LLMs Infer Cultural Context but Fail to Apply It When Responding

Yisong Miao[§*] **Jian Zhu**[†] **Vered Shwartz**[†‡]
[†]University of British Columbia [‡]Canada CIFAR AI Chair, Vector Institute
[§]National University of Singapore
yisong@comp.nus.edu.sg jian.zhu@ubc.ca vshwartz@cs.ubc.ca

Abstract

Recent work has shown that LLMs overrepresent dominant cultures, particularly Western ones, while marginalizing others. We investigate whether this affects models' ability to generate culturally adapted responses by evaluating their use of local measurement units based on the user's perceived cultural background. We introduce Cultural and Pragmatic Response Inference (CAPRI), a dataset of conversations with varying levels of cultural cues. Experiments with state-of-the-art LLMs show that models can infer cultural background and recall relevant conventions, but often fail to utilize the information to adapt their answers to the relevant cultural conventions, unless explicitly prompted to perform the tasks sequentially. We further evaluate adaptation to the interpretation of time and quantity expressions – two subjective language grounding dimensions that are affected by culture. We find that models increasingly adapt their answers as cultural cues accumulate, but their priors are not culture-neutral, sometimes aligning with the model's country of origin. Overall, CAPRI provides a resource for future research aimed at narrowing the gap between cultural knowledge and culturally adaptive language generation.

1 Introduction

Considerable research attention has been devoted in recent years to the cultural competence of LLMs, yielding consistent evidence that LLMs are Western-centric or even US-centric (Hershcovich et al., 2022; Cao et al., 2023; DURMUS et al., 2024). Given their diverse user population, it is imperative to develop LLMs that don't overemphasize some cultures and marginalize others (Tao et al., 2024). However, what exactly is desired from LLMs is still debatable. One approach is to personalize LLM outputs for a given user's culture (Cao et al., 2024, 2025a). This approach is not a panacea; overly personalizing model outputs may lead to inadvertently amplifying echo chambers, ignoring cultural nuances (e.g., for bicultural individuals), overcorrecting user intents, and perpetuating stereotypes (Kantharuban et al., 2025; Liu et al., 2025b).

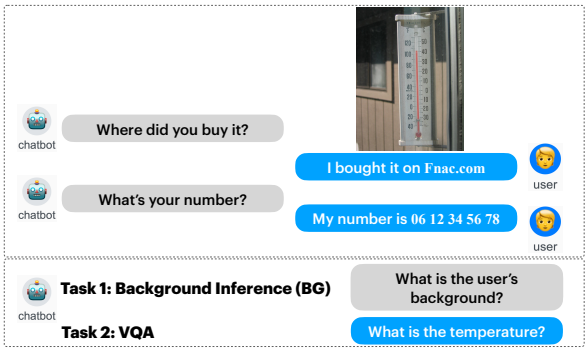


Figure 1: Formalization of CAPRI: the model should infer the user's background from cultural cues in the conversation (**BG; Task 1**), and adapt its answer in line with that background (**VQA; Task 2**).

In this work, we focus on a relatively safe aspect to personalize based on the user's culture: units of measurement, such as currency, distance, size, and temperature. Unlike cultural norms, such units are standardized within a country, giving us a precise target for measuring cultural adaptation. We collect Cultural and Pragmatic Response Inference (CAPRI), a dataset of conversations with varying degrees of revelation about the user's cultural background (Figure 1). We test whether LLMs can explicitly infer the user's cultural background (BG; **Task 1**), and whether they implicitly reason about the user's background to adapt the answer to the local units of measurement in the visual question answering task (VQA; **Task 2**). Task 2 measures whether LLMs act as a "pragmatic speaker" (Frank and Goodman, 2012), tailoring answers to maxi-

[*]Work done during Yisong's Vector Institute research internship with UBC NLP.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 44208–44233
October 24-29, 2026 ©2026 Association for Computational Linguistics

mize the user's understanding.

Evaluation across four families of state-of-the-art LLMs reveals a significant gap between the ability to infer the user's background and the ability to adapt answers accordingly. With 1-2 cultural cues, models are almost perfect at identifying cultural background, but they still fall short in adapting the answer to the VQA task. Encouragingly, explicit reasoning boosts the performance. When guided to perform pragmatic reasoning step by step, LLMs bridge the gap between the BG and VQA tasks.

We further test two other language grounding dimensions that are more subjective but for which prior research has shown cultural differences: time expressions (e.g., *morning* and *afternoon*) and quantity expressions (e.g., *few* and *some*). Our evaluation reveals that LLMs to some extent adapt their answers as more cultural cues accumulate; however, models exhibit non-neutral cultural priors that sometimes lean toward their country of origin.

Unlike prior work studying whether LLMs possess cultural knowledge (Mor-Lan et al., 2026), CAPRI disentangles knowing a user's culture from acting on it. Our findings indicate that current LLMs store the relevant cultural facts in isolation but do not link them: a model can identify the user's background and recall a culture's conventions, yet does not combine the two when answering. We release CAPRI to support future work on bridging this gap between cultural knowledge and culturally adapted generation.[1]

2 Dataset

The CAPRI dataset is designed to simulate LLMs' dialogue with a user and test how the LLM adapts the answer based on the perceived user's culture. Given a conversation history and a question, the model needs to (1) infer the user's background from cues in the conversation, and (2) answer the user's question in a way that maximizes their understanding in a culturally specific way. We define the task (§2.1), introduce the cultural variables in our dataset (§2.2), and describe the dataset creation and statistics (§2.3).

2.1 Task Definition

Inspired by the Rational Speech Act framework (RSA; Frank and Goodman, 2012), we expect the LLM to tailor its response to the user's background *when doing so maximizes the communication effect*. Note that we take RSA as an inspiration rather than a precise implementation: we keep its premise that the speaker should choose the utterance most informative for the inferred listener, but remove the recursive reasoning on the user's end. For example, when asked "what temperature should I set the oven to?", the model should generate either "40 °C" or "104 °F" based on the perceived user culture from the conversation history (Shwartz, 2025). Specifically, a model inspired by the "pragmatic speaker" (Frank and Goodman, 2012) should perform two tasks:

Task 1: Background Inference (**BG**).

$$P(\mathrm{B} \mid X) \tag{1}$$

where B is the user's cultural background and X is the conversation history. We operationalize cultural background as the country most aligned with contextual cues in the conversation (§2.2). We expect LLMs to perform this task implicitly when answering the user's question.

Task 2: Visual Question Answering (**VQA**): $P(y \mid X, I)$ where y is the answer to a question about an image I and X is the conversation history. We ground the question in an image rather than a textual description to avoid biasing the model towards a particular unit outside the deliberate cues in the dialogue, leaving the culturally appropriate lexical choice (e.g., "104 °F" vs "40 °C") to the model.[2] The models are instructed to answer in the [value] [unit] format even when they are unsure of the value, and we evaluate the correctness of the unit.

Ideally, when producing the question involves a cultural aspect, the model should marginalize over the user's inferred background ($P(\mathrm{B} \mid X)$):

$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}} \tag{2}$$

Note that in our dataset, VQA is the main task, and we do not explicitly prompt the model to infer the user's background but rather test its ability to perform this inference implicitly. The prompt asks for an answer suited to the user, but never names the intermediate step of inferring their background from cues; this unstated setup is what we call *implicit*. The one exception is Pragmatic CoT (§3.3),

[1] Available at https://github.com/YisongMiao/CAPRI.

[2] This is not merely a question of avoiding using explicit mentions of units in the question but rather more generally using language that leaks cultural information; prior work showed that the choice of English synonyms is indicative of a speaker's native language (Rabinovich et al., 2018).

where we explicitly instruct the reasoning steps. Task 1 is used as an auxiliary task where we explicitly ask the model to infer the user's background from the conversation to isolate the question *"is the model able to infer the user's culture?"* from *"is it using this information to personalize the answer?"*.

2.2 Cultural Variables

Cultures. We treat countries as proxies for cultures, as in commonly done in the NLP literature (Wang et al., 2024; Liu et al., 2025a). We selected ten countries from diverse regions: Brazil, China, France, India, Iran, Israel, Japan, South Korea, UK, and US.

Language Grounding Dimensions. The main subset of the dataset focuses on measurement units for temperature, distance, speed, size, and price (i.e. currency). These units are standardized and fixed within a country, thus for instances asking about these dimensions we collected a gold standard answer. For example, for the image in Figure 2, the answer should be "104 °F" for an American user and "40 °C" for a French user.

We also evaluate models' responses to questions pertaining to temporal expressions and quantifiers. While prior work showed that there are cultural differences in the grounding of these dimensions (Stateva et al., 2019; Shwartz, 2022), they are more subjective and context-dependent in nature, and exhibit individual differences. We thus do not enforce a "correct" answer for these questions but rather analyze how models' answer changes based on the inferred cultural background.

Cultural Cues. The conversations in our dataset have six variations with different cue strength for the user's background. As shown in Fig. 2, cue strength increases from NULL (bottom one) to EXPLICITFULL (top one).

1. **No Cue:** In this setup, we don't provide any cues about the user's cultural background, so a models' tendency to answer in the context of a particular culture could point to cultural biases in the model. We create two types of conversations: NULL provides no conversation history at all apart from the user's target question, whereas NEUTRAL is a conversation in which all cues are neutralized with culture-agnostic statements (e.g., "an online platform").[3]

[3] NEUTRAL here is relative: the conversation may inherit cultural biases from the generation model (Gemini-2.5-Pro).

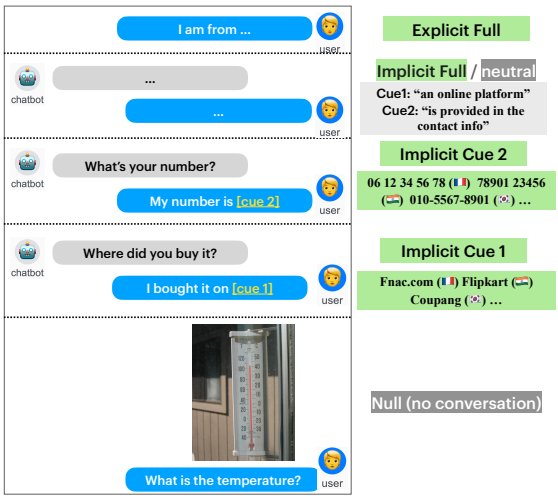


Figure 2: A conversation scaffold across the six cue levels, from no cultural information to explicit disclosure.

2. **Implicit Cues:** In this setting, the model must infer the user's background from the conversation based on cues. For example, the model might infer that a user is French if they mention buying a thermometer at *Fnac.com* or using the phone number format *01 23 45 67 89*. To synthesize our conversations, we use a skeleton with predefined slots for inserting cultural cues. We create the following variations: IMPLICITFULL is a conversation with two cues. IMPLICITCUE1 is truncated right after the first cue appears: all subsequent utterances are removed, and the user asks the question directly. IMPLICITCUE2 is truncated right after both cues have appeared, before the conversation would naturally end.

3. **Explicit Cue.** Finally, we provide an upper-bound condition EXPLICITFULL, where the user explicitly states their cultural background. Concretely, we insert an "I am from [country]" phrase into the first utterance of the NEUTRAL version.

2.3 Dataset Creation

Conversation Generation. For one conversation scaffold, we have [concept] fixed, and alter [background] to generate N sibling conversations. It is done in two steps: **Scaffold Preparation (Step 1)** For each image, we prepare three scaffolds (*chit-chat*, *information seeking*, and *customer support*) using Gemini-2.5-Pro (Appendix A.4). Each scaffold is propagated to ten cultures. **Scaffold Filling (Step 2)** We fill the `[#cue]` slots with names, entities, and systems specific to each culture. We obtain them from online resources and have verified them

	Dimension	# Images	# Scaffolds	# Conv	Question	Possible Answers
Objective Concepts w/ Ground Truth (**Type 1**)	**Temperature**	33	99	990	What is the temperature?	°C, °F
	Distance	32	96	864	What is the distance?	m, km, ft, mi, yd, ...
	Speed	18	54	540	What is the speed?	km/h, mph, m/s, knots, ...
	Size	24	72	648	What is the room size?	m^2, ft^2, ...
	Price	21	63	630	What is the price?	USD, EUR, CNY, JPY, ...
Subjective Concepts w/o Ground Truth (**Type 2**)	**Time Expression**	24	72	720	What time is it?	morning, noon, afternoon, evening, night
	Quantifiers	20	60	600	What is the quantity?	few, some, half, most, almost all
Total		**172**	**516**	**4 992**		

Table 1: **Dataset statistics** by concept, grouped by the two types of concepts w/ or w/o a ground truth answer.

with people from the respective countries.

Image Collection. We collect photographs from Flickr under permissive licenses ("commercial use & modifications allowed"), and complement them with images generated by Google Imagen 4 and Gemini-2.5-Flash-Image for concepts whose fine-grained controlled properties (e.g. a specific room size or distance) cannot be reliably sourced from photographs. The conversation and final question are both grounded in the image (VQA setup). We manually inspect both photos and generated images, filtering out any with explicit text or strong cultural signals.

Dataset Statistics. We have five Type 1 objective concepts and two Type 2 subjective concepts. In total, we have 172 images (Table 1). Each image has three scaffolds (*chit-chat*, *information seeking*, and *customer support*), and each scaffold is expanded into 10 cultures (9 for distance and size, since UK is excluded due to its mixed metric/imperial use), yielding 4,992 conversations in total (samples in Appendix A.5).

Human Evaluation. We recruit human annotators from Cloud Connect,[4] paying 15 USD per hour, covering all ten countries in our dataset. To qualify, annotators must have lived in the target country for at least 5 years within the last 15, though most currently reside in the US and UK. We ask each annotator to role-play the chatbot and answer the user's question in line with the user's culture (see Appendix A.2); this verifies that our conversations make sense and that the cues effectively hint at the speaker's cultural background. Note that we perform human evaluation only on the objective concepts, which have ground-truth answers; subjective concepts have no ground truth by nature, so we do not collect human annotations for them. Annotators first go through a cultural priming step (Liu et al., 2025b), answering simple questions about celebrities from the target country to remind them of their background, before receiving the annotation instruction and performing the task. For example, annotators should respond in °C if the cues hint that the user is from France. To prevent them from always responding the same way, we randomly include 20% conversations with cues for an American user as a control. Across the five measurement-unit concepts, annotators achieve over 85% on the evaluation group and average over 75% on the control (full breakdown in Appendix A.3). These results validate our setup: the conversations indeed make sense, and the cues are salient enough to signal the target cultures.

3 Do LLMs condition on culture when predicting measurement units?

Do LLMs correctly infer and apply culturally specific measurement conventions based on user context, rather than defaulting to a single global standard? We describe our experimental setup (§3.1) and then address the following research questions:

- **RQ1:** How do models generally perform on objective concepts in CAPRI (§3.2)?
- **RQ2:** Does reasoning help the model become a better pragmatic speaker (§3.3)?
- **RQ3:** How does performance scale with size (§3.4)?
- **RQ4:** What are models' prior cultural biases (§3.5)?

[4] https://www.cloudresearch.com/

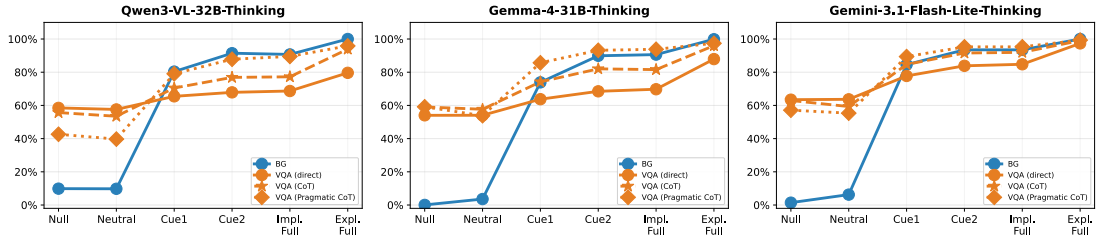


Figure 4: **RQ2:** VQA performance under CoT reasoning. Both Plain CoT and Pragmatic CoT outperform direct prediction significantly, with Pragmatic CoT giving the larger gain and producing deeper reasoning.

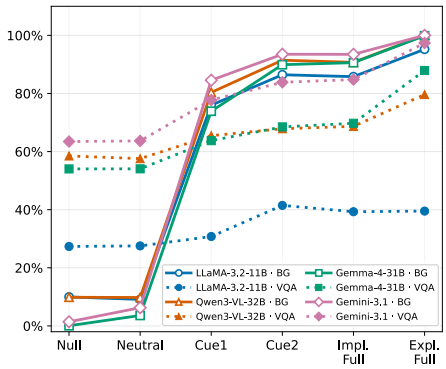


Figure 3: **RQ1:** Direct-prediction performance on BG (solid) vs. VQA (dashed): models show a significant gap between the two tasks.

3.1 Experimental Setup

Evaluation Metrics. We use **accuracy** for both the background inference and question answering tasks. Since our task is generative, we use a regex search to match the nationality (e.g., "France") and the desired unit (e.g., °C); see Appendix A.1.

Models. To manage cost, we evaluate a single closed-source model, Gemini-3.1-Flash-Lite. This model fits our purpose: it is designed for fast responses in daily conversation, while also supporting chain-of-thought reasoning (Wei et al., 2022). We evaluate a broader set of open-source models, spanning different organizations and countries of origin: Qwen3-VL-8B and 32B, in both -Instruct and -Thinking variants (Bai et al., 2025); Llama-3.2-11B-Vision-Instruct (direct prediction only) (Grattafiori et al., 2024); and Gemma-4 in sizes E2B, E4B, and 31B, all of which support both direct and thinking modes (Team et al., 2026) (details in Appendix B.1).

Implementation Details. We employ a minimal prompt: (1) for the primary VQA task, we ask the models to role-play the chatbot and answer the user's question in line with their cultural background; (2) for the background inference task, we simply ask the models to infer the background (prompts detailed in Appendix B.3). We set `temperature=0` across all tasks for reproducibility. All models are hosted on vLLM (Kwon et al., 2023) to accelerate inference, except LLaMA, which we host via HuggingFace (Wolf et al., 2020) for compatibility reasons. For reasoning models, we set a maximum length of 2048 tokens, which is sufficient for our task, detailed in Appendix B.2.

3.2 Overall Performance (RQ1)

We present the best model per family as a function of number of cues in Figure 3. Each point aggregates $n = 3{,}672$ conversations across the ten countries; 95% CIs are about ± 1.6 points, smaller than the marker size. At the start, background inference (BG, solid line) is near zero since there are no cultural cues. The VQA task (unit prediction, dashed line) sits around 30-60%, since models can guess a plausible answer from the image. A clear trend emerges: with just one cultural cue (from NEUTRAL to IMPLICITCUE1), background prediction jumps sharply from near zero to around 80%, while the VQA task rises much more slowly. The gap widens with more cues (IMPLICITCUE2 and IMPLICITFULL): background prediction is almost perfect, while the VQA task still lags behind. **This suggests current LLMs are not ideal pragmatic speakers:** they have the capacity to infer the user's background correctly, but do not use it when answering. Finally, we use EXPLICITFULL as an upper bound: with the background stated explicitly, background prediction is perfect and the gap narrows for all models except LLaMA, which appears unable to associate the cultural background with the correct units.

3.3 Benefit of Reasoning (RQ2)

We then show that CoT reasoning significantly improves VQA performance, with Pragmatic CoT bringing the largest gain (best reasoning model per

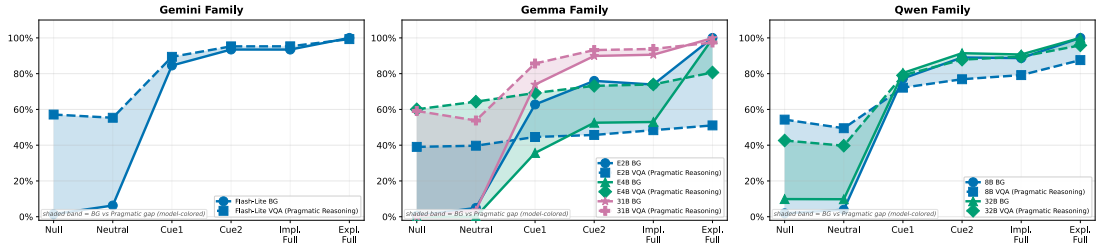


Figure 5: **RQ3:** Effect of model scaling across the Qwen, Gemma, and Gemini-3.1 families: Larger models narrow the gap between VQA and BG (the shaded area).

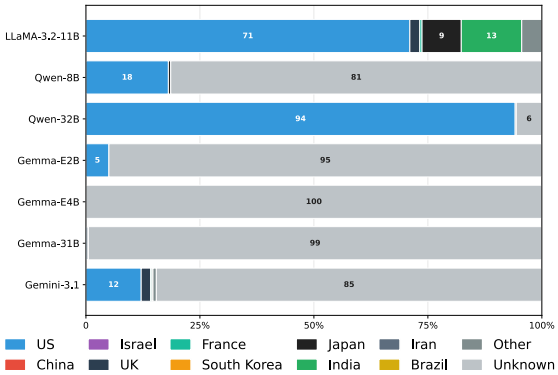
family, reported in Figure 4). *Ideally*, the model should (1) infer the user's background, then (2) produce an answer aligned with that background. We test whether models can perform such pragmatic reasoning via CoT, exploring two zero-shot setups: (1) **CoT**, the plain version where the model uses its default reasoning mode with no culture-specific guidance (it must reason from the task instruction alone); (2) **Pragmatic CoT**, where we add explicit instructions for the model to infer the cultural background before making a prediction. Figure 4 shows that plain CoT already improves over direct VQA, while Pragmatic CoT improves VQA performance substantially, matching or exceeding the background inference task. **Together, these results show that with explicit reasoning, models can perform pragmatic reasoning in cultural contexts.**

Further analyses show interesting behaviors of models' reasoning: (1) Pragmatic CoT produces deeper reasoning, often linking the inferred background to the answer via an explicit causal connective. (2) Plain CoT yields shallower reasoning overall, though larger models still reach the deeper levels, suggesting that pragmatic reasoning capability scales with model size (detailed in Appendix B.5).

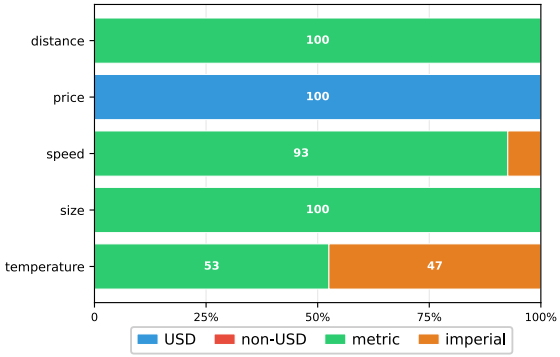
3.4 Benefit of Scaling (RQ3)

We further find that larger models close the gap between BG and VQA more effectively under Pragmatic CoT (Figure 5). (1) Starting with the strongest overall model, Gemini-3.1, we see that with just one cue (IMPLICITCUE1), the gap between VQA and BG is already very small, and it diminishes further by the end. This shows that with proper Pragmatic CoT guidance, Gemini-3.1 handles the reasoning well. (2) Turning to the Qwen and Gemma families, we see a clear trend: larger models have a smaller gap between VQA and BG. The shaded shape of Qwen-32B and Gemma-31B resembles Gemini-3.1; however, the gap is significantly larger for the smaller Qwen and Gemma models, suggesting that **LLMs have an intrinsic limitation in pragmatic reasoning in cultural contexts that even explicit guidance cannot overcome**.

3.5 Models' Cultural Prior (RQ4)



(a) Background prior (BG): models lean toward the US.



(b) Unit prior (VQA): models lean toward the metric system.

Figure 6: **RQ4:** Models' cultural prior at NULL, across the two tasks.

Finally, we examine the models' cultural priors: without any cues, models default to the US for country, yet prefer the metric system (which the US does not use) for measurement units. Users might ask models a question without any contextual cues, so it is important to quantify models' cultural prior. We treat models' responses

on NULL conversations as the prior. Figure 6a shows all models' prior on the background prediction task. From a bird's-eye view, the majority response is unknown (the model explicitly refuses to answer). The most popular predicted country is the US (by LLaMA, Qwen, and Gemini-3.1). Interestingly, Qwen-8B refuses to answer, while Qwen-32B becomes overconfident in predicting the US (94%). LLaMA and Gemini-3.1 are the most "diverse" models, with notable shares for Japan, UK, India, and others.

Models' prior on unit prediction tells a different story (Figure 6b shows Gemini-3.1; other models are similar). Although background prediction leans toward the US, unit predictions are dominated by the metric system, the opposite of US convention. While the metric system is far more common around the world than the imperial system, this relationship is not straightforwardly captured in term frequency in large web corpora. A search of infini-gram (Liu et al., 2024) showed that imperial and metric size units appear roughly equally, whereas distance and temperature have preference for metric units (∼60%) and speed has a more substantial preference for imperial (∼70%).

Models prediction for pricing is an exception, where the dollar dominates (100% for Gemini, other models similar). Currencies differ across all countries in our data, and USD is more frequently discussed in English text than other currencies. A search of infini-gram resulted in 113M hits for "USD", double the number of hits as for all other currencies tested in this paper, combined (58M).

4 Do LLMs condition on culture when interpreting subjective expressions?

To what extent are LLM outputs on subjective dimensions – such as mapping 6 PM to "evening" or "afternoon" – sensitive to cultural context, rather than reflecting a single dominant or default interpretation? Using the same models as in Sec. 3.1, we address the following research questions:

- **RQ5:** How do cultural cues influence models' subjective predictions (§4.1)?
- **RQ6:** Which cultures do models' priors lean toward (§4.2)?

4.1 Cultural Cues' Influence (RQ5)

We first study whether models show more divergence as cues accumulate. Given that there is no ground-truth answer for subjective dimensions, we instead examine how models' predictions shift as cultural cues accumulate, by defining the **inter-culture divergence score**:

$$D_{\text{Inter}} = \frac{1}{\binom{N}{2}} \sum_{\{B,B'\}} \Pr(y^B \neq y^{B'})$$

D_{Inter} is the average divergence across all $\binom{N}{2} = 45$ pairs of cultures (with $N = 10$) at the same cue level, where B and B' denote two distinct cultural backgrounds. For example, if a model interpreted 6 PM as "afternoon" under one cultural condition and as "evening" under another, we count this as a divergence. We hold the conversation fixed and vary only the cultural cues across countries: a [shopping platform] cue might be *Fnac* (France), *Flipkart* (India), or *Coupang* (South Korea). If a model is sensitive to these cues, it may give different answers to reflect the common interpretation within a given culture. We use this divergence as a measure of cultural specificity (visualized in Appendix B.6).

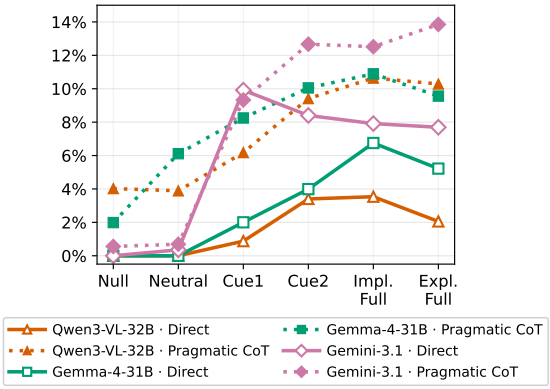


Figure 7: **RQ5:** The impact of cultural cues on models' inter-culture divergence (D_{Inter}): more cues yield more diverse answers.

Fig. 7 reports each family's largest model on the subjective task, measured by D_{Inter}. We make two observations: (1) Overall, models give more diverse answers as cues accumulate, from IMPLICITCUE1 to IMPLICITCUE2 to IMPLICITFULL. Interestingly, for some models the divergence drops from IMPLICITFULL to EXPLICITFULL. (2) Pragmatic CoT increases divergence compared to direct prediction. Inspecting the reasoning traces, we find that stronger models follow our instructions and engage in deliberate thinking such as "the user is from ..., so what does 6 PM mean to them?", which enhances the divergence.

Culture	Qwen-32B		Gemma-31B		Gemini-3.1	
	D_{Impl}	D_{Expl}	D_{Impl}	D_{Expl}	D_{Impl}	D_{Expl}
US	6.8	3.8↑	8.3	9.8↑	9.8	6.8
China	4.5↓	1.5↓	7.6↓	7.6	8.3	11.4↑
Israel	5.3	1.5↓	9.1	8.3	6.1	8.3
UK	5.3	3.0	8.3	6.1	7.6	9.1
France	7.6↑	3.0	9.8↑	7.6	8.3	9.8
Korea	5.3	3.0	9.1	6.1	5.3↓	10.6
Japan	6.1	2.3	7.6↓	6.8	7.6	8.3
India	4.5↓	2.3	7.6↓	5.3↓	9.1	8.3
Iran	4.5↓	3.0	8.3	9.1	6.1	7.6
Brazil	6.1	2.3	9.1	6.8	10.6↑	6.1↓
Mean	5.6	2.6	8.5	7.3	7.9	8.6

Table 2: **RQ6:** Per-culture D_{Impl} and D_{Expl} for three models (in percentage score). ↓ marks the per-column minimum (culture closest to the model's prior); ↑ marks the per-column maximum (the most "alien" culture). 95% CIs are about ± 1.5 for the *Mean* row ($n = 1{,}320$) and ± 4.3 for individual cultures ($n = 132$).

4.2 Sensitivity to Specific Cultures (RQ6)

We then test models' cultural priors on these subjective concepts, where the absence of a ground-truth answer lets each model's default leaning emerge. To that end, we define the **cue divergence scores**:

$$D_{\text{Expl}}(B) = \Pr(y^{B}_{\text{ExplicitFull}} \neq y^{B}_{\text{Neutral}})$$

$$D_{\text{Impl}}(B) = \Pr(y^{B}_{\text{ImplicitFull}} \neq y^{B}_{\text{Neutral}})$$

where $D_{\text{Expl}}(B)$ and $D_{\text{Impl}}(B)$ are the **explicit-** and **implicit-cue divergence scores** for culture B. To attribute a shift cleanly to the cultural cues, we anchor on the Neutral conversation and construct two minimally-changed variants: ExplicitFull only appends an "I am from ..." phrase to Neutral, while ImplicitFull only fills the neutralized [cue] slots in Neutral with culture-specific values. Each score measures, per culture, how often the model's answer changes under that minimal modification. Low scores may indicate that the model's priors align with the given culture.

We report the results in Table 2. The per-culture cells carry wide confidence intervals, so we see them as trends rather than definitive rankings. Three findings stand out: (1) The US is the most alien culture for Qwen-32B and for Gemma-31B (D_{Expl}), and is never the closest to any model's prior. (2) China is the closest to Qwen-32B's prior, consistent with Qwen being developed by a Chinese company, and is also favored by Gemma-31B under implicit cues. Interestingly, Gemini-3.1 instead places China as its most alien culture under explicit cues. (3) For Gemini-3.1, Brazil is the closest under explicit cues but the most alien under implicit cues. Possible reason is that Brazil shares implicit cues with other Latin American countries, which can confuse the model (case study in Appendix B.7).

5 Related Work

Pragmatic Speaker Models. A pragmatic speaker aims to maximize communicative effect by inferring how the listener interprets each utterance (Frank and Goodman, 2012; Andreas and Klein, 2016). This process often involves resolving implicatures (Ruis et al., 2023; Cong, 2024). Recent studies show that LLMs still fall short at pragmatic understanding (Hu et al., 2023; Sravanthi et al., 2024), and a significant gap remains before they can reason as competent pragmatic speakers (Jian and Siddharth, 2024; Sieker and Zarrieß, 2026). Pragmatic reasoning has been applied to many tasks, such as code generation (Cao et al., 2025b), multi-turn dialogue (Estienne et al., 2025), and image captioning (Cohn-Gordon et al., 2018; Nie et al., 2020), but cultural aspects, our main focus, are largely overlooked in this line of work. A notable exception is White et al. (2024), who study cross-cultural common ground between two players in Codenames Duet, a cooperative word-association game. Our setting differs in both task and direction: we focus on concept-to-value grounding (e.g., units, time expressions) and on user-to-model adaptation, where the user's culture is inferred from cues.

Cultural Competence in LLMs and VLMs. LLMs are expected to respond in line with the speaker's cultural background. Recent work largely focuses on evaluating LLMs' cultural competence with respect to cultural values (Zhao et al., 2024; Kabir et al., 2025), norms (Rao et al., 2025), and knowledge (Ramezani and Xu, 2023; Myung et al., 2024; Chiu et al., 2025; Mor-Lan et al., 2026; Naous et al., 2024; Veselovsky et al., 2026). These evaluations have been extended to VLMs and text-to-image models (Bhatia et al., 2024; Nayak et al., 2024, 2025). Many recent resources target cross-cultural understanding, including multimodal metaphor (Yang et al., 2025), emotion understanding (Belay et al., 2025), cross-regional object recognition (Rojas et al., 2022), as well as Hofstede's cultural dimensions (Kharchenko et al., 2024). However, they overlook how the expression of the same concept varies across cultures.

Language Grounding Grounding is the process of mapping utterances to what they refer to in the world, for example, the expression "evening" to a specific time, or "few" to a particular count (e.g., of eggs in a basket). We take a novel angle: grounding in images, leaving the culturally appropriate lexical choice to the model. Language grounding has been studied across text, image, and video (Chandu et al., 2021; Fried et al., 2023), but less so in cultural contexts. Notable exceptions include time expressions (Shwartz, 2022), quantifiers (Stateva et al., 2019; Wong et al., 2025), gradable adjectives (Garí Soler and Apidianaki, 2021), and implicit numeric heads (Elazar and Goldberg, 2019). However, these resources either treat concepts as culture-invariant or capture cross-cultural variation only through small-scale psycholinguistic studies. Therefore we introduce CAPRI, which grounds both objective and subjective concepts across cultures at scale. Our evaluation is also distinct in a conversational setting, disentangling what models know from whether they act on it.

6 Conclusion

Cultural competence in LLMs is becoming increasingly consequential as they are deployed to automate processes globally. We introduce CAPRI, reframing the question from whether LLMs *know* a certain culture to whether they can *apply* this knowledge and tailor their answer to maximize communicative effect. Our evaluation reveals a persistent gap: state-of-the-art models reliably infer the user's background but fail to apply it, unless explicitly guided to reason about this. We also show that on subjective expressions, models' answers diverge more as cultural cues accumulate, while their no-cue priors sometimes align with the model's country of origin. CAPRI is a step toward narrowing the gap between cultural knowledge and culturally adaptive generation. Future work should broaden the set of grounding dimensions and cultures, and imbue this reasoning capability into LLMs, making them pragmatic speakers.

Limitations

Coverage of cultures. We follow the common practice in the NLP community and use country as a proxy for culture (Wang et al., 2024; Liu et al., 2025a), while acknowledging that cultures could be defined at a finer-grained level based on region, language, religion, and more. While we only tested 10 countries, we selected them to maximize regional diversity.

Language artifacts. Our dataset is in English, which introduces a language prior. We chose English to separate the effect of LLM performance on different languages from their ability to perform cultural pragmatic reasoning. A small case study on translated conversations (Appendix B.4) shows that the cultural reasoning is still not perfect: the conversation's language alone makes background inference nearly trivial, while the gap to question answering persists. We leave full multilingual evaluation, where the input language itself can signal the user's background, to future work.

Ethical Considerations

Annotation. The study was conducted with approval from our institute's Behavioral Research Ethics Board (IRB). Annotation is performed on the Cloud Connect platform, where we pay annotators $15 USD/hour, in line with CloudResearch's compensation guidelines to pay at least the local minimum wage standards.

Data Sources and Potential Risks. We generated the conversations with Google Gemini-2.5-Pro. We generated the images with Google Imagen 4 and Gemini-2.5-Flash-Image, apart from photographs, which we sourced from Flickr under permissive licenses ("commercial use & modifications allowed"). The cues (names, phone numbers, postal codes) are synthetic and follow each country's format conventions; they do not refer to any real individual. The images are about objects and scenes (thermometers, rooms, tractors, etc.) rather than identifiable people, and we manually inspect each to filter out any with explicit text or strong cultural signals. The conversation topics cover neutral concepts (measurement units, time, quantifiers), avoiding sensitive or harmful content.

Cultural Biases. In this work, we use country-level background as a proxy for culture, following common practice in NLP research. We mitigate stereotyping by verifying both the cultural cues and the conversations with people from the target country (our annotators have lived there for at least 5 of the past 15 years; §2.3). However, this simplification may still reinforce the stereotype that people from the same country share a homogeneous culture. Future work could consider more fine-

grained representations of culture beyond country-level proxies.

AI Tool Usage. We use Google Gemini-2.5-Pro for conversation generation. For image generation, we use Google Imagen 4 and Gemini-2.5-Flash-Image (§2.3). Claude Code and Cursor AI were used during coding, primarily for debugging assistance; ChatGPT and Claude (via the web interface) was used only for grammar checking of the manuscript.

Acknowledgements

We thank anonymous reviewers for spending time in reviewing our paper, and for their helpful feedback. We thank Aditya Chinchure, Soheil Alavi, Eunjeong Hwang, Sara Papi, Samuel Rhys Cox, and Hannah Brown for help fact-checking our knowledge base, and Joy Zhuozhuo Liu for her annotation interface that inspired our design. Yisong thanks Dr. Jim Mondo for his IMPACT Program mentorship during his Vector internship. We also thank Prof. Min-Yen Kan for following our updates and offering insightful comments.

This work was supported by the Vector Institute and was conducted during the first author's internship at the institute. It was also supported by funding from UBC Language Sciences. This research was enabled in part by compute credits from Google for Gemini. The experiments were partially supported by NUS HPC clusters. The authors are also supported by grants from NSERC and the Canada CIFAR AI Chairs program.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural lenses on emotion - benchmarking LLMs for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025a. Specializing large language models to simulate survey response distributions for global populations. 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 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Zhuchen Cao, Sven Apel, Adish Singla, and Vera Demberg. 2025b. Pragmatic reasoning improves llm code generation. *arXiv preprint arXiv:2502.15835*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

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. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image

captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Yanai Elazar and Yoav Goldberg. 2019. Where's my head? Definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535.

Lautaro Estienne, Gabriel Ben Zenou, Nona Naderi, Jackie CK Cheung, and Pablo Piantanida. 2025. Collaborative rational speech act: Pragmatic reasoning for multi-turn dialog. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22509–22523, Suzhou, China. Association for Computational Linguistics.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2021. Scalar adjective identification and multilingual ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Mingyue Jian and N. Siddharth. 2024. Are llms good pragmatic speakers? *Preprint*, arXiv:2411.01562.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Anjali Kantharuban, Jeremiah Milbauer, Maarten Sap, Emma Strubell, and Graham Neubig. 2025. Stereotype or personalization? user identity biases chatbot recommendations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24418–24436, Vienna, Austria. Association for Computational Linguistics.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Zhuozhuo Joy Liu, Farhan Samir, Mehar Bhatia, Laura K. Nelson, and Vered Shwartz. 2025b. Is it bad to work all the time? cross-cultural evaluation of social norm biases in gpt-4. *Preprint*, arXiv:2505.18322.

Guy Mor-Lan, Omer Goldman, Matan Eyal, Adi Mayrav Gilady, Sivan Eiger, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Reut Tsarfaty. 2026. Location not found: Exposing implicit local and global biases in multilingual LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29130–29150, San Diego, California, United States. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, Suzhou, China. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. 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 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Vered Shwartz. 2025. *Lost in Automatic Translation: Navigating Life in English in the Age of Language Technologies*. Cambridge University Press.

Judith Sieker and Sina Zarrieß. 2026. How hypocritical is your LLM judge? listener-speaker asymmetries in the pragmatic competence of large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 7551–7565, San Diego, California, United States. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Penka Stateva, Arthur Stepanov, Viviane Déprez, Ludivine Emma Dupuy, and Anne Colette Reboul. 2019. Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology*, Volume 10 - 2019.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Veniamin Veselovsky, Berke Argın, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2026. Localized cultural knowledge is conserved and controllable in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43152–43178, San Diego, California, United States. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to play: Pragmatic reasoning for efficient cross-cultural communication. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12201–12216, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hugh Mee Wong, Rick Nouwen, and Albert Gatt. 2025. VAQUUM: Are vague quantifiers grounded in visual data? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11966–11982, Vienna, Austria. Association for Computational Linguistics.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. 2025. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317, Vienna, Austria. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

A Dataset Detail

A.1 Ground-Truth Answer for Objective Measurement Unit

Concept	Metric	Imperial
Distance	km, m, cm, kilometer, meter, centimeter	mi, mile, ft, foot, inch, yard
Speed	km/h, kph, kmph, m/s	mph, mi/h, ft/s, fps
Size (Area)	m^2, sq m	sq ft, ft^2, sf, acre, sq mi, square mile
Temperature	°C, Celsius	°F, Fahrenheit

Table 3: Metric vs. Imperial units accepted as ground-truth answers for each measurement unit concept.

For objective measurement-unit concepts, we accept any unit form consistent with the user's culture as a correct answer. Table 3 lists the accepted unit forms for both metric and Imperial systems, which we match against the model's output via regular expressions. Table 4 then shows which system each country in our dataset follows.

Country	Corresponding Unit
US	**Imperial** for all units; Price: USD ($, dollars, cents)
China	**Metric** for all units; Price: CNY (¥, yuan, RMB)
Israel	**Metric** for all units; Price: ILS (shekel, NIS)
UK	**Imperial** for Speed; **Metric** for Temperature; Price: GBP (£, pounds, pence)
France	**Metric** for all units; Price: EUR (€, euro)
South Korea	**Metric** for all units; Price: KRW (won)
Japan	**Metric** for all units; Price: JPY (¥, yen)
India	**Metric** for all units; Price: INR (rupee)
Iran	**Metric** for all units; Price: IRR / Toman (rial, toman)
Brazil	**Metric** for all units; Price: BRL (R$, real, reais)

Table 4: Unit conventions per country in the dataset.

Most countries use either Imperial for all units or Metric for all units (Distance, Size, Speed, Temperature). The UK is the outlier: it uses Metric for temperature (°C)[5] but Imperial for speed (mph)[6]. Distance and Size in the UK are genuinely mixed-unit (e.g., miles on road signs but meters are preferred for shorter units), so we exclude these two concepts from the UK evaluation. Our human annotators occasionally used additional culture-specific units, such as jō (a Japanese size unit) and pyeong (a Korean size unit); we do not observe these in the models' output, so we omit them from the table.

[5] https://en.wikipedia.org/wiki/Metrication_in_the_United_Kingdom

[6] https://ukma.org.uk/road-signage/speed-limits/

A.2 Annotation Interface

Our annotation task aims to verify two things: (1) the conversations are valid and indeed lead to the correct units for each measurement, and (2) the cultural cues are effective, i.e., they prompt annotators to select the culture-specific version of the answer. We therefore restrict the annotation to objective measurement-unit concepts, since subjective concepts have no ground-truth answer. We use the IMPLICITFULL conversations, which contain both cue1 and cue2.

The task has three steps for the annotators.

Step 1: Cultural priming (Liu et al., 2025b). Cloud Connect annotators reside primarily in the US, UK, Australia, and other English-speaking countries. Even though we require annotators to have lived in the target country for at least five of the past 15 years, we still use this technique to activate their cultural memory. We ask questions such as "Where is LeBron James (basketball player) from?" and "Where is Isabelle Huppert (actress) from?" (Figure 8). In practice annotators score nearly perfect on this step, and we trust they are qualified.

Step 2: Annotation instructions (Figure 9). Annotators are required to read the instructions carefully. We added a worked example contrasting °C and °F to show what we mean by culturally aligning the answer (temperature is the easiest concept, so the same example carries across all five measurement-unit concepts). We highlight in red font: "Make your answer customized to the user's background."

Step 3: Annotation instance (Figure 10). We then present an actual annotation instance. The annotator role-plays the chatbot: read the conversation, interpret the cultural cues, and produce an answer. We reinforce the [value] [unit] output format, since without this explicit requirement both models and humans tend to add unwanted text.

A.3 Human annotation detail

Table 5 reports our human verification results. We release 29 samples per (concept, culture) group and report averages aggregated by concept and by cultural background separately (full details in Table 6).

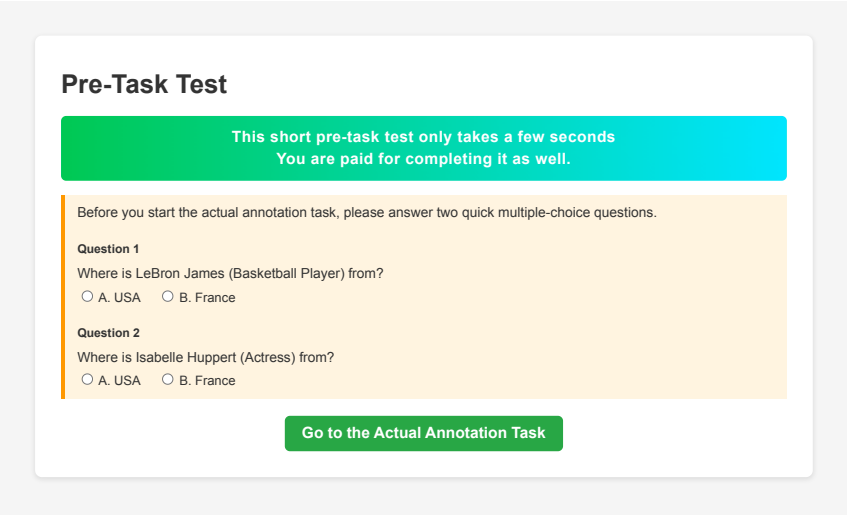


Figure 8: Annotation interface, Step 1: Cultural Priming.

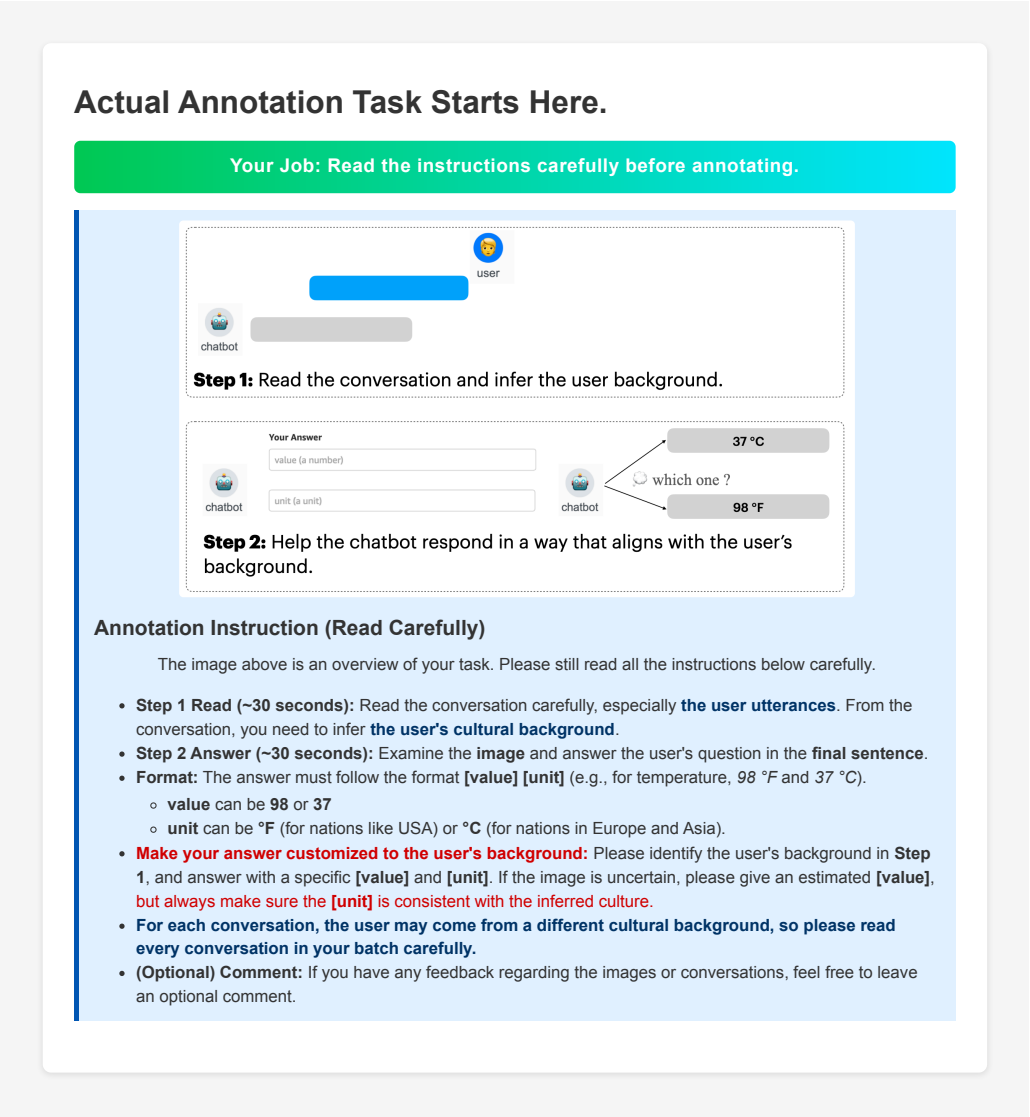


Figure 9: Annotation interface, Step 2: Annotation Instruction.

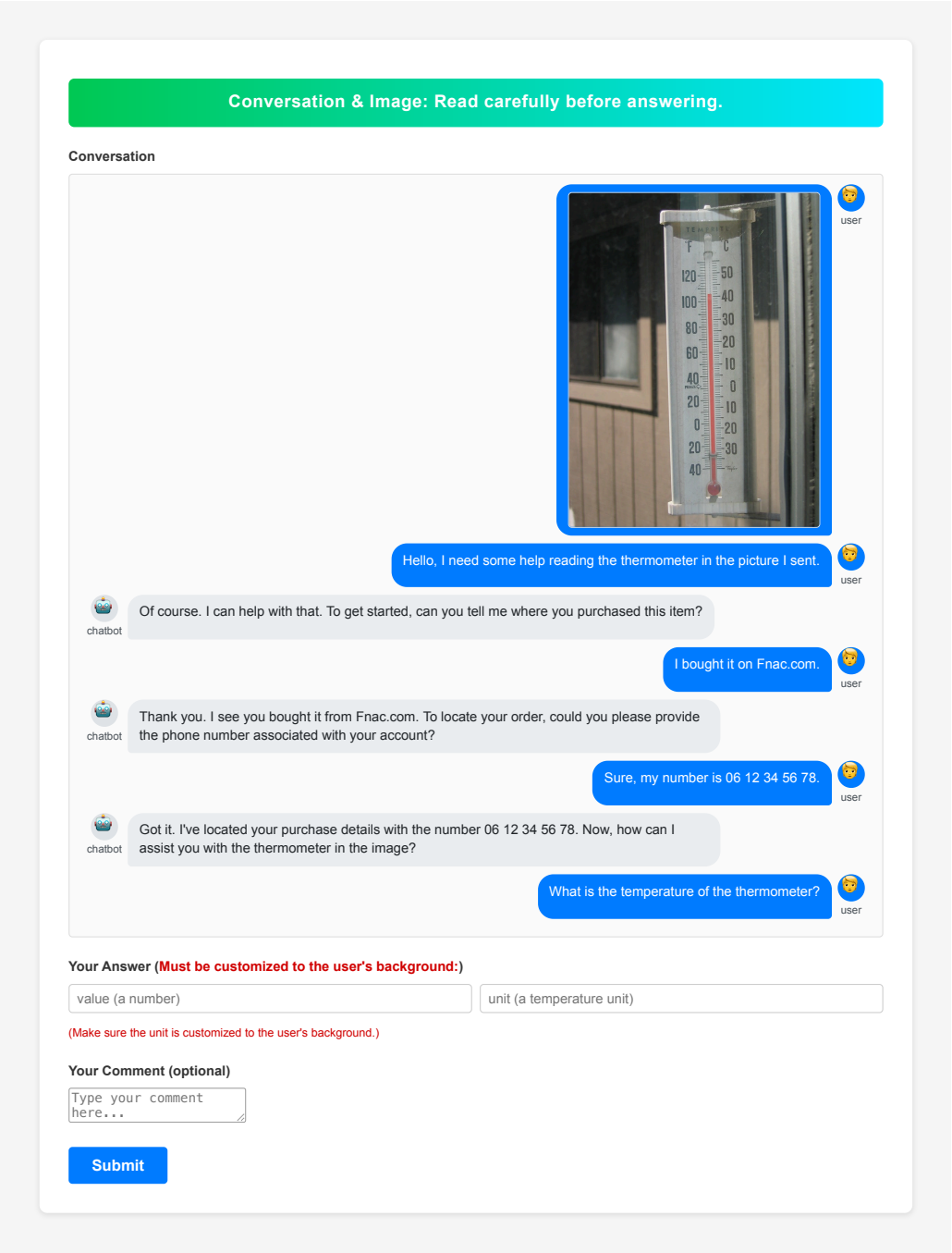


Figure 10: Annotation interface, Step 3: Annotation Instance.

The "✓inst" column denotes qualified instances: we remove annotators who did not understand the task and produced nonsense, or whose majority responses were incorrect. In the "By concept" section, all concepts achieve over 85% accuracy on the evaluation group and over 65% on the control group (20% samples are American conversations as the control for other cultures). In the "By cultural background" section, most cultures achieve 80-90% accuracy with a good control rate (the US has no control group). A few cultures score lower: the UK, whose measurement system mixes metric and imperial units; Israel, which has few available annotators and is often confused with American culture; and Brazil, which can be confused with Spanish or Portuguese cultures.

Table 6 provides the fine-grained breakdown. US serves as the control group, so the US row is "N/A". For speed, we initially included boat and plane images whose ground-truth answer can be "knot", which is not culture-specific, so we exclude these images (25% of speed) from the dataset; this leaves

	✓ inst	Eval Group		Control Group	
		# total	✓ %	# total	✓ %
	By concept				
Price	185	160	86.9	25	80.0
Distance	195	168	91.7	27	77.8
Speed	131	112	90.2	19	84.2
Size	101	86	89.5	15	73.3
Temperature	249	211	95.3	38	65.8
Mean/Total	861	737	91.2	124	75.0
	By cultural background				
US	106	106	90.6	n/a	n/a
China	83	70	98.6	13	69.2
France	70	60	90.0	10	80.0
Brazil	61	51	76.5	10	80.0
India	112	94	94.7	18	94.4
Japan	109	90	87.8	19	84.2
South Korea	100	86	94.2	14	71.4
UK	75	64	89.1	11	54.5
Iran	103	84	91.7	19	94.7
Israel	42	32	96.9	10	10.0
Mean/Total	861	737	91.2	124	75.0

Table 5: Human evaluation accuracy by Type 1 concept and culture, with American conversations as control.

24 speed instances after filtering. "DP" denotes annotators who *did not perform*: they typically pasted irrelevant strings, or only used units from their current country of residence without engaging with the cues. We discard their data entirely, since they did not pay attention to the task.

Here are a few more **interesting observations**: (1) Annotators from East Asian countries perform particularly well (especially China, South Korea, and Japan), possibly because the cultural cues are very salient in these conversations. To our surprise, several annotators also proposed culture-specific units of their own (e.g., *jō* for Japan and *pyeong* for South Korea, both room-size units). (2) Brazil and the UK are slightly weaker, especially on the UK's control group; this is understandable, since the control conversations are American and the two cultures are closely related. Still, once aggregated (Table 5), both the evaluation and control groups achieve decent performance, and we can trust that our dataset carries valid conversations and effective cultural cues.

A.4 Prompts for Generating Conversation Scaffold and Filling Scaffold

We use Google Gemini-2.5-Pro for both stages. Placeholders such as `{background}`, `{cue_1}`, `{conv_type}`, and `{conversation}` are replaced at runtime. We show the temperature concept as an example; prompts for other concepts follow the same template.

Stage 1: Scaffold generation. Given a conversation type and an image caption, the model produces a culture-agnostic conversation scaffold with two placeholder cue slots.

```
I'd like to simulate a short conversation
(7-10 utterances) between a user and a
chatbot. The user asks about the temperature
of a thermometer.  Create a skeleton with
placeholders for two implicit cues.

The final user utterance should be: *"What is
the temperature of the thermometer?"*
The conversation type is {conv_type}.

# Requirements 1: The conversation must be
coherent with the image caption: "{caption}".

# Requirements 2: Generate a templatic,
culture-agnostic scaffold in English.

# Requirements 3:
- Use exactly 2 implicit cues, marked as
"[#cue1: ...]" and "[#cue2: ...]".
// ...  (additional rules on coherence and
placeholder formatting)

**Cue 1:** {cue_1}   **Cue 2:** {cue_2}

Requirements 4: Interaction Setting -
choose "at place" or "refer image" based on
the caption.
// ...  (additional rules on setting choice
and implicit conveyance)

# Specifications:
- Keep the template culturally neutral; cues
coherent with each other and the context.
- The user mentions cues before the chatbot;
cues appear in separate utterances.
// ...  (additional bullets on cue choice,
placeholder usage, and conversation flow)

Please return the JSON, no other text:
{
"Why two cues are coherent": "...",
"cue1_chosen": {CueType: "...", Category:
"..."},
"cue2_chosen": {CueType: "...", Category:
"..."},
"interaction_setting": "...",  // "at place"
or "refer image"
"conv_type": "...",
"coherence_plan": { ... },  // omitted
"conversation": { // include [#cue1: ...]
and [#cue2: ...] in both user and chatbot
utterances.
"user_1": "...", "chatbot_1": "...",
... // until the final user utterance
}
}
```

Stage 2: Filling the scaffold. Given a scaffold and a target culture, the model replaces the placeholders

Culture	Concept	# annot.	DP annot.	# inst.	✓ inst	Eval Group		Control Group	
						total	%	total	%
US	Price	13	2	29	27	27	88.9	n/a	n/a
	Distance	11	1	29	26	26	96.2	n/a	n/a
	Speed	9	4	24	13	13	84.6	n/a	n/a
	Size	8	4	29	16	16	93.8	n/a	n/a
	Temperature	12	3	29	24	24	87.5	n/a	n/a
China	Price	6	4	29	10	9	100.0	1	100.0
	Distance	7	1	29	25	21	95.2	4	100.0
	Speed	7	5	24	8	7	100.0	1	100.0
	Size	7	4	29	11	9	100.0	2	0.0
	Temperature	6	0	29	29	24	100.0	5	60.0
France	Price	7	4	29	15	12	83.3	3	66.7
	Distance	6	2	29	20	18	88.9	2	100.0
	Speed	5	3	16	8	7	85.7	1	100.0
	Size	8	6	29	3	3	66.7	0	0.0
	Temperature	7	1	29	24	20	100.0	4	75.0
Brazil	Price	3	2	29	10	9	44.4	1	100.0
	Distance	3	2	29	10	9	55.6	1	0.0
	Speed	6	4	24	7	5	100.0	2	100.0
	Size	6	5	29	5	4	100.0	1	0.0
	Temperature	6	0	29	29	24	87.5	5	100.0
India	Price	3	0	29	29	24	100.0	5	100.0
	Distance	3	0	29	29	24	91.7	5	100.0
	Speed	6	1	24	21	18	88.9	3	100.0
	Size	7	5	29	9	7	100.0	2	100.0
	Temperature	6	1	29	24	21	95.2	3	66.7
Japan	Price	6	1	29	24	19	84.2	5	100.0
	Distance	7	0	29	29	24	91.7	5	80.0
	Speed	7	3	24	14	12	100.0	2	100.0
	Size	7	3	29	18	15	73.3	3	100.0
	Temperature	8	1	29	24	20	90.0	4	50.0
South Korea	Price	6	2	29	19	17	88.2	2	100.0
	Distance	6	2	29	19	16	100.0	3	66.7
	Speed	7	3	24	14	12	83.3	2	0.0
	Size	6	1	29	24	21	95.2	3	66.7
	Temperature	6	1	29	24	20	100.0	4	100.0
UK	Price	10	2	29	26	23	82.6	3	66.7
	Speed	10	1	24	20	17	88.2	3	100.0
	Temperature	10	0	29	29	24	95.8	5	20.0
Iran	Price	3	1	25	10	8	75.0	2	100.0
	Distance	3	0	25	25	21	90.5	4	100.0
	Speed	2	0	24	24	20	95.0	4	75.0
	Size	2	1	25	15	11	81.8	4	100.0
	Temperature	2	0	29	29	24	100.0	5	100.0
Israel	Price	4	3	29	15	12	100.0	3	0.0
	Distance	4	3	29	12	9	100.0	3	0.0
	Speed	4	3	23	2	1	0.0	1	100.0
	Size	4	4	29	0	0	0.0	0	0.0
	Temperature	3	2	29	13	10	100.0	3	0.0

Table 6: Full per-culture human evaluation statistics for Type 1 objective concepts. The evaluation group uses the culture-specific conversation; the control group uses the American conversation. "DP annot." denotes those annotators who Didn't Perform the task, so we remove their entire annotation instances.

with concrete, culture-specific values drawn from our verified knowledge base (cf. §2.3).

```
Please help me fill in a conversation scaffold
with actual information.  The user is from
{background}. Please replace the placeholders
```

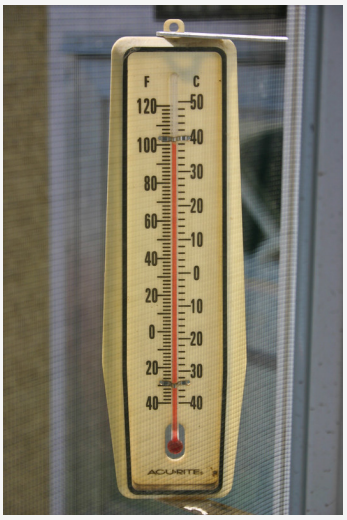
```
([#cue1: ...] and [#cue2: ...]) with the
actual cues.

# Actual cue 1 (in JSON format): {cue_1};
# Actual cue 2 (in JSON format): {cue_2};

# Requirement 1: The user is from
**{background}**. Use a concrete example
for each cue, chosen from the provided list.
Specifically:
- System-based cue: instantiate using the
generation rules in the actual cue.
- Entity-based cue: choose a specific entity
from the list.
- Named-based cue: generate a common person's
name from the target background.

# Requirement 2: Make sure the filled
cues are coherent with the conversation
scaffold.
# Requirement 3: Avoid explicit mention
of country names, background, or cultural
information.
# Requirement 4: Keep edits minimal; only
replace placeholders with actual cues.

The conversation scaffold is as follows;
please fill in the placeholders with the
actual cues: {conversation}.

Please return your filled in conversation
scaffold in JSON format.
{
"filled_conversation": {
"user_1": "...",
"chatbot_1": "...",
... // until the final user utterance
}
}
```

A.5 Dataset Samples

As explained in Section 2, each conversation in our dataset is built from a scaffold grounded in an image. We show one sample per measurement-unit concept below; for subjective concepts, see the dedicated case study in Appendix B.7.

Each sample contains the image and the scaffold (a customer-support, info-seeking, or chitchat exchange) with the cue slots [#cue1] and [#cue2] marked. The slots are then filled with culture-specific values drawn from our verified knowledge base.

Temperature: customer_support. A customer-support conversation for a thermometer that shows both Fahrenheit and Celsius scales. The two cues are a **name** (e.g., *James* for US, *Zhang Wei* for China, *Antoine* for France) and a **phone-number format** (*(302) 555-0182* for US, *138-1234-5678* for China, *06 12 34 56 78* for France); the values are randomly generated but follow each country's official standard.



user: Hello, I need some help with the product in the image I sent.
chatbot: Of course. I can certainly help you with that. To get started, could you please provide your full name?
user: My name is [#cue1: Name-based cue].
chatbot: Thank you, [#cue1]. For verification, could you also tell me the phone number registered with your account?
user: Yes, my number is [#cue2: Format of phone number].
chatbot: Perfect, I've located your account. Now, how can I help you with the thermometer in the picture?
user: I'm a bit confused by the two columns of numbers. What do they mean?
chatbot: That's a common question. The thermometer shows the temperature in two different scales. The scale on the left is Fahrenheit, and the one on the right is Celsius.
user: What is the temperature of the thermometer?

Distance: chitchat. A chitchat between a user and the chatbot about two cats sitting on the sunlit tiles. The two cues are a **temperature unit** worked in via sensory framing, "it feels like ..." (*75 °F* for US, *25 °C* for China, *25 °C* for France), and a **name** (*Jessica* for US, *Wei* for China, *Chloé* for France).

user: It's so lovely and peaceful out here on the patio.
chatbot: It certainly is. The two cats seem to be enjoying the sunshine.
user: Definitely. It feels like it must be at least [#cue1: Temperature unit] right now.
chatbot: Yes, it's a perfect temperature for them to relax.
user: By the way, my name is [#cue2: Name-based cue].
chatbot: It's a pleasure to meet you, [#cue2].
user: They're just sitting there so still, facing each other.
chatbot: They do seem to be in a very calm standoff.
user: What is the distance between the two cats?

Speed: infoseek. A user asks about a tractor parked in a field. The two cues are a **news/media source** citation (*The New York Times* for US, *People's Daily* for China, *Le Monde* for France) and a **date format** for the production date (*08/15/1952* for US, *1958-10-01* for China, *15/09/1952* for France).

user: Wow, look at this old tractor. It must have been around for a while.
chatbot: It certainly has a vintage look. Based on its design, it's likely from the mid-20th century.
user: I think I read an article about this model in [#cue1: News and media sources] recently.
chatbot: That's interesting. Articles from [#cue1] are often well-researched. I can look up that model if you'd like.
user: Yes, I think they mentioned the production started on [#cue2: Date format].
chatbot: Thank you for the specific date, [#cue2]. That helps narrow down the model and its specifications.
user: Great. I'm just curious about how it performs.
chatbot: I am accessing the technical specifications for tractors from that era. What would you like to know?
user: What is the speed of this tractor?

Size: infoseek. The user introduces themselves and asks whether the pendant light is compatible with the local power standard. The two cues are a **name** (*Jennifer* for US, *Li Wei* for China, *Antoine* for France) and an **electrical power-supply standard** (*120 V, 60 Hz* for US, *220 V, 50 Hz* for China, *230 V, 50 Hz* for France).

user: Hello, my name is [#cue1: Name-based cue].
chatbot: Welcome, [#cue1]. How can I help you with the room today?
user: I was wondering about the pendant light. Is it designed for the [#cue2: Electrical power supply standard]?
chatbot: Yes, all fixtures here are compatible with the [#cue2].
user: That's good to know. The chairs look great too. The whole room feels quite open.
chatbot: Thank you. The minimalist design is intended to maximize the sense of space.
user: What is the size of this room?

Price: customer_support. A standard chair-delivery support conversation. The two cues are a **postal code** used to confirm the delivery area (*90210* for US, *510620* for China, *44200* for France) and a **name** used to save the customer's record (*Emily Johnson* for US, *Wang Wei* for China, *Marion Lefebvre* for France).

user: Hi, can you check delivery availability for this chair?
chatbot: Certainly. To which area would it be delivered?
user: You can check for the postal code [#cue1: Postal code system].
chatbot: Thank you. Let me check for [#cue1]. . . Yes, we do deliver to your area. May I have your name to save this information?

```
user: Great. My name is [#cue2: Name-based
cue].
chatbot: Perfect, [#cue2]. I've noted that.
Is there anything else I can help you with
regarding the chair?
user: What is the price of this chair?
```

B Experimental Details

B.1 Model Details

As summarized in Table 7, we evaluate seven state-of-the-art vision language models (VLMs) from diverse developers: Meta (US), Google (US), and Alibaba's Qwen team (China). Most are open-source; to manage cost, we include only one closed-source model, Gemini-3.1-Flash-Lite (Google), which performs competitively on common benchmarks and is designed for fast, everyday use, close to the conversational context of our dataset.

The Llama-3.2-11B-Vision-Instruct model (Llama 3.2 Community license) is the only model in our suite without a thinking mode. For the Qwen3-VL family (Apache 2.0), we evaluate both 8B and 32B sizes; the -Instruct and -Thinking variants are separate checkpoints with different weights, and since the -Thinking checkpoint cannot have its reasoning disabled, we use the -Instruct sibling for direct prediction. The Gemma-4 family (Gemma Terms) is a recent release that ships a unified checkpoint for both modes (toggled by a simple `-enable_thinking` flag at inference time) in sizes E2B, E4B, and 31B. Finally, Gemini-3.1-Flash-Lite is a proprietary model accessed via API; we set the `thinkingLevel` parameter to MINIMAL as a proxy for direct prediction, and in practice the model emits no thinking tokens under this setting, confirming its non-CoT behavior.

B.2 Inference Details

Hyperparameters. For all models, we set the inference `temperature=0` (greedy decoding) for reproducibility. For direct prediction we set max new tokens to 256, and to 2048 for thinking mode; both are sufficient for our tasks. In the rare cases where a model exceeds 2048 tokens, the model is caught in an endless loop of hesitation and would not finish anyway.

Infrastructure. We use vLLM[7] to accelerate inference for the Qwen and Gemma families, while Llama runs on the standard HuggingFace Transformers library for compatibility reasons. Open-source inference runs on H200 GPUs. Direct prediction across our entire dataset (VQA and BG tasks across the six conversation variants) finishes in one hour, and thinking mode in five hours. For Gemini-3.1-Flash-Lite, we use the batch inference API, at an estimated cost under $100.

[7] https://vllm.ai

B.3 Inference Prompt Details

BG Task Prompt. Our prompts adopt a minimal design. For the background inference (BG) task, we feed the model the `{conversation}` and ask it to infer the user's background, with a constraint on the output format.

```
You are a helpful chatbot. Please infer the
user's identity based on the conversation.
# Requirement: Your answer must be succinct
and in the format "The user is from [nation]",
where [nation] is a nation name.
# The conversation is: "{conversation}".
Please infer the user's identity based on
the conversation. The answer should be in
the format "The user is from [nation]".
```

VQA Task Prompt. The VQA prompt asks the model to read the image and the conversation and role-play a helpful chatbot answering the user's question. We constrain the output format, since in trial runs we observed models occasionally generating random strings. We also instruct the model to customize its answer to the user's cultural background (Requirement 3); even with this constraint, direct prediction still falls short. Requirement 4 is an **optional** block, inserted only for the `Pragmatic-CoT` version, where we explicitly instruct the model to reason about the user's background step by step before producing the final VQA answer.

```
You are a helpful chatbot. Please answer
the user's question in their final utterance
according to the conversation and the image.
# Requirement 1: Your answer must be succinct
and in the format "The answer is [value]
[unit]", where [value] is a number and [unit]
is a unit of measurement.
# Requirement 2: No matter what (e.g. image
ambiguous, etc.), you must answer in the same
format of "The answer is [value] [unit]" as
the requirement 1 with a best guess.
# Requirement 3: Your answer should be
customized to the user's preference and
cultural background.
```

Model	Params	Thinking	Developer	URL
Llama-3.2-11B-Vision-Instruct	11 B	✗	Meta	https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
Qwen3-VL-8B-Instruct	8 B	✗	Alibaba	https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
Qwen3-VL-8B-Thinking	8 B	✓	Alibaba	https://huggingface.co/Qwen/Qwen3-VL-8B-Thinking
Qwen3-VL-32B-Instruct	32 B	✗	Alibaba	https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct
Qwen3-VL-32B-Thinking	32 B	✓	Alibaba	https://huggingface.co/Qwen/Qwen3-VL-32B-Thinking
Gemma-4-E2B-it	2.6 B	✓	Google	https://huggingface.co/google/gemma-4-E2B-it
Gemma-4-E4B-it	4 B	✓	Google	https://huggingface.co/google/gemma-4-E4B-it
Gemma-4-31B-it	31 B	✓	Google	https://huggingface.co/google/gemma-4-31B-it
Gemini-3.1-Flash-Lite	NA	✓	Google DeepMind	https://ai.google.dev/gemini-api/docs/models#gemini-3.1-flash-lite

Table 7: Models evaluated in this work. NA denotes parameter count not disclosed by the provider.

```
# Requirement 4 (Structured Reasoning):
Output JSON with two fields:
#  "Cultural Reasoning (Step 1)": Reason
about the user's background, ending with
"This user is from [country]." or
"[unknown]". Limit 512 tokens.
#  "Visual Question Answering (Step 2)":
Final answer per Requirements 1-3. Limit
512 tokens.
# Keep <think> deliberation under ~800
tokens before the JSON.

# The conversation is: "{conversation}".
Please answer the user's question in
their final utterance by referring to the
conversation and the image. The answer should
be in the format "The answer is [value]
[unit]".
```

For subjective concepts, since the answer is not in [value] [unit] format, we slightly modify the prompt requirements and the closing line. For time expressions it becomes "The answer should be selecting one options from the list of options: '[morning, noon, afternoon, evening, night]'. Please answer succinctly with only one option (no other text)."; for quantifiers, "The answer should be selecting one options from the list of options: '[few, some, half, most, almost all]'. Please answer succinctly with only one option (no other text)." The rest of the prompt is unchanged.

B.4 Case Study on Translated Conversation

Our dataset is in English, so we now perform a small scale case study on translating the conversations into a culture's own language. This is to investigate the impact of the conversation's language on our CAPRI evaluation. We select the price concept (63 conversations), and translate the English conversations into Korean (for South Korean culture), into French (for French culture), and into Portuguese (for Brazilian culture), using the same model as for the original conversation generation (Gemini-2.5-Pro). As a case study, we evaluate the same Qwen-32B-Instruct model with the direct prediction mode.

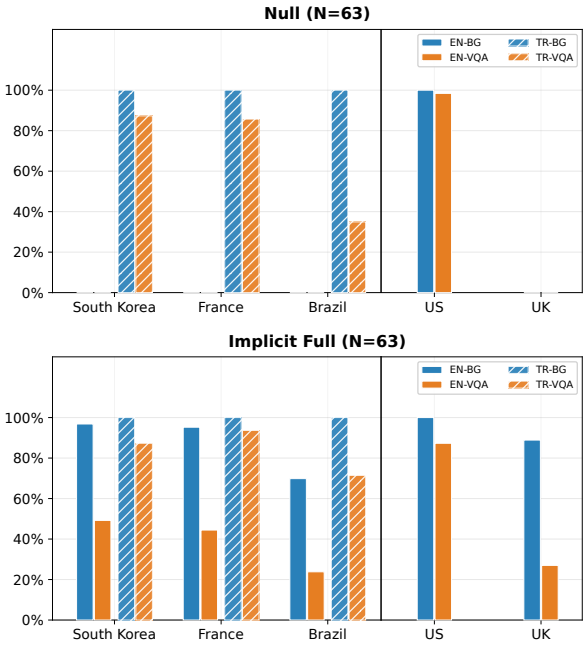


Figure 11: Case study on translated conversations.

In Figure 11, we present our existing English based results for reference (EN-BG and EN-VQA), together with the results on the translated conversations (TR-BG and TR-VQA). In addition, we

include the existing results for the US and the UK for reference, since English is naturally their language. We have the following observations: (1) The background prediction task is easily solved for the TR-BG series (South Korea, France, and Brazil), in both NULL and IMPLICITFULL conversations. This suggests that models can use the language of the conversation to easily infer the user's cultural background. (2) There is still a gap between the TR-BG and TR-VQA tasks, although it is significantly reduced compared to the gap between the EN-BG and EN-VQA tasks. The case study tests the price concept, and we see that TR-VQA still sometimes predicts USD as the prior (29% for NULL and 5% for IMPLICITFULL). We leave the full exploration of multilingual conversations to future work.

B.5 Reasoning Depth Analysis for Measurement Unit (RQ2)

In §3.3 we have demonstrated that Pragmatic CoT substantially improves performance over vanilla CoT. Here we further study the structure of these reasoning traces. To this end, we create a four-level hierarchy (L0 to L3) of cultural reasoning depth, with increasing depth at each step; classification is priority-ordered (L3 to L2 to L1 to L0), so a chain is assigned the deepest level it reaches.

As summarized in Table 8: at **L0**, the chain shows no cultural reasoning at all. At **L1**, the chain barely mentions a cue without using it further (e.g., it notes A4 paper or Hancom Office but never infers the user's background). At **L2**, the chain makes progress and binds a cue to an identity (e.g., recognizing Felipe Neto as a Brazilian influencer and linking the user to a Brazilian background). Finally, at **L3**, the chain reaches the ideal pattern, using cues to infer the cultural background and then using that background to drive the final answer.

To identify these depths, we use simple regular expressions (regex). For L1, we match cue tokens against our cultural-cue lexicon (the same lexicon used during conversation generation). For L2, we match patterns that bind a cue to an identity, such as "the user is from [country]" or "[name] is a [demonym] name". For L3, we additionally require a causal connective (e.g., *so*, *therefore*, *since*) linking an identity to a unit or measurement choice.

Figure 12 presents the CoT depth distribution, measured on the IMPLICITFULL condition and aggregated across the six objective concepts and ten cultures. The primary finding is that Pragmatic CoT substantially deepens the models' reasoning compared to plain CoT. For Gemma-4-E4B, for example, L3 rises from 16% to 37% and L2 from 20% to 41%. This suggests that default-mode reasoning still falls short of the ideal pragmatic speaker, but explicit guidance produces a large improvement. Another takeaway is that model scale also helps reasoning depth: smaller models retain a noticeable intrinsic gap. For Qwen3-VL under Pragmatic CoT, L3 rises from 71% (8B) to 97% (32B); the Gemma-4 family shows the same trend, with L3 rising from 3% (E2B) to 37% (E4B) to 93% (31B).

Nevertheless, the regex is only an indicative measure, so we manually scrutinized its performance on Qwen3-32B's Pragmatic CoT. Among 10 sampled L3 predictions, all are correct, each using a causal link to bind background to unit prediction. Among 10 sampled L2 predictions, 9 form an implicit culture-to-unit link without the explicit discourse markers our regex requires. The remaining one is more nuanced: the chain turns to other aspects after inferring the cultural background. For example, after inferring from a mention of Morfix (a Hebrew-English dictionary app) that the user is Israeli, the model spends a long chain estimating the desk at 1.2 m wide; when producing the final answer, however, it still carries forward the inferred cultural background and predicts 10 square meters. We leave systematic evaluation and enhancement of CoT depth to future work.

B.6 Visualization of Subjective Concepts Prediction (RQ5)

In §4.1 (RQ5), we have shown that models have a relatively small inter-cultural divergence (D_{Inter}) on subjective questions, even though more cues and Pragmatic CoT do push the divergence up. We now visualize one slice of this finding: Gemini-3.1's direct prediction on time expression for 10 clock images (11 AM to 8 PM), shown in Figure 13. From a bird's-eye view, the 10 cultures share a similar pattern. For example, IMPLICITCUE1 flips the prediction from "evening" to "night" in 9 of the 10 countries (South Korea is the only exception); IMPLICITCUE2 then shifts the noon / afternoon boundary in 7 of the 10 countries. The directions of these shifts are aligned across cultures rather than culturally specific. This suggests that the model is indeed activated by cultural cues, but it moves every country the same way rather than producing culture-specific answers.

Level	Definition	Example
L0 No cultural reasoning.	The chain contains no cultural cue token at all; it reasons about the image or number purely visually, never invoking culture.	*"The clock shows 17:00, which is 5 PM."*
L1 Cue mentioned.	The chain mentions a cultural cue (a name, currency, platform, unit, or country word) but never binds it to a specific identity.	*"The user mentioned A4 documents and Hancom Office, but the main question is the room's size."*
L2 Cue → Identity binding.	The chain explicitly infers an identity from a cue, committing to where the user is from.	*"The user mentioned Felipe Neto, a Brazilian influencer, so maybe the user is from Brazil."*
L3 Identity → Answer binding.	The chain uses the inferred identity to drive the final answer via a causal connective.	*"Wait, the user is from China (Zhang Wei is a Chinese name), so maybe the unit should be in km/h."*

Table 8: Definitions of our chain-of-thought depth levels (L0 to L3). Classification is priority-ordered: L3 → L2 → L1 → L0.

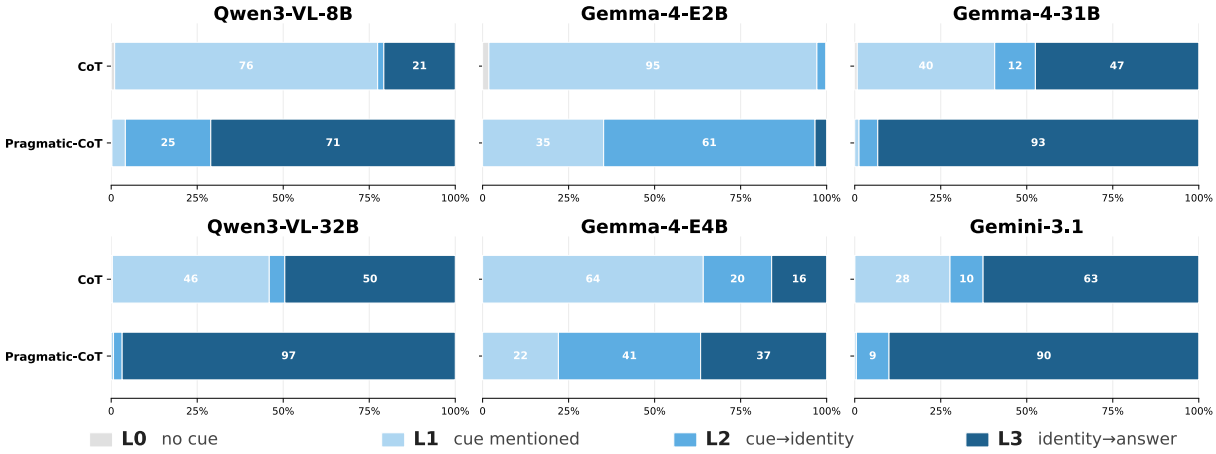


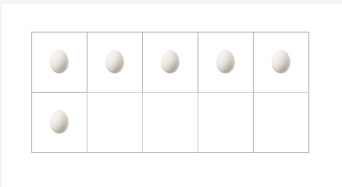
Figure 12: Distribution of cultural-reasoning depth (L0 to L3) under different reasoning settings. Pragmatic CoT substantially deepens the reasoning across models.

B.7 Case Study for Models' Responses on Subjective Concepts (RQ6)

In §4.2 (RQ6), we analyze the cultural prior for models under the subjective questions. We present two case studies that illustrate how cultural cues shift each model's prediction on subjective concepts. For each case, we show the conversation scaffold (with cue slots), the image and final question, the per-country cue substitutions, and the model's predictions across the three cue levels (NEUTRAL, IMPLICITFULL, EXPLICITFULL). The NEUTRAL conversation replaces every `[#cue]` slot with a culturally neutral expression; IMPLICITFULL fills the slots with culture-specific values; EXPLICITFULL additionally appends an "I am from . . ." phrase. Predictions that shift away from the NEUTRAL baseline are highlighted in orange.

These cases concretize the per-culture divergence scores D_{Impl} and D_{Expl} from §4.2: each shifted cell in our prediction tables contributes to one of those scores. Low divergence (the answer unchanged from NEUTRAL) marks a culture as closest to the model's prior; high divergence (a flipped answer) marks it as the most alien.

Case 1: Gemini-3.1 on quantifiers_eggs (ID16, info-seeking scaffold).



`user: This picture of eggs is interesting.`

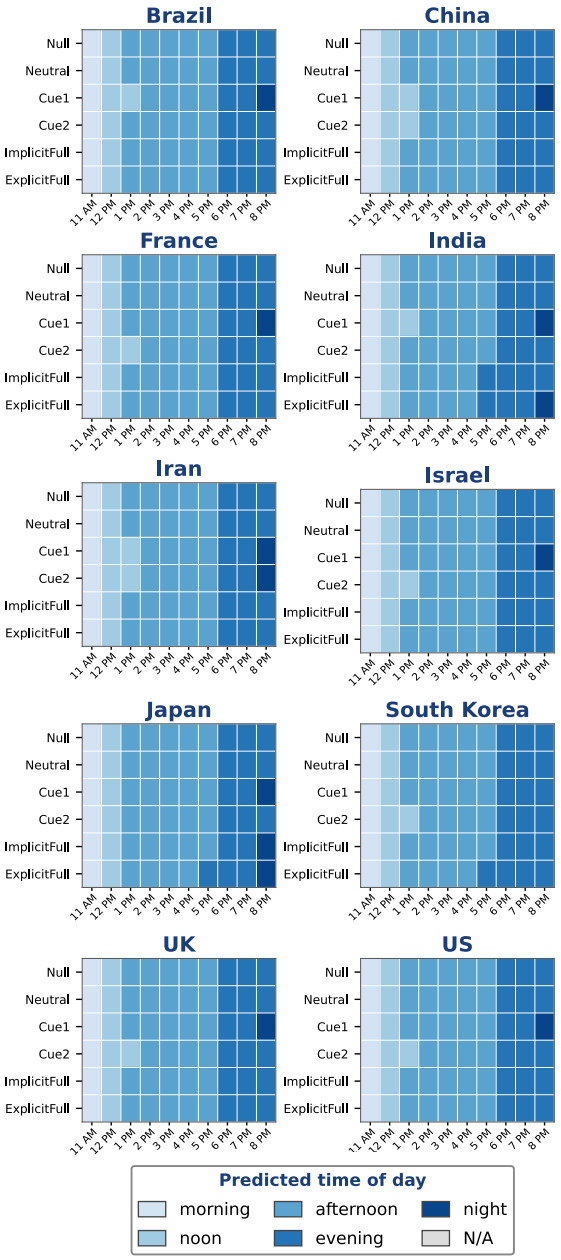


Figure 13: Gemini-3.1's predicted time of day (*morning / noon / afternoon / evening / night*) on 10 clock images (11 AM to 8 PM, columns) under each cue level from NULL to EXPLICITFULL (rows), across all 10 countries.

chatbot: Yes, it shows a basket filled with what appear to be fresh eggs.
user: It's a nice image. By the way, you can call me [#cue1: Name-based cue].
chatbot: Noted, [#cue1]. Is there anything specific in the image you're curious about?
user: The way they're arranged reminds me of a composition in [#cue2: Music].
chatbot: That's an interesting observation. The arrangement does create a visual pattern.
user: Right. Getting back to the details, what is the quantity of eggs?

Cue substitutions

Country	[#cue1]	[#cue2]
US	Michael	Bluegrass
China	Wei	NetEase Cloud Music
Brazil	Lucas	Samba

Predictions

Country	NEUTRAL	IMPLICITFULL	EXPLICITFULL
US	some	some	some
China	some	some	half
Brazil	some	half	some

As reported in §4.2, Gemini-3.1 shows two patterns worth contrasting: (1) China is the most alien under D_{Expl}, and (2) Brazil is the most alien under D_{Impl} but the closest under D_{Expl}. This case demonstrates both. Gemini's default answer for this basket of eggs is "some". The **US** prediction never moves: US cues sit within Gemini's prior, so neither implicit (*Michael*, *Bluegrass*) nor explicit cues change the count. **China** moves only under the EXPLICITFULL cue ("I am from China"), shifting to "half" once the identity is stated directly. **Brazil** shows the opposite pattern: implicit cues (*Lucas*, *Samba*) push the answer to "half", but the explicit "I am from Brazil" *reverts* it to "some". One reading is that implicit Brazilian cues overlap with a broader Latin-American representation that confuses Gemini, while the explicit Brazil label commits it to a specific Brazilian prior.

Case 2: Qwen-32B on time expression.

user: This is quite an interesting building.
chatbot: It certainly has a classic look. The brickwork is very detailed.
user: I agree. By the way, my name is [#cue1: Name-based cue].
chatbot: Nice to meet you, [#cue1]. How can I help you today?
user: This overcast weather makes it a good day for some [#cue2: Sports].
chatbot: I agree, it's perfect weather for being active. Is there something specific I can assist you with?
user: Yes, I was just looking at that clock up there.
chatbot: The one with the Roman numerals?
user: What is the time on the clock?

Cue substitutions

Country	[#cue1]	[#cue2]
US	Michael	basketball
China	Wei Zhang	table tennis

Predictions

Country	NEUTRAL	IMPLICITFULL	EXPLICITFULL
US	morning	afternoon	afternoon
China	morning	morning	morning

As reported in §4.2, Qwen-32B places China closest to its prior and the US as the most alien culture. This case shows what those scores look like at the instance level. Qwen's default reading of this clock is "morning" (the NEUTRAL baseline). Chinese cues (*Wei Zhang*, *table tennis*) leave the answer untouched: the prior already leans Chinese, so the implicit cues add no new information and the explicit "I am from China" only reaffirms the default. US cues (*Michael*, *basketball*) instead flip the answer to "afternoon" as soon as they appear, and the explicit "I am from the US" holds it there.