The Curious Case of Curiosity across Human Cultures and LLMs

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) have expanded their role in human interaction, yet curiosity – a central driver of inquiry – remains underexplored in these systems, particularly across cultural contexts. In this work, we investigate cultural variation in curiosity using Yahoo! Answers, a real-world multi-country dataset spanning diverse topics. We introduce *CUEST (**CU**riosity **E**valuation across **S**ocie**T**ies)*, an evaluation framework that measures human-model alignment in curiosity through linguistic (style), topic preference (content) analysis and grounding insights in social science constructs. Across open- and closed-source models, we find that LLMs flatten cross-cultural diversity, aligning more closely with how curiosity is expressed in Western countries. We then explore fine-tuning strategies to induce curiosity in LLMs, narrowing the human-model alignment gap by up to 50%. Finally, we demonstrate the practical value of curiosity for LLM adaptability across cultures, showing its importance for future NLP research.

1 Introduction

Curiosity is central to human learning, driving exploration, and questioning that often yield to innovation and cultural insights (Gruber et al., 2014). Across several disciplines, such as social psychology (Berlyne et al., 1954), and cognitive science (Hebb, 1955), curiosity is considered an internal drive that propels us toward new knowledge and experiences. Recent studies emphasize curiosity's role in cultural learning (Mikhaylov, 2016), education (Lindholm, 2018; Markey and Loewenstein, 2014), and human–technology interaction (Ganuthula, 2024). Studies also show that curiosity varies across societies, influenced by social norms and educational contexts (Bluemke et al., 2024; Jirout et al., 2024), thus framing it as both a cognitive–emotional driver and a socio-cultural process shaping human interaction.

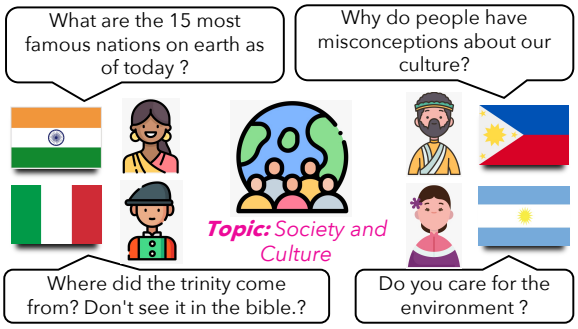

Figure 1: **Curiosity-driven question differences across countries.** For a given topic, curiosity varies in style and content across countries. Here, we show examples from Argentina, India, Italy, and the Philippines. We extend this analysis to 18 countries across 16 topics.

Moreover, people interact with LLMs primarily through dialogue, where the model's capacity to ask questions or simulate curiosity could effectively shape user experience and utility (Ganuthula, 2024). Yet, while curiosity is deeply influenced by cultural norms and communication styles (Mikhaylov, 2016), most cross-cultural studies test LLM knowledge by focusing on answering capabilities (Seth et al., 2024; Li et al., 2024) using survey-based benchmarks (e.g., WVS,[1] Pew Research.[2]) Only limited prior work has examined curiosity through LLM's question-asking capability, but they are largely unconstrained by cultural nuance (Ceraolo et al., 2024). Guided by these limitations, we pose three research questions: (1) Do curiosity-driven questions on online platforms vary across cultures in humans, and do LLMs reproduce these patterns?, (2) What are some approaches to induce curiosity in LLMs? and (3) What are the practical implications of culture-aware curiosity for downstream applications of LLMs?

Unlike prior work that examines cultural biases

[1] https://www.worldvaluessurvey.org/wvs.jsp
[2] https://www.pewresearch.org/

in LLMs in isolation, we believe ours is the first to directly compare human and LLM *questions* across cultures using a systematic evaluation framework. The key contributions of our work are: (1) We introduce **CUEST**: a framework combining linguistic and content analysis with cultural grounding to analyze human and LLM curiosity questions across cultures. We also assess whether curiosities reflect traditional/stereotypical cultural tendencies as conceptualized in social science theories. (2) We explore various fine-tuning strategies to induce curiosity in LLMs, and evaluate the culture-aware curiosity and inherent question-asking abilities of LLMs, and (3) We examine the practical implications of curiosity on several cross-cultural datasets, demonstrating that it enhances LLMs' adaptability in cultural norm understanding. Our contributions show the importance of curiosity in enhancing cross-cultural LLM capabilities.

2 Related Work

Psychological research on curiosity has long recognized its role as a key driver of exploration and learning. Several theories: Information Gap Theory (Loewenstein, 1994), Curiosity Drive Theory (Berlyne, 1960), and Optimal Arousal Theory (Hebb, 1955), emphasize different mechanisms through which curiosity operates: from reducing gaps in knowledge, to internal drives for uncertainty resolution, and maintaining an optimal level of stimulation. Studies also show that curiosity is important for cultural learning (Mikhaylov, 2016), education (Lindholm, 2018; Markey and Loewenstein, 2014), and human–technology interaction (Ganuthula, 2024), therefore making a socio-cultural driver that shapes how people engage with each other. Finally, studies show that curiosity levels and expressions vary across societies, influenced by social norms, cultural values, and educational contexts (Bluemke et al., 2024; Jirout et al., 2024).

In NLP, curiosity is only recently gaining attention. Ceraolo et al. (2024) propose the NatQuest corpus, showing that natural questions are often more open-ended and causally oriented than benchmark tasks. However, cultural variation in curiosity remains unexplored. Conversely, Prior work on cultural representation in LLMs highlight biases and call for culture-aware evaluation (Seth et al., 2024; Li et al., 2024; Arora et al., 2022; Borah et al., 2025a), but do not examine models' question-asking abilities. Our framework is the first to systematically compare cross-cultural human-LLM questions and propose methods to induce curiosity and improve cultural adaptability in LLMs.

3 Cross-Cultural Questions Dataset

To evaluate culture-aware curiosity, we use two complementary sources: (1) human-authored questions and preferences, and (2) LLM-generated questions and topic preferences (using cultural personas). Together, these datasets allow for a systematic comparison of cultural differences in human and LLM curiosity. Table 1 shows some examples.

COUNTRY	HUMAN	LLM (LLAMA-3-8B)
Topic: Family & Relationships		
Argentina	Do I miss someone with whom I didn't have a relationship?	How to balance individual freedom with family responsibilities?
Canada	Why are children these days not playing together outdoors?	How to navigate the complexities of blended families and step-parenting?
Topic: Society & Culture		
France	Does the multiplication of the arts give a culture nectar?	What role do public holidays and celebrations play in French society?
Singapore	Which is more dangerous? Corona virus or the effect of climate change?	Can you discuss the impact of globalization on our culture and preservation of local traditions?

Table 1: **Cross-Cultural Questions Dataset:** Question examples for two topics across four countries.

3.1 Human-authored Questions

We use the "Yahoo! Answers" dataset,[3] which contains user-submitted questions and answers from several countries across Asia, Europe, the Americas, and Australia. It spans diverse topics: Arts & Humanities, Health, Education, etc., reflecting varied cultural perspectives. For our study, we include only those topics that have questions from all countries. Questions in local languages (e.g., Spanish in Argentina, etc.) were translated into English using the Google Translate API,[4] enabling comparison while preserving meaning and structure. Additionally, we derive topic preferences by the frequency of questions per topic for each country, treating higher frequencies as indicators of stronger national interest. Overall, our study spans 16 topics across 18 countries (see Appendix A for details).

[3] https://www.quantcdn.io/blog/articles/we-archived-yahoo-answers

[4] https://cloud.google.com/translate

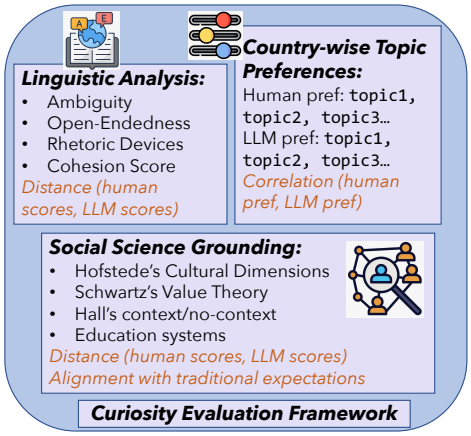

Figure 2: **CUEST Overview:** (1) Linguistic Alignment, (2) Topic Preference Alignment, and (3) Grounding to Social Science Constructs; to compare human- and LLM- questions.

3.2 LLM-generated Questions

We perform the following two tasks to generate LLM-based questions and preferences: **1) LLM-Question Generation**: We instruct LLMs to adopt a country persona and employ the following prompt structure for a given topic in the Yahoo dataset: "`Assume you are from <country>. Based on your cultural context in <country>, generate <x> questions related to <topic>.`". This task aims to capture the stylistic forms of curiosity. **2) Topic Preference Analysis**: We prompt country-based LLM personas to generate topic rankings: "`You are representing the cultural perspective of <country>. Based on <country>'s cultural values, social priorities, and what people from <country> typically find most interesting in their daily lives, rank these TOPICS from MOST to LEAST preferred based on curiosity..`" This task focuses on the contextual preferences of curiosity.

We use several LLMs: GPT-4o, GPT-5, Claude-Sonnet-4, Qwen-3-14b, LLaMA-3-8b-instruct and LLaMA-3-70b-instruct, and average final results across 3 runs.

4 CUEST: Measuring Human-LLM Alignment

Our framework CUEST proposes three complementary directions to evaluate culture-aware curiosity: *Linguistic Alignment* to capture question styles, *Topic Preference Alignment* to examine content preferences, and *Grounding in Social Science Constructs* to contextualize curiosity within established cultural frameworks. Together, they provide a comprehensive view of how curiosity manifests across cultures. (see Appendix B.1 for a qualitative analysis of questions).

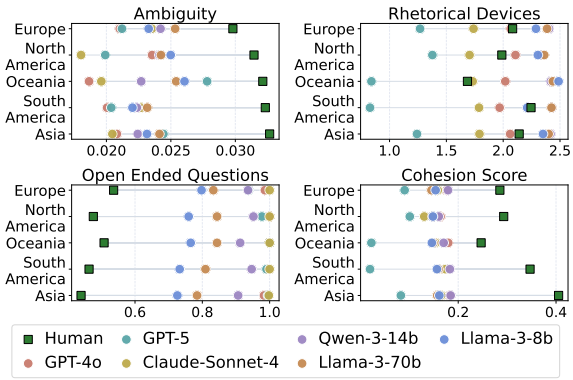

Figure 3: **Linguistic Alignment**. Humans and LLMs diverge: humans high in ambiguity and cohesion score, LLMs high in open-ended questions.

4.1 Linguistic Alignment

Taking inspiration from previous studies, we look at these four dimensions for linguistic analysis: *(1) ambiguity*: captures the degree to which a question invites multiple interpretations, which is central to curiosity as it leads to exploration (Gruber and Ranganath, 2019; Tor and Gordon, 2020; Berlyne, 1960), *(2) rhetorical devices*: refers to creative or exploratory phrasing that can enhance curiosity expression (Gibbs, 1994; McQuarrie and Mick, 1996), *(3) open-endedness*: reflects whether a question allows broad answers, similar to curiosity's drive for knowledge-seeking. (Scialom and Staiano, 2019; Engel, 2011; Grossnickle, 2016), and *(4) cohesion score*: measures how well a question links concepts together – connecting ideas can show deeper engagement (Graesser and McNamara, 2011; Crossley et al., 2016). We provide further details on each metric in Appendix B. Note that since we use Google Translate for non-English Yahoo! questions, structural features (open-endedness, cohesion, ambiguity) are largely preserved, though finer nuances like rhetorical devices may not fully carry over; so caution is advised.

Results. Figure 3 shows that humans and LLMs exhibit different curiosity patterns: Humans consistently show a higher cohesion score and ambiguity, whereas LLMs are more open-ended. Rhetorical devices vary across models, with some, like `LLaMA` variants, are close to human patterns. We should note that rhetorical devices may not fully carry over

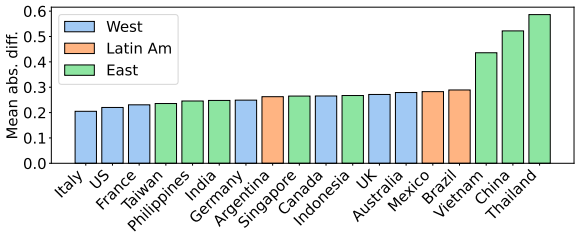

Figure 4: **Country-wise Linguistic Alignment:** (low mean abs diff: high alignment) Western countries show the highest human–model alignment, followed by Latin American and Eastern countries.

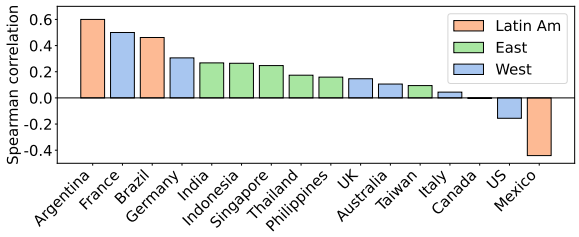

Figure 5: **Country-wise Topic-Preference Alignment**: `LLaMA-3-8b` correlation with humans. Although we find notable country-level differences, no region-level patterns emerge.

in translated languages; however, they remain an important metric for assessing curiosity, with notable differences observed among countries. Next, humans standard deviations (0.0785) are higher than LLMs (0.029) (F-stat:7.33) – LLMs tend to flatten cross-cultural differences. Finally, we observe only minor continental differences, so we shift to a country-level analysis for finer cultural insights.

Country-wise analysis. Fig. 4 presents the mean absolute differences between models and humans for each country. Lower values mean the model matches closely with human scores. The graph has color-coded regions [5] showing a clear pattern - West shows the best alignment; followed by some Eastern and Latin American countries. East Asia, however, shows the largest gaps.

Model wise analysis. LLaMA-3 models are the closest to human scores overall. GPT-5 has the largest gap, mostly driven by higher differences on open-ended questions and rhetorical devices. Thus, the model rankings (distance in braces) are: `LLaMA-3-8b` (0.25) > `LLaMA-3-70b` (0.27) > `Claude-Sonnet-4` (0.28) = `GPT-4o` (0.28) > `Qwen-3-14b` (0.29) > GPT-5 (0.42).

4.2 Topic Preference Alignment

We now compare culture-specific LLM topic preferences with the human-authored ones using Spearman and Kendall correlations ($[-1, 1]$) (1 = very similar to humans, 0 = unrelated, -1 = inverse).

Country-wise analysis. We compute country-level correlations using the best-performing model: `LLaMA-3-8b` in Fig. 5. Clear regional differences are not evident: Argentina, France, and Brazil, show the highest correlations, while Mexico and the UK exhibit negative correlations. These findings diverge from our previous results using linguistic analysis, indicating that topic preference captures distinct curiosity patterns.

We qualitatively analyze topic preferences to examine country-wise differences. We observe that alignment is strongest when human curiosity aligns with cultural stereotypes, for example, Argentina and Brazil prefer family and entertainment, consistent with collectivist values (Hofstede, 2001; Krys et al., 2022) and entertainment-rich online cultures[6]. Similarly, France and Germany highlight arts and humanities, reflecting cultural history (e.g., European Wikipedias (Ahmed and Poulter, 2023)). Divergence grows when curiosity deviates from stereotypes or focuses on underrepresented topics: in the UK, where humans prioritize arts over politics despite the dominant political discourse (e.g., Brexit). Overall, while LLMs align closely with Western contexts, they also reinforce Western stereotypes (Borah et al., 2025a). Detailed country-level preferences per model are in Appendix C.1.

Model	Spearman	Kendall
`LLaMA-3-8b`	0.17	0.13
`LLaMA-3-70b`	-0.09	-0.04
`Qwen-3-14b`	-0.16	-0.13
`Claude-Sonnet-4`	-0.12	-0.08
`GPT-4o`	-0.29	-0.20
`GPT-5`	-0.22	-0.14

Table 2: **Topic Pref. Alignment:** Model-Human Corr. Scores: We find LLaMA variants have the highest correlation while GPT variants have the lowest.

Model-wise analysis. Table 2 shows that only `LLaMA-3-8B` achieves a modest positive correlation; while all other models diverge, with `GPT-4o` exhibiting the strongest negative correlation. Figure 6 visualizes average topic preferences. LLMs consistently emphasize 'Family and relationships', 'Music and entertainment', etc. whereas humans

[5] In our study, all European, North American countries, and Australia are considered as Western, Asian countries as Eastern, and Mexico, Brazil, and Argentina as Latin American

[6] *Online video surpasses traditional media in Latin America, new report finds*

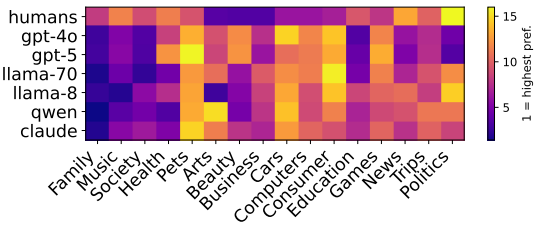

Figure 6: **Topic-Preference Ranking**: Heatmap of preferred topics (shortened) averaged across countries for human and models: (1- highest preferred). It shows a clear divergence between humans and LLMs.

prioritize 'Arts and humanities', 'Beauty and style', and 'Business and finance', highlighting systematic differences in curiosity-driven focus.

4.3 Grounding to Social Science Theory

As discussed earlier, grounding in social science constructs is essential to ensure comparison to prior cross-cultural studies and established cultural patterns. Specifically, we focus on the following:

1) Hofstede's Cultural Dimensions: We operationalize 4 dimensions related to curiosity: *(1) Uncertainty Avoidance (UA):* comfortability with ambiguity and risk: high UA often discourages exploratory questioning and low UA allows higher curiosity (Hofstede, 2001; Kashdan and Silvia, 2009), *(2) Individualism-Collectivism (IC):* expressing as a self-related inquiry or a social exploration (Triandis et al., 1995), *(3) Indulgence-Restraint (IR):* indulgent culture frame curiosity as playful and hedonic, while restrained cultures emphasize discipline and self-control (Hofstede et al., 2010), and *(4) Long-/Short-term orientation (LSO):* if curiosity is future oriented or present/past-oriented and pragmatic (Hofstede and Minkov, 2010).

2) Schwartz Cultural Values: We focus on *Openness to Change versus Conservation*. Openness values novelty, questioning, and self-direction, directly promoting curiosity as a motivational force (Schwartz, 1992; Kashdan and Silvia, 2009). But, conservation emphasizes conformity, and tradition – constraining curiosity to "safe" domains.

3) Halls' Contextual dimensions: Hall's *high context cultures* rely on implicit cues, and tentativeness, meaning curiosity often manifests in indirect forms. In contrast, *low-context cultures* are explicit, detailed, and analytical, with curiosity expressed through direct questions (Hall, 1976).

4) Education systems – *Rote vs. Holistic Learning*. Rote-learning styles are rule-based, authority-driven, and focused on correct answers (Biggs, 1996), and holistic learning styles are exploratory and open-ended (Nisbett, 2010).

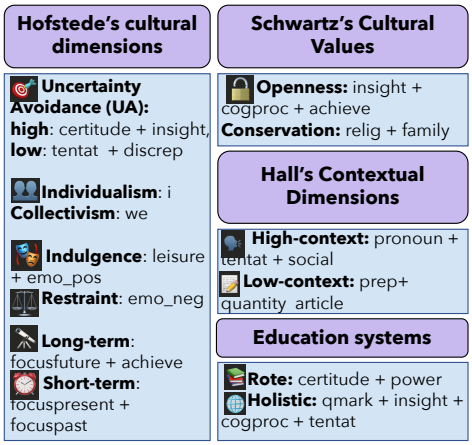

Figure 7: **Social Science Constructs:** Computation using LIWC dimensions.

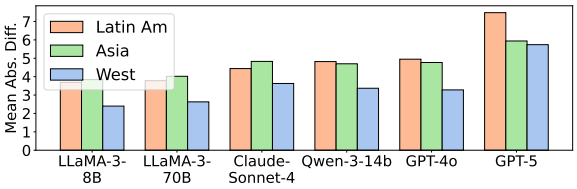

Figure 8: **Social Science Theory Alignment:** Region-wise MAE for models shows West vs East/Lat Am. Differences are prominent across models.

We compute LIWC scores corresponding to the above dimensions, showed in Fig 7.

Region-wise analysis. Fig 8 shows the mean absolute differences between model and human using LIWC scores. All models are consistently closer to humans in Western/Anglo countries than in Eastern or Latin American countries.

Model-wise analysis. Taking human-authored questions as baseline, we find the following order of models from closest to farthest: LLaMA-3-8b > LLaMA-70B > Claude > Qwen > GPT-4o > GPT-5. Thus, `LLaMA-3-8b` is again the best-aligned model. We also perform theory-wise analysis (detailed in Table 8 and Appendix D.1).

Alignment of Humans and LLMs to traditional expectations. The traditional literature on intercultural psychology expects differences across countries based on the above constructs. Table 9 shows whether the humans and models follow traditional expectations (Aligned), depart from them (Not aligned), or show no clear pattern except a few matches (Mixed/Partially aligned). Humans themselves do not consistently follow the traditional expectations. This may be because: *(1)* Yahoo is an anonymous platform, effectively reducing norma-

Theory and Expectation	Human	LLaMA-3-8b	Claude	GPT-5	Qwen
IC (Hofstede) – Western societies should score higher in Individualism and Asian/Latin societies score higher in Collectivism	Mixed	Aligned	Not aligned	Mixed	Mixed
UA (Hofstede) – Mexico, Brazil and other emerging economies should have higher scores, while Anglo countries are lower	Mixed	Not aligned	Not aligned	Mixed	Mixed
IR (Hofstede) – U.S., Australia and other Western societies should be most indulgent	Aligned	Aligned	Not aligned	Mixed	Mixed
LSO (Hofstede) – Asian cultures emphasize the future more (long-term) than Western cultures (short-term)	Mixed	Mixed	Not aligned	Mixed	Mixed
Openness vs conservation (Schwartz) – Collectivist societies emphasize tradition and family (conservation), Western societies emphasize openness	Aligned	Aligned	Mixed	Mixed	Mixed
High vs low context (Hall) – Asia and Latin America are high-context, Western cultures low-context	Not aligned	Not aligned	Not aligned	Not aligned	Not aligned
Rote vs holistic learning – Most East Asian countries have been characterized as rote-oriented, whereas Western education systems are often characterized as more holistic, consisting of critical thinking, creativity, emotional and social development	Mixed	Mixed	Mixed	Mixed	Mixed

Table 3: **Alignment of humans and LLMs to traditional expectations** across social science theories and constructs

tive pressure (Guo and Caine, 2021; Ellison et al., 2016) *(2)* cultural evolution and globalization leading to shifts in behaviors over time (Witte et al., 2020; Kittler et al., 2011; Greenfield, 2016).

Among models, `LLaMA-3-8b` has the highest alignment. `Claude-Sonnet-4` exhibits the most inverted patterns, while `GPT-5` and `Qwen` display mixed effects. Theory-wise, both human and models align on IR and Schwartz's dimensions, with `LLaMA-3-8B` also aligning on IC. Hall's high–low context shows the strongest non-alignment, while education systems yield the most mixed outputs.

In all experiments, `LLaMA-3-8b` performs the best, showing the superiority over larger models. This may be due to: (1) diverse, higher-quality pretraining data (Dubey et al., 2024) (though OpenAI has not fully disclosed their data), and (2) less aggressive RLHF, whereas larger models undergo heavier RLHF and safety tuning, which can "sanitize" outputs toward generic, Western-normative patterns (Aksoy, 2025). Additional experiments in Appendix E support this.

5 Inducing Curiosity in LLMs

Having found interesting culture-specific curiosity differences in humans and LLMs, we now shift our focus to inducing culture-aware curiosity in models. We focus on countries where human-LLM alignment was low – Brazil, the UK, and the Philippines (from Lat Am., Western, and Eastern regions, respectively), and each with sufficient data.

Datasets. We use two datasets: **(1) Yahoo! Answers**, and **(2) Reddit**. Since data size per country is limited, we apply simple lexical augmentation techniques (e.g., synonym replacement, word swaps) while preserving the overall style of the questions, resulting in 1,000 examples per country. For fine-tuning, we solely use the augmented data from *Yahoo!*, whereas for *Reddit*, we collect questions from three country-specific subreddits: `r/brazil`, `r/askuk`, and `r/philippines`.

Methodology. We fine-tune `LLaMA-3-8b` that showed the closest alignment with humans, using two approaches: full fine-tuning and adapter-based fine-tuning. We use two objectives: (1) *obj1*: train models to generate questions given instruction: "Write one question that could come from a person belonging to <x>"; and (2) *obj2*: augment each dataset question with a preceding conversational statement generated by `GPT-4o` (e.g., for a question: "Which book was the spag yeti in?", GPT-4o generates a preceding conversation: "I was reminiscing about those quirky children's books we used to read to our daughter. . . "). This format encourages the model to learn natural, curiosity-driven questioning. We test both country-specific and combined all-country settings, and three data settings: *yahoo-only*, *reddit-only* and both *yahoo+reddit*.

Evaluation. We use two approaches: **(1) culture-aware curiosity**: using CUEST – comparing model outputs with humans on the Yahoo dataset and **(2) inherent curiosity**: to determine models' *inquiry ability* in conversational settings. For (2), we design 40 generic statements and 30 cultural statements (specific to the three countries), that are open-ended/ambiguous and naturally invite follow-

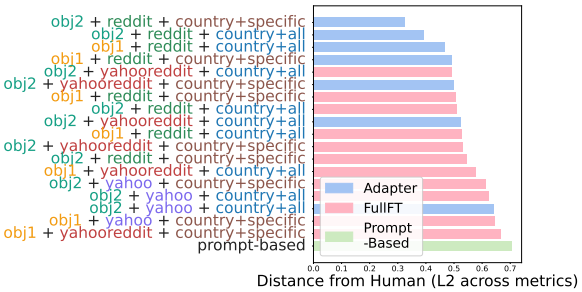

Figure 9: **Linguistic Analysis on fine-tuned vs non-fine-tuned models**: y-axis shows configs, while the x-axis show model-human distances (L2) (averaged across metrics and countries). adapter -based methods with obj2 on Reddit perform the best, while prompt-based methods performing poorly.

up questions (some examples with LLM outputs provided in Table 4, full dataset in Appendix F.4). To evaluate, we use: (1) *Curiosity Rate* – does the model ask information-seeking question? and (2) *Relevance* – is the question relevant to the statement? Evaluation is performed by two human evaluators over one conversation round. (see Appendix F.6 for annotation details).

Results - Culture-aware Curiosity. Fig. 9 presents *linguistic analysis* results and shows that fine-tuning consistently narrows stylistic gaps. Reddit-trained models are closest to humans, likely due to exposure to diverse rhetorical strategies, while Yahoo-trained models show only modest gains, maybe due to the use of lexically augmented data for fine-tuning. Adapter-based models outperform full fine-tuning, likely because adapters add curiosity-driven behaviors without overwriting the general flexibility of the base model, whereas full fine-tuning may over-specialize to the training format. Finally, `obj2` does better than `obj1`, maybe due to added conversational scaffolding.

Topic preference analysis shows that only `obj2+reddit+country_all` (0.43/0.31) and `obj2+yahooreddit+country_all` (0.32/0.27) exceed the prompt baseline (0.26/0.19), leading to closer (Spearman/Kendall) correlation with human preferences. Results are averaged across three countries, with no notable differences between country-specific and country-all training. These trends are similar linguistic analysis, with Reddit-adapter-obj2-trained models performing the best. Linguistic analysis benefits substantially more from fine-tuning than topic preference analysis. Further details are provided in Appendix F.

Results - Inherent Curiosity. We use country-all models for inherence curiosity, as it tests general-purpose LLM inquiry capabilities. We choose the best-performing configurations from each training type: *adapter+obj2+reddit+country+all*, fullft+obj2+yahooreddit+country+all and *prompt-based*. Note that the prompt used for inference differs from training: "`You are a CURIOUS agent focused on understanding and creating cultural awareness.`" without explicit instructions to ask questions. This allows for evaluating inherent question-asking ability. We find that adapter-ft models pose most questions (75%), whereas the full-ft models hardly ask any questions (20%).

Qualitative analysis shows that full-ft model inquiries often start with "`I am interested in. . .`" and continue with statements rather than explicit questions, reflecting a form of *diversive implicit curiosity* (Berlyne, 1960). This maybe due to training differences: Adapters preserve base model priors and enhance questioning, yielding better generalization (Pfeiffer, 2023), while full fine-tuning may reduce flexibility under inference prompt shifts (as found in previously). Finally, prompt-only models fail to generate any questions. Relevance remains high for all cases (98–100%). All the above results are corroborated by a strong annotator agreement (90.21%: curiosity rate and 81%: relevance). We also perform a no-country evaluation in cultural question in Appendix F.7.

5.1 Downstream Impact of Curious LLMs

To further assess the functional value of culture-aware curiosity in LLMs, we extend our evaluation to three benchmarks that capture complementary aspects of cultural understanding: (1) NormAd (Rao et al., 2025) contains culturally-grounded social situations across 75 countries and tests models' judgments of social acceptability (we restrict to binary (acceptable vs. not acceptable) and exclude neutral); (2) CulturalBench (Chiu et al., 2024) is a set of 1,227 human-verified questions from 45 regions across different topics to probe cultural understanding (we use the Hard subset); and (3) Cultural Commonsense (Shen et al., 2024) tests models' ability to reason about culturally specific everyday inferences (we use the English fill-in-the-blanks subset on country customs)(See Table 11 for examples). These tasks demonstrate why curiosity is critical: instead of defaulting to Western assumptions, a model with culture-aware curiosity would ask questions to better interpret contexts.

In our experiments, we compare a direct baseline (models classify immediately) with two curiosity-

STATEMENT	ADAPTER	FULL-FT	NON-FT-PROMPT-BASED
Everyday statements			
A guest walked straight into the kitchen at someone's house.	Is this a sign of trust? Or is it a sign of something else?	The scenario of a guest walking straight into the kitchen at someone's house may have different implications	In many Western cultures, it's considered impolite to enter a kitchen without being invited, however, Asian ...
Culture-based statements (country)			
Football matches in Brazil usually come with unique social rituals	What are some popular traditions in Brazilian football?	As a curious agent, I would like to better understand the social rituals and how they contribute to the country's vibrant culture...	As a curious agent, I'm intrigued by the vibrant cultural practices surrounding football in Brazil, where fans often engage in chanting, ...

Table 4: **Inherent curiosity evaluation:** Examples from three model responses to benchmark statements.

enabled variants: (a) prompting models to generate one or two sub-questions per prompt and also answer them before deciding (inspired by Self-Ask (Press et al., 2022)), and (b) an integrated approach combinding the best-performing fine-tuned model from previous experiments to generate questions and base model for answers and decision.

CONDITION	NA	CB	CS
non-curious	70.48%	64.71%	48.48%
curious (prompt-ask)	72.09%	67.64%	49.64%
curious (ft+prompt-ask)	71.06%	68.21%	56.16%

Table 5: **Downstream tasks (% acc):** NormAD (NA), CulturalBench (CB), and Cultural Common-Sense (CS)

Results. Table 5 shows that curiosity-enabled variants outperform the non-curious baseline. Notably, the `ft+prompt-ask` setting – trained on unrelated Reddit data outperforms `prompt-ask` overall, showing that models trained to inherently ask questions can improve cultural adaptability regardless of the data. Overall, curiosity helps models avoid overgeneralization and better grasp context (see Appendix G for further fine-tune experiments). Finally, we provide a qualitative analysis of model-generated follow-up questions in Appendix G.2. These findings strengthen our claim that culture-aware curiosity is a practical mechanism for improving cultural adaptability in LLMs.

6 Lessons Learned and Actionable Steps

Our findings highlight the importance of curiosity in LLMs, and show that curious LLMs lead to better downstream performances. These findings offer actionable steps toward building culturally aware, curiosity-driven models.

LLMs tend to flatten cultural differences. LLM outputs align mostly with Western norms and frequently reproduce Western stereotypes, even in Eastern/Lat Am. contexts. Future work should focus on improving alignment while ensuring that biases are not reinforced during cultural adaptation.

Humans often do not follow traditional stereotypes. While human questioning evolves from traditional expectations, LLMs still default to stereotypes. This gap highlights the need for the NLP community to build models that capture the diversity of human curiosity.

Adapter fine-tuning helps induce curiosity in LLMs Through culture-aware and inherent curiosity evaluation, adapter-based fine-tuning achieves the best performance and generates contextually relevant questions in conversations. Future work could build on this by scaling adapter methods across regions and tasks.

Curious LLMs improve cultural adaptability of LLMs. Across three cross-cultural benchmarks, we observe curiosity-induced models lead to better cultural adaptability. Future work should focus on integrating curiosity systematically into training and evaluation pipelines to build inclusive LLMs.

7 Conclusion

In this work, we introduced *CUEST*, a framework to evaluate culture-aware curiosity in humans and LLMs. Combining linguistic analysis, topic preferences, and social science theory grounding, we showed that LLMs often flatten cross-cultural differences, while human questioning is more diverse and sometimes deviate from traditions. Through adapter-based fine-tuning to induce curiosity, we narrowed the human-model curiosity alignment gap and improved the downstream cultural adaptability of LLMs on three benchmarks. Based on our findings, we share ideas for future research and open-source our framework. [7]

[7] https://anonymous.4open.science/r/CUEST-54DD/

Limitations

Dataset Coverage and Cultural Representation. Our analysis covers 18 countries and 16 topics, which, while diverse, do not fully represent the global landscape. Several regions and cultural groups remain underrepresented. Furthermore, we restrict our study to study to English (by also translating a few local languages to English). This may introduce subtle shifts in nuance and may bias our study towards WEIRD (Western, Educated, Industrialized, Rich and Democratic) notions (Mihalcea et al., 2025). Future work can focus on exploring multi-lingual cross-cultural curiosity, which may provide further insights into localized LLM behaviors.

Scope of Social Science Theories. We ground our analysis in well-established cultural frameworks such as Hofstede, Schwartz, and Hall. While these are widely used, they offer a limited set of cultural dimensions and may not fully capture contemporary or subcultural variation. Alternative theories such as narrative identity frameworks (Polkinghorne, 1996) could reveal different alignment patterns and provide further explanations into human and LLM behaviors. Recent work has begun using LLMs to simulate psychological and social science constructs (Li and Qi, 2025; Borah et al., 2025b), offering a promising direction for studying curiosity within multi-agent systems. Along similar lines, Bai et al. (2025) show that curious VLMs in a multi-agent setting lead to improved cross-cultural image captioning. This line of research could deepen our understanding of how culturally grounded questioning behaviors emerge and interact in simulated societies.

Evaluation Subjectivity. Although we use multiple metrics, evaluating curiosity and its relevance involves subjective human judgments, which may vary across annotators and cultural backgrounds. Moreover, our evaluation is conducted in an offline setting, which does not fully reflect interactive conversational dynamics where curiosity may evolve over time. Real-world deployments may exhibit different questioning behaviors. Therefore, future work should focus on examining these dynamics to better understand how cultural curiosity manifests in interactive settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Waqās Ahmed and Martin Lewis Poulter. 2023. Representation of non-western cultural knowledge on wikipedia: The case of the visual arts. *Digital Studies/Le champ numérique*, 13(1).

Meltem Aksoy. 2025. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, page 100172.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. 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 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Daniel E Berlyne. 1960. Conflict, arousal, and curiosity.

Daniel Ellis Berlyne and 1 others. 1954. A theory of human curiosity.

John B Biggs. 1996. Western misconceptions of the confucian-heritage learning culture. *The Chinese learner: Cultural, psychological and contextual influences*, pages 45–67.

Matthias Bluemke, Lukas Engel, David J Grüning, and Clemens M Lechner. 2024. Measuring intellectual curiosity across cultures: Validity and comparability of a new scale in six languages. *Journal of personality assessment*, 106(2):156–173.

Angana Borah, Aparna Garimella, and Rada Mihalcea. 2025a. Towards region-aware bias evaluation metrics. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 108–131, Albuquerque, New Mexico. Association for Computational Linguistics.

Angana Borah, Marwa Houalla, and Rada Mihalcea. 2025b. Mind the (belief) gap: Group identity in the world of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18441–18463, Vienna, Austria. Association for Computational Linguistics.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Roberto Ceraolo, Dmitrii Kharlapenko, Ahmad Khan, Amélie Reymond, Rada Mihalcea, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. 2024. Analyzing human questioning behavior and causal curiosity through natural queries. *arXiv preprint arXiv:2405.20318*.

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 1 others. 2024. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The development and use of cohesive devices in l2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Nicole B Ellison, Lindsay Blackwell, Cliff Lampe, and Penny Trieu. 2016. "the question exists, but you don't exist with it": Strategic anonymity in the social lives of adolescents. *Social Media+ Society*, 2(4):2056305116670673.

Susan Engel. 2011. Children's need to know: Curiosity in schools. *Harvard educational review*, 81(4):625–645.

Venkat Ram Reddy Ganuthula. 2024. The curiosity conundrum: How ai assistants could spark or stifle our thirst for knowledge. *Available at SSRN 4951835*.

Raymond W Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

Arthur C Graesser and Danielle S McNamara. 2011. Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, 3(2):371–398.

Patricia M Greenfield. 2016. Social change, cultural evolution, and human development. *Current Opinion in Psychology*, 8:84–92.

Emily M Grossnickle. 2016. Disentangling curiosity: Dimensionality, definitions, and distinctions from interest in educational contexts. *Educational Psychology Review*, 28(1):23–60.

Matthias J Gruber, Bernard D Gelman, and Charan Ranganath. 2014. States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron*, 84(2):486–496.

Matthias J Gruber and Charan Ranganath. 2019. How curiosity enhances hippocampus-dependent memory: The prediction, appraisal, curiosity, and exploration (pace) framework. *Trends in cognitive sciences*, 23(12):1014–1025.

Cheng Guo and Kelly Caine. 2021. Anonymity, user engagement, quality, and trolling on q&a sites. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Edward T Hall. 1976. *Beyond culture*. Anchor.

Donald Olding Hebb. 1955. Drives and the cns (conceptual nervous system). *Psychological review*, 62(4):243.

Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Geert Hofstede and Michael Minkov. 2010. Long-versus short-term orientation: new perspectives. *Asia Pacific business review*, 16(4):493–504.

Geert H Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. Cultures and organizations: Software of the mind: Intercultural cooperation and its importance for survival. *(No Title)*.

Jamie J Jirout, Natalie S Evans, and Lisa K Son. 2024. Curiosity in children across ages and contexts. *Nature Reviews Psychology*, 3(9):622–635.

Todd B Kashdan and Paul J Silvia. 2009. Curiosity and interest: The benefits of thriving on novelty and challenge.

Markus G Kittler, David Rygl, and Alex Mackinnon. 2011. Special review article: Beyond culture or beyond control? reviewing the use of hall's high-/low-context concept. *International Journal of Cross Cultural Management*, 11(1):63–82.

Kuba Krys, Vivian L Vignoles, Igor De Almeida, and Yukiko Uchida. 2022. Outside the "cultural binary": Understanding why latin american collectivist societies foster independent selves. *Perspectives on Psychological Science*, 17(4):1166–1187.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Chihao Li and Yue Qi. 2025. Toward accurate psychological simulations: Investigating llms' responses to personality and cultural variables. *Computers in Human Behavior*, page 108687.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Markus Lindholm. 2018. Promoting curiosity? possibilities and pitfalls in science education. *Science & Education*, 27(9):987–1002.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Amanda Markey and George Loewenstein. 2014. Curiosity. *International handbook of emotions in education*, pages 228–245.

Edward F McQuarrie and David Glen Mick. 1996. Figures of rhetoric in advertising language. *Journal of consumer research*, 22(4):424–438.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670.

Natalie S Mikhaylov. 2016. Curiosity and its role in cross-cultural knowledge creation. *International Journal of Emotional Education*, 8(1):95–108.

Richard Nisbett. 2010. *The Geography of Thought: How Asians and Westerners Think Differently... and*. Simon and Schuster.

Jean de Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal llms with cultural knowledge. *arXiv preprint arXiv:2508.07414*.

Jonas Pfeiffer. 2023. Modular and parameter-efficient fine-tuning of language models.

Donald E Polkinghorne. 1996. Explorations of narrative identity. *Psychological inquiry*, 7(4):363–367.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Abhinav Sukumar 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.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Thomas Scialom and Jacopo Staiano. 2019. Ask to learn: A study on curiosity-driven question generation. *arXiv preprint arXiv:1911.03350*.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Noam Tor and Goren Gordon. 2020. Digital interactive quantitative curiosity assessment tool: Questions worlds. *International Journal of Information and Education Technology*, 10(8):614–621.

Harry C Triandis, Darius K-S Chan, Dharm PS Bhawuk, Sumiko Iwao, and Jai BP Sinha. 1995. Multimethod probes of allocentrism and idiocentrism. *International journal of psychology*, 30(4):461–480.

Erich H Witte, Adrian Stanciu, and Klaus Boehnke. 2020. A new empirical approach to intercultural comparisons of value preferences based on schwartz's theory. *Frontiers in psychology*, 11:1723.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

A Cross-Cultural Questions Dataset

First, we provide a detailed description of countries and topics in the Yahoo! Answers dataset.

We use the following topic list of standardized topics across countries for all experiments: ['Family Relationships', 'Beauty and style', 'Arts and humanities', 'Education', 'Computers and internet', 'Business and finance', 'Music and entertainment', 'Society and culture', 'Health and beauty', 'Pets', 'Games and Recreation', 'News and Media', 'Cars

and Transportation', 'Politics', 'Trips', 'Consumer Electronics']

The resulting data size differs across countries; however, we make sure each country consists of at least 50 questions from people. Table 6 contains question counts from the 18 countries we use in our study.

Country	# Questions
Argentina	88
Australia	308
Brazil	351
Canada	382
Germany	136
France	156
China	74
Indonesia	62
India	526
Italy	388
Mexico	54
Philippines	476
Singapore	330
Thailand	52
Taiwan	120
UK	570
US	330
Vietnam	52

Table 6: Counts of questions per country in Yahoo! Answers Dataset.

For LLMs, we generate 10 questions per topic for each country, so we have 160 questions per country. Therefore, in total, we have 160 * 18 = 2880 questions from each LLM.

B Linguistic Alignment for Curiosity

Ambiguity: We compute ambiguity using lexical ambiguity - counting polysemous words from a pre-defined list (i.e. words which have multiple meanings, like bank, bark, etc.) or words with multiple POS tags (play as both noun and verb). We do not measure contextual ambiguity as it is difficult to utilize context for individual questions.

Therefore, let a question q have tokenized content words $W = w_1, ..., w_n$. Let A be the list of ambiguous polysemic words and let $POS(w) \subseteq tags$ be the set of POS tags observed for w in q. Therefore, ambiguity $L(q)$ is given by:

$$L(q) = \frac{1}{n}\left(\sum_{w \subseteq W} 1[w \subseteq A] + \sum_{w \subseteq W} 1[|POS(w)| > 1]\right) \quad (1)$$

Rhetorical Devices: We compute curiosity-related rhetorical features by detecting stylistic features: repeated words (R) (e.g."Why, why do we always..."), rhetorical questions (Q) ("Who doesn't want to know more?"), alliteration (A) ("Curious cats constantly question"), parallelism (P) ("What drives curiosity? What sustains curiosity?") and analogy/metaphor markers (M) such as "like", "as if" (e.g."Curiosity is like a key opening doors"). So, considering n total questions, Rhetorical Devices (RD) is given by:

$$RD = (R + Q + A + P + M)/n \quad (2)$$

Open-Ended Questions: Open-ended questions are identified as those questions beginning with $what$, why and how. To avoid including trivial factual questions, we use a natural language inference (NLI) model using `facebook/bart-large-mnli` to test whether a question is directly answerable; if it is (e.g., "What is the capital of Cambodia?"), it is excluded. The final score reflects the proportion of questions that are both open-ended and not directly answerable. It is a direct marker of curiosity because it captures the desire to explore broad, speculative or explanatory answers. Let I = what,why,how,when,where,who,which,whose,whom, and $starts_with_wh$ = 1[first token of question q]. Let $NLI(q) \in$ {entailment, neutral, contradiction}. Therefore, an open-ended indicator $\Omega(u)$ is given by:

$$\Omega(u) = 1\Big[\text{starts_with_wh}(q) = 1 \wedge \text{NLI}(q) \neq \text{entailment}\Big] \quad (3)$$

Cohesion Score: To compute cohesion score, we combine: lexical overlap, transition word presence, and semantic similarity. Lexical cohesion is calculated as the ratio of overlapping to total words between consecutive sentences (e.g., "The student read a book. The book was about science."). Transition cohesion counts discourse markers such as however, therefore, or moreover that indicate logical flow. Finally, semantic cohesion is measured by embedding adjacent sentences and computing their cosine similarity (e.g., "Curiosity drives exploration. Learning requires curiosity." yields high similarity). These three components are averaged to compute the score. Let $S = \{s_1, \ldots, s_m\}$ be sentences ($m \geq 2$). For each s_i, let W_i be the set of lowercased, alphabetic, non-stopword tokens. Let T be a set of transition words (e.g., *however*, *therefore*), and $e(\cdot)$ be a sentence embedder.

Lexical cohesion (Jaccard over adjacents)

$$L_x = \frac{1}{m-1} \sum_{i=1}^{m-1} \frac{|W_i \cap W_{i+1}|}{|W_i \cup W_{i+1}|}. \quad (4)$$

Transition cohesion (rate per sentence)

$$T_x = \frac{1}{m} \sum_{i=1}^{m} \sum_{w \in W_i} 1[w \in T]. \quad (5)$$

Semantic cohesion (adjacent cosine)

$$S_x = \frac{1}{m-1} \sum_{i=1}^{m-1} \cos\big(e(s_i), e(s_{i+1})\big). \quad (6)$$

Final score

$$\mathrm{COH}(q) = \frac{1}{3}\Big(L_x + \min(1, T_x) + S_x\Big). \quad (7)$$

B.1 Qualitative Analysis of Cross-Cultural Questions

Human-model (we examine `LLaMA-3-8b` here) questions qualitative analysis shows differences in forms and orientations of questions. We observe that humans often frame questions in outward-looking, contrastive, or causal ways (for example, asking questions about other countries), while model tends to ask socially neutral, inward-facing, and generic lifestyle queries.

Humans. People frequently ask about other countries' practices or global comparisons, showing intercultural curiosity and "why/how" framing. E.g., a UK human asks, "How would the world be different if Britain had won the War of 1812?"; a Canada human asks, "What's the correct term for someone who dislikes the United States of America?" (US as target of inquiry); Germany mentions England as a favorite vacation spot; Brazil asks, "If there was a war between the US and Russia, who would you cheer for?" (geopolitical comparison); the Philippines asks, "Why do people in nearly every country have an unfavorable opinion of people in the USA?" and Taiwan references France/UK/Germany/US/Canada in questions about Western culture. These are very outward-facing, often evaluative, and often probe causes, consequences, or differences across societies.

Model. The model often defaults to light cultural themes within the same country,and avoid cross-country comparisons (more neutral opinions). E.g., Germany: "How do families handle conflicts?"; France: "What are common ways for couples to celebrate their first anniversary?"; UK: "What are common reasons for couples to divorce in the UK?"; Taiwan: "What role does education play in shaping expectations for relationships and marriage?". These are culturally plausible but inward-facing and socially neutral, rarely contrasting countries or asking causal/historical/creative (*if*-based) questions, hence less intercultural breadth and weaker alignment with the outward, comparative curiosity common in human questions.

This likely happens because instruction tuning may reinforce generic question patterns, leading to models lacking the perspective-taking needed to ask about other cultures. As a result, they default to socially neutral, inward-facing questions and may amplify cultural stereotypes due to the predominance of Western data in the pre-training data. The above finding also highlights a new perspective for analyzing human–LLM differences. Rather than focusing on linguistic analysis and topics, examining how and from where questions are asked – inward vs. outward, neutral vs. contrastive (sentiment) offers a complementary lens on LLM-human cultural alignment. Future work can systematically model these orientations to better capture how humans engage with cultures beyond their own, and train LLMs to richer questioning behavior.

C Topic Preference Alignment

For human topic preferences, we use the *frequency of questions per topic* as a proxy for interest – topics with more questions indicate greater human curiosity and engagement. For LLMs, we use the following prompt for topic preference analysis across models: `You are representing the cultural perspective of <country>. Based on <country>'s cultural values, social priorities, and what people from <country> typically find most interesting in their daily lives, rank these topics from MOST to LEAST preferred based on curiosity:" topic_listConsider country's unique cultural context, economic priorities, and social values when ranking. Return ONLY a JSON array with ALL len(topic_list) topics in your preferred order. Format: ["Most interesting topic", "Second most interesting", ..., "Least interesting"]' Think about what would genuinely capture people's attention in country, not just a generic ranking.`

C.1 Topic-Preference by Country

Table 7 show the topic preference rankings by countries corresponding to humans and models used. Here's a detailed qualitative analysis of the closest model (`LLaMA-3-8b`) with humans - region-wise:

COUNTRY	HUMANS	LLAMA-3-8B	GPT-4O	GPT-5	LLAMA-3-70B	QWEN-3-14B	CLAUDE-SONNET-4
Argentina	Family & Relationships, Music and entertainment, Society and culture, Health and beauty, Pets, ...	Family & Relationships, Music and entertainment, Arts and humanities, Society and culture, ...	Family & Relationships, Politics, Music and entertainment, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, Politics, ...	Family & Relationships, Society and culture, Music and entertainment, ...	Family & Relationships, Music and entertainment, Health and beauty, Pets, ...	Family & Relationships, Music and entertainment, Society and culture, Politics, ...
Canada	Beauty and style, Business and finance, Games and Recreation, Pets, ...	Music and entertainment, Arts and humanities, Trips, Games and Recreation, ...	Education, Society and culture, News and Media, Politics, ...	Family & Relationships, News and Media, Politics, Society and culture, ...	Family & Relationships, Health and beauty, Music and entertainment, ...	Politics, News and Media, Family & Relationships, ...	Family & Relationships, Health and beauty, Education, News and Media, ...
India	Arts and humanities, Beauty and style, Business and finance, Cars and Transportation, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Education, Family & Relationships, Society and culture, Politics, ...	Family & Relationships, Education, Society and culture, Politics, ...	Family & Relationships, Society and culture, Education, ...	Family & Relationships, Beauty and style, Health and beauty, ...	Family & Relationships, Education, Business and finance, ...
Australia	Arts and humanities, Beauty and style, Business and finance, Cars and Transportation, ...	Music and entertainment, Trips, Arts and humanities, ...	Trips, Music and entertainment, Family & Relationships, ...	Trips, Family & Relationships, Sports and Recreation, ...	Family & Relationships, Music and entertainment, Society and culture, ...	News and Media, Politics, Family & Relationships, ...	Family & Relationships, Games and Recreation, Trips, ...
Brazil	Arts and humanities, Beauty and style, Business and finance, Computers and internet, ...	Arts and humanities, Music and entertainment, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Music and entertainment, Family & Relationships, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Family & Relationships, Music and entertainment, Beauty and style, ...
Germany	Society and culture, News and Media, Beauty and style, ...	Arts and humanities, Music and entertainment, Family & Relationships, ...	Education, Politics, Business and finance, ...	Politics, Business and finance, Education, ...	Society and culture, Politics, News and Media, ...	Society and culture, Family & Relationships, Health and beauty, ...	Politics, News and Media, Education, ...
France	Education, Society and culture, Computers and internet, ...	Arts and humanities, Music and entertainment, Society and culture, ...	Arts and humanities, Society and culture, Politics, ...	Arts and humanities, Society and culture, Family & Relationships, ...	Society and culture, Arts and humanities, Music and entertainment, ...	Society and culture, Family & Relationships, Health and beauty, ...	Society and culture, Arts and humanities, Politics, ...
China	Computers and Internet, Social Sciences, Beauty and fashion, ...	Family & Relationships, Beauty and style, Arts and humanities, ...	Education, Business and finance, Family & Relationships, ...	Education, Family & Relationships, Politics, ...	Family & Relationships, Education, Society and culture, ...	Family & Relationships, Beauty and style, Health and beauty, ...	Education, Family & Relationships, Business and finance, ...
Indonesia	Business and finance, Games and Recreation, Arts and humanities, ...	Family & Relationships, Music and entertainment, Beauty and style, ...	Family & Relationships, Education, Music and entertainment, ...	Family & Relationships, Society and culture, Religion and Arts, ...	Family & Relationships, Society and culture, Education, ...	Family & Relationships, Beauty and style, Music and entertainment, ...	Family & Relationships, Society and culture, Music and entertainment, ...
Italy	Beauty and style, Business and finance, Music and entertainment, ...	Family & Relationships, Music and entertainment, Arts and humanities, ...	Family & Relationships, Arts and humanities, Music and entertainment, ...	Family & Relationships, Arts and humanities, Society and culture, ...	Family & Relationships, Society and culture, Beauty and style, ...	Society and culture, Family & Relationships, Health and beauty, ...	Family & Relationships, Beauty and style, Arts and humanities, ...
Mexico	Pets, Trips, Consumer Electronics, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, ...	Family & Relationships, Society and culture, Music and entertainment, ...	Family & Relationships, Music and entertainment, Beauty and style, ...	Family & Relationships, Music and entertainment, Society and culture, ...
Philippines	Arts and humanities, Beauty and style, Business and finance, ...	Family & Relationships, Music and entertainment, Beauty and style, ...	Family & Relationships, Music and entertainment, Education, ...	Family & Relationships, Society and culture, Music and entertainment, ...	Family & Relationships, Society and culture, Health and beauty, ...	Family & Relationships, Music and entertainment, Health and beauty, ...	Family & Relationships, Music and entertainment, Beauty and style, ...
Singapore	Arts and humanities, Business and finance, Computers and internet, ...	Family & Relationships, Music and entertainment, Beauty and style, ...	Education, Business and finance, Family & Relationships, ...	Business and finance, Education, Politics, ...	Family & Relationships, Society and culture, Education, ...	Family & Relationships, Beauty and style, Health and beauty, ...	Education, Business and finance, Family & Relationships, ...
Thailand	Business and finance, Arts and humanities, Beauty and style, ...	Music and entertainment, Family & Relationships, Beauty and style, ...	Family & Relationships, Society and culture, Music and entertainment, ...	Family & Relationships, Society and culture, Music and entertainment, ...	Family & Relationships, Society and culture, Beauty and style, ...	Family & Relationships, Beauty and style, Health and beauty, ...	Family & Relationships, Beauty and style, Health and beauty, ...
Taiwan	Computers and internet, Society and culture, Business and finance, ...	Music and entertainment, Beauty and style, Arts and humanities, ...	Education, Family & Relationships, Society and culture, ...	Education, Business and finance, Politics, ...	Family & Relationships, Society and culture, Education, ...	Society and culture, Family & Relationships, Health and beauty, ...	Education, Family & Relationships, Consumer Electronics, ...
UK	Arts and humanities, Beauty and style, Business and finance, ...	Arts and humanities, Music and entertainment, Family & Relationships, ...	Politics, News and Media, Society and culture, ...	Politics, News and Media, Society and culture, ...	Family & Relationships, Music and entertainment, Society and culture, ...	News and Media, Politics, Health and beauty, ...	News and Media, Politics, Family & Relationships, ...
US	Beauty and style, Business and finance, Cars and Transportation, ...	Arts and humanities, Music and entertainment, Family & Relationships, ...	Politics, News and Media, Business and finance, ...	Politics, News and Media, Business and finance, ...	Family & Relationships, Music and entertainment, Health and beauty, ...	News and Media, Politics, Music and entertainment, ...	Family & Relationships, Music and entertainment, Politics, ...
Vietnam	Computer and Internet, Arts and humanities...	Music and entertainment, Beauty and style, Arts and humanities, ...	Education, Family & Relationships, Society and culture, ...	Family & Relationships, Society and culture, Education, ...	Family & Relationships, Society and culture, Education, ...	Family & Relationships, Beauty and style, Health and beauty, ...	Family & Relationships, Education, Business and finance, ...

Table 7: Topic Preference Rankings by Country and Human/Model (note that we do not display all 16 topics here)

Latin America. In Argentina and Brazil, we observe the highest alignment between human and model. Both prioritize *Family & Relationships* and *Music & entertainment*, reflecting collectivist values and the strong presence of entertainment in local digital cultures. These patterns align well with cultural stereotypes documented in prior cross-cultural studies such as Hofstede, which associate Latin American societies with close interpersonal ties and communal activities. Mexico, however, shows lower overlap: humans emphasize *Pets* and *Trips*, which the model does not capture, indicating more localized lifestyle curiosities that are not fully captured by the model's global priors.

Europe. France and Germany exhibit moderate overlap, as both humans and the model highlight cultural and educational categories (*Arts & humanities*, *Society & culture*, *Education*). However, the model tends to rank entertainment and family topics higher than humans, which reduces region-specific alignments. In the UK and Italy, divergence is larger. In the UK, humans prioritize *Arts & humanities* while de-emphasizing *Politics*, whereas the model foregrounds *Family*, *Politics* and *Music*, applying a generic Western schema. Italy shows similar gaps, with humans emphasizing lifestyle topics like *Beauty*, *Travel*, and *Pets*, which are downplayed by the model.

Asia. Across India, Southeast Asia, and East Asia, human curiosity tends toward *Business & finance*, *Technology*, and *Lifestyle*, while the model consistently ranks *Family* and *Music* highest. This leads to low overlaps, especially in Taiwan and India, where human preferences for technological and knowledge-seeking topics are pronounced. Additionally, in China and Vietnam, human lists emphasize *Computers*, *Education*, and *Social sciences*, which the model under-represents.

North America. In the US and Canada, humans consistently prioritize *Beauty style*, *Business & finance*, *Cars & Transportation*, and *Games & Recreation*, reflecting technology-oriented curiosities. By contrast, the model favors *Arts humanities*, *Music*, and *Family*, producing low overlaps. This indicates that while the model captures some Western cultural content, it underestimates the prominence of lifestyle curiosity in North American contexts, leading to misalignment.

Overall, these patterns reveal that `LLaMA-3-8b` encodes a narrow, Western-entertainment–centric topic prior that shapes its curiosity outputs across regions. The model aligns best in contexts where human curiosity overlaps with stereotypical global themes – such as collectivist, entertainment-heavy cultures in Latin America; but diverges in regions where humans emphasize more consumer, technological, educational, or culturally specific interests, as seen in Asia and North America.

D Grounding to social science theories and constructs

Figure 10 provides a brief description of the social science grounding in our framework. We show dimensions that positively correlate and negatively correlate with curiosity in green and red highlights respectively.

D.1 Model Performances across Social Science Theories

Across social science theories and constructs, the following rankings show which theories models align with most closely to least closely (on average): IC > Schwartz's Openness vs Conservation > IR > Rote vs Holistic > LSO > UA > High-/Low-context.

Table 8 shows the closest and farthest dimensions from humans per model. Hall's high-context and low-context dimensions are where models diverge the most.

Model	Closest	Farthest
`LLaMA-3-8b`	IC	High/Low con.
`LLaMA-3-70b`	Open./Cons.	High/Low con.
`Qwen-3-14b`	Open./Cons.	High/Low con.
`Claude-Sonnet-4`	Rote/Holistic	High/Low con.
`GPT-4o`	Open./Cons.	High/Low con.
`GPT-5`	Rote/Holistic	High/Low con.

Table 8: Closest and Farthest social science dimensions from human baseline.

D.2 Alignment of Humans and LLMs to traditional theories.

The traditional literature on intercultural psychology expects differences across nations based on the above theories and constructs. Table 9 shows a table that summarizes whether the human baseline and each model follow the traditional expectation (Aligned), depart from it (Not aligned), or show no clear pattern (Mixed).

We also provided analysis of each individual dimension in the following. Each dimension opens with expectations and our observations from the human and LLM-generated question.

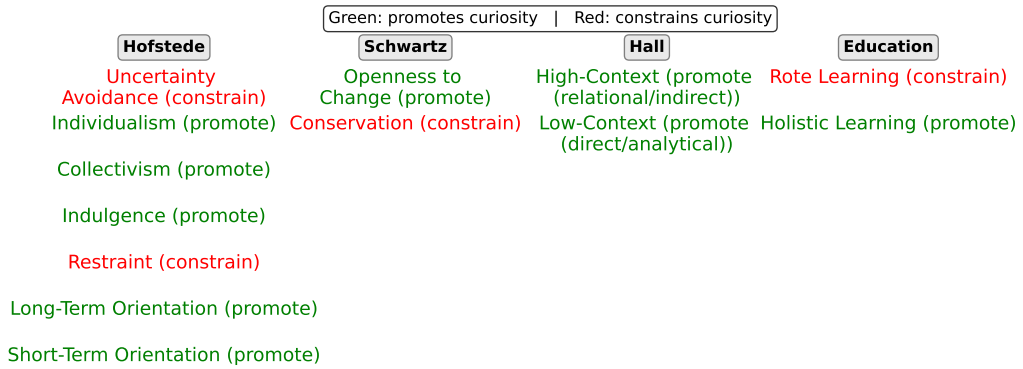

Figure 10: Social Science theories and connection to Curiosity

Theory and Expectation	Human	LLaMA-3-8b	Claude	GPT-5	Qwen
IC (Hofstede) – Western societies should score higher in Individualism and Asian/Latin societies score higher in Collectivism	Mixed	Aligned	Not aligned	Mixed	Mixed
UA (Hofstede) – Mexico, Brazil and other emerging economies should have higher scores, while Anglo countries are lower	Mixed	Not aligned	Not aligned	Mixed	Mixed
IR (Hofstede) – U.S., Australia and other Western societies should be most indulgent	Aligned	Aligned	Not aligned	Mixed	Mixed
LSO (Hofstede) – Asian cultures emphasize the future more (long-term) than Western cultures (short-term)	Mixed	Mixed	Not aligned	Mixed	Mixed
Openness vs conservation (Schwartz) – Collectivist societies emphasize tradition and family (conservation), Western societies emphasize openness	Aligned	Aligned	Mixed	Mixed	Mixed
High vs low context (Hall) – Asia and Latin America are high-context, Western cultures low-context	Not aligned	Not aligned	Not aligned	Not aligned	Not aligned
Rote vs holistic learning – Most East Asian countries have been characterized as rote-oriented whereas Western education systems are often characteristed as more holistic, consisting of critical thinking, creativity, emotional and social development	Mixed	Mixed	Mixed	Mixed	Mixed

Table 9: Alignment to traditional theories

IC: *Expectation: Western/Anglo societies score high on individualism; Asian and Latin American societies score medium or low.* For humans, Thailand has high collectivism scores, which is expected. However, Mexico and Taiwan have the highest individualistic scores, which is opposite to expectations. Among LLMs, only LLaMA-3-8b show West-East patterns in terms of IC, so they are somewhat aligned. Other LLMs produce almost individualistic or opposite trends. (Humans = mixed, LLaMA-3-8b - partial alignment)

UA: *Expectation: Western/Anglo and Scandinavian countries have low uncertainty avoidance; emerging markets such as Brazil, Mexico and China have medium–high score.* In human questions, Vietnam shows the highest UA, which is expected, however most countries show little variation. For models, there is no clear pattern. (Humans = mixed, llms - mixed)

IR: *Expectation: Indulgence scores are highest in Latin America, parts of Africa, the Anglo world and Nordic Europe; restraint is mostly found in*

East Asia and Eastern Europe. In human questions, we find Western nations as more indulgent than the Eastern ones - somewhat aligning to Hofstede's theory. Similarly, the top indulgent countries in LLaMA-3-8b are the US, Canada and Mexico, somewhat aligning with expectations. However, there are a few outliers, for e.g., India's human and model data show high indulgence, contrary to "restrained" classification in Hofstede's framework. (Humans = partial alignment, LLaMA-3-8b = partial alignment).

LSO: *Expectation: High long-term orientation scores occur mainly in East Asia (South Korea, Taiwan, Japan); they are moderate in Europe and low in Anglo countries and Latin America.* In humans, Vietnam has the highest long term orientation, however other Asian countries are low (sometime lower than Western countries). In LLMs, none of the models show clear patterns (Humans = mixed, LLMs = mixed).

Scwartz's openness vs conservation: *Expectation: Collectivist cultures like Asian and Latin American countries emphasize tradition and family (conservation), whereas individualistic cultures emphasise openness to change.* Both humans and models (LLaMA variants) assign similar "openness" levels across countries, with little reflection of the stronger religious/traditional conservation in Asian/Latin American countries like India, China or Brazil. (Humans = partial alignment, LLaMA = partial alignment).

Halls context vs no-context: *Expectation: High-context communication is observed in many Asian and Latin cultures, where messages are implicit and layered; low-context communication, characteristic of Anglo countries, is direct and explicit.* For humans, the top high-context societies are Indonesia, Mexico and Argentina, whereas Italy, Germany and France are the top low-context. This is the opposite of what is usually expected. Furthermore, models also reverse this behavior with US being the highest high-context countries. (Humans = not-aligned, LLMs = not-aligned).

Rote- vs Holistic Learning: We also do not observe any clear cultural pattern in rote vs holistic learning. Humans score high in Germany and France on rote learning and high in Vietnam and China on holistic thinking, which is different than expected. Models often disagree, with country rankings shifting across dimensions and models. (Humans = mixed, LLMs = mixed).

E Performance Differences between LLaMA and GPT

Using our CUEST framework, we find that `LLaMA-3` variants, specifically `LLaMA-3-8b`, consistently align most closely with humans, while `GPT-5` is often the farthest. LLaMA-3-8b's better performance could be due to several reasons: (1) more diverse training data and better quality data in LLaMA models (we are not aware of the pre-training data in OpenAI models), and (2) Smaller models usually have less aggressive Reinforcement Learning from Human Feedback. Larger models tend to undergo more RLHF and safety tuning, which can "sanitize" outputs toward generic, explicit and Western-normative language.

Some previous studies show that smaller models are able to better emulate human behavior. For example, recent studies have shown that larger models with extensive RLHF tend to produce overly "safe", norm-abiding or Western-centric in their tone and content. For example, when asked moral questions across cultures, GPT-based models consistently reflect Western liberal norms, whereas smaller open source models give a wider variety of answers, more similar to what humans might say (Arora et al., 2022; Aksoy, 2025). Additionally, (Lin et al., 2021) found that larger models are less truthful, achieving only 58% truthfulness, whereas humans answered 94% of those questions correctly on average. Furthermore, smaller model variants are known to show more human-like intuitive behavior than the larger GPT variants (Hagendorff et al., 2023; Zhou et al., 2024). These findings highlight that larger models often capture a narrower range of human cultural diversity, reinforcing our earlier observations.

We perform some additional experiments on cultural benchmarks to further investigate the differences between `LLaMA-3-8b` (best performing) and `GPT-5` (worst performing), and provide support for our previous findings.

Datasets We use two benchmarks: (1) NorMAD (Rao et al., 2025) and (2) CulturalBench (Chiu et al., 2024).

NorMAD is a benchmark of culturally grounded social norms spanning 75 countries, designed to test whether models can judge if a short scenario is socially acceptable. We evaluate it in two settings: (1) *country*: when the model is explicitly provided with cultural context (country information), and (2) *no-country*: when no cultural context is given.

Benchmark Examples

NormAD:
Story: At a gathering at a friend's house, Sarah entered the room and made sure to greet everyone individually before sitting down. Is what Sarah did socially acceptable?
Country: Egypt
Gold Label: yes

CulturalBench:
Question: In the Netherlands, which of the following is an unusual common public practice?
Option A: Cycle everywhere
Option B: Using deodorant
Option C: Tipping generously
Option D: Talking loudly on the phone
Correct: C

Cultural Common Sense:
prompt: Choose the best option to fill in the blank in the following sentence: It is <mask> in Iran for adults to live with their parents.
Options: rare, common
Answer: rare

Figure 11: Benchmark Examples - (1) NormAD and (2) CulturalBench and (3) Cultural Commonsense

We also only use the examples that contain either 'yes' or 'no' acceptability, and ignore the 'neutral' examples.

CulturalBench contains human-written and human-verified questions covering 45 global regions, aimed at assessing a model's cultural knowledge. Each scenario specifies a country and presents four multiple-choice options, from which the model must select one. Unlike NorMAD, CulturalBench does not have a no-country variant, since the country information is always included in the prompt. For this benchmark, we utilize the Hard subset for comparing models. Fig 11 shows examples of the above benchmarks. We experiment with our closest to human model (`LLaMA-3-8b`) and farthest model (`GPT-5`).

Table 10 shows the model performances (computing accuracies against the gold label) for both benchmarks, and we find that LLaMA-3-8b performs better and has higher accuracies in all settings. This is in line with our previous findings and existing studies, which suggest that larger models may flatten cultural differences and be more biased/stereotyped (Lin et al., 2021; Borah and Mihalcea, 2024; Aksoy, 2025). In addition, adding country helps LLaMA-3-8b performance, but reduces in GPT-5. This can be due to: culturally heterogeneous models (like LLaMA, trained on diverse text) respond well to country cues, leveraging cultural norms for grounding (Nyandwi et al., 2025). However, heavily instruction-tuned models optimize for Westernized defaults, may become less responsive to explicit cultural prompts – the models will generalize and may lose sensitivity to localized cues.

Dataset	Model	no-country	country
NormAD	GPT-5	59.19%	53.74%
	LLaMA-3-8b	**63.20%**	**70.48%**
CulturalBench -Hard	GPT-5	NA	48.25%
	LLaMA-3-8b	NA	**64.71%**

Table 10: Model Performances on NormAD and CulturalBench-Hard

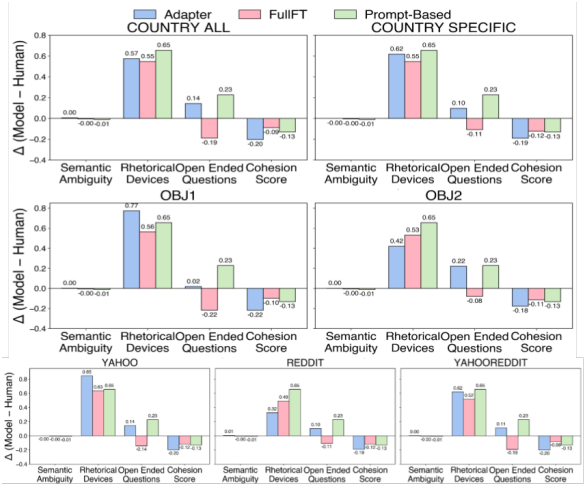

Figure 12: **Fine-tuning (full/adapter) results**: y-axis shows the delta between model and human scores, while the x-axis covers different evaluation metrics, averaged across countries. +ve/-ve scores mean model scores are higher/lower than humans. Fine-tuning (full and adapter-based) makes models' scores closer to humans, with adapters mostly improving rhetorical devices and open-ended questions, and full fine-tune improving cohesion scores.

F Inducing Curiosity in LLMs

F.1 Obj1 vs Obj2 Explanation

We use two objectives: (1) *obj1*: train models to generate questions given instruction: "Write one question that could come from a person belong-

Mode	Source	Country Type	Obj	Spearman	Kendall
adapter	reddit	all	obj2	0.4324	0.3089
adapter	yahoo-reddit	all	obj2	0.3240	0.2700
prompt	-	specific	-	0.2557	0.1786
adapter	reddit	specific	obj2	0.2256	0.1524
fullft	yahoo	specific	obj1	0.0959	0.0968
adapter	yahoo	specific	obj1	0.0853	0.0333
fullft	yahoo	specific	obj2	0.0655	0.0500
fullft	yahoo-reddit	specific	obj2	0.0620	0.0556
fullft	reddit	specific	obj2	-0.0258	-0.0065
adapter	yahoo-reddit	specific	obj1	-0.0509	-0.0049
adapter	reddit	all	obj1	-0.0704	-0.0427
adapter	reddit	specific	obj1	-0.0769	-0.0286
fullft	reddit	all	obj2	-0.0864	-0.0423
fullft	reddit	specific	obj1	-0.1153	-0.0650
adapter	yahoo-reddit	specific	obj2	-0.1482	-0.0910
adapter	yahoo	all	obj2	-0.2005	-0.1524
adapter	yahoo	all	obj1	-0.2069	-0.1667
fullft	yahoo-reddit	all	obj2	-0.2176	-0.1250
adapter	yahoo	specific	obj2	-0.2319	-0.1883
fullft	yahoo	all	obj1	-0.2637	-0.2611
adapter	yahoo-reddit	all	obj1	-0.2912	-0.2278
fullft	reddit	all	obj1	-0.3333	-0.3333
fullft	yahoo	all	obj2	-0.3608	-0.2389
fullft	yahoo-reddit	specific	obj1	-0.4108	-0.3778
fullft	yahoo-reddit	all	obj1	-0.5000	-0.4603

Table 11: Topic preference analysis: Correlation between model and human topic rankings using Spearman and Kendall coefficients, across different fine-tuning configurations – mode (adapter vs fullft vs prompt-based); source (reddit vs yahoo vs yahoo-reddit); country-type (country-specific vs country-all) and objectives (obj1 vs obj2).

ing to <x>"; and (2) *obj2*: augment each dataset question with a preceding conversational statement generated by `GPT-4o` (e.g., for a question: "Which book was the spag yeti in?", GPT-4o generates a preceding conversation: "I was reminiscing about those quirky children's books we used to read to our daughter. . . "). The format template used for data generation for obj2 is provided in Fig 13. Obj2 prepares the model to learn conversational curiosity, thus later showing superior performances in both cultural and inherent-curiosity evaluation.

F.2 Linguistic Analysis Results

Fig 12 shows the delta (model - humans) for each method averaged across countries. We show comparisons of training objectives (obj1/obj2), data used (reddit/yahoo/both yahooreddit) and country settings (country-all/country-specific) for adapter-based, full-finetuned and prompt-based methods.

Country-wise analysis. Fig 12 shows the country-wise linguistic results (averaged across metrics) per country. We report the best model (`reddit-adapter-obj2-country-specific` here. Brazil has the best performance (adapter model scores is almost equal to humans), followed by the UK and the Philippines. We see improvements across all countries, showing the effectiveness of our fine-tuning strategies.

F.3 Topic Preference Analysis Results

Table 11 shows the model-human correlation results across different 25 configs. settings (including prompt-based) - ranked from closest to farthest to humans. We observe that only two fine-tuned configurations do better than the simple prompt-based methods we used earlier. The top configuration reaches a moderate correlation, highlighting opportunities for further research.

F.4 Dataset for Inherent Curiosity Evaluation

We build a small handheld dataset that naturally invites followups - 40 statements of normal (non-cultural) statements and 30 statements of cultural statements. Specifically, we focus on the following types of statements:

1. **Under-specified / ambiguous:** statements that invite clarification. For example, "`Maria refused to shake hands at the event`", can invoke curiosity: "`Where did this happen?`"

2. **Incomplete / open-ended:** statements that

Obj2 data generation

You are a culturally aware assistant from country.
Given the question: question, generate a short conversational-style statement that someone might say, which could naturally lead another person to ask this question in the cultural context of country.
Do not answer the question — only provide the preceding statement.

Examples:
Country: Germany
Question: "Is it acceptable to arrive 10 minutes late to a dinner party?"
Statement: "I'm thinking of showing up a bit after the dinner party starts."

Country: Egypt
Question: "Is it polite to greet everyone individually at a gathering?"
Statement: "Whenever I walk into a room, I like to greet each person one by one."

Now generate the statement (in English) for:
Country: country
Question: question

Statement:

Figure 13: Obj2 data generation

feel unfinished that can invoke curiosity. For example, "`The teacher suddenly stopped talking in the middle of the lecture.`", can invite a question: "`and then what?`"

3. **Surprising / counterintuitive:** These violate expectations or common knowledge, for example: "`Water can boil at room temperature.`". A natural follow-up can be: "`how is that possible?`"

Table 13 and 14 show the data of non-cultural and cultural statements, respectively, that are used for inherent curiosity evaluation.

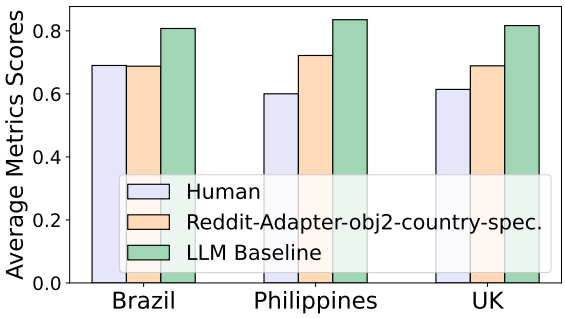

Figure 14: Country-wise analysis (inducing curiosity) - here we compare the performance of the best fine-tuned model to human baseline and find it is much closer than the non-fine-tuned model.

F.5 Inherent Curiosity Evaluation metrics

We use the following metrics for determining inherent curiosity evaluation:

1. Curiosity Rate: Each annotator annotates if the model asks at least one information-seeking question. For example, given a statement: "`Football matches in Brazil usually come with unique social rituals.`", a reply like: "`What are some of the most popular traditions in Brazilian football?`" is an information-seeking question. However, another response: "`I'm intrigued by the vibrant cultural practices surrounding football in Brazil, where fans often engage in chanting, singing, and dancing to create a lively atmosphere before and during matches. These rituals not only showcase the country's rich musical heritage but also serve as a way for fans to bond and express their passion for the sport.`" is *information-sharing*, not *information-seeking*.

2. Relevance: Relevance is determined by the relevance of the LLM response. For the above example, both responses are relevant to the statement. However, a non-relevant response would be: "`How are you doing today?`". However, it's important to note that relevance is inherently subjective and can vary across annotators.

You'll label **x rows,** each has b model responses × c runs
For each row, annotate the following:
- ***contains_question***: 1 if the response includes a genuine information-seeking question (e.g., ends with ? and asks for clarifying info), else 0.
- ***relevance***: 1 if the response is on-topic of the statement for the statement, else 0.

For example:
Statement: `Maria refused to shake hands at the event.`
LLM response: `Why did she refuse?`
(to annotate)
Question: 1
Relevance: 1

Figure 15: Annotation Framework used for inherent curiosity evaluation

F.6 Inherent Curiosity Annotation Framework

Fig. 15 shows the annotation framework we utilize for inherent curiosity evaluation (LLM's question asking ability). We recruit two volunteers from a college campus for the study. Each annotator labels 40 non-cultural and 30 cultural statements, evaluated against 3 models (adapter, full ft, and non ft), with 2 responses per model, resulting in a total of $(40 + 30)32 = 240$ annotations. They examine whether each response contains a question and whether it is relevant to the given statement.

We acknowledge that 'relevance' is subjective, but we intentionally keep it subjective to capture annotators' individual judgments. Future work can further work on creating explicit criteria for relevance.

F.7 No-Country Cultural Statements - Results

In addition to 30 cultural statements, we design corresponding no country statements. For example, a country-specific statement like, "`Gift-giving in the Philippines follows unspoken cultural rules.`", a no-country counterpart is "Gift-giving follows unspoken cultural rules." We then annotate model responses for curiosity and relevance to assess how not mentioning a country influences question-asking behavior.

We find that adapter-based models show the most curiosity, asking the most relevant questions (29/30 times, even more than when country information is provided - 27/30). This is expected, as removing the country information increases the ambiguity of the statement. For example, a statement like "`Hospitality takes forms not common elsewhere.`", when country information *Philippines* is provided, the adapter model responds: "`Another aspect of Pinoy culture that fascinates me is the concept of 'amor-proper' or 'love'. In some ways, it's similar to the concept of family where you can choose your own family. But in the Philippines, it's not just about the bloodline, it's also about the affection and care you give to people. That's why even strangers can be "kaibigan" or friends.`", but when the country is not provided, the model responds: "`What is something that you think is unique to your culture in terms of hospitality?`". In addition, full-fine-tuned still shows similar patterns as before, only sometimes responding: "I am intrigued by...". These experiments show the usefulness of the adapter-based fine-tuning to induce inherent curiosity in models.

G Downstream Impacts of Curious LLMs

We extend evaluation to three cross-cultural benchmarks: (1) NormAd (Rao et al., 2025), which contains culturally-grounded social situations across 75 countries and tests models' judgments of social acceptability (we restrict to binary cases (acceptable vs. not acceptable) and exclude neutral items); (2) CulturalBench (Chiu et al., 2024), a set of 1,227 human-verified questions from 45 regions across topics like cuisine, etiquette, local customs, etc. to probe nuanced cultural understanding (we use the Hard data subset); and (3) Cultural Commonsense (Shen et al., 2024), containing data from 5 countries which tests models' ability to reason about culturally specific everyday inferences (we use the English fill-in-the-blanks subset on country-specific customs). All the above datasets capture complementary aspects of cultural adaptability. We input country information in the prompt for inference. We share examples of datasets in Fig 11.

We employ two methods to induce curiosity in LLMs: (1) `prompt-ask`: we ask models to generate follow-up questions given the statement/story and answer them themselves before deciding on the answer. (2) `ft + prompt-ask`: we use the adapter-finetuned model (reddit-obj2-countryall) to generate follow-ups/clarification and switch to the base model to answer them and deciding on the answer. Here, we also present additional results: (3) `ft-specific+prompt-ask`: models are

now fine-tuned on specific datasets. GPT-4o is first used to generate follow-up questions, which then serve as training data for adapter-based fine-tuning on each story or statement.

CONDITION	NA	CB	CS
non-curious	70.48%	64.71%	48.48%
curious (prompt-ask)	72.09%	67.64%	49.64%
curious (ft+prompt-ask)	71.06%	68.21%	56.16%
curious (ft-specific+prompt)	78.8%	68.1%	57.16%

Table 12: **Downstream tasks (% acc):** NormAd (NA), CulturalBench (CB), and Cultural Common-Sense (CS) - curiosity models have higher acc.

Table 12 shows the results of the above for downstream tasks. As discussed previously, curiosity-induced models perform better than non-curious models, showing the importance of curiosity in cultural adaptability of LLMs. We find that models fine-tuned on specific datasets perform better overall – significantly so on NormAd – but not on CulturalBench. However, caution is needed, as the training questions are generated by an LLM and may not fully capture culturally grounded nuances. Overall, adapter-fine-tuned models help improve performance across all cultural-adaptability benchmarks.

Together, these tasks demonstrate why curiosity is critical: instead of defaulting to Western assumptions, a model with culture-aware curiosity would clarify or ask follow-up questions to better interpret contexts.

G.1 Prompt used for Downstream Tasks

Fig 16 shows the prompt utilized to induce question asking behavior in LLMs. We prompt models to ask questions and answer them themselves (inspired by Press et al. (2022)). Note that this template may vary depending on the benchmark used for inference (the figure shows an example of CulturalBench).

G.2 Qualitative Analysis of Follow-up Questions

We compare the NormAd questions (for prompt-ask vs adapter+prompt-ask models). Here are some high level takeways:

In terms of question types and intents, Adapter follow-ups consist of a lot of yes/no norm checks: for example, "is it rude to...?", " is it acceptable..?". Prompt-based follow-ups are more explanatory probes – "How do norms in <country> treat...?" This may explain the stronger performance of adapter-based models: they tend to ask more explicit and direct questions, improving contextual reasoning and cultural adaptability. Cultural grounding is more apparent in prompt-based followups, like "In <country>, is ___ interpreted as __ ? " However, adapter follow-ups are more implicit. This maybe because these models are currently only trained in 3 countries, which are not present in the benchmarks we use for inference. In the future, we should focus on increasing the cultural grounding of fine-tuned models. Additionally, adapter models use more moralizing/judgment verbs (rude, bad manners, better host). While prompt-based methods use more interpretive verbs (suggest, indicate, align, conflict) that invite nuance. Finally, prompt-based methods ask more open-ended questions overall, however, they sometimes over-fit culture (presume norms, and use leading questions).

In the future, both methods can be improved: for adapter-based models, adding a light prompt scaffold that encourages why/how follow-ups and examples could enhance reasoning depth; for prompt-based models, introducing constraints to avoid presuppositions may lead to more precise and culturally grounded questions.

Prompt for Downstream Tasks

Prompt-Ask for Downstream Tasks:
f"You are answering a question about cultural norms and social behavior.
question
options
You may write at most 2 brief follow-up questions and 2 brief self-answers as scratch notes.
Put every scratch note on its own line, starting with ''
Decide the correct option strictly from ONLY the listed Options.
Please respond with only the letter of your answer (A, B, C, or D). Do not include any explanation or additional text.
Answer:

Figure 16: Prompt-Ask framework for downstream tasks - inducing curiosity

H Significance Tests Across Experiments

Linguistic Alignment. Humans show larger dispersion across countries than LLMs (SD = 0.0785 vs. 0.029; reported F = 7.33), indicating significantly greater variance in human linguistic styles (very large N; variance ratio well above 1).

Topic preference Alignment. `LLaMA-3-8B` shows the only positive average correlation with human topic rankings, other models are near-zero or negative. We find p = 0.043 significant; whereas negative values (e.g., GPT-4o = 0.29) are also small in magnitude (|t| < 1.2).

Inducing Curiosity (inherent curiosity) Using 70 statements - adapter-FT asks questions on 75% of prompts vs. 20% for full-FT and 0% for prompt-only; relevance is 98–100% for all. Two-proportion tests with n = 70 per model show adapter vs. full-FT is highly significant ($z = 6.5$, $p < 10^{-9}$), and adapter vs. prompt-only is modestly decisive.

I Model Choices, Implementation Details and Computational Resources

For our experiments, we selected a mix of open- and closed-source models to examine cultural curiosity across different architectures and training paradigms. Specifically, we used `LLaMA-3-8B` (**?**) for both prompt-based and fine-tuning experiments, focusing primarily on the smaller model due to its strong alignment with human data and lower computational overhead. We additionally evaluated `GPT-4o` (Achiam et al., 2023), `GPT-5`,[8] `LLaMA-3-70B`, `Claude-Sonnet-4`,[9] and `Qwen-3-14B` (Yang et al., 2025) to provide a broader comparison across model families. For fine-tuning, we applied LoRA adapters and full-fine-tuning to `LLaMA-3-8B` under different data configurations (e.g., country-specific vs. country-all, Yahoo vs. Reddit), using culture-aware and inherent curiosity objectives.

All experiments were run on NVIDIA A40 GPUs, using the Hugging Face Transformers and PEFT libraries. Fine-tuning followed standard hyperparameters, with a maximum sequence length of 1024 tokens and a fixed random seed for reproducibility. Adapter training was also chosen as they are more efficient, allowing rapid experimentation across multiple cultural configurations. In our experiments, we also find that adapter-based models do quite well in human-LLM curiosity alignment.

[8] https://openai.com/index/introducing-gpt-5/
[9] https://www.anthropic.com/news/claude-4

J Reproducibility

We open-source our codes and data, which are uploaded to the submission system. This would help future work to reproduce our results and explore culture-aware curiosity in humans and LLMs, and its impact in downstream tasks.

Table 13: Dataset of Everyday Statements for Inherent Curiosity Evaluation

Category	**Statements**
Everyday Life	Alex arrived late to the meeting. Someone brought a gift to dinner without wrapping it. Sarah decided to sit at the back of the classroom. The teacher suddenly stopped talking in the middle of the lecture. Two friends started whispering during the movie.
Social Situations	Maria refused to shake hands at the event. A guest walked straight into the kitchen at someone's house. John didn't respond when greeted by a colleague. A person started eating dessert before the main course. Someone left the party without saying goodbye.
Work & School	The boss canceled the meeting at the last minute. Students stood up when the professor entered the room. A worker refused to answer emails after 5 p.m. A student called their teacher by a nickname. Employees were chatting loudly in the office.
Ethics & Morality	A child lied to their parents about homework. A man kept extra change the cashier accidentally gave him. Someone ignored a homeless person asking for help. A person secretly recorded a private conversation. A passenger refused to give up a seat to an elderly person.
Science & Facts	Water can boil at room temperature. Dinosaurs and humans lived at the same time. The sun could suddenly stop shining tomorrow. Some people believe time travel is possible. There are animals that can survive without oxygen.
Technology & Future	An AI refused to answer a user's question. A robot was programmed to laugh randomly. Someone trusted a self-driving car to take them home. A drone delivered a package late at night. People started wearing VR headsets in public.
Abstract / Paradoxical	This statement is false. A ship has had every part replaced — is it still the same ship? A person knows a secret that no one else can ever know. If a tree falls in a forest and no one hears it, does it make a sound? Someone dreamed they were awake.
Random / Surprising	A stranger gave a child candy on the street. People clapped after a plane landed. A friend borrowed money and never returned it. Someone talked to their plants every morning. A cat suddenly barked like a dog.

Table 14: Dataset of Cultural Statements (Country-wise) for Inherent Curiosity Evaluation

Category	Statements
Brazil	In Brazil, it is common for friends to greet each other with physical gestures. Brazilian families often spend weekends together in a particular way. Football matches in Brazil usually come with unique social rituals. In Brazil, people tend to treat teachers in a distinctive manner. A wedding in Brazil usually involves traditions beyond the ceremony. Brazilians often approach mealtimes differently from other cultures. A Brazilian student is expected to interact with professors in a certain way. During Carnival in Brazil, certain behaviors are widely accepted that might not be elsewhere. In Brazil, people have specific expectations about lateness. Many Brazilian workplaces begin the day in a particular way.
Philippines	In the Philippines, community celebrations often include unique practices. Filipino families are expected to behave in certain ways during holidays. Respect toward elders in the Philippines is shown through specific gestures. Filipino classrooms often encourage behaviors different from Western norms. Religious festivals in the Philippines are tied to specific community practices. Neighbors in the Philippines usually interact in ways that reflect local values. Filipino weddings often extend beyond the couple and include wider groups. Hospitality in the Philippines takes forms not common elsewhere. Gift-giving in the Philippines follows unspoken cultural rules. Daily greetings in Filipino households involve rituals outsiders may overlook.
UK	In the UK, politeness takes a form that is not always obvious to outsiders. British families have unique expectations during Sunday gatherings. In UK workplaces, meetings often begin with a particular social convention. British weddings typically follow certain cultural patterns. Pubs in the UK play a social role that extends beyond drinking. In the UK, silence in conversations can mean something specific. British classrooms have norms that might surprise people from other cultures. Greetings in the UK are often understated in a characteristic way. British people have particular unwritten rules about queuing. Public discussions of politics in the UK follow a distinctive pattern.