Evaluating Cross-Cultural Information-Seeking Questions in Humans and LLMs

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) have expanded their role in human interaction, yet information-seeking behavior, particularly across cultural contexts, remains underexplored in these systems. In this work, we study cross-cultural variation in information-seeking using question-asking as a mechanism. We introduce *IQUEST (**I**nformation-seeking **QU**estion **E**valuation across **S**ocie**T**ies)*, an evaluation framework that measures human-model alignment in information-seeking question behavior 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 question-asking patterns expressed in Western countries. We also explore fine-tuning strategies, which narrow the human-model alignment gap by 43% overall. Finally, we demonstrate that stronger information-seeking question behavior improves downstream cultural adaptability, highlighting its importance for inclusive NLP research.

1 Introduction

When people seek to understand the world, they ask questions, and increasingly they do so in dialogues with LLMs (y Arcas, 2022; Steyvers et al., 2025). Conversely, a model's ability to ask appropriate follow-up questions in return can shape usefulness, engagement and cultural sensitivity (Rahmani et al., 2024; Srinivas et al., 2025; Deva et al., 2025). This has further implications in high-stakes domains such as education (Wang et al., 2024; Nair et al., 2024), healthcare (Wang et al., 2025), and so on. Because information-seeking is often expressed through question-asking (Graesser et al., 1993; Raz and Kenett, 2024), studying question behavior offers a tractable way to compare humans and models across cultures. Prior work in psychology and education has linked question-asking to curiosity (Berlyne, 1957; Loewenstein, 1994; Kidd and Hayden, 2015), but in computational settings it is especially valuable because it is directly observable without requiring access to latent internal states in models (Graesser and Olde, 2003; Chin and Osborne, 2008).

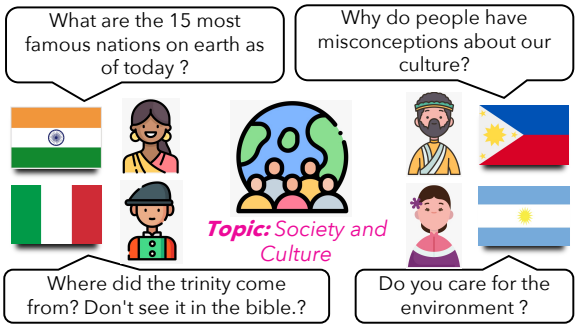

Figure 1: **Information-seeking differences across countries.** For a given topic, questions vary in style and content across countries. Here, we show examples from Argentina, India, Italy, and the Philippines. We perform this analysis in 18 countries and 16 topics.

Prior work shows that question-asking and exploratory behavior vary across societies (Fig. 1), influenced by social norms and educational contexts (Bluemke et al., 2024; Jirout et al., 2024), framing it as a socio-cultural process shaping human interactions. However, most LLM-based cross-cultural studies do not investigate question-asking and focus on answering capabilities (Seth et al., 2024; Li et al., 2024) and use survey-based benchmarks (e.g., WVS,[1] Pew Research[2], etc.). Only limited prior work has examined LLMs' question-asking capability, but they are largely unconstrained by cultural nuance (Ceraolo et al., 2024). Guided by these gaps, we address three research questions that connect information-seeking and culture: (1) Do information-seeking questions vary across human cultures, and do LLMs repro-

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

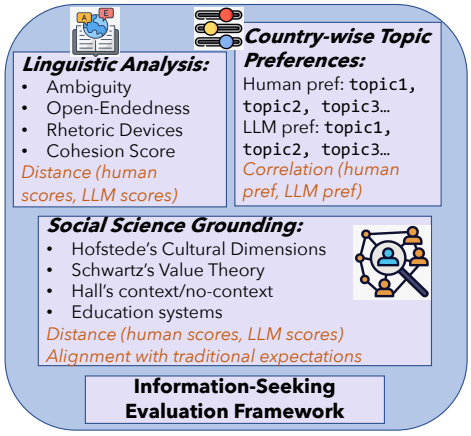

Figure 2: **IQUEST Overview:** (1) Linguistic Alignment, (2) Topic Preference Alignment, and (3) Grounding to Social Science Constructs; to compare human- and LLM- questions.

duce these patterns?, (2) Can we align LLMs with human information-seeking question-asking patterns? and (3) What are the practical implications of cross-cultural information-seeking for downstream applications such as cultural adaptability?

Unlike prior work that examines cultural adaptability in LLMs in isolation, we believe ours is the first to directly compare human and LLM *question-asking*, as an operationalization of *information-seeking* behavior using a systematic evaluation framework. The key contributions of our work are: (1) We introduce **IQUEST**: a three-dimensional evaluation framework to analyze human and LLM information-seeking across cultures. (2) We explore fine-tuning strategies to improve information-seeking behaviors in LLMs, and (3) We examine practical implications of this behavior on cultural tasks, demonstrating its effectiveness in improving LLMs' adaptability.

2 Related Work

Information seeking is a core mechanism of learning and exploration, and question asking is one of its most observable behavioral forms (Graesser et al., 1993; Raz and Kenett, 2024). Prior research in psychology and cognitive science has also connected question asking to curiosity (Berlyne, 1957; Loewenstein, 1994; Kidd and Hayden, 2015). Several works in tutoring systems, science education and human–LLM interactions (Graesser and Person, 1994; Scialom and Staiano, 2019; Ceraolo et al., 2024) have explored question-asking as a behavioral manifestation of exploratory knowledge acquisition. In computational settings, question-asking is particularly useful because it is observable without requiring access to internal mental states (Chin and Osborne, 2008; Graesser and Olde, 2003) and enables scalable analysis.

Studies show that information seeking is crucial for cultural learning, often varying across societies (Bluemke et al., 2024; Jirout et al., 2024). This makes it especially important to study information-seeking behavior in LLMs, which are increasingly used in culturally diverse settings where asking the right follow-up questions can shape understanding, usefulness, and sensitivity to context (Kim et al., 2022; Deva et al., 2025). In NLP, previous work on cultural representation in LLMs mostly focus on culture-aware evaluation that does not examine question-asking (Seth et al., 2024; Arora et al., 2022; Borah et al., 2025a). Cross-cultural understanding of information-seeking behavior in LLMs is critical, as models that fail to reflect diverse curiosity patterns may misrepresent intent and homogenize information-seeking (Mihalcea et al., 2025). Furthermore, homogeneous LLM questioning patterns may suppress exploration and collaboration among users (Engel, 2015; Chi and Wylie, 2014). To our knowledge, IQUEST is the first framework to systematically compare cross-cultural question-asking behaviors and evaluate its practical impacts in LLMs.

3 Cross-Cultural Questions Dataset

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

3.1 Human-authored Questions

We use the "Yahoo! Answers" dataset,[3] which contains user-submitted questions from countries across Asia, Europe, the Americas, and Australia. For our study, we include only those topics that have questions from all countries. Questions in local languages are translated into English using the Google Translate API.[4] Translation is a practical necessity due to the lack of annotators for several

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

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

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 stepparenting?
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.

native languages. We perform experiments showing translation largely preserves linguistic trends (Sec 4.1). We derive topic preferences by treating higher frequencies of questions as indicators of stronger interest. Overall, our study spans 16 topics across 18 countries. We also perform some quality control experiments (see Appendix A).

3.2 LLM-generated Questions

We use two strategies: 1) LLM-Question Generation: We instruct LLMs to adopt a country persona and generate n questions related to a given topic. 2) Topic Preference Analysis: We prompt country-based LLM personas to generate topic rankings based on preference. Prompt details are provided in Appendix A.2. We use: GPT-4o, GPT-5, Claude-Sonnet-4, Qwen-3-14b, LLaMA-3-8b-instruct and LLaMA-3-70b-instruct, and average final results across three runs to account for variability. We additionally include a no-persona baseline to evaluate information-seeking in a more naturalistic setting.

4 IQUEST: Measuring Human-LLM Alignment

Our framework IQUEST consists of three complementary directions to evaluate culture-aware information-seeking: *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 (see Appendix B.3 for a qualitative analysis of questions).

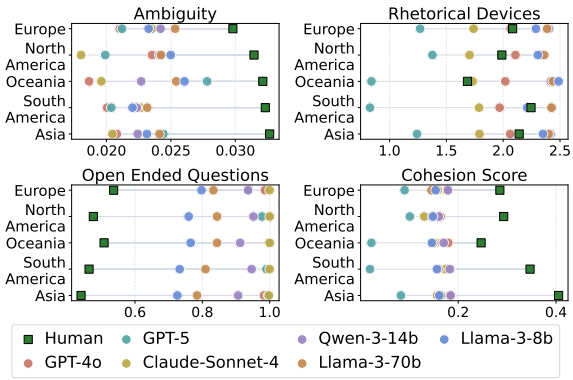

Figure 3: **Linguistic Alignment**. Humans and LLMs diverge: humans are high in ambiguity/cohesion scores, LLMs are high in number of open-ended questions.

4.1 Linguistic Alignment

Taking inspiration from previous studies, we consider four dimensions for linguistic analysis: *(1) ambiguity*: captures the degree to which a question invites multiple interpretations, which is central to information-seeking as it leads to exploration (Gruber and Ranganath, 2019; Tor and Gordon, 2020; Berlyne, 1960) (we measure this using a pre-defined list of polysemous and multiple POS tag words), *(2) rhetorical devices*: refers to creative or exploratory phrasing that can enhance information-seeking expression (Gibbs, 1994; McQuarrie and Mick, 1996) (we use stylistic features: repeated words, alliteration, parallelism, etc.), *(3) open-endedness*: reflects whether a question allows broad answers in general, similar to drive for knowledge-seeking. (Scialom and Staiano, 2019; Engel, 2011; Grossnickle, 2016) (note this metric is intrinsic to the question, not on whether it could be answered by different types of experts.), 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 combine lexical overlap, transition word presence and semantic similarity for this, details are in Appendix B).

Results. Fig. 3 shows that humans and LLMs exhibit different question-asking 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, being closer to human patterns. Interestingly, humans standard deviations (0.0785) are higher than LLMs (0.029) (F-stat:7.33), indicating that 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. We also run prompt-ablation analyses (Appendix E), showing that our findings remain consistent under prompt variations.

Country-wise analysis. Fig. 4 presents the mean absolute differences (MAD) 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 with the West showing the best alignment; followed by some Eastern and Latin American countries. East Asia, however, shows the largest gaps. For the no-persona baseline condition, we observe a similar pattern where gaps are: East > Latin Am > West. Overall, using cultural prompting (as we do in current experiments) reduces these gaps, showing its benefits in question-asking alignment (details in Appendix B.2).

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 (distances) are: `LLaMA-8b` (0.25) > `LLaMA-70b` (0.27) > `Claude` (0.28) = `GPT-4o` (0.28) > `Qwen-14b` (0.29) > GPT-5 (0.42).

Impact of translation. We compute Spearman correlations between non-translated and translated versions for countries where we have native-language data (e.g., Spanish in Argentina). The results show high agreement for Open-endedness ((ρ=0.81)) and Ambiguity ((ρ=0.78)), and decent agreement for Cohesion ((ρ=0.65)) and Rhetorical Devices ((ρ=0.55)). Furthermore, the core trends are stable under translation: human questions exhibit higher ambiguity and cohesion, while LLM questions remain more open-ended. We also perform a robustness check by analyzing only the English-native countries subset. The macro MAD in this subset (0.255) is comparable to the full-dataset result (0.220). This indicates that even when translation is removed, the observed differences remain. Since translation largely preserves trends across metrics, we use translated data for the study (further analyses in Appendix B.1).

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).

[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

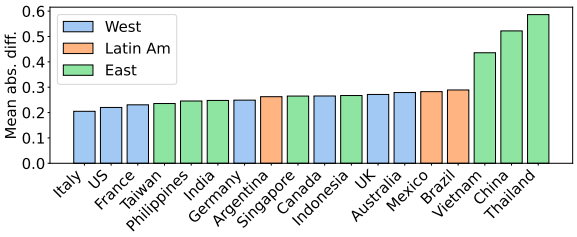

Figure 4: **Country-wise Linguistic Alignment:** (low mean abs diff: high alignment) Western countries show the highest human–model alignment, followed by Latin American and Eastern countries.

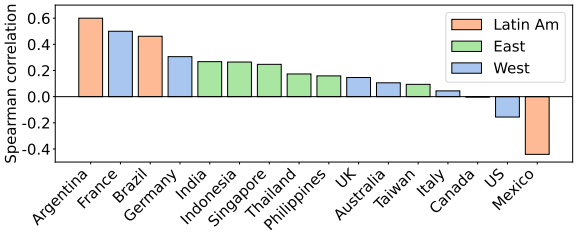

Figure 5: **Country-wise Topic-Preference Alignment**: `LLaMA-3-8b` correlation with humans.

This experiment is largely invariant to translation, as it uses Yahoo!'s defined high-level topic clusters.

Country-wise analysis. We compute country-level correlations using the best-performing model: `LLaMA-3-8b` in Fig. 5. Among countries, Argentina, France, and Brazil show the highest correlations, while Mexico and the US exhibit negative correlations. These findings diverge from our previous results using linguistic analysis, indicating that topic preference captures distinct information-seeking patterns. Note that this experiment measures relative topic-preference similarity, not absolute cultural alignment, so negative correlations reflect divergence in ranked topics. Detailed discussions are in Appendices C.1 and C.2.

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. Fig 6 visualizes average topic preferences. LLMs consistently emphasize 'Family and relationships', 'Mu-

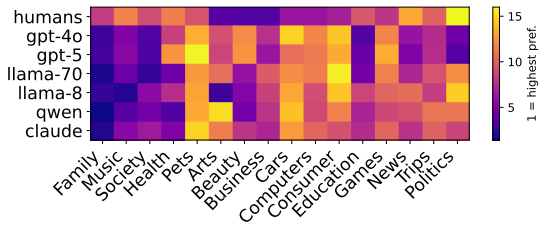

Figure 6: **Topic-Preference Ranking**: Heatmap of preferred topics (shortened) averaged across countries for human and models: (1- highest preferred). It shows a clear divergence between humans and LLMs.

sic and entertainment', etc. whereas humans prioritize 'Arts and humanities', 'Beauty and style', and 'Business and finance', highlighting systematic differences in information-seeking focus.

4.3 Grounding to Social Science Theories

Grounding to 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 four information-seeking-relevant dimensions: *(1) Uncertainty Avoidance (UA):* high UA discourages exploratory questioning, while low UA encourages it (Hofstede, 2001; Kashdan and Silvia, 2009). *(2) Individualism–Collectivism (IC):* whether inquiry is self-focused or socially oriented (Triandis et al., 1995). *(3) Indulgence–Restraint (IR):* indulgent cultures frame information-seeking as hedonic, whereas restrained cultures emphasize discipline (Hofstede et al., 2010). *(4) Long/Short-term Orientation (LSO):* whether question-asking is future-oriented or present-focused (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 information-seeking as a motivational force (Schwartz, 1992; Kashdan and Silvia, 2009). But, conservation emphasizes conformity, and tradition – constraining information-seeking to "safe" domains.

3) Halls' contextual dimensions: Hall's *high context cultures* rely on implicit cues, meaning information-seeking often manifests in indirect forms. In contrast, *low-context cultures* are explicit with informatio-seeking expressed through direct questions (Hall, 1976).

4) Education systems: *Rote-learning* styles are

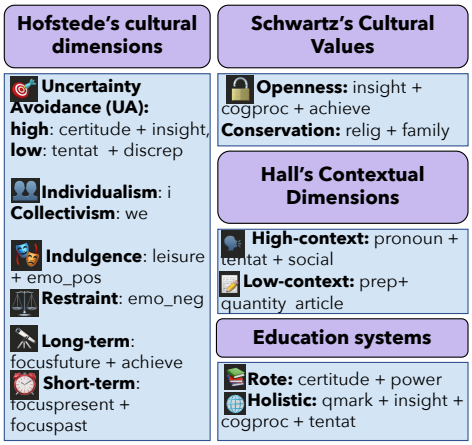

Figure 7: **Social Science Constructs:** Computation using LIWC dimensions.

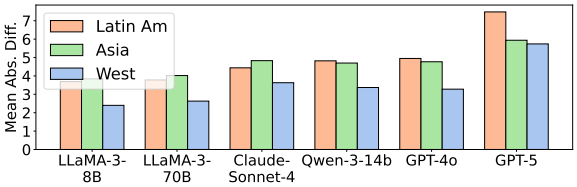

Figure 8: **Social Science Theory Alignment:** Region-wise MAE for models shows West vs East/Lat Am. Differences are prominent across models.

rule-based, authority-driven, and focused on correct answers (Biggs, 1996), and *holistic learning* styles are exploratory (Nisbett, 2010).

For each of these dimensions, we compute LIWC (Linguistic Inquiry Word Count) (Boyd et al., 2022) scores (Fig 7); LIWC is a tool that maps text to psychologically and socially meaningful categories to quantify information-seeking-driven social dimensions.

Region-wise analysis. Fig 8 shows the MAD between model and human using LIWC scores. All models are consistently closer to Western countries than Eastern/Lat Am. 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 10 and Appendix D.1).

Alignment of humans and LLMs to traditional expectations. The traditional literature on intercultural psychology expects differences across countries based on social constructs described earlier (Table 3 'Theory and Expectation'). We categorize LIWC scores by alignment with traditional expectations as Aligned, Not aligned, or Mixed/Partially

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. A dimension is labeled Aligned if a majority of countries (> total countries/3) match traditional expectations, Not aligned if none match, and Mixed/Partially aligned if only a minority of countries (<= total countries/3) match, therefore, no other clear pattern.

aligned (details in the caption).

Humans themselves do not consistently follow the traditional expectations, maybe because: *(1)* Yahoo is an anonymous platform, effectively reducing normative pressure (Guo and Caine, 2021; Ellison et al., 2016) *(2)* cultural evolution and globalization (Witte et al., 2020; Kittler et al., 2011; Greenfield, 2016). Among models, `LLaMA-3-8b` has the highest alignment and `Claude-Sonnet-4` has the most inverted patterns. Theory-wise, both humans and models align on IR and Schwartz's dimensions. Hall's high–low context shows the strongest non-alignment, while education systems yield the most mixed outputs (details in Appendix D.2).

Across experiments, `LLaMA-3-8b` performs the best, likely due to higher-quality pretraining data (Dubey et al., 2024) and lighter RLHF; heavier safety tuning in larger models may "sanitize" outputs toward Western norms (Aksoy, 2025). Supporting results are in Appendix G.

4.4 Construct Validity of Metrics

To validate that our metrics capture cross-cultural variation in information-seeking, we use two benchmarks:

Hofstede's UA correlation (n=18) (Hofstede, 2001) We correlate country-level human linguistic metrics with country-level UA scores in the benchmark, and observe a significant negative correlation between UA and rhetorical device use (Spearman $\rho = -0.521$, $p = 0.027$) (as expected), consistent with high-UA countries such as Argentina and France (UA = 86) using fewer rhetorical markers than low-UA countries like Singapore (UA = 8) and the UK (35). Ambiguity shows a weak, non-significant negative trend ($\rho = -0.139$, $p = 0.584$). Open-endedness shows a marginal positive correlation ($\rho = 0.456$, $p = 0.047$), suggesting that high-UA cultures may use explicit interrogative framing as a strategy to resolve uncertainty (Loewenstein, 1994). Together, these correlations validate that our metrics capture theoretically grounded information-seeking expression.

Directional consistency with psychometric IC scores (n=3) Bluemke et al. (2024) provides validated intellectual curiosity scale (ICS) scores (using a questionnaire) for 3 countries (USA, Germany, and France) that are also present in the Yahoo dataset. Germany shows relatively lower psychometric ICS score ($\delta = -0.27$ relative to USA), and a relatively higher mean absolute human-LLM linguistic distance in our study (Fig 4). This is consistent with the view that lower psychometric IC populations express information-seeking in ways less aligned with LLM defaults. While n=3 is statistically limiting, this pattern is directionally consistent with our metrics. These results support our linguistic metrics as proxies for information-seeking question-asking (insights in Appendix F).

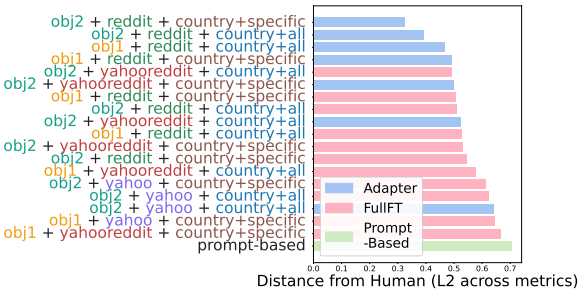

Figure 9: **Linguistic Analysis on fine-tuned vs non-fine-tuned models**: y-axis shows configs, while the x-axis show model-human distances (L2) (averaged across metrics and countries). adapter -based methods with obj2 on Reddit perform the best, while prompt-based methods performing poorly.

5 Aligning Human-LLM Question-Asking

Having found differences in question-asking patterns by humans and LLMs, we now shift our focus to reduce these human-LLM gaps. We use three countries where alignment was low: Brazil, the UK, and the Philippines (from Lat Am., Western, and Eastern regions), and each with sufficient data.

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

Methodology. We fine-tune `LLaMA-3-8b`, the model most aligned with humans, using full and adapter-based methods. We use two objectives: *obj1*, generating a question for a specified country ("Write one question from someone in <x>"), and *obj2*, adding a `GPT-4o`-generated conversational preface to each dataset question to encourage natural queries (e.g., for a Yahoo 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…"). We evaluate both country-specific and country-all (including data from all countries) training under three data settings: *yahoo-only*, *reddit-only*, and combined.

Evaluation. We use two approaches: **(1) culture-aware evaluation**: using IQUEST on the Yahoo dataset and **(2) inherent evaluation**: to determine models' *natural inquiry ability*. For (2), we design 70 generic and cultural statements that are ambiguous and naturally invite follow-up questions (example provided in Table 4, full dataset in Appendix H.5). To evaluate, we use: (1) *Question-Asking Rate* – does the model ask an 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 H.7 for annotation details).

5.1 Results

Culture-aware Evaluation. Fig. 9 shows that fine-tuning consistently narrows linguistic gaps between humans and LLMs, overall by 42.8%, with improvements on semantic ambiguity (−73%), rhetorical devices (−57%), and open-endedness (−56%). Cohesion shows the opposite trend, suggesting that while models adopt questioning patterns, structural discourse properties may shift differently under fine-tuning. We also compute Kendall's coefficient of concordance[6], which increases from 0.19 (baseline) to 0.56 (fine-tuned), indicating stronger agreement among behavioral signals after training.

Across datasets, reddit-trained models are closest to humans, likely due to exposure to diverse rhetorical discourses. Adapter-based models outperform full fine-tuning, likely because adapters shift output patterns without overwriting the general flexibility of the base model, whereas full fine-tuning may over-specialize to the training format. Finally, `obj2` is better than `obj1`, maybe due to added conversational scaffolding. (trends for topic preference analysis are similar – Appendix H).

Inherent Evaluation. We use country-all models for inherent information-seeking, as it tests general-purpose inquiry capabilities, and choose the best-performing configurations from each training type. We find that adapter-ft models pose most questions (75%), whereas the full-ft models hardly ask any questions (20%) (Example is in Table 4).

Qualitative analysis shows that the full-FT model often produces statements (e.g., "I am interested in…") rather than explicit questions, reflecting *implicit curiosity* (Berlyne, 1960). This may stem from training differences: adapters preserve base model priors and better support questioning and generalization (Pfeiffer, 2023), while full fine-tuning may reduce flexibility under prompt shifts. Prompt-only models generate no questions. Relevance remains high across (98–100%). These are corroborated by strong annotator agreement (Co-

[6] https://en.wikipedia.org/wiki/Kendall%27s_W

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 ...

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

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

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

hen's κ: 0.81 for questions, 0.62 for relevance). Overall, these findings reflect behavioral alignment with human information-seeking questioning, without claims about internal model states (see Appendices H.8, H.9).

5.2 Downstream Impact of Information-Seeking in LLMs

To further assess the functional value of culture-aware information-seeking in LLMs, we extend our evaluation to three complementary benchmarks that capture culturally grounded social reasoning, cultural knowledge, and commonsense inference: (1) NormAd (Rao et al., 2025); (2) CulturalBench (Hard subset) (Chiu et al., 2024); and (3) Cultural Commonsense (Shen et al., 2024) (See Table 12 for examples). These tasks highlight contexts where information-seeking may assist in avoiding default Western assumptions (details in Appendix I.1).

In our experiments, we compare a direct baseline (models classify immediately) with two information-seeking variants: (a) prompting models to generate 1-2 sub-questions and answer before deciding (inspired by Self-Ask (Press et al., 2022)), and (b) an integrated approach combining our best-performing fine-tuned model to generate questions and base model for answers and decision.

Results. Table 5 shows that information-seeking-enabled variants outperform the non-information-seeking baseline. Notably, the `ft+prompt-ask` setting, trained on Reddit outperforms `prompt-ask` overall, showing that models trained to inherently ask questions can improve cultural adaptability. Overall, information-seeking helps models avoid overgeneralization and better grasp context (see Appendix I for further experiments and Appendix I.3 for qualitative analysis of questions).

6 Lessons Learned and Actionable Steps

Our experimental findings highlight the importance of information-seeking in LLMs and offer actionable steps toward building human-aligned 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. Human questioning evolves from traditional expectations. These changes highlight the need for the NLP community to build models that capture the diversity of human question-asking.

Adapter fine-tuning helps with human-LLM alignment. Adapter-based fine-tuning generates contextually relevant questions in conversations. This shows that information-seeking questioning is best introduced by augmenting rather than overwriting base model priors. Future work could build on this by scaling adapter methods.

Information-seeking in LLMs improve cultural adaptability. Across three cross-cultural benchmarks, we observe information seeking in models lead to better cultural adaptability. Future work should focus on integrating information-seeking systematically into training and evaluation pipelines to build inclusive LLMs.

7 Conclusion

In this work, we introduced *IQUEST* to evaluate culture-aware information-seeking in humans and LLMs. Through complementary analyses, 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, we narrowed the human-model alignment gap and improved the downstream cultural adaptability of LLMs. Based on our findings, we share ideas for future research and open-source our framework. [7]

[7] https://anonymous.4open.science/r/ICUEST/

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 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.

Question-Asking as Information-Seeking. We operationalize information-seeking through question-asking, which is its most frequent and observable behavioral manifestation across both human and computational settings (Berlyne, 1957; Kidd and Hayden, 2015; Graesser and Person, 1994). This choice is theoretically grounded in information-seeking and inquiry research, and is consistent with established practice in education and cognitive science (Chin and Osborne, 2008; Dillon, 1988). Crucially, question-asking is tractable for large-scale computational analysis without requiring access to internal mental states. Future work can further investigate other manifestations of information-seeking, such as sustained attention, elaborative search (Grossnickle, 2016), and follow-up behavior in extended interactions. Furthermore, most LLM question-asking is prompted rather than intrinsically motivated, meaning our results reflect behavioral alignment rather than internal information-seeking drive. Future work could broaden this proxy to provide complementary validation of behavioral proxies.

Evaluation Subjectivity. Although we use multiple metrics, evaluating information-seeking 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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Daniel E Berlyne. 1957. Uncertainty and conflict: a point of contact between information-theory and behavior-theory concepts. *Psychological Review*, 64(6p1):329.

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

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

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*.

Michelene TH Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.

Christine Chin and Jonathan Osborne. 2008. Students' questions: a potential resource for teaching and learning science. *Studies in science education*, 44(1):1–39.

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.

Roshini Deva, Dhruv Ramani, Tanvi Divate, Suhani Jalota, and Azra Ismail. 2025. " kya family planning after marriage hoti hai?": Integrating cultural sensitivity in an llm chatbot for reproductive health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

James T Dillon. 1988. *Questioning and discussion: A multidisciplinary study.* Ablex Publishing.

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.

Susan Engel. 2015. *The hungry mind: The origins of curiosity in childhood*. Harvard University Press.

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, Mark C Langston, and William B Baggett. 1993. Exploring information about concepts by asking questions. In *Psychology of learning and motivation*, volume 29, pages 411–436. Elsevier.

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

Arthur C Graesser and Brent A Olde. 2003. How does one know whether a person understands a device? the quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, 95(3):524.

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal*, 31(1):104–137.

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 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.

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.

Celeste Kidd and Benjamin Y Hayden. 2015. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.

Junhan Kim, Jana Muhic, Lionel Peter Robert, and Sun Young Park. 2022. Designing chatbots with black americans with chronic conditions: Overcoming challenges against covid-19. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

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.

Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657.

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*.

Hossein A Rahmani, Xi Wang, Mohammad Aliannejadi, Mohammadmehdi Naghiaei, and Emine Yilmaz. 2024. Clarifying the path to user satisfaction: An investigation into clarification usefulness. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1266–1277.

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.

Tuval Raz and Yoed N Kenett. 2024. Question-asking as a mechanism of information seeking. *Behavioral & Brain Sciences*, 47.

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.

Vidya Srinivas, Xuhai Xu, Xin Liu, Ayush Kumar, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, and Tim Althoff. 2025. Substance over style: Evaluating proactive conversational coaching agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20848–20880.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231.

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.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Ziyu Wang, Hao Li, Di Huang, Hye-Sung Kim, Chae-Won Shin, and Amir M Rahmani. 2025. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, page 100570.

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.

Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus*, 151(2):183–197.

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.

A.1 Quality Control on the Yahoo Dataset

Before using Yahoo for IQUEST evaluation, we perform quality-control on the Yahoo! Answers corpus to quantify noise and ensure the text is suitable for linguistic analysis. We first: (i) remove obvious navigation/scrape artifacts (e.g., "All Categories", "Discover" topics etc.), (ii) measure duplicate inflation at the exact-string level (Country–Category–Text), etc.

Furthermore, to evaluate if the data is stable, we perform split-half reliability tests for each country. To do this, we randomly divide the set of Yahoo! questions into two equally sized subsets multiple times and compute the cosine similarity between the resulting distributions. High similarity across splits indicates that the observed patterns are reproducible and not sensitive to random sampling variation. This analysis serves as a sanity check that the dataset captures consistent question-asking behavior rather than artifacts of sampling or noise.

We observe very high stability for most countries, except Taiwan, which provides a decent stability (0.855). Average similarity across countries is 0.91. This is a strong sanity check that our country topic distributions are reproducible under resampling, i.e., not dominated by random noise.

A.2 Prompt details for LLM-generated questions

We generate LLM-based questions and preferences by using two strategies: **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.

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. Note that we do not include contextual ambiguity due to dataset-level operational constraints. We note that it is important to consider contextual ambiguity and leave it to future work to investigate.

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"). To compute the above, we use both rule-based detection and LLMs to operationalize them. We also perform error analysis on these sets on a set of randomly chosen 10 questions from each country. We will make them explicit in the revision. So, considering n total questions, Rhetorical Devices (RD) is given by:

$$RD = (R + Q + A + P + M)/n \tag{2}$$

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:

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

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)

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

Transition cohesion (rate per sentence)

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

Semantic cohesion (adjacent cosine)

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

Final score

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

B.1 Impact of Translation on Linguistic Analysis

To enable cross-country comparison, we translate non-English questions to English using the Google Translate API. Translation is applied only to countries whose Yahoo! Answers data are not primarily in English: Argentina (Spanish), Mexico (Spanish), Brazil (Portuguese), France (French), Germany (German), Italy (Italian), Indonesia (Indonesian), Thailand (Thai), Vietnam (Vietnamese), Hong Kong (Chinese), and Taiwan (Chinese). Others countries (US, UK, Canada, Australia, Singapore, Philippines, India) are in English and hence, not translated.

Methodology. We evaluate whether translation preserves relative patterns in our linguistic analysis by comparing *non-translated* vs. *translated* versions for countries where we have native-language data.

For each country, we compute the four metrics on (i) the original-language questions using language-appropriate equivalents of the heuristics (e.g., interrogative markers and polysemous word lists), and (ii) their English translations, using the same metric definitions. We then compute Spearman rank correlation ρ between the two versions and summarize results across regions.

Additionally, we compare the overall trends per metric in linguistic analysis, comparing human and LLM questions.

Results. The results show strong agreement for Open-endedness ((ρ=0.81)) and Ambiguity ((ρ=0.78)), with moderate agreement for Cohesion ((ρ=0.65)) and Rhetorical Devices ((ρ=0.55)). Overall, translation largely preserves relative patterns across metrics (to varying degrees), thus, we use translated data for the remaining analyses, with

Region	ρ_{Open}	ρ_{Amb}	ρ_{Coh}	ρ_{Rhet}
West	0.91	0.89	0.74	0.69
East	0.82	0.69	0.61	0.49
Lat Am	0.71	0.75	0.60	0.49

Table 7: Per-region Spearman correlations between non-translated and translated metrics (computed within each region).

caution about translation-induced shifts in finer-grained metrics like Rhetorical Devices.

Overall, agreement is strongest in the West for all metrics, with similar numbers for the East and Latin Americas.

In terms of overall trends, they are consistent with translated variants. Considering Fig 3, we find that LLMs are more open-ended and humans output higher ambiguous/cohesion scores. This trend is upheld in non-translated variants. Furthermore, region-wise variations are also consistent, that is, Europe (0.769) and Americas (0.660) have higher open-endedness than Asia (0.612). Cohesion and ambiguity trends are also consistent, with Asia being the highest in both translated and non-translated sets. Rhetorical devices are also directionally stable but noisier (as expected).

Since there is no major degradation of the metrics, we proceed with using the English translations for all countries in our analyses. Furthermore, translation may impact linguistic features (one eval dimension), while other dimensions like topic preferences and cultural grounding are largely preserved, as they rely on semantic content rather than surface form. While evaluation with multilingual data is important, cross-cultural human validation of the data and computational methods is also important. However, it is not feasible for us to conduct at scale due to cost constraints, and we leave broader multilingual validation as future work.

B.1.1 Translation Robustness Analysis

To further examine whether translation artifacts influence our results, we repeat the linguistic alignment analysis using only the seven English-native countries in our dataset, i.e., US, UK, Canada, Australia, Singapore, Philippines and India, where no translation was applied.

The macro mean absolute difference (MAD) between human and model scores in this English-only subset is 0.255 and the full datatse has a MAD of 0.220. This shows that restricting the analysis to non-translated data does not reduce the human–LLM alignment gap; if anything, the divergence is slightly larger. Thus, even without translation, the gap still remains.

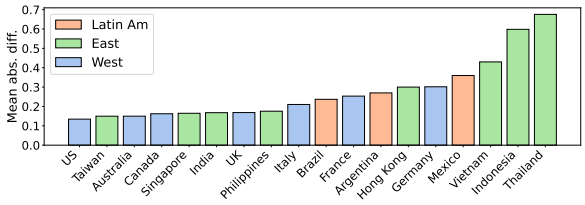

Figure 10: **Country-wise Linguistic Alignment: on baseline LLM vs human questions**: Western countries show the higher human-model alignment on average, followed by Latin American and Eastern countries. This gap is consistent with previous findings, however, cultural personas reduces these gaps substantially across regions.

At the metric level, rhetorical devices drive much of the English-only gap (MAD = 0.717 compared to 0.237 in translated countries), while open-endedness and cohesion show smaller differences. This is notable because rhetorical devices were the most translation-sensitive metric in our earlier analysis ($\rho = 0.55$ in Appendix B.1). The larger gap in English-native countries suggests that translation may suppress stylistic divergence rather than artificially inflate it.

Overall, these results support treating our cross-cultural alignment findings as robust to potential translation artifacts.

B.2 Baseline comparison to human results

To contextualize existing results that use persona prompting for cultural grounding, we evaluate a no-persona baseline condition using the prompt: "`You are browsing the internet and curious about [topic]. What question would you ask?`" Fig 10 shows country-wise MAD for both conditions. Table 8 reports regional averages.

Region	Baseline	Cultural LLM	Reduction (%)
West	0.201	0.153	−23.7%
Latin Am	0.270	0.151	−44.1%
East	0.230	0.158	−31.3%

Table 8: Regional MAD for baseline (no-persona) and cultural LLM (persona-prompted) conditions, averaged across six models. Cultural persona prompting reduces human-model distance across all regions, with largest gains for non-Western contexts.

We observe that alignment follows a West < Latin Am < East pattern (0.20, 0.27, 0.29), con-

sistent with LLMs defaulting to Western question-asking styles. Introducing cultural personas reduces gaps substantially across all regions, but benefits non-Western regions more, after which regional differences nearly decrease. Therefore, we use cultural grounding in our main experiments, which also shows similar West < Latin Am < East patterns in alignment.

B.3 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:" 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 9 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 to determine divergence in country wise results:

Latin America. In Argentina and Brazil, we observe the highest alignment between human and

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

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 information-seeking that is 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 information-seeking 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 information-seeking. 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 information-seeking 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 information-seeking outputs across regions. The model aligns best in contexts where human information-seeking 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.

C.2 Topic Preference Divergence from previous findings

To deep dive into this divergence from previous findings, we qualitatively analyze topic preference rankings and interpret human-LLM similarity through established cultural and historical frameworks (e.g. collectivism-individualism, media ecosystems, etc.). We observe that similarity is strongest when human information-seeking aligns with widely documented cultural associations, for example, Argentina and Brazil prefer family and entertainment, consistent with collectivist values (Hofstede, 2001; Krys et al., 2022) and are entertainment-rich online cultures[8]. Similarly, France and Germany highlight arts and humanities, reflecting cultural history (e.g., European Wikipedias (Ahmed and Poulter, 2023)). Human-LLM similarity decreases when human information-seeking deviates from stereotypes or focuses on underrepresented topics. For example, in the UK, humans prioritize arts over politics despite the dominant political discourse (Brexit). Overall, while LLMs align closely with Western contexts, they also reinforce Western stereotypes (Borah et al., 2025a). These findings suggest that LLMs systematically reproduce traditional cultural stereotypes, favoring historically dominant topic–country associations. It is also important to note that this experiment measures relative topic-preference similarity, not absolute cultural alignment; thus, negative correlations (e.g., for the US in Fig 5) reflect divergence in ranked topics rather than lack of Western alignment.

D Grounding to social science theories and constructs

Figure 11 provides a brief description of the social science grounding in our framework. We show dimensions that positively correlate and negatively correlate with information-seeking in green and red highlights respectively.

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

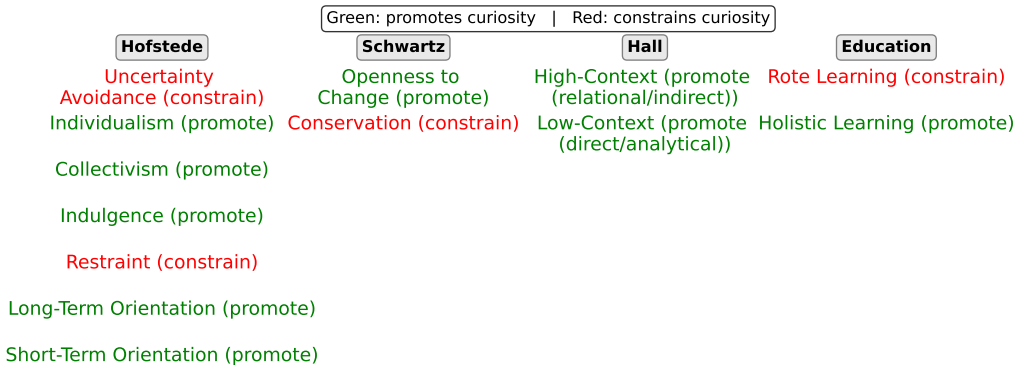

Figure 11: Social Science theories and connection to information-seeking

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 10 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 10: 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 11 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.

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, how-

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 11: Alignment to traditional theories

ever 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 Prompt Ablation Experiments

We perform some experiments to determine prompt sensitivity using prompt ablations for the linguistic analysis task. We also perform additional analyses with neutral (no-persona) LLM instantiation.

Specifically, we test out three different prompts in addition to the original one for the : (1) "You grew up in country and are familiar with the cultural norms there. From this cultural perspective, write <n> questions related to the topic: topic." (2) "Answer as someone living in country, drawing on typical cultural perspectives. Produce <n>

questions that you would naturally ask about the topic: topic." and (3) "Generate <n> questions related to the topic: topic. Do not adopt any particular cultural background; answer as a neutral respondent." Here, (1) and (2) include cultural perspective, whereas (3) does not include any persona. Averaging across countries and topics, we find that (1) and (2) generate closer outputs to original prompt, with cosine similarity among metrics averaging around 0.6. The neutral persona is farther away from the original prompt with cosine similarity around 0.04. On doing country-level analysis, we find the neutral persona is closest to Western countries Italy, US, and France (similarity of 0.6). This shows that LLM inherently are closer to output questions that are similar to the ones in Western.

F Detailed Construct Validity Analysis

A central concern is whether our linguistic metrics: *ambiguity*, *open-endedness*, *cohesion*, and *rhetorical devices* capture information-seeking-driven question-asking behavior or merely surface-level stylistic variation. To address this, we conduct a multi-source construct validity analysis, comparing our metrics against two independent external benchmarks: (1) Hofstede's Uncertainty Avoidance (UA) index (Hofstede, 2001) and (2) the Intellectual Curiosity Scale (ICS) validated by Bluemke et al. (2024).

In this section, we investigate whether they produce country-level patterns consistent with established cultural and psychological theory if they are capturing information-seeking-related expression. The theoretical rationale connects UA to information-seeking as follows: Uncertainty Avoidance measures the degree to which members of a culture feel threatened by ambiguous or unknown situations (Hofstede, 2001). High-UA cultures tend to suppress ambiguity, whereas low-UA cultures are more tolerant of exploratory thinking. If our metrics capture information-seeking questioning, we predict that: (a) rhetorical devices and lexical ambiguity should correlate negatively with UA; and (b) open-endedness may behave differently, depending on whether high-UA cultures employ explicit interrogative framing as a strategy for resolving uncertainty. Furthermore, we examine directional consistency with psychometric scores from Bluemke et al. (2024).

F.1 Hofstede Uncertainty Avoidance Correlation

We obtained published UA scores for all 18 countries from Hofstede (2001) and correlated them with per-country mean human linguistic scores computed from the Yahoo! Answers dataset. Table 12 lists the country-level values sorted by UA.

Correlation results using Spearman's ρ and Kendall's τ are shown in Table 13.

Rhetorical devices. The strongest result is the significant negative correlation between UA and rhetorical device use ($\rho = -0.521$, $p = 0.027$). High-UA countries such as Argentina (UA=86) and France (UA=86) show fewer rhetorical markers than low-UA countries such as Singapore (UA=8) and Australia (UA=51). Because rhetorical devices capture expressive and stylistically elaborate questioning patterns, this alignment supports the interpretation that these markers reflect information-seeking expression.

Ambiguity. Ambiguity trends in the predicted negative direction ($\rho = -0.139$) but is not statistically significant. The relatively narrow variance of lexical ambiguity across countries limits statistical power, suggesting potential for future research.

Open-endedness. Open-endedness shows a marginal positive correlation with UA ($\rho = 0.456$, $p = 0.057$). This suggests that high-UA cultures may use explicit interrogative framing as a strategy for resolving uncertainty rather than as an expression of open-ended exploration.

Cohesion. Cohesion does not show a significant correlation with UA. This is expected, as cohesion measures structural coherence within questions rather than exploratory or uncertainty-driven inquiry.

F.2 Directional Validity using Intellectual Curiosity Scale

Bluemke et al. (2024) validated a six-item Intellectual Curiosity Scale (ICS) across six countries: USA, Germany, France, Spain, Poland, and Japan. Of these, three countries overlap with our Yahoo dataset (USA, Germany, France). With only three overlapping countries, statistical testing is not meaningful. However, we perform a directional comparison: Germany has the lowest ICS score among the three ($\Delta = -0.27$ relative to USA) and

Country	UA	Ambiguity	Open-Ended	Rhetorical Dev.	Cohesion
Singapore	8	0.0293	0.739	2.478	0.145
China	30	0.0251	0.756	2.400	0.167
Vietnam	30	0.0235	0.694	2.369	0.169
UK	35	0.0227	0.750	2.356	0.138
India	40	0.0191	0.706	2.269	0.155
Philippines	44	0.0208	0.727	2.447	0.147
USA	46	0.0230	0.720	2.255	0.116
Canada	48	0.0236	0.727	2.342	0.165
Indonesia	48	0.0240	0.756	2.375	0.167
Australia	51	0.0261	0.765	2.488	0.146
Thailand	64	0.0222	0.708	2.298	0.160
Germany	65	0.0268	0.826	2.117	0.145
Taiwan	69	0.0211	0.720	2.168	0.182
Italy	75	0.0225	0.788	2.325	0.168
Brazil	76	0.0203	0.700	2.350	0.160
Mexico	82	0.0283	0.831	2.319	0.164
Argentina	86	0.0238	0.765	2.072	0.153
France	86	0.0211	0.827	2.356	0.165

Table 12: Per-country Hofstede UA scores and human linguistic metric means from IQUEST. Countries sorted by UA (ascending).

Metric	ρ	p	τ	p(τ)
Rhetorical Devices	-0.521	0.027**	-0.356	0.040**
Ambiguity	-0.139	0.584	-0.079	0.649
Open-Endedness	0.456	0.047**	0.321	0.068†

Table 13: Correlation between Hofstede UA and human linguistic scores ($n = 18$). ** $p < 0.05$

also exhibits higher mean human–LLM linguistic distance than both France and the USA in our dataset (Fig. 4 in main paper). This directional alignment suggests that populations with lower psychometric curiosity scores also deviate more strongly from LLM-generated questioning patterns. We emphasize that this observation is only directional and should not be interpreted as statistical evidence.

F.3 Summary

Table 14 summarizes the construct validity evidence across the two external benchmarks. Taken together, the above analyses suggest that IQUEST's linguistic metrics function as behavioral proxies for information-seeking question-asking while capturing different aspects of the broader information-seeking construct.

G Performance Differences between LLaMA and GPT

Using our IQUEST 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

Source	Result	Interpretation
Hofstede UA × Rhetorical	$\rho = -0.521$, $p = 0.027$	Significant negative
Hofstede UA × Ambiguity	$\rho = -0.139$, n.s.	Trend in predicted direction
Hofstede UA × OE	$\rho = 0.456$, $p = 0.057$	Positive; closure-seeking
ICS (Bluemke et al.)	Directional check	Lower ICS → larger deviation

Table 14: Summary of construct validity evidence.

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

`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. 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 12 shows examples of the above benchmarks. We experiment with our closest to human model (`LLaMA-3-8b`) and farthest model (`GPT-5`).

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 15: Model Performances on NormAD and CulturalBench-Hard

Table 15 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.

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 16: 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).

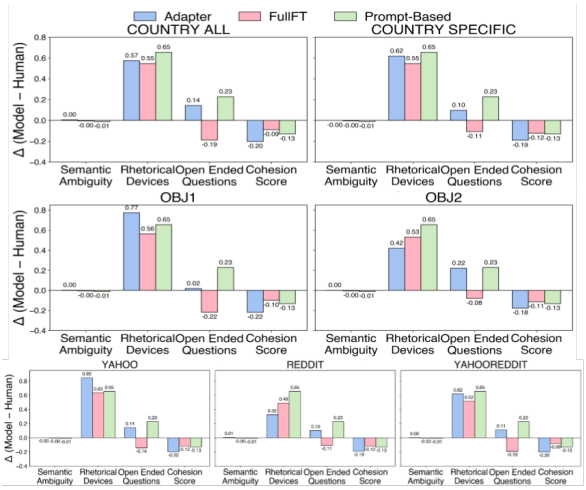

Figure 13: **Fine-tuning (full/adapter) results**: y-axis shows the delta between model and human scores, while the x-axis covers different evaluation metrics, averaged across countries. +ve/-ve scores mean model scores are higher/lower than humans. Fine-tuning (full and adapter-based) makes models' scores closer to humans, with adapters mostly improving rhetorical devices and open-ended questions, and full fine-tune improving cohesion scores.

H Fine-tuning for Human-LLM alignment

H.1 Behavioral Alignment Analysis of Fine-tuned Models

To further analyze the behavioral impact of fine-tuning, we measure the distance between model outputs and human linguistic metrics for the three countries used in training (Brazil, UK, and Philippines). For each metric m, we compute the mean absolute distance:

$$d_m = \frac{1}{N} \sum_{i=1}^{N} |x_{model,i} - x_{human,i}| \quad (8)$$

where i indexes countries and N is the number of countries considered. Lower values indicate closer alignment with human question-asking patterns.

Table 17 reports the results. Fine-tuning reduces distance to the human baseline for three of the four linguistic metrics: semantic ambiguity (-73.2%), rhetorical device usage (-57.4%), and open-ended question rate (-55.8%). Cohesion scores increase in distance, suggesting that discourse structuring properties may shift differently under fine-tuning. Averaging across all metrics and countries, the

macro-distance decreases from 0.256 to 0.146, corresponding to a 42.8% reduction in overall human–model divergence.

Finally, we measure agreement across the four linguistic metrics using Kendall's coefficient of concordance (W). Concordance increases from 0.19 in the baseline model to 0.56 after fine-tuning, indicating stronger agreement among behavioral signals following training.

H.2 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 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…"). The format template used for data generation for obj2 is provided in Fig 14. Obj2 prepares the model to learn conversational information-seeking, thus later showing superior performances in both cultural and inherent evaluation.

H.3 Linguistic Analysis Results

Fig 13 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 13 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.

H.4 Topic Preference Analysis Results

Table 16 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.

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.

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 14: Obj2 data generation

Metric	Baseline Dist.	FT Dist.	Reduction
Semantic Ambiguity	0.0103	0.0028	−73.2%
Rhetorical Devices	0.6547	0.2789	−57.4%
Open-ended Questions	0.2269	0.1002	−55.8%
Cohesion Score	0.1310	0.2029	+54.9%
Macro-average	0.256	0.146	−42.8%

Table 17: Behavioral alignment of baseline and fine-tuned models with human linguistic metrics. Distances represent MAD averaged across the three countries used for fine-tuning (Brazil, UK, Philippines). Lower values indicate closer alignment with human question patterns.

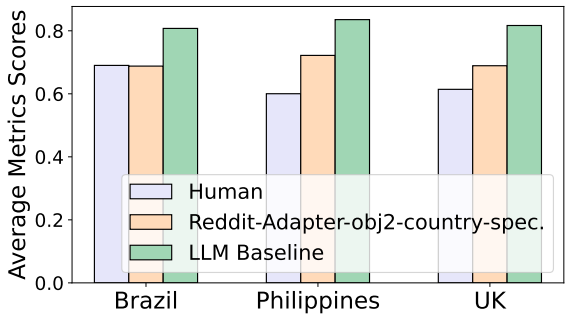

Figure 15: Country-wise analysis (inducing informationg-seeking) - here we compare the performance of the best fine-tuned model to human baseline and find it is much closer than the non-fine-tuned model.

H.5 Dataset for Inherent 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 information-seeking: "`Where did this happen?`"

2. **Incomplete / open-ended:** statements that feel unfinished that can invoke information-seeking. 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 19 and 20 show the data of non-cultural and cultural statements, respectively, that are used for inherent evaluation.

H.6 Inherent Evaluation metrics

We use the following metrics for determining inherent evaluation:

1. Question-Asking 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.

H.7 Inherent Annotation Framework

Fig. 16 shows the annotation framework we utilize for inherent evaluation (LLM's question-asking

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 16: Annotation Framework used for inherent evaluation

ability). We recruit two volunteers from a college campus for the study.[9] 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.

H.8 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 information-seeking questions and relevance to assess how not mentioning a country influences question-asking behavior.

We find that adapter-based models show the most information-seeking, 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 information-seeking in models.

H.9 Fine-tuning Results Implications

Behavioral vs. psychological curiosity. Our fine-tuning results are behavioral and not psychological. We do not claim models "experience" information-seeking internally (Bender et al., 2021). After fine-tuning, we observe output patterns that function like curiosity expressions — inviting exploration, reducing information gaps, and reflecting cultural context — consistent with how information-seeking is operationalized in human behavioral research (Berlyne, 1957; Graesser and Person, 1994).

Generalizable lessons. Three lessons transfer beyond our specific datasets and countries. First, adapter-based fine-tuning outperforms full fine-tuning, suggesting information-seeking questioning is best induced by augmenting rather than overwriting base model priors. Second, conversational scaffolding (obj2) outperforms direct instruction (obj1), consistent with how information-seeking is acquired through contextually grounded examples in human learning (Engel, 2011). Third, the asymmetric metric improvements — strong gains on ambiguity, rhetorical devices, and open-endedness but increased distance on cohesion — suggest different aspects of information-seeking questioning respond differently to fine-tuning, pointing to the need for multi-objective training in future work.

Internal consistency. The increase in Kendall's W from 0.19 to 0.56 indicates fine-tuning produces more internally consistent information-seeking be-

[9] Our task is designed to be understandable to non-experts and is objective, therefore we do not require experts

havior overall, suggesting the model learns a coherent questioning style rather than optimizing isolated surface features.

I Downstream Impacts of information-seeking in 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 12.

We employ two methods to elicit information-seeking question-asking behaviors 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.

Table 18 shows the results of the above for downstream tasks. As discussed previously, information-seeking-induced models perform better than non-information-seeking models, showing the importance of information-seeking 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. Overall, adapter-fine-tuned models help improve performance across all cultural-adaptability benchmarks.

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

Table 18: **Downstream tasks (% acc):** NormAd (NA), CulturalBench (CB), and Cultural Common-Sense (CS) - information-seeking models have higher acc.

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

I.1 Choice of Downstream Evaluation Benchmarks

The three benchmarks we use capture three complementary aspects of cultural adaptability where information-seeking-driven questioning is most consequential: judging social acceptability requires sensitivity to implicit cultural norms; cultural knowledge probes require moving beyond Western defaults and commonsense inference requires contextual reasoning rather than pattern matching. Together, these benchmarks provide a controlled, human-verified setting to isolate the effect of information-seeking on cultural reasoning across structurally different task types and hence reducing the risk of benchmark-specific artifacts. Our results show consistent gains across all three benchmarks. These findings strengthen the generalizability of our findings beyond any single task, as the benchmarks have different formats and cultural coverages.

Future evaluation directions. In the future, complementary evaluation settings could further establish the practical value of information-seeking-aligned LLMs. First, interactive dialogue evaluations measuring user satisfaction and task completion would directly test whether information-seeking-driven questioning improves real-world interaction quality (Ganuthula, 2024). Second, education-focused benchmarks such as measuring student engagement or learning outcomes in LLM-assisted tutoring (Wang et al., 2024) would test information-seeking's value in high-stakes

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 17: Prompt-Ask framework for downstream tasks - inducing information-seeking

pedagogical contexts (Lindholm, 2018; Markey and Loewenstein, 2014). Third, healthcare dialogue benchmarks (Wang et al., 2025) could assess whether information-seeking-prompted clarification reduces miscommunication in clinical interactions. Fourth, extending evaluation to multilingual settings and recently proposed culturally grounded benchmarks (Borah et al., 2025a; Nyandwi et al., 2025) would be important to evaluate. Finally, multi-agent settings where information-seeking-aligned LLMs interact with culturally diverse users (Bai et al., 2025) is also a promising direction to study information-seeking propagation and collaborative cultural understanding.

I.2 Prompt used for Downstream Tasks

Fig 17 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).

I.3 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.

J 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 information-seeking (inherent information-seeking) 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.

K Model Choices, Implementation Details and Computational Resources

For our experiments, we selected a mix of open- and closed-source models to examine cultural information-seeking across different architectures and training paradigms. Specifically, we used `LLaMA-3-8B` (Dubey et al., 2024) 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`,[10] `LLaMA-3-70B`, `Claude-Sonnet-4`,[11] 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 information-seeking alignment.

L 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 information-seeking in humans and LLMs, and its impact in downstream tasks.

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

Table 19: Dataset of Everyday Statements for Inherent information-seeking 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 20: Dataset of Cultural Statements (Country-wise) for Inherent 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.