34.8 _{2.13}	[30.8, 39.1]
InternVL-1.5	-	-	15.4 _{1.61}	[12.5, 18.8]	-	-	23.0 _{1.88}	[19.5, 26.9]

Model	IN-hi		IN-en		JP-ja		JP-en	
Gemini 2.5 Flash	65.4 _{2.13}	[61.1, 69.4]	42.4 _{2.21}	[38.1, 46.8]	54.2 _{2.23}	[49.8, 58.5]	58.0 _{2.21}	[53.6, 62.3]
Claude Sonnet 4	54.0 _{2.23}	[49.6, 58.3]	39.0 _{2.18}	[34.8, 43.3]	47.4 _{2.23}	[43.1, 51.8]	53.6 _{2.23}	[49.2, 57.9]
GPT-4o	65.8 _{2.12}	[61.5, 69.8]	45.6 _{2.23}	[41.3, 50.0]	59.4 _{2.20}	[55.0, 63.6]	60.0 _{2.19}	[55.7, 64.2]
Qwen2.5-VL-32B	40.2 _{2.19}	[36.0, 44.6]	45.0 _{2.22}	[40.7, 49.4]	47.6 _{2.23}	[43.3, 52.0]	46.8 _{2.23}	[42.5, 51.2]
Llama-3.2-11B	35.4 _{2.14}	[31.3, 39.7]	40.2 _{2.19}	[36.0, 44.6]	27.2 _{1.99}	[23.5, 31.3]	35.2 _{2.14}	[31.1, 39.5]
Kimi-VL-A3B	-	-	34.4 _{2.12}	[30.4, 38.7]	-	-	37.2 _{2.16}	[33.1, 41.5]
Qwen2.5-Omni-7B	19.4 _{1.77}	[16.2, 23.1]	33.6 _{2.11}	[29.6, 37.9]	25.6 _{1.95}	[22.0, 29.6]	39.2 _{2.18}	[35.0, 43.6]
Deepseek-VL	-	-	25.0 _{1.94}	[21.4, 29.0]	-	-	26.0 _{1.96}	[22.4, 30.0]
GLM-4.1V-9B	50.8 _{2.24}	[46.4, 55.2]	43.4 _{2.22}	[39.1, 47.8]	44.0 _{2.22}	[39.7, 48.4]	44.8 _{2.22}	[40.5, 49.2]
InternVL-1.5	-	-	39.2 _{2.18}	[35.0, 43.6]	-	-	40.6 _{2.20}	[36.4, 45.0]

Model	KR-ko		KR-en		MN-mn		MN-en	
Gemini 2.5 Flash	72.0 _{2.01}	[67.9, 75.8]	72.8 _{1.99}	[68.7, 76.5]	41.2 _{2.20}	[37.0, 45.6]	49.2 _{2.24}	[44.8, 53.6]
Claude Sonnet 4	71.8 _{2.01}	[67.7, 75.6]	74.4 _{1.95}	[70.4, 78.0]	35.0 _{2.13}	[31.0, 39.3]	49.0 _{2.24}	[44.6, 53.4]
GPT-4o	69.2 _{2.06}	[65.0, 73.1]	65.8 _{2.12}	[61.5, 69.8]	41.4 _{2.20}	[37.2, 45.8]	53.0 _{2.23}	[48.6, 57.3]
Qwen2.5-VL-32B	63.2 _{2.16}	[58.9, 67.3]	67.0 _{2.10}	[62.8, 71.0]	20.6 _{1.81}	[17.3, 24.4]	45.2 _{2.23}	[40.9, 49.6]
Llama-3.2-11B	30.0 _{2.05}	[26.2, 34.2]	47.8 _{2.23}	[43.5, 52.2]	5.8 _{1.05}	[4.1, 8.2]	21.6 _{1.84}	[18.2, 25.4]
Kimi-VL-A3B	-	-	56.0 _{2.22}	[51.6, 60.3]	-	-	-	-
Qwen2.5-Omni-7B	18.6 _{1.74}	[15.4, 22.3]	49.2 _{2.24}	[44.8, 53.6]	3.2 _{0.79}	[2.0, 5.1]	36.6 _{2.15}	[32.5, 40.9]
Deepseek-VL	-	-	30.0 _{2.05}	[26.2, 34.2]	-	-	-	-
GLM-4.1V-9B	61.0 _{2.18}	[56.7, 65.2]	65.4 _{2.13}	[61.1, 69.4]	28.0 _{2.01}	[24.2, 32.1]	39.4 _{2.19}	[35.2, 43.8]
InternVL-1.5	-	-	56.6 _{2.22}	[52.2, 60.9]	-	-	-	-

Model	SG-zh		SG-en		SG-ms		SG-ta	
Gemini 2.5 Flash	53.0 _{2.23}	[48.6, 57.3]	62.2 _{2.17}	[57.9, 66.3]	59.8 _{2.19}	[55.4, 64.0]	60.8 _{2.18}	[56.5, 65.0]
Claude Sonnet 4	34.6 _{2.13}	[30.6, 38.9]	57.2 _{2.21}	[52.8, 61.5]	55.2 _{2.22}	[50.8, 59.5]	49.8 _{2.24}	[45.4, 54.2]
GPT-4o	30.2 _{2.05}	[26.3, 34.4]	70.6 _{2.04}	[66.5, 74.4]	68.4 _{2.08}	[64.2, 72.3]	62.4 _{2.17}	[58.1, 66.5]
Qwen2.5-VL-32B	40.8 _{2.20}	[36.6, 45.2]	52.8 _{2.23}	[48.4, 57.1]	45.6 _{2.23}	[41.3, 50.0]	22.6 _{1.87}	[19.2, 26.5]
Llama-3.2-11B	32.2 _{2.09}	[28.3, 36.4]	39.4 _{2.19}	[35.2, 43.8]	35.0 _{2.13}	[31.0, 39.3]	15.0 _{1.60}	[12.1, 18.4]
Kimi-VL-A3B	38.6 _{2.18}	[34.4, 42.9]	43.6 _{2.22}	[39.3, 48.0]	-	-	-	-
Qwen2.5-Omni-7B	30.0 _{2.05}	[26.2, 34.2]	43.4 _{2.22}	[39.1, 47.8]	38.6 _{2.18}	[34.4, 42.9]	14.2 _{1.56}	[11.4, 17.5]
Deepseek-VL	16.4 _{1.66}	[13.4, 19.9]	27.2 _{1.99}	[23.5, 31.3]	-	-	-	-
GLM-4.1V-9B	44.8 _{2.22}	[40.5, 49.2]	52.8 _{2.23}	[48.4, 57.1]	41.8 _{2.21}	[37.6, 46.2]	34.8 _{2.13}	[30.8, 39.1]
InternVL-1.5	28.0 _{2.01}	[24.2, 32.1]	44.8 _{2.22}	[40.5, 49.2]	-	-	-	-

Model	VN-vi		VN-en	
Gemini 2.5 Flash	76.6 _{1.89}	[72.7, 80.1]	65.4 _{2.13}	[61.1, 69.4]
Claude Sonnet 4	73.2 _{1.98}	[69.2, 76.9]	67.4 _{2.10}	[63.2, 71.4]
GPT-4o	75.8 _{1.92}	[71.9, 79.4]	63.8 _{2.15}	[59.5, 67.9]
Qwen2.5-VL-32B	65.0 _{2.13}	[60.7, 69.1]	64.6 _{2.14}	[60.3, 68.7]
Llama-3.2-11B	48.6 _{2.24}	[44.3, 53.0]	49.6 _{2.24}	[45.2, 54.0]
Kimi-VL-A3B	-	-	50.0 _{2.24}	[45.6, 54.4]
Qwen2.5-Omni-7B	59.8 _{2.19}	[55.4, 64.0]	58.2 _{2.21}	[53.8, 62.4]
Deepseek-VL	-	-	41.6 _{2.20}	[37.4, 46.0]
GLM-4.1V-9B	59.4 _{2.20}	[55.0, 63.6]	59.8 _{2.19}	[55.4, 64.0]
InternVL-1.5	-	-	54.0 _{2.23}	[49.6, 58.3]

Table 21: Rephrase VQA (Text-Only) modality performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500. “-” means “not support”. Results are presented as Mean $_{SE}$ [95% CI], where the bolded figure is the primary metric, followed by its standard error in subscript and the confidence interval in brackets. The better-performing result among different languages within the same country is **bolded**.

Model	CN-zh		CN-en		ID-id		ID-en	
Gemini 2.5 Flash	45.2 _{2.23}	[40.9, 49.6]	37.2 _{2.16}	[33.1, 41.5]	52.0 _{2.23}	[47.6, 56.4]	48.8 _{2.24}	[44.4, 53.2]
Claude Sonnet 4	52.8 _{2.23}	[48.4, 57.1]	48.0 _{2.23}	[43.7, 52.4]	53.8 _{2.23}	[49.4, 58.1]	52.6 _{2.23}	[48.2, 56.9]
GPT-4o	41.0 _{2.20}	[36.8, 45.4]	46.2 _{2.23}	[41.9, 50.6]	57.8 _{2.21}	[53.4, 62.1]	55.2 _{2.22}	[50.8, 59.5]
Qwen2.5-VL-32B	45.0 _{2.22}	[40.7, 49.4]	34.4 _{2.12}	[30.4, 38.7]	45.0 _{2.22}	[40.7, 49.4]	44.4 _{2.22}	[40.1, 48.8]
Llama-3.2-11B	18.4 _{1.73}	[15.3, 22.0]	15.6 _{1.62}	[12.7, 19.0]	35.2 _{2.14}	[31.1, 39.5]	30.4 _{2.06}	[26.5, 34.6]
Kimi-VL-A3B	-	-	21.6 _{1.84}	[18.2, 25.4]	-	-	24.6 _{1.93}	[21.0, 28.6]
Qwen2.5-Omni-7B	31.4 _{2.08}	[27.5, 35.6]	22.8 _{1.88}	[19.3, 26.7]	23.4 _{1.89}	[19.9, 27.3]	35.8 _{2.14}	[31.7, 40.1]
Deepseek-VL	-	-	15.6 _{1.62}	[12.7, 19.0]	-	-	30.6 _{2.06}	[26.7, 34.8]
GLM-4.1V-9B	46.4 _{2.23}	[42.1, 50.8]	30.6 _{2.06}	[26.7, 34.8]	37.6 _{2.17}	[33.5, 41.9]	33.6 _{2.11}	[29.6, 37.9]
InternVL-1.5	-	-	16.8 _{1.67}	[13.8, 20.3]	-	-	27.0 _{1.99}	[23.3, 31.1]

Model	IN-hi		IN-en		JP-ja		JP-en	
Gemini 2.5 Flash	67.0 _{2.10}	[62.8, 71.0]	53.8 _{2.23}	[49.4, 58.1]	62.4 _{2.17}	[58.1, 66.5]	65.6 _{2.12}	[61.3, 69.6]
Claude Sonnet 4	68.6 _{2.08}	[64.4, 72.5]	51.8 _{2.23}	[47.4, 56.2]	67.8 _{2.09}	[63.6, 71.8]	68.2 _{2.08}	[64.0, 72.1]
GPT-4o	69.0 _{2.07}	[64.8, 72.9]	59.6 _{2.19}	[55.2, 63.8]	67.0 _{2.10}	[62.8, 71.0]	66.4 _{2.11}	[62.2, 70.4]
Qwen2.5-VL-32B	51.1 _{2.24}	[46.8, 55.6]	47.0 _{2.23}	[42.7, 51.4]	56.0 _{2.22}	[51.6, 60.3]	58.4 _{2.20}	[54.0, 62.6]
Llama-3.2-11B	39.8 _{2.19}	[35.6, 44.2]	36.6 _{2.15}	[32.5, 40.9]	39.2 _{2.18}	[35.0, 43.6]	45.6 _{2.23}	[41.3, 50.0]
Kimi-VL-A3B	-	-	32.2 _{2.09}	[28.3, 36.4]	-	-	43.8 _{2.22}	[39.5, 48.2]
Qwen2.5-Omni-7B	28.4 _{2.02}	[24.6, 32.5]	42.8 _{2.21}	[38.5, 47.2]	35.0 _{2.13}	[31.0, 39.3]	51.4 _{2.24}	[47.0, 55.8]
Deepseek-VL	-	-	29.2 _{2.03}	[25.4, 33.3]	-	-	40.4 _{2.19}	[36.2, 44.8]
GLM-4.1V-9B	51.4 _{2.24}	[47.0, 55.8]	33.8 _{2.12}	[29.8, 38.1]	49.0 _{2.24}	[44.6, 53.4]	53.0 _{2.23}	[48.6, 57.3]
InternVL-1.5	-	-	37.4 _{2.16}	[33.3, 41.7]	-	-	40.9 _{2.20}	[36.6, 45.2]

Model	KR-ko		KR-en		MN-mn		MN-en	
Gemini 2.5 Flash	76.4 _{1.90}	[72.5, 79.9]	76.4 _{1.90}	[72.5, 79.9]	47.8 _{2.23}	[43.5, 52.2]	49.2 _{2.24}	[44.8, 53.6]
Claude Sonnet 4	79.6 _{1.80}	[75.9, 82.9]	77.0 _{1.88}	[73.1, 80.5]	38.8 _{2.18}	[34.6, 43.1]	56.0 _{2.22}	[51.6, 60.3]
GPT-4o	73.2 _{1.98}	[69.2, 76.9]	75.2 _{1.93}	[71.2, 78.8]	43.2 _{2.22}	[38.9, 47.6]	59.8 _{2.19}	[55.4, 64.0]
Qwen2.5-VL-32B	66.8 _{2.11}	[62.6, 70.8]	72.0 _{2.01}	[67.9, 75.8]	19.0 _{1.75}	[15.8, 22.7]	56.2 _{2.22}	[51.8, 60.5]
Llama-3.2-11B	43.8 _{2.22}	[39.5, 48.2]	54.6 _{2.23}	[50.2, 58.9]	10.2 _{1.35}	[7.8, 13.2]	29.0 _{2.03}	[25.2, 33.1]
Kimi-VL-A3B	-	-	53.8 _{2.23}	[49.4, 58.1]	-	-	-	-
Qwen2.5-Omni-7B	29.6 _{2.04}	[25.8, 33.7]	60.6 _{2.19}	[56.3, 64.8]	2.6 _{0.71}	[1.5, 4.4]	43.4 _{2.22}	[39.1, 47.8]
Deepseek-VL	-	-	58.8 _{2.20}	[54.4, 63.0]	-	-	-	-
GLM-4.1V-9B	66.6 _{2.11}	[62.4, 70.6]	67.0 _{2.10}	[62.8, 71.0]	26.6 _{1.98}	[22.9, 30.6]	43.8 _{2.22}	[39.5, 48.2]
InternVL-1.5	-	-	48.6 _{2.24}	[44.3, 53.0]	-	-	-	-

Model	SG-zh		SG-en		SG-ms		SG-ta	
Gemini 2.5 Flash	54.6 _{2.23}	[50.2, 58.9]	63.6 _{2.15}	[59.3, 67.7]	59.6 _{2.19}	[55.2, 63.8]	52.4 _{2.23}	[48.0, 56.7]
Claude Sonnet 4	46.0 _{2.23}	[41.7, 50.4]	65.2 _{2.13}	[60.9, 69.2]	61.0 _{2.18}	[56.7, 65.2]	51.8 _{2.23}	[47.4, 56.2]
GPT-4o	30.2 _{2.05}	[26.3, 34.4]	68.2 _{2.08}	[64.0, 72.1]	66.4 _{2.11}	[62.2, 70.4]	53.6 _{2.23}	[49.2, 57.9]
Qwen2.5-VL-32B	36.4 _{2.15}	[32.3, 40.7]	57.6 _{2.21}	[53.2, 61.9]	50.8 _{2.24}	[46.4, 55.2]	25.4 _{1.95}	[21.8, 29.4]
Llama-3.2-11B	25.4 _{1.95}	[21.8, 29.4]	45.2 _{2.23}	[40.9, 49.6]	40.2 _{2.19}	[36.0, 44.6]	20.0 _{1.79}	[16.7, 23.7]
Kimi-VL-A3B	40.4 _{2.19}	[36.2, 44.8]	36.6 _{2.15}	[32.5, 40.9]	-	-	-	-
Qwen2.5-Omni-7B	21.4 _{1.83}	[18.0, 25.2]	43.6 _{2.22}	[39.3, 48.0]	29.4 _{2.04}	[25.6, 33.5]	9.8 _{1.33}	[7.5, 12.7]
Deepseek-VL	25.2 _{1.94}	[21.6, 29.2]	39.2 _{2.18}	[35.0, 43.6]	-	-	-	-
GLM-4.1V-9B	42.0 _{2.21}	[37.8, 46.4]	51.0 _{2.24}	[46.6, 55.4]	40.4 _{2.19}	[36.2, 44.8]	34.8 _{2.13}	[30.8, 39.1]
InternVL-1.5	36.0 _{2.15}	[31.9, 40.3]	41.6 _{2.20}	[37.4, 46.0]	-	-	-	-

Model	VN-vi		VN-en	
Gemini 2.5 Flash	76.0 _{1.91}	[72.1, 79.5]	71.6 _{2.02}	[67.5, 75.4]
Claude Sonnet 4	76.0 _{1.91}	[72.1, 79.5]	73.6 _{1.97}	[69.6, 77.3]
GPT-4o	79.4 _{1.81}	[75.6, 82.7]	72.8 _{1.99}	[68.7, 76.5]
Qwen2.5-VL-32B	69.2 _{2.06}	[65.0, 73.1]	70.0 _{2.05}	[65.8, 73.9]
Llama-3.2-11B	39.4 _{2.19}	[35.2, 43.8]	51.2 _{2.24}	[46.8, 55.6]
Kimi-VL-A3B	-	-	40.0 _{2.19}	[35.8, 44.4]
Qwen2.5-Omni-7B	58.4 _{2.20}	[54.0, 62.6]	60.0 _{2.19}	[55.7, 64.2]
Deepseek-VL	-	-	58.4 _{2.20}	[54.0, 62.6]
GLM-4.1V-9B	66.2 _{2.12}	[61.9, 70.2]	62.8 _{2.16}	[58.5, 66.9]
InternVL-1.5	-	-	46.0 _{2.23}	[41.7, 50.4]

Table 22: Speech modality (speech question & text options) performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500. “-” means “not support”. Results are presented as Mean $_{SE}$ [95% CI], where the bolded figure is the primary metric, followed by its standard error in subscript and the confidence interval in brackets. The better-performing result among different languages within the same country is **bolded**.

Model	CN-zh		CN-en		CN-en	
	-		Acc		NoAcc	
Qwen2.5-Omni-7B	26.2 _{1.97}	[22.5, 30.2]	21.2 _{1.83}	[17.8, 25.0]	21.0 _{1.82}	[17.7, 24.8]
Gemini 2.5 Flash	40.6 _{2.20}	[36.4, 45.0]	37.6 _{2.17}	[33.5, 41.9]	36.0 _{2.15}	[31.9, 40.3]
GPT-4o	28.8 _{2.03}	[25.0, 32.9]	27.8 _{2.00}	[24.1, 31.9]	26.6 _{1.98}	[22.9, 30.6]

Model	ID-id		ID-en		ID-en	
	-		Acc		NoAcc	
Qwen2.5-Omni-7B	29.4 _{2.04}	[25.6, 33.5]	29.4 _{2.04}	[25.6, 33.5]	32.0 _{2.09}	[28.1, 36.2]
Gemini 2.5 Flash	36.6 _{2.15}	[32.5, 40.9]	40.2 _{2.19}	[36.0, 44.6]	40.4 _{2.19}	[36.2, 44.8]
GPT-4o	37.0 _{2.16}	[32.9, 41.3]	39.2 _{2.18}	[35.0, 43.5]	38.2 _{2.17}	[34.0, 42.5]

Model	IN-hi		IN-en		IN-en	
	-		Acc		NoAcc	
Qwen2.5-Omni-7B	33.2 _{2.11}	[29.2, 37.4]	33.2 _{2.11}	[29.2, 37.4]	30.4 _{2.06}	[26.5, 34.6]
Gemini 2.5 Flash	53.4 _{2.23}	[49.0, 57.7]	25.4 _{1.95}	[21.8, 29.4]	24.2 _{1.92}	[20.6, 28.1]
GPT-4o	52.0 _{2.23}	[47.6, 56.4]	46.6 _{2.23}	[42.3, 51.0]	46.2 _{2.23}	[41.9, 50.6]

Model	JP-ja		JP-en		JP-en	
	-		Acc		NoAcc	
Qwen2.5-Omni-7B	45.4 _{2.23}	[41.1, 49.8]	41.8 _{2.21}	[37.6, 46.2]	-	-
Gemini 2.5 Flash	52.4 _{2.23}	[48.0, 56.7]	60.2 _{2.19}	[55.9, 64.4]	57.0 _{2.21}	[52.6, 61.3]
GPT-4o	51.0 _{2.24}	[46.6, 55.4]	55.6 _{2.22}	[51.2, 59.9]	55.4 _{2.22}	[51.0, 59.7]

Model	KR-ko		KR-en		KR-en	
	-		Acc		NoAcc	
Qwen2.5-Omni-7B	53.6 _{2.23}	[49.2, 57.9]	51.4 _{2.24}	[47.0, 55.8]	-	-
Gemini 2.5 Flash	73.6 _{1.97}	[69.6, 77.3]	70.2 _{2.05}	[66.0, 74.0]	72.0 _{2.01}	[67.9, 75.8]
GPT-4o	72.2 _{2.00}	[68.1, 76.0]	72.2 _{2.00}	[68.1, 76.0]	72.0 _{2.01}	[67.9, 75.8]

Model	MN-mn		MN-en		MN-en	
	-		Acc		NoAcc	
Qwen2.5-Omni-7B	43.8 _{2.22}	[39.5, 48.2]	47.8 _{2.23}	[43.5, 52.2]	-	-
Gemini 2.5 Flash	32.0 _{2.09}	[28.1, 36.2]	48.0 _{2.23}	[43.6, 52.4]	50.4 _{2.24}	[46.0, 54.8]
GPT-4o	19.0 _{1.75}	[15.8, 22.7]	42.6 _{2.21}	[38.3, 47.0]	40.8 _{2.20}	[36.6, 45.2]

Model	SG-zh		SG-en		SG-en		SG-ms	
	-		Acc		NoAcc		-	
Qwen2.5-Omni-7B	39.4 _{2.19}	[35.2, 43.8]	39.4 _{2.19}	[35.2, 43.8]	37.0 _{2.16}	[32.9, 41.3]	-	-
Gemini 2.5 Flash	47.0 _{2.23}	[42.7, 51.4]	55.2 _{2.22}	[50.8, 59.5]	54.4 _{2.23}	[50.0, 58.7]	41.4 _{2.20}	[37.2, 45.8]
GPT-4o	44.4 _{2.22}	[40.1, 48.8]	53.2 _{2.23}	[48.8, 57.5]	53.2 _{2.23}	[48.8, 57.5]	45.2 _{2.23}	[40.9, 49.6]

Model	SG-ta		VN-vi		VN-en		VN-en	
	-		-		Acc		NoAcc	
Qwen2.5-Omni-7B	-	-	-	-	54.6 _{2.23}	[50.2, 58.9]	52.0 _{2.23}	[47.6, 56.4]
Gemini 2.5 Flash	39.6 _{2.19}	[35.4, 44.0]	60.6 _{2.19}	[56.2, 64.8]	68.6 _{2.08}	[64.4, 72.5]	67.8 _{2.09}	[63.6, 71.8]
GPT-4o	36.8 _{2.16}	[32.7, 41.1]	64.0 _{2.15}	[59.7, 68.1]	67.6 _{2.09}	[63.4, 71.5]	65.6 _{2.12}	[61.3, 69.6]

Table 23: Evaluated under the MCQ+Explanation metric, Text-only modality performance is reported as Accuracy (%), defined as the number of items where the model’s choice exactly matches the correct option, divided by 500.

Model	CN-en	ID-en	IN-en	JP-en	KR-en	MN-en	SG-en	VN-en
Gemini 2.5 Flash	27.2	32.1	56.0	33.9	53.0	63.7	32.1	71.0
Claude Sonnet 4	41.8	43.1	69.2	45.0	52.2	65.7	34.3	70.8
GPT-4o	43.1	46.1	65.0	48.3	50.6	64.1	36.3	69.4
GLM-4-9B-chat	15.0	27.2	32.4	14.4	24.2	35.8	12.2	42.4
Mistral-8B-Instruct	9.5	14.0	25.5	11.0	19.0	27.9	11.6	33.6
Baichuan-M2-32B	25.3	25.7	45.2	29.9	31.1	55.4	22.7	61.2
Qwen3-30B-A3B	22.8	21.4	37.6	18.6	26.3	38.9	12.9	58.3

Table 24: Evaluated under the MCQ+Explanation metric, VQA modality performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500.

Model	CN-en	ID-en	IN-en	JP-en	KR-en	MN-en	SG-en	VN-en
Gemini 2.5 Flash	29.8	35.9	30.5	43.7	66.4	43.9	47.1	53.8
Claude Sonnet 4	23.8	28.2	25.9	38.4	67.2	40.5	41.4	55.9
GPT-4o	29.9	31.9	27.4	40.0	60.1	37.2	53.2	55.8
Qwen2.5-VL-32B	25.0	22.6	26.1	34.2	62.4	41.0	36.8	51.6
Llama-3.2-11B-Vision	7.0	12.3	14.8	13.8	40.0	15.2	19.4	28.8
Kimi-VL-A3B-instruct	11.8	9.9	13.1	16.6	49.9	29.2	21.7	30.6
Qwen2.5-Omni-7B	15.6	21.5	14.7	23.5	43.9	31.6	29.5	44.3
GLM-4.1V-9B-Thinking	19.3	19.1	22.6	28.6	57.6	33.2	36.8	46.7
InternVL-Chat-V1-5	6.3	7.6	13.0	16.1	47.6	20.6	22.6	30.5

Table 25: Evaluated under the MCQ+Explanation metric, Rephrase VQA (Text-Only) modality performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500.

Model	CN-en	ID-en	IN-en	JP-en	KR-en	MN-en	SG-en	VN-en
Gemini 2.5 Flash	30.2	27.1	27.9	58.3	75.1	45.0	53.7	62.1
Claude Sonnet 4	39.5	37.2	43.3	59.7	75.1	51.7	55.0	63.0
GPT-4o	39.8	35.7	48.4	61.1	68.8	53.5	62.6	66.4
Qwen2.5-VL-32B	28.0	27.1	30.0	47.5	70.1	51.1	44.3	57.8
Llama-3.2-11B-Vision	7.6	13.8	17.6	24.5	52.6	21.6	27.1	35.5
Kimi-VL-A3B-Instruct	12.5	12.3	15.2	26.5	51.7	32.2	23.5	28.5
Qwen2.5-Omni-7B	18.0	25.6	25.5	39.4	59.5	39.6	35.3	49.9
GLM-4.1V-9B-Thinking	22.9	19.8	20.8	42.6	64.6	36.3	39.3	51.8
InternVL-Chat-V1.5	9.3	11.0	16.9	21.2	45.4	28.0	25.1	31.3

Table 26: Evaluated under the MCQ+Explanation metric, Speech modality (speech question & text options) performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500.

Model	CN-en	ID-en	IN-en	JP-en	KR-en	MN-en	SG-en	VN-en
Qwen2.5-Omni-7B	21.0	32.0	30.4	41.8	51.4	47.8	37.0	52.0
Gemini 2.5 Flash	36.0	40.4	24.2	57.0	72.0	50.4	54.4	67.8

Table 27: Speech modality (speech question & options) performance with exact numbers, measured by Accuracy (%): the number of items where the model’s choice exactly matches the correct option, divided by 500. “-” means “not support”. The better-performing result among different languages within the same country is **bolded**.

Model	CN-zh	CN-en	CN-en	ID-id	ID-en	ID-en	IN-hi	IN-en	IN-en
	-	Acc	NoAcc	-	Acc	NoAcc	-	Acc	NoAcc
Qwen2.5-Omni-7B	28.4	14.2	15.2	-	23.8	27.0	-	22.2	20.8
Gemini 2.5 Flash	27.4	13.8	15.2	25.0	31.0	27.0	34.6	34.2	29.4
Model	JP-ja	JP-en	JP-en	KR-ko	KR-en	KR-en	MN-mn	MN-en	MN-en
	-	Acc	NoAcc	-	Acc	NoAcc	-	Acc	NoAcc
Qwen2.5-Omni-7B	-	35.2	32.8	-	34.0	35.0	-	23.6	38.4
Gemini 2.5 Flash	35.0	42.8	36.2	38.6	20.8	49.4	10.6	18.2	28.6
Model	SG-zh	SG-en	SG-en	SG-ms	SG-ta	VN-vi	VN-en	VN-en	
	-	Acc	NoAcc	-	-	-	Acc	NoAcc	
Qwen2.5-Omni-7B	27.0	26.0	23.6	-	-	-	39.2	37.0	
Gemini 2.5 Flash	28.8	42.2	37.6	16.0	15.8	52.0	46.2	45.8	

A.19 Attention heatmaps for incorrect model predictions

Figure 20 shows attention heatmaps over image regions when the model produces incorrect answers. In Case 2, the model concentrates on the subject’s clothing and surrounding scene but overlooks the footwear beside the person and the key text in the upper-left corner that identifies the

subject. Similarly, in Figure 20 the model focuses excessively on attire and playing posture while ignoring the crucial detail that the instrument’s body is wrapped in snakeskin. All of these support our finding: the model’s excessive focus on items mentioned in the prompt causes it to overlook critical cultural details.

A.20 Error Case Study

As illustrated in Figure 21, when queried with “What is the mountain in the picture called?”, the model exhibits hallucinations under both English and Chinese prompts; however, the Chinese reasoning more closely approximates the ground truth. When the VQA query is reformulated into a corresponding text-only question, errors under the English prompt are primarily attributed to cross-lingual alignment failures. Specifically, the model erroneously maps “the world’s number one mountain” to the Chinese phrase “天下第一奇山”, resulting in an incorrect prediction. Conversely, the Chinese prompt elicits the correct answer and explanation. This observation substantiates our hypothesis that image tokens play a more significant role in inducing hallucinations during reasoning. Furthermore, for non-English languages with adequate representation, employing the native language for culture-specific queries consistently enhances performance.

Figure 22 depicts the divergence in model responses across Text-Only, VQA, and Speech modalities for the same inquiry. VQA results indicate that the model lacks specific visual grounding for “Chongqing Guoyuan Port,” rendering it unable to recognize the location. Consequently, it relies on generic scene-level inference, which leads to erroneous outputs. In the Speech modality, failures stem from token misinterpretation or transcription errors—for instance, the phonetic misrecognition of “Chongqing Guoyuan” as “Chongqing Guoyan”, which similarly yields an incorrect response. In other words, the lack of relevant image training data and the disturbances introduced by speech input both make the model more prone to errors in the VQA and speech modalities than in the text-only modality.

A.21 Analysis of cultural and linguistic biases

Building upon the phenomena observed in our experiments, we conducted a systematic and in-depth analysis of model biases. Given the intrinsic inextricability of culture and language, we structured our analysis along two dimensions: **Language Bias** and **Culture Bias**, while also exploring the underlying causes of these phenomena.

Language Bias Experimental results reveal a discernible bias across different models regarding linguistic scripts. Specifically, the Llama-3.2-11B model tends to exhibit superior performance for

the same cultural context when prompts are formulated in Latin-based languages (e.g., English, Indonesian, and Malay), whereas performance degrades noticeably in non-Latin languages.

As illustrated in Fig.23, when presented with the prompt “*In the city where Qian Xuesen’s ancestral home is located, you hear two people having a conversation: A: Qin ah zong ah peng! B: Qin ah zong ah peng! What are they doing?*” in English and Chinese respectively, the model correctly identifies the answer and provides an explanation in the English context. However, in the Chinese context, the model ignores the critical premise regarding “*the city where Qian Xuesen’s ancestral home is located,*” resulting in severe hallucinations.

Similarly, Qwen3-30B-A3B-Thinking-2507 demonstrates robust performance on non-low-resource languages but experiences a sharp performance decline in Mongolian and Tamil. Specifically, under identical cultural contexts, performance drops by 20.4% and 6.4% respectively compared to English prompts, highlighting a bias in the distribution of training languages.

Concurrently, we observed that while errors occur in non-low-resource languages, they primarily manifest as inference errors regarding adherence to problem premises. Conversely, in low-resource languages, model outputs often show weak correlation with the question, frequently devolving into mere elaborations of the options. This indicates a marked deficiency in contextual integration for low-resource languages, limiting the models’ capacity for cultural comprehension and analysis.

These phenomena not only uncover biases in language distribution under identical cultural contexts but also expose the limitations of cross-lingual knowledge transfer within multilingual Large Language Models (LLMs). Since the next-token prediction task serves as the primary pre-training objective, the majority of data relies on predicting the subsequent token within a monolingual sequence, lacking explicit objectives for cross-lingual semantic space alignment. Consequently, knowledge tends to remain siloed within single languages, restricting cross-lingual transfer and contextual grounding in low-resource languages. Future research should explore methods to leverage knowledge inherent in high-resource languages to aid understanding and generation in low-resource languages via semantic alignment.

Culture Bias Prior literature has established that most models, due to training data distribution, ex-

hibit stronger capabilities in mainstream Western cultures while underperforming in Asian contexts. In our study, we observed similar cultural biases across eight Asian nations. Our analysis reveals that models are most attuned to Korean culture (63.98%), supported by global diffusion and data richness (Jang et al., 2024; Dal Yong, 2018). Vietnamese culture (62.96%) shows similarly high performance, hypothetically driven by the abundance of digital content resulting from the country’s 79.8% social media penetration rate (DataReportal et al., 2025). However, distinct challenges emerge for China and India, where larger performance gaps consistent with cultural and regional heterogeneity are observed. The results for Mongolia fall further behind, corroborating the impact of data scarcity in low-resource language domains.

In terms of data distribution, cultural popularity correlates positively with the quality and quantity of training data (e.g., Korea, Vietnam). However, cultural diversity and complexity serve as significant confounding factors. Although Chinese culture possesses a vast volume of data, its high complexity and regional diversity make it difficult for models to fit; for instance, the distinction between Northern “Tangyuan” and Southern “Yuanxiao” is subtle yet non-negligible. Similarly, India encompasses 22 official languages, with distinct regional customs for the same festivals. Conversely, Mongolia exhibits a significant “long-tail effect”: it is not only a typical low-resource language domain but also suffers from data sparsity exacerbated by the concurrent use of Cyrillic and Traditional Mongolian scripts.

From a model training perspective, the local languages of Indonesia, Vietnam, and Malaysia share the Latin alphabet with English, facilitating the cross-lingual transfer of English knowledge during training. Furthermore, shared tokens reduce the compression rate of these languages in the vocabulary, thereby enhancing model performance.

These findings highlight that cultural bias in current models is jointly driven by uneven data distribution and cultural–linguistic complexity (e.g., China, India, Mongolia), beyond simple “data volume” effects seen in Korea and Vietnam. Going forward, a key direction is to curate higher-quality, regionally diverse data and develop training strategies that better support cross-script transfer and fine-grained, intra-cultural distinctions, especially for low-resource and long-tail cultures.

Ethics, Governance & Compliance

A.22 Data Sensitivity and Privacy Discussion

In this section, we discuss data sensitivity and privacy for three modalities.

Text. We collect culture-related textual facts from the public internet and filter out any content containing racism or hate speech. Our “knowledge points” are a few sentences manually summarized by annotators, and all questions are authored from scratch, so no infringement issues are involved.

Images. We use only images under Creative Commons licenses and strictly for research purposes. We apply an automated face-blurring tool (Ren et al., 2023) to protect privacy, followed by manual review to catch any misses or false positives. Please note that faces of historical figures or cartoon characters are not masked. Our images do not contain pornography, violence, or other harmful content.

Speech. We obtained voice samples from individuals outside the data team for speech synthesis, without disclosing any personal information. Consent was obtained prior to recording, and the audio is used solely for research.

A.23 Potential Risks of Our Work

Our work primarily focuses on constructing a dataset of cultural facts. Throughout the dataset development process, we avoid using uncertain factual information and any materials that may contain hate speech. All images are licensed under Creative Commons (CC) licenses and are used in compliance with their terms for research purposes. In addition, after each country team completes its portion of the dataset, we conduct a second-round expert review to prevent factual inaccuracies, hate speech, improper or unauthorized image usage, and translation errors.

A.24 License of Models

Table 28 lists all the models involved in this work and their corresponding licenses. According to the license terms, all of these models may be used for research purposes.

Table 28: Summary of Evaluation Models and Licenses

Model Name	License / Terms of Use
Qwen2.5-VL-32B-Instruct	Apache 2.0
Llama-3.2-11B-Vision-Instruct	Llama 3.2 Community License
Kimi-VL-A3B-Instruct	MIT License
DeepSeek-VL2-Small	MIT (repo) + DeepSeek Model License (weights/usage)
GLM-4.1V-9B-Thinking	MIT License
InternVL-Chat-V1-5	MIT License
Qwen2.5-Omni-7B	Apache 2.0
Qwen3-30B-A3B-Thinking-2507	Apache 2.0
Baichuan-M2-32B	Apache 2.0
GLM-4-9B-Chat	GLM-4-9B custom license (commercial use requires registration)
Minstral 8B-Instruct	Mistral Research License (MRL)
GPT-4o	Proprietary (OpenAI Terms)
Claude Sonnet 4	Proprietary (Anthropic Terms)
Gemini 2.5 Flash	Proprietary (Google Terms)

User Guideline for Cultural Dataset Creation

Target categories:

Daily Life/Culture, Food/Cuisine, Transportation, Buildings, History, Geographical Location and Climate, Education, Fashion/Clothing, and Language/Ethnicity.

Please follow the steps below to create the cultural dataset.

1) Cultural Prompt Creation

Hold a group discussion and brainstorm prompts. For **both tracks—Text-Only and Multi-Modal**—propose **at least 56** cultural keywords or short phrases **for each category** above. Ensure diversity and broad regional coverage across the country. Prefer culturally distinctive, locally specific items, not only well-known ones.

2) Data Collection

- Using the prompts, search the web for relevant **texts and images**, and selectively excerpt passages to serve as the **content base** for question construction.
- Guidelines:
- **Verify uncertain content** via multiple sources; do not use unverifiable information.
- **Images must be under a CC license** and permitted for research use. If no usable image can be found for a prompt, assign it to the **Text-Only** track.
- For **Language/Ethnicity** in the Text-Only track, you may include **code-switching** scenarios or **dialects** where locally relevant.

3) Question Construction


- Build questions from the collected cultural materials. Ensure **over 60%** of your questions require **multi-step cultural reasoning**. Start drafting in the **local language**. Keep in mind:
- Wording must be **clear and grammatical**, with correct spelling.
- If a referent is used, its **coreference must be unambiguous**.
- For **image-based** questions, the answer **must depend on the image**.
- Ensure that image-based questions are **answerable from visual evidence** in the image.
- Provide **four options** per question. Include **at least one correct answer**; **distractors** must be in the **same category** and **reasonably similar** to prevent trivial elimination. **No blank options**.
- Record the **knowledge points** used to create each question; these will be used to assess answer faithfulness. Provide a **2–4 sentence** summary of the **minimal knowledge** required to answer.
- Avoid repeating the **same query style** (e.g., repeatedly asking about “taste”). Vary both **knowledge points** and **question angles**.
- The **question stem** must not exceed **300 words**. Each **option** should be **≤ 50 words** and as concise as possible.
- **Randomize** the position of the correct option; do not always use the same letter.

4) Translation

Translate all questions into **English** using a **closed-source LLM**, then **manually check** each translation for accuracy and correct any errors. For terms without official English renderings, use **transliteration** or a **widely accepted** translation.

Please fill in the questions and options according to the **template provided in the email attachment**.

Figure 10: Annotator Guideline



Question: In the official history, which of the following generals was killed by the person in the picture? A. Hua Xiong B. Yan Liang C. Wen Chou D. Yu Jin

Baseline Output: The image depicts a statue of Guan Yu, one of the most famous and revered generals from the Three Kingdoms period in Chinese history...Hua Xiong: **this is a key figure in the story of the "Battle of Changban"** (or "Battle of Changping"). According to historical records and the novel Romance of the Three Kingdoms, Guan Yu famously defeated Hua Xiong in battle... **but this happened during the Battle of Xiapi**, which came after the Battle of Changban. D. Yu Jin: **Yu Jin was a general of Cao Wei who surrendered to Guan Yu during the Battle of Fancheng**. He was not killed by Guan Yu. Final Answer: The person in the picture (Guan Yu) is most famously known for killing Hua Xiong. Therefore, the correct answer is: **A**.

Vision-ablated Prefix Replay: [Replay Prefix]: The image depicts a statue of Guan Yu **----** Full result **----** The image depicts a statue of Guan Yu...Hua Xiong (华雄):Hua Xiong was a general serving under Dong Zhuo during the late Eastern Han dynasty. He was killed by Guan Yu in the Battle of Hulao Gate, as described in the Romance of the Three Kingdoms...Yan Liang (颜良):Yan Liang was a general serving under Yuan Shao. He was killed by Guan Yu during the Battle of Baima (白马之战) in 200 CE... Wen Chou (文丑):Wen Chou was another general serving under Yuan Shao. He was also killed by Guan Yu shortly after the Battle of Baima, during the same campaign against Yuan Shao. Yu Jin (于禁)... He was not killed by Guan Yu... Correct Answer: Based on the historical context and the question, the general who was killed by Guan Yu is Yan Liang. Final Answer: **B**.

Figure 11: Comparison of baseline vs. "VPR" outputs, with hallucinations highlighted in green.

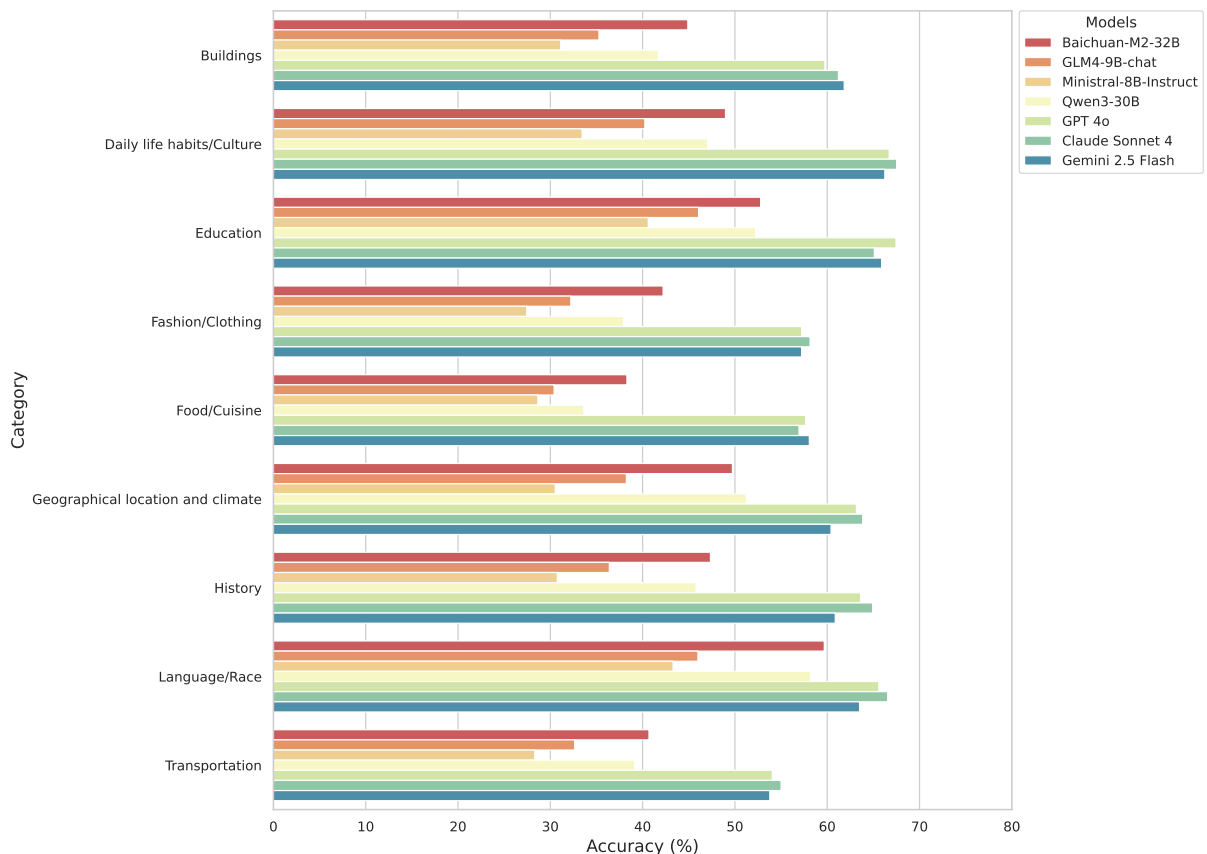


Figure 12: Performance of different models on different categories in the text-only dataset.

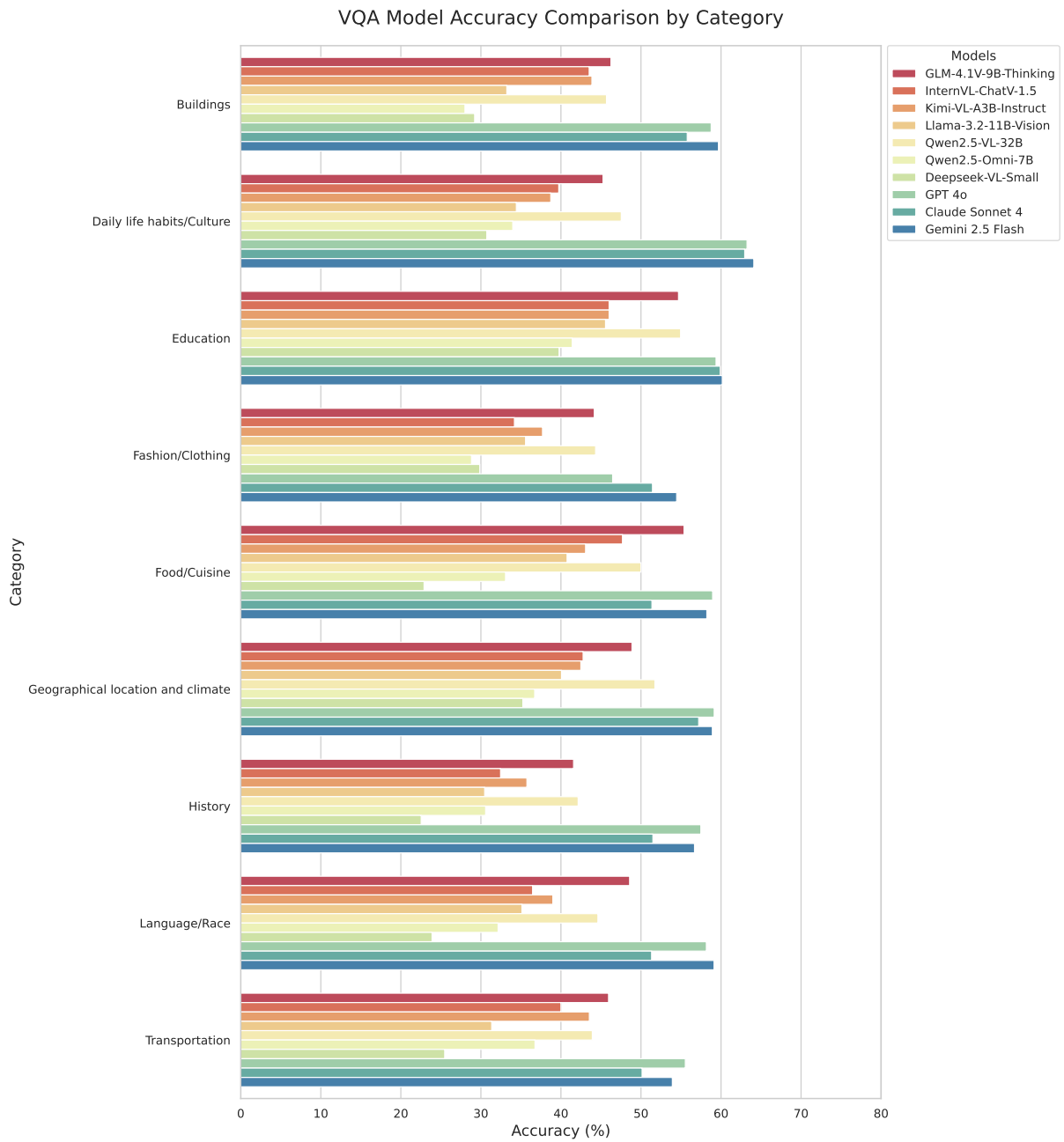


Figure 13: Performance of different models on different categories in the VQA dataset.

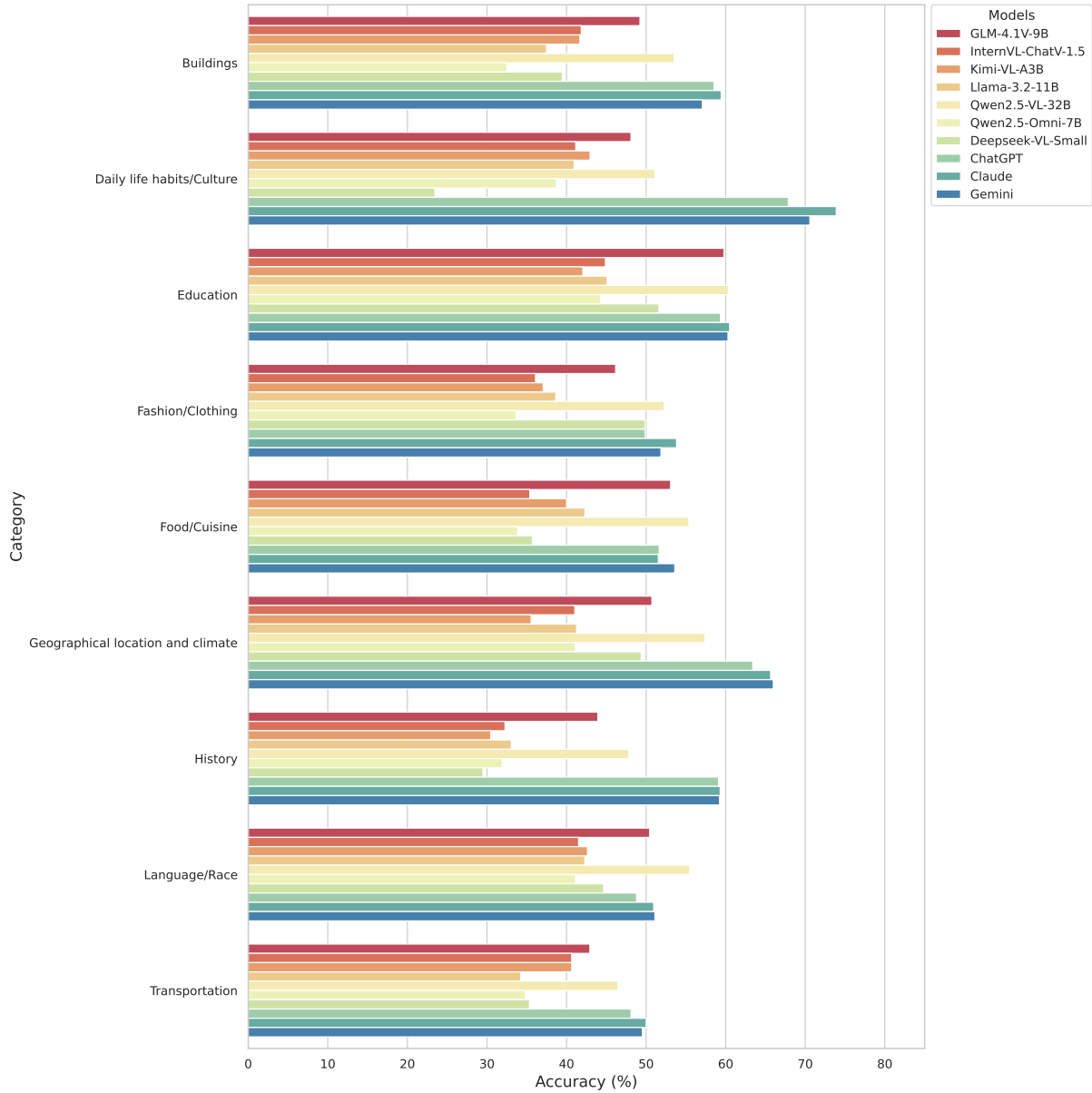


Figure 14: Performance of different models on different categories in the rephrased VQA dataset.

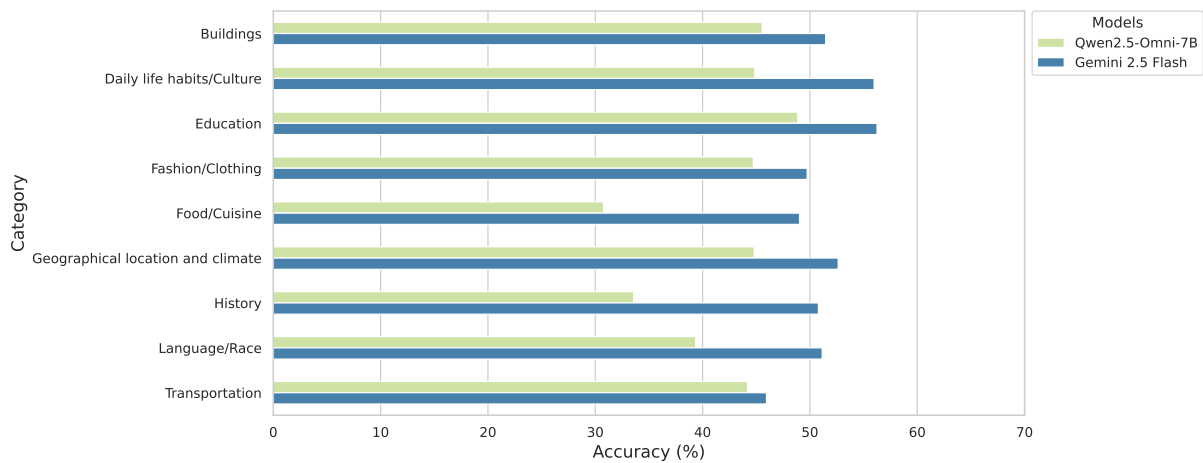


Figure 15: Performance of different models on different categories in the Speech dataset.

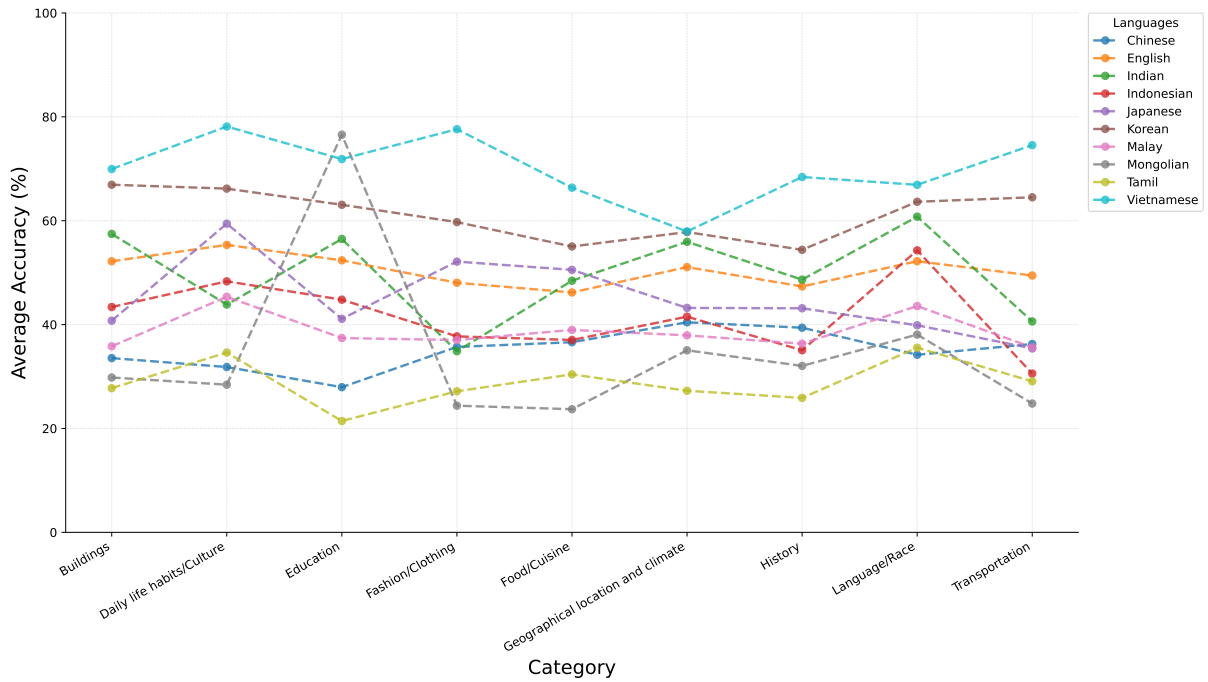


Figure 16: Average performance of the models on different question categories under different language settings.

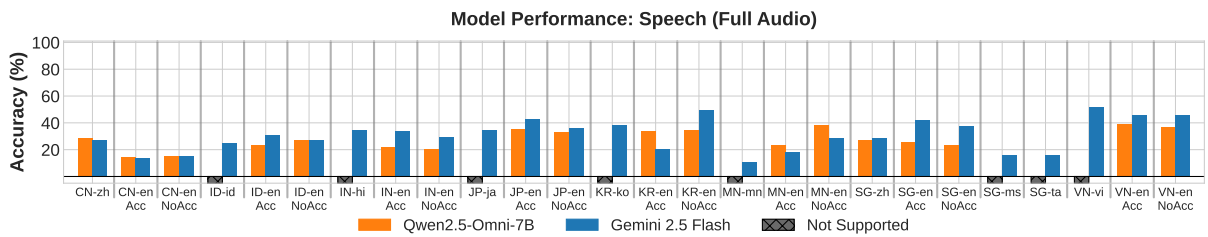


Figure 17: Speech modality (speech question & options) performance with exact numbers, measured by Accuracy (%): the number of items where the model's choice exactly matches the correct option, divided by 500.

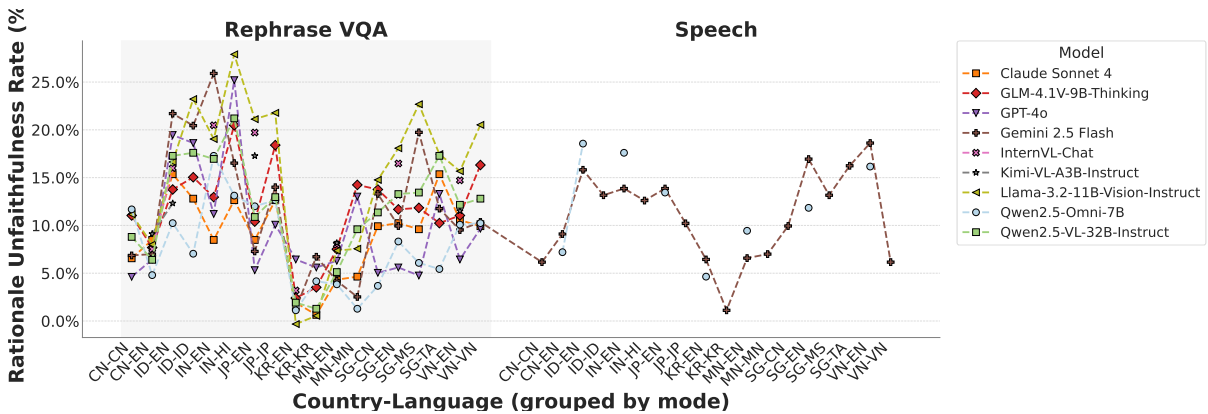


Figure 18: Rationale Unfaithfulness Rates of LLMs across Rephrase VQA and Speech.

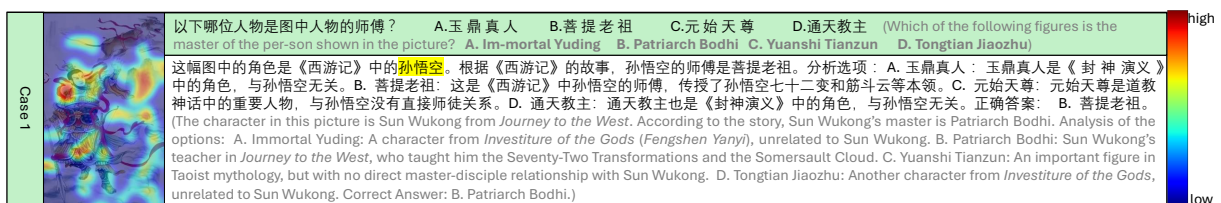


Figure 19: Attention heatmap visualization over image regions during incorrect model answers. Color scale from blue (low) to red (high) indicates increasing model attention.

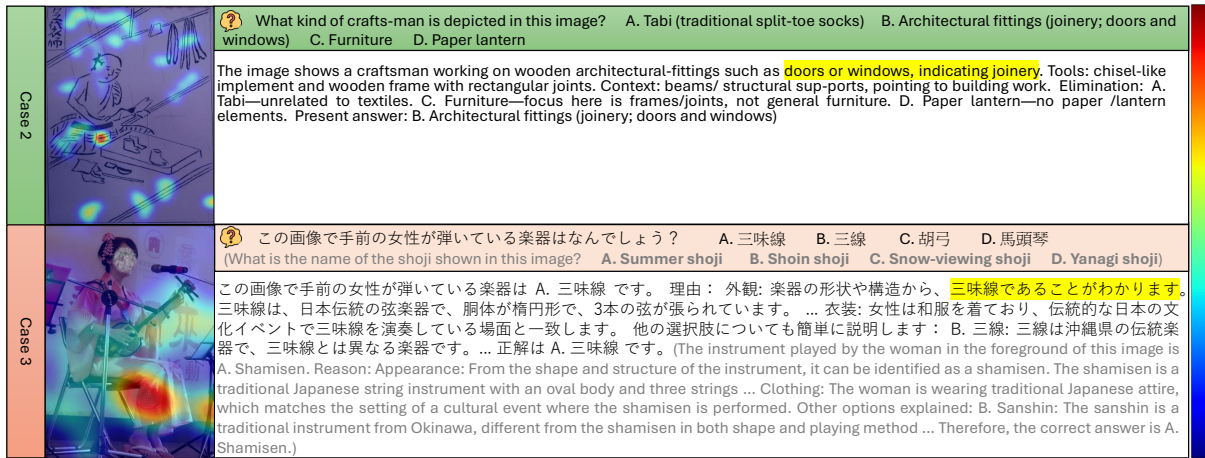


Figure 20: Attention heatmap visualization over image regions during incorrect model answers. Color scale from blue (low) to red (high) indicates increasing model attention.

Question	Model Answer	CA
<p>What is the mountain in the picture called?</p> <p>A. The world's most unique mountain B. The world's number one mountain C. Central Mountain D. The most revered of the Five Great Mountains</p> 	<p>The mountain in the image is Mount Huangshan (Yellow Mountain), famous for its granite peaks, ancient pine trees, and sea of clouds. None of the options directly name Mount Huangshan. However, if the question is about reverence and cultural significance, D. The most revered of the Five Great Mountains could be seen as the closest, as it usually refers to Mount Tai (Taishan), the most sacred of the Five Great Mountains in Chinese culture. If the image is indeed Huangshan, strictly speaking, none of the options are fully correct.</p>	A
<p>图中的大山被称为什么？(What is the mountain in the picture called?)</p> <p>A. 天下第一奇山 (The world's most unique mountain) B. 天下第一山 (The world's number one mountain) C. 中岳 (Central Mountain) D. 五岳独尊 (The most revered of the Five Great Mountains)</p> 	<p>A, D 这张图片展示的是中国著名的黄山。黄山以奇松、怪石、云海、温泉和冬雪“五绝”闻名，被誉为“天下第一奇山”（选项A）。同时，黄山也有“五岳归来不看山，黄山归来不看岳”的美誉，因此也被称为“五岳独尊”（选项D）(This image shows China's famous Mount Huangshan. Huangshan is renowned for its “five wonders” — strange pines, peculiar rocks, seas of clouds, hot springs, and winter snow — and is therefore known as “the world's most unique mountain” (Option A). It is also praised with the saying “After visiting the Five Great Mountains, one need not see other mountains; after visiting Huangshan, one need not see the Five Great Mountains,” so it is also called “the most revered of the Five Great Mountains” (Option D).)</p>	A
<p>What are the other names for Huangshan?</p> <p>A. The world's most unique mountain B. The world's number one mountain C. Central Mountain D. The most revered of the Five Great Mountains</p>	<p>The correct answers are A. The world's most unique mountain and B. The world's number one mountain. Explanation: Huangshan (黄山), also known as Yellow Mountain, is one of China's most famous and iconic mountain ranges. It is celebrated for its stunning natural beauty, including granite peaks, ancient pine trees, and seas of clouds. Over the centuries, Huangshan has inspired many Chinese poets, painters, and philosophers. A. The world's most unique mountain: Huangshan is often considered one of the most unique mountains in the world because of its extraordinary landscape and cultural significance. Its beauty is regarded as unparalleled, so this title is appropriate. B. The world's number one mountain: In Chinese culture, Huangshan is sometimes called “the world's number one mountain” (天下第一奇山), highlighting its exceptional status and reputation. C. Central Mountain: This usually refers to Mount Song (嵩山), one of the Five Great Mountains of China in Henan Province. It is not another name for Huangshan. D. The most revered of the Five Great Mountains: This title is typically used for Mount Tai (泰山), which has great historical and spiritual significance. Huangshan is not one of the Five Great Mountains. Therefore, the correct options are A and B.</p>	A
<p>图中的大山被称为什么？(What are the other names for Huangshan?)</p> <p>A. 天下第一奇山 (The world's most unique mountain) B. 天下第一山 (The world's number one mountain) C. 中岳 (Central Mountain) D. 五岳独尊 (The most revered of the Five Great Mountains)</p>	<p>正确答案：A 解释：黄山是中国著名的山岳，以奇松、怪石、云海、温泉和冬雪闻名，被誉为“天下第一奇山”（选项A）。被称为“天下第一山”的是泰山（选项B）。选项C“中岳”是嵩山的别称，与黄山无关。选项D“五岳独尊”是泰山的著名称号，也与黄山无关。(Correct answer: A Explanation: Mount Huangshan is a famous mountain in China, known for its unique pine trees, oddly shaped rocks, seas of clouds, hot springs, and winter snow. It is honored as “the world's most unique mountain” (Option A). The title “the world's number one mountain” refers to Mount Tai (Option B). Option C, “Central Mountain,” is another name for Mount Song and has nothing to do with Huangshan. Option D, “the most revered of the Five Great Mountains,” is also an honorary title for Mount Tai and is unrelated to Huangshan.)</p>	A

Figure 21: A comparison of the model's responses to the question “What are the other names for Huangshan?” across languages and modalities. The first two examples are VQA questions in English and Chinese, and the last two are text-only questions in English and Chinese.



Question	Model Answer	CA
<p>What multimodal transport functions are demonstrated by Chongqing Guoyuan Port?</p> <p>A. Water-to-water transfer B. Rail-water intermodal transport C. Road-water intermodal transport D. Air-water intermodal transport</p>	<p>The possible options are A, B, C. Chongqing Guoyuan Port can demonstrate multimodal transport functions including water - to - water transfer, rail - water intermodal transport, and road - water intermodal transport. It doesn't have air - water intermodal transport function. So the answer is A, B, and C. If you have any other questions about Chinese culture or anything else, feel free to let me know.</p>	A, B, C
<p>What multimodal transport functions are reflected in the transportation hub shown in the picture?</p> <p>A. Water-to-water transfer B. Rail-water intermodal transport C. Road-water intermodal transport D. Air-water intermodal transport</p> 	<p>The possible options are B, C. The image shows a large number of containers, which are typically used for rail and road transport. There's also a waterway nearby, indicating rail-water intermodal transport and road-water intermodal transport. So, the correct answer is B and C. What do you think about this? Do you have any other questions related to this?</p>	A, B, C
<p> Audio Content: (What multimodal transport functions are demonstrated by Chongqing Guoyuan Port?)</p> <p>A. Water-to-water transfer B. Rail-water intermodal transport C. Road-water intermodal transport D. Air-water intermodal transport</p>	<p>The correct options are B, C, and D. Chongqing Guoyuan Port demonstrates rail - water intermodal transport, road - water intermodal transport, and air - water intermodal transport. Well, that's the answer. If you have any other questions or want to discuss more about multimodal transport, feel free to let me know.</p>	A, B, C

Figure 22: Comparison of model responses across text-only, VQA, and speech modalities to the English question “What multimodal transport functions are demonstrated by Chongqing Guoyuan Port?”. The first, second, and third examples correspond to text-only, VQA, and speech, respectively, with red highlights marking misinterpretations of the speech input.

Question	Category	Language	Model Answer
<p>In the city where Qian Xuesen's ancestral home is located, you hear two people having a conversation:</p> <p>A: Qin ah zong ah peng! B: Qin ah zong ah peng! What are they doing? A. Sawing wood B. Tug of war C. Fluffing cotton D. Rock-paper-scissors</p>	Language/Race	English	<p>In the conversation, the two people are saying "Qin ah zong ah peng" which is a Wu Chinese phrase used as a signal to start a game of rock-paper-scissors.</p>
<p>在钱学森祖籍所在的城市，你听到两个人正在对话：</p> <p>甲：琴啊棕啊棚！ 乙：琴啊棕啊棚！ 请问，他们这是在做什么呢？ A.锯木头 B.拔河 C.弹棉花 D.石头剪刀布</p>	Language/Race	Chinese	<p>他们这是在弹棉花。“琴啊棕啊棚”是一句传统的中国谐音谜语，谐音为“弹棉花”。</p>
<p>What characteristics of Mongolian people's production and daily life do traditional saddles reflect?</p> <p>A. Agriculture-based, horse tack primarily used for farm work. B. Nomadic lifestyle and military requirements are equally emphasized. C. Urban trade is frequent, and saddlery primarily focuses on aesthetics. D. Relying on donkey and mule transport, saddle culture had limited influence.</p>	Transportation	English	<p>Possible Option Letters: B Explanation: Traditional Mongolian saddles mainly reflect two key aspects of Mongolian life: Nomadic lifestyle Mongolians have long lived as nomads on the grasslands and relied heavily on horses for moving, herding animals, and daily travel. Their saddles are strong, stable, and suitable for long-distance riding across rough terrain. Military needs The Mongols were famous warriors. Horses were essential in battle, so the saddles were designed to help riders stay steady and fight while riding, especially for shooting arrows on horseback. Why the other options are wrong: Option A: The saddle reflects a mobile, horse-based lifestyle, not mainly agriculture. Option C: It is about practicality and function, not urban trade or decoration. Option D: Mongolians mainly relied on horses, not donkeys or mules. So Option B is the most accurate choice.</p>
<p>Уламжлалт эмээл нь Монгол ард түмний үйлдвэрлэл, өдөр тутмын амьдралын ямар шинж чанаруудыг тусгасан байдаг вэ?</p> <p>A.Хөдөө аж ахуй нь голлон, морин тоног төхөөрөмж нь ихэвчлэн талбайн ажилд ашиглагддаг. B. Нүүдэлчин амьдрал ба цэргийн хэрэгцээ тэнцүү эрхэмлэгддэг. C. Хотын худалдаа ихтэй, эмээл гоё сайхныг эрхэмлэдэг. D. Илжиг, луусанд тулгуурласан тээвэр, эмээлний соёлын нөлөө хязгаарлагдмал.</p>	Transportation	Mongolian	<p>Бүцах боломжит сонголтууд: A, B, D Тайлбарла: A. Хөдөө аж ахуй нь голлон, морин тоног төхөөрөмж нь ихэвчлэн талбайн ажилд ашиглагддаг. * Уламжлалт эмээл нь Монгол ард түмний үйлдвэрлэлийн үндсэн шинж чанаруудыг тусгасан байдаг. Монгол ард түмний үйлдвэрлэл нь хөдөө аж ахуйг эрхэмлэдэг байсан бөгөөд энэ нь талбайн ажил болон мориныг ашигласан хөдөлмөрөөр илэрдэг. Морины тоног төхөөрөмжийг талбайн ажилд голчлон ашигладаг. Зарим тохиолдолд морийг нүүдэл, өдөр тутмын хөдөлгөөн, аж ахуйн бусад үйл ажиллагаатай холбоотой ажилд ашигладаг. B. Нүүдэлчин амьдрал ба цэргийн хэрэгцээ тэнцүү эрхэмлэгддэг * Монгол ард түмний үйлдвэрлэл нь нүүдэлчин амьдралд тулгуурладаг. Нүүдэлчин амьдрал нь тал нутгаар нүүх, мал маллах, байршлаа байнга өөрчлөх зэрэгтэй холбоотой. Үүнтэй зэрэгцэн, Монголчууд цэргийн хэрэгцээг маш өндөрт тавьдаг байсан. Цэргийн хэрэгцээ нь нүүдэлчин амьдралтай салшгүй холбоотой бөгөөд энэ нь Монгол ард түмний үйлдвэрлэлийн үндсэн шинж чанаруудын нэг юм. D. Илжиг, луусанд тулгуурласан тээвэр, эмээлний соёлын нөлөө хязгаарлагдмал * Монгол ард түмний үйлдвэрлэлд илжиг, луусанд тулгуурласан тээврийг ашиглах явдал байсан ч энэ нь үндсэн гол хэсэг нь биш юм. Илжиг, луусны тээвэр нь тодорхой нөхцөлд ашиглагддаг ч морьтой харьцуулахад хязгаарлагдмал нөлөөтэй. Энэ нь эмээлний соёл Монголын нүүдэлчин, морин соёлд илүү гүн шингэснийг харуулдаг. Сонголт: A, B, D</p>

Figure 23: Language and Cultural Bias Cases. The first and second examples are the English and Chinese versions of the same question; the third and fourth examples are the English and Mongolian versions of the same question.