

Country Is Not Enough: Recognition Action Gaps in Cultural Answerability

Aditya Channa Madhav Kapoor Jastej Singh

Maharaja Agrasen Institute of Technology, Delhi, India

{adityachannadelhi,madhav.kapoor13,jastej28.singh}@gmail.com

Abstract

Large language models increasingly answer culturally situated questions whose surface form specifies a country but omits the subnational, community or situational context needed for a reliable answer. We study *cultural answerability*: whether the cultural scope supplied by a user supports a direct answer without presenting one situated practice as a country-wide norm. Using region-specific evidence from INDICA, we construct a bilingual English–Hindi evaluation with 50 culturally stable and 50 culturally dependent held out questions. We evaluate two multilingual LLMs across three stages: explicit recognition of missing context, natural response policy, and paired regional counterfactual grounding. Models identify 89.5% of dependent conditions but correctly classify only 28.0% of stable conditions, revealing strong over detection of cultural dependence. Under natural assistance, the appropriate policy rate on dependent questions is 69.0%, while 30.5% of responses make an unscoped cultural commitment. Moreover in 28.0% of dependent model language conditions a model explicitly recognises missing context yet still makes such a commitment when answering naturally. Human adjudicated regional grounding succeeds in only 13.5% of paired conditions, showing that explicit regional context does not reliably produce grounded adaptation. These results show that cultural reliability is not primarily a knowledge problem: it is a calibration problem spanning recognition, response policy and grounded adaptation.

1 Introduction

Many everyday cultural questions appear answerable at the country level: what people wear to a wedding, how elders are greeted, which foods are avoided on ritual days, or what gifts are common at a family ceremony. For a general-purpose assistant this framing is tempting: asked “In India, what is the most common gift given to new par-

ents?”, a fluent model can produce a plausible answer immediately. Yet that plausibility may be local rather than national. A response that fits one region, community, institution, or household can become misleading once presented as what people in the country usually do.

We study this as a problem of *cultural answerability*. A culturally answerable question supports a reliable response at the scope the user supplied. A culturally dependent question instead requires qualification, clarification, or additional context such as region, religion, community, institution, household practice, occasion, or relationship. We call a response an *unscoped cultural commitment* (UCC) when it presents one situated practice as a country-wide norm without meaningful qualification.

We focus on India as a high-variation setting where country-level phrasing is common but often too coarse. We derive a bilingual English–Hindi benchmark from region-specific references in INDICA (Madhusudan et al., 2026), balanced between STABLE questions, whose references share a stable answer-bearing core, and DEPENDENT questions, where an unqualified country-level answer can materially mislead. We evaluate two hosted multilingual LLMs through three complementary experiments. Experiment A asks the model to decide explicitly whether more within-country context is needed. Experiment B gives the model the original country-level question and measures natural assistant behaviour. Experiment C supplies two different Indian regions in fresh conversations and tests whether outputs satisfy the predeclared region-specific relation.

These experiments separate three behaviours often conflated: classifying a question as context-dependent, selecting a responsible response policy, and binding substantive cultural content to supplied context. We find a persistent recognition–behaviour disconnect: models are sensitive to cultural dependence under explicit diagnosis but poorly cal-

ibrated, rarely ask for clarification during natural assistance, and, even once regional context is supplied, often localise the surface form of an answer without adapting its substantive claim.

Our contributions are:

- We introduce cultural answerability as an evaluation target linking context recognition, response policy and grounded cultural adaptation.
- We construct a bilingual English–Hindi benchmark with stable controls, dependent questions and counterfactual regional pairs derived from INDICA.
- We operationalise UCC as a natural-assistance scope failure and PRAF as the same-condition disconnect between explicit dependence classification and unscoped natural assistance.
- We define Regional Counterfactual Grounding (RCG) as the primary Experiment C metric, computed from final human-adjudicated labels, and report automated judge-order diagnostics separately.

2 Related Work

Cultural evaluation and the granularity of culture. Culturally aware NLP spans knowledge, values, norms, pragmatics, and adaptation, but lacks a single operational definition of culture (Liu et al., 2025). National or linguistic labels are convenient proxies rather than homogeneous cultural units: Zhou et al. (2025) argue that national boundaries can erase local, community-specific meaning and that benchmarks should avoid reducing culture to static trivia. Broad resources such as BLEND, CULTURALBENCH, and CULTUREBANK show that LLM performance varies across cultural settings and that multiple plausible answers may co-exist (Myung et al., 2024; Chiu et al., 2025; Shi et al., 2024). We address a complementary question: before evaluating *which* cultural answer an LLM produces, can it determine whether the supplied context supports a single direct answer at all?

Subnational variation and cultural commonsense. A smaller body of work examines variation within countries. INDOCULTURE evaluates geographically influenced commonsense across eleven Indonesian provinces and shows that location context improves reasoning (Koto et al., 2024).

DOSA represents social artifacts from 19 Indian subcultures through participatory collection (Seth et al., 2024). Most directly, INDICA collects region-specific answers across five Indian regions and documents within-country disagreement and geographic defaults in LLM responses (Madhusudan et al., 2026). We use INDICA’s regional references as evidence but shift the focus from regional knowledge retrieval to *cultural answerability*: stable controls test whether models avoid requesting unnecessary context when regional references share an answer-bearing core, while dependent items test whether situated answers are incorrectly promoted to national norms.

Ambiguity, clarification, and response policy.

Prior work examines whether LLMs recognise underspecified information needs and decide when interaction is beneficial. CLAMBER finds models unreliable at identifying ambiguity and generating effective clarifying questions (Zhang et al., 2024); other work frames clarification as a decision balancing benefit against interaction cost (Zhang and Choi, 2025), or uses output uncertainty to improve ambiguity detection (Shi et al., 2025). Most closely, Su and Cardie (2026) identify a recognition–action gap in which LLMs detect ambiguity under explicit diagnosis but rarely seek clarification during normal answering. Cultural underspecification differs from lexical or referential ambiguity: answers may be locally valid yet misleading at national scope, and appropriate responses may involve qualification or requests for non-regional context rather than clarification alone. We therefore compare explicit dependence judgments with matched natural responses to test whether recognised limitations constrain the final answer.

From cultural knowledge to grounded behaviour.

Recent work distinguishes possessing cultural knowledge from applying it in context. NORMAD shows that LLMs struggle even when relevant cultural-norm information is supplied (Rao et al., 2025), while SOCIALCC reports a similar gap between stored knowledge and appropriate behaviour in interactive cross-cultural settings (Wu et al., 2025). Explicit cultural prompts can surface localised knowledge but may also reduce diversity or amplify stereotypes (Veselovsky et al., 2026). Multilingual studies show that LLMs recover only part of the adaptation induced by explicit cultural instructions (Nasim et al., 2026), and may infer users’ cultural context without consistently apply-

Split	STABLE	DEPENDENT	Total
Development	10	10	20
Held-out test	50	50	100
Total	60	60	120

Table 1: Dataset composition. The resource contains 120 matched English–Hindi question pairs, including 20 development items and 100 held-out test items.

ing it in responses (Miao et al., 2026). We extend these findings to subnational counterfactual grounding: matched English and Hindi questions fix the information need, while paired regional prompts test whether answer-bearing content changes, remains invariant, or stays qualified when regional context alone is insufficient.

3 Dataset and Annotation

3.1 Source and Data Splits

We derive our evaluation from the region specific references in INDICA (Madhusudan et al., 2026). Each source question is associated with benchmark supported answers from up to five Indian regions: North, South, East, West and Central. We treat these references as evidence of reported regional practices, not descriptions of cultural groups.

The final resource contains 120 matched English–Hindi question pairs. Twenty form a development set used only for prompt design, guideline refinement, parser testing, and judge calibration. The held-out test set contains 100 questions, balanced between 50 STABLE and 50 DEPENDENT items. All prompts, policy labels, regional pairs, and evaluation rules are frozen before test generation; only the held-out set is used for reported model results.

3.2 Cultural Answerability Labels

Each retained question receives one of two labels. A question is STABLE when the regional references preserve a sufficiently consistent answer-bearing core, so a concise country-level response can be reliable without selecting one situated practice as nationally representative. A question is DEPENDENT when an unqualified country-level answer could materially mislead because the answer-bearing content varies across regions, or because another variable—religion, community, institution, city, household, or occasion—remains necessary.

This distinction concerns the reliability of an answer’s *scope*, not whether cultural variation exists in any form: minor differences in wording or

Pair type	<i>n</i>	Expected behaviour
CONTRAST	34	Adapt the answer-bearing content to the supplied region.
INVARIANCE	7	Preserve the shared core and avoid inventing differences.
REGION-INSUFFICIENT	9	Remain qualified and avoid treating region as fully determinative.

Table 2: Predeclared Experiment C pair types.

implementation do not make an item dependent when the core answer remains stable, while a locally plausible answer may still be inappropriate at country scope if it constitutes a UCC. Candidates are excluded when references are contradictory, too incomplete to support a label, or fail to instantiate the target construct.

3.3 Acceptable Response Policies

Because responsible assistance does not always require clarification, labels are paired with a set of response policies: {DIRECT, QUALIFY, ASK-REGION, ASK-OTHER-CONTEXT}, and multiple policies may be approved for the same item, a dependent question may permit either a qualified answer or a targeted request for regional context.

Policy success is computed from observable response behaviour rather than surface wording. A dependent response succeeds when it follows at least one approved scope-control policy, stays relevant, and avoids UCC. A stable response succeeds when it provides a substantive answer without unnecessarily refusing or demanding additional context; a brief harmless caveat is not penalised.

3.4 English–Hindi Parallel Construction

Every English question is first translated into Hindi using Claude Opus 4.8, then reviewed across the full 120-item resource by two bilingual human reviewers for semantic equivalence, naturalness, and preservation of the same missing context—no resolving variable may be introduced or removed, and culture-specific terms may not be replaced with unnecessarily North- or Hindi-specific forms. Reviewer corrections, not the machine draft, define the released Hindi version. We therefore interpret English–Hindi differences as changes in observable model behaviour under matched prompts, not as evidence that either language uniquely identifies a user’s cultural background.

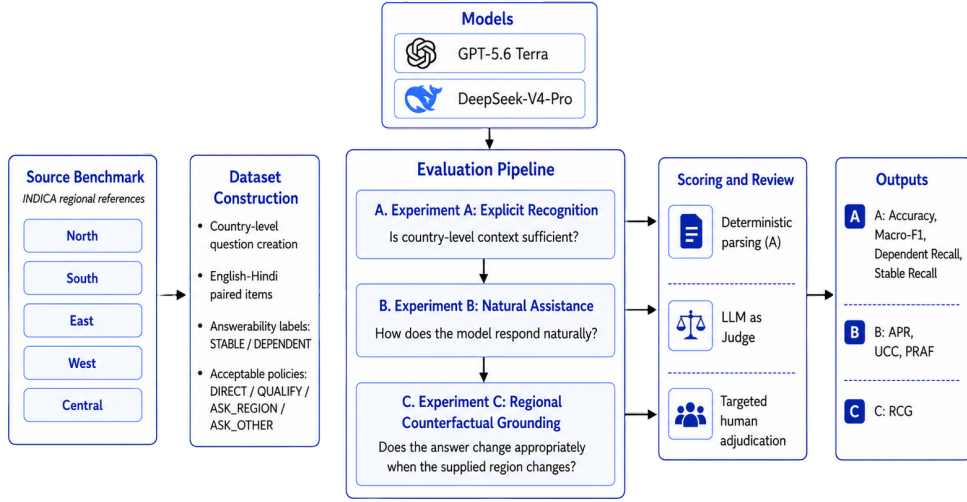


Figure 1: **Overview of the cultural answerability evaluation pipeline.** Country-level English–Hindi questions derived from INDICA are evaluated through explicit recognition, natural assistance and regional counterfactual grounding.

3.5 Regional Counterfactual Pairs

For every DEPENDENT item we preselect two valid regional contexts from the corresponding INDICA references; the region pair, reference answers, and expected relation are fixed before model generation. The 50 templates comprise 34 CONTRAST items, where answer-bearing content should change; 7 INVARIANCE controls, where the shared core should be preserved; and 9 REGION-INSUFFICIENT items, where the model should stay qualified because region alone does not resolve the question. Exposure across the 100 regional prompt slots is balanced over all five regions.

3.6 Annotation and Adjudication

Dataset validation precedes test generation and remains separate from model evaluation. Two bilingual Indian annotators from distinct regional and cultural backgrounds independently review every item’s answerability label, acceptable policy set, regional evidence, Experiment C metadata, and final English–Hindi pair. Disagreements are resolved through discussion using the frozen guidelines and INDICA references; thus, the released dataset is fully human-reviewed rather than checked only on flagged cases.

Experiment A is scored deterministically from the first-line YES/NO output. For Experiment B, GPT-5.6 Sol assigns a structured behavioural label to each of the 400 natural-assistance responses. For

Experiment C, the 400 regional responses form 200 matched pairs, judged in both A–B and B–A order under a frozen schema and stored as one pair-level record.

Targeted human adjudication covers records with elevated risk, including potential UCC, unsupported generalisation, partial reference compatibility, assignment reversal, language drift, malformed output, explicit judge uncertainty, or disagreement between judge orders. The adjudicator confirms the structured label or replaces it with a corrected judgment. On the held-out test set, 62 Experiment B responses and 126 Experiment C pair records are adjudicated; all remaining records retain their structured labels. All qualitative examples are manually verified.

GPT-5.6 Sol evaluates all Experiment B responses and Experiment C pairs. The reduction from 1,200 model calls to 600 judge records follows from deterministic Experiment A scoring and pair-level Experiment C evaluation, where each record contains two region-conditioned responses.

4 Experimental Design

Figure 1 summarises the complete evaluation. The three experiments isolate distinct stages of cultural answerability: recognising whether the supplied scope is sufficient, selecting an appropriate natural response policy, and grounding answer-bearing content to explicit regional context.

4.1 Models and Runtime Controls

We evaluate GPT-5.6 Terra with medium reasoning and DeepSeek-V4-Pro with thinking enabled. Each request is issued in a fresh conversation with tools, web search, retrieval, memory, and prior conversational context disabled. The system instruction asks the model to respond concisely in the language of the user. Prompts and model settings are versioned and frozen before held-out test execution.

The test benchmark contains 1,200 generation calls: 400 for Experiment A, 400 for Experiment B, and 400 for Experiment C. Experiment C responses are grouped into 200 matched regional pairs.

4.2 Experiment A: Explicit Recognition

Experiment A uses an explicit recognition prompt. Given the stored country-level question, the model is asked whether a reliable answer materially requires additional within-country context, such as region, community, religion, institution, city or venue, household, occasion, or a similar variable. The model must output only YES or NO and is explicitly instructed not to answer the original question. We map YES to DEPENDENT and NO to STABLE and score the output deterministically. We report accuracy, Macro-F1, dependent recall, and stable recall. Macro-F1 is the primary summary because high dependent recall can be obtained by predicting that nearly every cultural question requires additional context.

4.3 Experiment B: Natural Assistance

Experiment B presents only the stored English or Hindi question; the prompt does not mention ambiguity, cultural variation, qualification, or clarification. GPT-5.6 Sol assigns structured behavioural flags to every response: substantive answer, scope qualification, request for region, request for another context variable, UCC, refusal or irrelevance, and language drift. The source-model identity and gold policy set are hidden from the judge.

A stable response succeeds when it provides a substantive answer without unnecessarily refusing or requesting additional context. A dependent response succeeds when it follows at least one item-approved scope-control policy, remains relevant, and avoids UCC. Appropriate Policy Rate (APR) is the proportion of responses satisfying these item-specific rules.

To connect explicit recognition with natural behaviour, we define Paired Recognition–Action Fail-

ure:

$$\text{PRAF} = P(\hat{y}_A = D, \text{UCC}_B = 1 \mid y = D),$$

where D denotes a gold DEPENDENT item. PRAF therefore identifies matched model–language conditions that explicitly diagnose missing context in Experiment A but nevertheless make an unscoped commitment in Experiment B. The corresponding conditional rate among recognised dependent conditions is reported as a secondary diagnostic.

4.4 Experiment C: Regional Counterfactual Grounding

Experiment C presents two fresh region-specific versions of every dependent question. For each template, the two regions, regional references, pair type, and expected relation are fixed before model generation. The intervention tests calibrated sensitivity rather than simply rewarding different answers: CONTRAST items should change, INVARIANCE items should preserve their shared core, and REGION-INSUFFICIENT items should remain qualified because region alone does not resolve the question.

GPT 5.6 Sol evaluates every regional pair in both A–B and B–A order. For a pair p , Regional Counterfactual Grounding succeeds only when both individual responses are reference-compatible and the predeclared relation is satisfied:

$$\text{RCG}(p) = \mathbf{1}[C_A \wedge C_B \wedge R_p \wedge \neg V_p \wedge \neg I_p \wedge Q_p \wedge \neg G_p].$$

where C_A and C_B denote compatibility with the two assigned references, R_p is the type-specific relation, V_p denotes reversed regional assignments, I_p denotes an ignored region, Q_p denotes adequate qualification when required, and G_p denotes unsupported generalisation. The type-specific relation is given in Table 3.

4.5 Final Labels and Uncertainty

For Experiments B and C, the final protocol label uses the human-adjudicated JSON when a row is reviewed and otherwise retains the frozen structured judgment. Thus, “final” refers to the output of the targeted-adjudication protocol, not exhaustive independent human annotation of every response.

Model and language differences are treated descriptively unless supported by the matched outcome analysis.

Type	Relation-specific requirement
CONTRAST	The answer-bearing difference between the two regional references is realised.
INVARIANCE	The shared answer-bearing core is preserved without fabricating a contrast.
REGION-INSUFFICIENT	Region is not treated as fully resolving the question; the remaining answer-changing context is preserved.

Table 3: Type-specific requirements for Regional Counterfactual Grounding.

Stage	Metric	Count	Rate (%)
A	Accuracy	235/400	58.8
A	Macro-F1	—	54.4
A	Dependent recall	179/200	89.5
A	Stable recall	56/200	28.0
B	Dependent APR	138/200	69.0
B	Dependent UCC	61/200	30.5
B	Qualification (dep.)	170/200	85.0
A+B	PRAF	56/200	28.0
C	Final RCG (adjud.)	27/200	13.5
C	CONTRAST RCG	13/136	9.6
C	INVARIANCE RCG	11/28	39.3
C	REGION-INSUFF. RCG	3/36	8.3

Table 4: Principal held-out results.

5 Results and Analysis

Table 4 reports the principal held-out results, summarised visually across the three evaluation stages in Figure 2. Across the three experiments, recognising possible cultural dependence does not reliably translate into an appropriate response policy or grounded use of supplied regional context. Table 5 provides representative cases for the failure patterns discussed below.

5.1 Recognition Is Sensitive but Poorly Calibrated

Experiment A identifies 89.5% of dependent conditions but correctly identifies only 28.0% of stable conditions. Overall, 80.8% of judgments predict that additional cultural context is required. The low Macro-F1 is therefore driven primarily by over-detection rather than cultural blindness. Eighteen stable questions are classified as dependent in all four model–language conditions, whereas no dependent question is missed in all four. Stable controls are essential here: without them, a system that recommends qualification for nearly every cultural question would appear highly capable.

The q_307 case in Table 5 illustrates this error. Although the regional references preserve silk as their shared answer-bearing core, every model–

language condition classifies the question as dependent. The model recognises that some surrounding practices may vary, but fails to distinguish such variation from variation that materially changes the requested answer.

5.2 Models Qualify but Almost Never Clarify

On dependent questions, APR falls to 69.0%, and 30.5% of responses make a UCC. Models overwhelmingly answer immediately: 85.0% of dependent responses contain a qualification, but none asks which region the user means and only one asks for another context variable. Moreover, 32 of the 62 policy failures contain both a qualification and a UCC. We call this a *caveat–commitment contradiction*: the response acknowledges variation while still centring one situated practice as the national answer. Counting hedging language alone would therefore overestimate responsible behaviour.

The model difference is substantial. GPT-5.6 Terra obtains 84% dependent APR with 15% UCC, compared with 54% APR and 46% UCC for DeepSeek-V4-Pro. Aggregate English-Hindi differences are much smaller (70% versus 68% dependent APR), indicating that language effects are model-specific rather than favouring one language.

5.3 Recognition Does Not Guarantee Responsible Action

PRAF occurs in 56 of 200 dependent model–language conditions (28.0%). Among the 179 conditions that explicitly recognise dependence, 31.3% still make a UCC under natural assistance. The gap is especially pronounced for DeepSeek-V4-Pro: its PRAF rate is 45%, compared with 11% for GPT-5.6 Terra. This extends the recognition–action pattern observed for generic ambiguity (Su and Cardie, 2026) to culturally situated questions, where the failed response may remain locally plausible rather than straightforwardly false.

The q_128 triplet makes this disconnect concrete. Terra classifies the item as DEPENDENT in Experiment A, but subsequently presents Diwali as the national answer under natural assistance. When regional context is supplied, the model produces the benchmark-supported East–Central distinction between Durga Puja and Diwali. Thus, the same model configuration exhibits appropriate scope sensitivity under explicit classification and regional prompting, but not under the natural country-level assistance framing.

Case	Experiment A: parsed outcome	Experiment B: natural assistance	Experiment C / diagnosis
q_128, Terra-EN	Predicted DEPENDENT (YES); gold DEPENDENT.	“In India, Diwali is the holiday most commonly associated with wearing new clothes.” The response makes an unscoped country-level commitment.	East: Durga Puja; Central: Diwali. The model realises the regional distinction once explicit regional context is supplied, despite defaulting nationally in Experiment B.
q_436, DeepSeek-EN	Predicted DEPENDENT (YES); gold DEPENDENT.	“The most common nonverbal gesture is touching the feet.” The response presents one practice as the national default.	Both West and Central responses give touching feet, although the West reference specifies joined palms. The substantive default therefore persists after the region changes.
q_307, Terra-EN	Predicted DEPENDENT (YES); gold STABLE. This is an over-contextualisation false positive.	“Silk is the most universally auspicious choice.”	Stable control: the regional references preserve silk as the shared answer-bearing core. Experiment C is not applicable because regional counterfactual pairs are constructed only for dependent items.

Table 5: Representative cross-stage diagnostic cases. Experiment A reports the parsed binary prediction rather than a free-text rationale. Experiment B excerpts are shortened without changing their answer-bearing claims. Experiment C reports the regional outcome or the relevant diagnosis.

5.4 Explicit Region Does Not Ensure Grounded Adaptation

Final RCG is 13.5%. The pair-type decomposition is more informative than the aggregate alone. Invariance controls are easiest at 39.3%, while contrast items achieve 9.6% and region-insufficient items 8.3%. Models can therefore preserve a shared cultural core more reliably than they can bind different answer-bearing content to different regions or recognise that another variable remains necessary.

The common failure is not a complete reversal of the two regions. Instead the model changes the regional framing while retaining the same substantive default gives broad lists compatible with multiple regions or adds unsupported region-wide claims. In q_436, for example DeepSeek produces touching feet for both West and Central India even though the West reference specifies joined palms. The regional label changes, but the answer bearing default persists. GPT-5.6 Terra succeeds on 24 of 100 conditions compared with 3 of 100 for DeepSeek-V4-Pro. English succeeds on 18 of 100 conditions and Hindi on 9 of 100; given the small success counts we treat this language difference descriptively.

6 Discussion

6.1 Recognition Does Not Reliably Control the Final Claim

Explicitly recognising cultural dependence is not sufficient if that diagnosis does not constrain the

generated answer. This disconnect appears in two forms. First, models that identify missing context in Experiment A can still make an unscoped country-level commitment under natural assistance. Second, even responses containing a qualification may continue to centre the same national default. The resulting caveat–commitment contradiction shows why detecting hedging language alone is an inadequate measure of culturally responsible assistance: a qualification is useful only when it narrows the scope of the answer-bearing claim. This extends the recognition–action gap observed for generic ambiguity (Su and Cardie, 2026) to cases where the resulting answer may remain locally plausible while being unreliable at the scope requested by the user.

6.2 Supplying Cultural Context Can Still Produce False Certainty

Experiment C treats region as a controlled intervention rather than a complete description of cultural identity. Contrast pairs test whether answer-bearing content changes with the supplied region, while region-insufficient pairs test whether models preserve the need for another resolving variable, such as religion, community, institution, household, relationship, or occasion. The low grounding rates on both types indicate two related failures: models may retain a substantive default after the regional label changes, or treat the added label as more determinative than the available evidence supports.

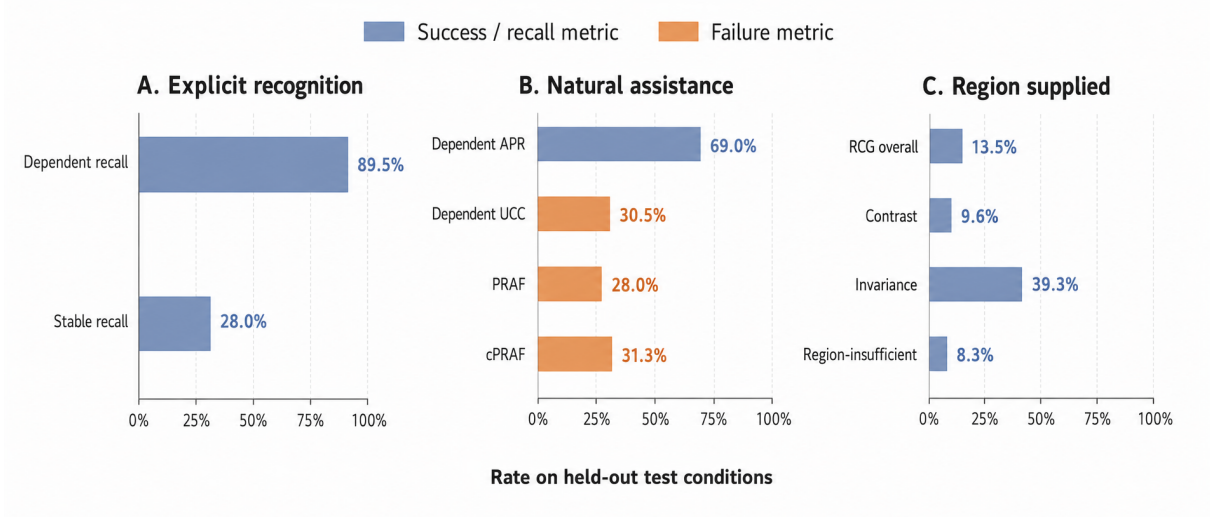


Figure 2: Summary of the main held-out test results across the three evaluation settings. Experiment A shows high dependent recall but low stable recall, indicating over detection of cultural dependence. Experiment B shows that dependent questions often still receive unscoped country level answers under natural assistance. Experiment C shows that supplying a region rarely yields successful answerable regional grounding under final human labels.

Reliable adaptation therefore requires identifying which context actually changes the answer, not merely adding demographic specificity. This is consistent with evidence that models may infer or receive cultural context without applying it appropriately in their final response (Miao et al., 2026).

6.3 From Diagnosis to Mitigation

Our evaluation primarily diagnoses failures in cultural answerability rather than evaluating corrective interventions. The recognition–action gap suggests one possible role for an explicit oversight stage that checks whether a final response is consistent with an earlier assessment that additional context is required. X-Agent, for example, uses an oversight reasoning framework to audit model behaviour and steer final outputs when undesirable behaviour is detected (Pucci and Ranaldi, 2025). While X-Agent targets sycophancy rather than cultural answerability, its oversight mechanism is complementary to the recognition–action failure identified here. A second direction is to make the comparison among competing contextual evidence explicit before answer generation. Dialectic-RAG structures multilingual retrieval around comparing, contrasting, and resolving conflicting evidence (Ranaldi et al., 2025). For cultural answerability, a related strategy could compare regional evidence before deciding whether answer-bearing content should change, remain invariant, or remain qualified. Neither intervention is evaluated in the present study;

testing whether such mechanisms reduce PRAF and improve RCG is left to future work.

7 Conclusion

We introduced cultural answerability as an evaluation of whether the cultural scope supplied by a user supports a reliable direct answer. Across 100 held-out English–Hindi questions, the evaluated models usually recognise dependent conditions but over-predict dependence on stable controls. Under natural assistance, they almost never request clarification, and qualifications frequently coexist with unscoped country-level commitments. Supplying regional context does not resolve the problem: Regional Counterfactual Grounding succeeds in only 13.5% of model–language pair conditions.

Stable controls are important. Without them, a model that qualifies nearly every cultural answer could appear reliably cautious despite failing to distinguish genuine dependence from harmless variation. Cultural answerability therefore requires both sensitivity and restraint: models should seek or apply additional context when it changes the answer-bearing claim, but should preserve a shared core when the available evidence supports one. More broadly, culturally grounded assistance cannot be evaluated through cultural knowledge or ambiguity recognition alone; evaluation must also test whether the model’s response policy and substantive claims remain calibrated to the scope supported by the user and the available evidence.

Limitations

The study evaluates two hosted model versions, two languages, and broad Indian regional references from INDICA. Results should not be generalised beyond these settings. Broad regions conceal substantial state, linguistic, religious, caste, and urban-rural variation, and benchmark compatibility does not imply universal cultural correctness. The references should therefore be treated as evidence of reported practices rather than exhaustive descriptions of regional or community behaviour.

The 100-item held-out test set limits fine-grained claims by domain and region. Although English and Hindi items were reviewed for equivalence, translation may still affect pragmatics, register, or model familiarity. Experiment A also enumerates possible context variables, which may encourage dependence over-prediction; neutral and counter-balanced prompts should be tested.

Experiments B and C use exhaustive judging with targeted human adjudication, so errors outside the routing criteria may remain. Unflagged records retain their frozen structured labels rather than receiving independent human review. Qualitative examples are manually verified, and results should be interpreted as protocol-guided evaluation rather than ethnographic validation.

Ethical Considerations

This work does not attempt to define a single authoritative representation of Indian culture. Regional labels are analytical abstractions and should not be treated as homogeneous identities. We do not infer a user’s region, religion, community, caste, or ethnicity from their language, and matched English–Hindi conditions are used only to study model behaviour. Model outputs and source references may reproduce stereotypes or simplifications; released materials should therefore retain qualifications, provenance, and annotation rationales. No personal user data are collected.

References

- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. [IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces](#). *Transactions of the Association for Computational Linguistics*, 12:1703–1719.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Sangmitra Madhusudan, Trush Shashank More, Steph Buongiorno, Renata Dividino, Jad Kabbara, and Ali Emami. 2026. [Common to whom? regional cultural commonsense and LLM bias in India](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5474–5519, San Diego, California, United States. Association for Computational Linguistics.
- Yisong Miao, Jian Zhu, and Vered Shwartz. 2026. [LLMs infer cultural context but fail to apply it when responding](#). *Preprint*, arXiv:2606.17688.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. [BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages](#). *Advances in Neural Information Processing Systems*, 37:78104–78146.
- Mehwish Nasim, Sanjeevan Selvaganapathy, Neel Ganapathi Sabhahit, Marie Griesbach, Pranav Bhandari, Janina Lütke Stockdiek, Lennart Schäpermeier, Usman Naseem, and Christian Grimme. 2026. [Do LLMs use cultural knowledge without being told? a multilingual evaluation of implicit pragmatic adaptation](#). *Preprint*, arXiv:2604.17718.
- Giulia Pucci and Leonardo Ranaldi. 2025. [Advancing oversight reasoning across languages for audit sycophantic behaviour via X-agent](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12949–12965, Suzhou, China. Association for Computational Linguistics.
- Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. 2025. [Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9064–9085, Suzhou, China. Association for Computational Linguistics.

- Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. **NormAd: A framework for measuring the cultural adaptability of large language models**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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.
- Weiyang Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rog rio Abreu de Paula, and Diyi Yang. 2024. **CultureBank: An online community-driven knowledge base towards culturally aware language technologies**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.
- Zhengyan Shi, Giuseppe Castellucci, Simone Filice, Saar Kuzi, Elad Kravi, Eugene Agichtein, Oleg Rokhlenko, and Shervin Malmasi. 2025. **Ambiguity detection and uncertainty calibration for question answering with large language models**. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 41–55, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jinyan Su and Claire Cardie. 2026. **Knowing but not showing: LLMs recognize ambiguity but rarely ask clarifying questions**. *Preprint*, arXiv:2605.25284.
- Veniamin Veselovsky, Berke Argin, Benedikt Stro bl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2026. **Localized cultural knowledge is conserved and controllable in large language models**. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43152–43178, San Diego, California, United States. Association for Computational Linguistics.
- Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. **SocialCC: Interactive evaluation for cultural competence in language agents**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271, Vienna, Austria. Association for Computational Linguistics.
- Michael JQ Zhang and Eunsol Choi. 2025. **Clarify when necessary: Resolving ambiguity through interaction with LMs**. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. **CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.
- Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. **Culture is not trivia: Sociocultural theory for cultural NLP**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

A Judge-Order Diagnostics

Experiment C is judged in both A–B and B–A order. The GPT-5.6 Sol diagnostic counts a model–language pair condition as consensus-only when both orders agree on compatibility, expected relation satisfaction, qualification adequacy, default persistence, assignment reversal, unsupported generalisation, and ignored region. In the current test run, 75/200 conditions reached this consensus set. The consensus-only success rate is 19/75 (25.33%), while the conservative all-conditions success rate is 19/200 (9.50%). We treat these as diagnostics and use human-adjudicated RCG as the primary Experiment C metric.

B Reproducibility Details

All requests are single-sample, fresh-conversation calls with tools, retrieval, web search, and memory disabled. The evaluated generator model names are GPT-5.6 Terra via Azure OpenAI and DeepSeek-V4-Pro via OpenRouter. The judge is GPT-5.6 Sol via Azure OpenAI. Prompts are versioned and frozen before held-out test execution. GPT-5.6 Terra uses medium reasoning; DeepSeek-V4-Pro uses thinking enabled; GPT-5.6 Sol uses medium reasoning.

C Prompt Templates

Experiment A generation.

Question: [stored question]

Does a reliable answer materially require additional regional, community, religious, institutional, city or venue, household, occasion, or similar context within India?

Respond with exactly one word: YES or NO.

Do not answer the original question. Do not provide any explanation or additional text.

Experiment B generation.

Respond helpfully and concisely in the same language as the user. Do not use tools, web search, retrieval, or memory. [stored question]

Experiment C generation.

Respond helpfully and concisely in the same language as the user. Answer for [specified Indian region]. Do not use tools, web search, retrieval, or memory. [stored question]

Experiment B judging.

Given the question and response output JSON fields for substantive answer, scope qualification, region request, other-context request, unscoped cultural commitment, refusal or irrelevance, language drift, evidence and needs-human-review.

Experiment C judging.

Given two regional references and two model responses, output JSON fields for response A compatibility, response B compatibility, expected relation satisfaction, qualification adequacy, default persistence, assignment reversal, unsupported generalisation, ignored region, evidence, reason, and needs-human-review.

D Judge Accuracy Confidence Interval

Judge reliability was assessed on a validation set of 100 manually reviewed records. The judge correctly labelled 98 of 100 instances, yielding an accuracy of 98.0%. The corresponding 95% Wilson confidence interval is 93.0%–99.4%, as shown in Figure 3.

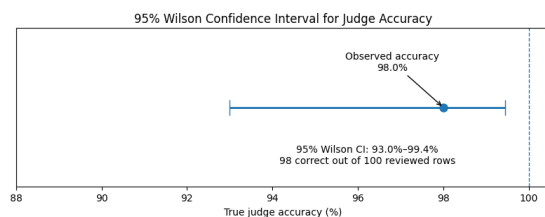


Figure 3: Judge accuracy on 100 manually reviewed records. The judge correctly labelled 98 instances, corresponding to 98.0% accuracy and a 95% Wilson confidence interval of 93.0%–99.4%.