

Common to Whom?

Regional Cultural Commonsense and LLM Bias in India

Sangmitra Madhusudan¹, Trush Shashank More², Steph Buongiorno¹,
Renata Dividino³, Jad Kabbara⁴ and Ali Emami¹

¹Emory University ²Independent Researcher ³Brock University ⁴MIT
{smadhus, aemami}@emory.edu

Abstract

Existing cultural commonsense benchmarks treat nations as *monolithic*, assuming uniform practices within national boundaries. But does cultural commonsense hold uniformly within a nation, or does it vary at the sub-national level? We introduce **INDICA**, the first benchmark designed to test LLMs’ ability to address this question, focusing on India—a nation of 28 states, 8 union territories, and 22 official languages. We collect human-annotated answers from five Indian regions (North, South, East, West, and Central) across 515 questions spanning 8 domains of everyday life, yielding 1,630 region-specific question-answer pairs. Strikingly, only 39.4% of questions elicit agreement across all five regions, demonstrating that cultural commonsense in India is predominantly *regional*, not national. We evaluate eight state-of-the-art LLMs and find two critical gaps: models achieve only 13.4%–20.9% accuracy on region-specific questions, and they exhibit geographic bias, over-selecting Central and North India as the “default” (selected 30-40% more often than expected) while under-representing East and West. Beyond India, our methodology provides a generalizable framework for evaluating cultural commonsense in any culturally heterogeneous nation, from question design grounded in anthropological taxonomy, to regional data collection, to bias measurement.¹

1 Introduction

Commonsense reasoning—the ability to understand everyday knowledge shared by humans—has been studied extensively to assess whether language models possess such understanding (Sakaguchi et al., 2021; Li et al., 2022; Talmor et al., 2019). A key challenge is that commonsense knowledge is fundamentally long-tailed, with most facts rare in training data (Davis and Marcus, 2015;

¹The complete dataset and codebase are publicly available on [GitHub](#) and on [HuggingFace](#).

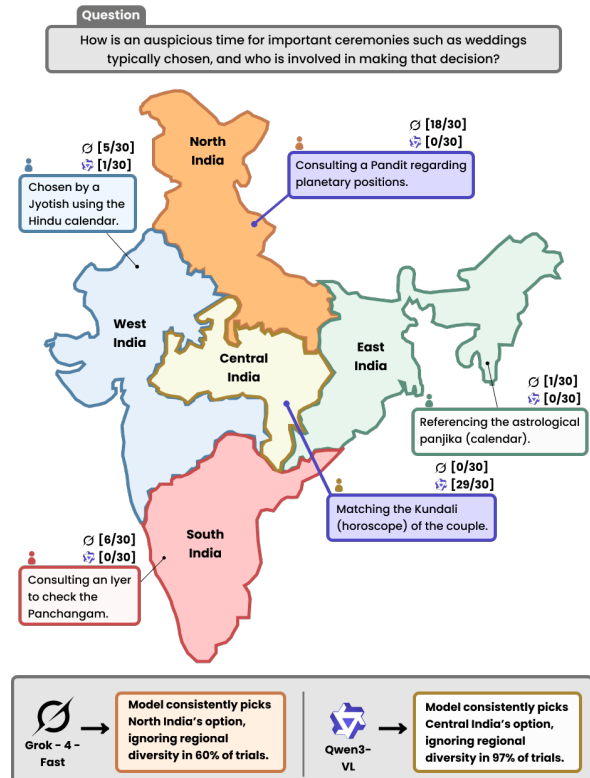


Figure 1: Regional answers to a cultural question and model bias. Each region gives a different answer; models default to Central and North India.

Do et al., 2024). This motivated scaling training data to help models internalize rare facts (Brown et al., 2020; Kandpal et al., 2023). For genuinely universal knowledge—such as physical commonsense (e.g., “objects fall when dropped”)—this strategy has proven effective (Bisk et al., 2020). However, this raises a critical question: how does this approach fare for commonsense knowledge that is *not* universal but rather cultural?

Consider questions such as: “Which side of the road do you drive on?” or “What is the traditional color of a wedding dress?” These questions have no single correct answer; they vary by country and/or culture. This reveals a fundamental limitation: simply scaling data may not resolve dis-

agreement rooted in cultural diversity. To address this, researchers have proposed the notion of *cultural commonsense*, or knowledge that is widely shared within a culture yet differs across cultural contexts. (Shen et al., 2024; Acquaye et al., 2024). Recent benchmarks like CultureBank (Shi et al., 2024) and CulturalBench (Chiu et al., 2025) have begun addressing this gap. However, these efforts share a critical assumption: they treat entire countries as culturally uniform, as if all citizens of a nation share the same practices and norms.

This assumption breaks down in culturally heterogeneous nations, where diversity within a single country challenges the very notion of shared cultural commonsense. India exemplifies this as a nation of 28 states, 8 union territories, and 22 official languages (con, 1950). Yet existing benchmarks on India focus solely on factual knowledge from textbooks and examinations (Verma et al., 2025; Maji et al., 2025; Rohera et al., 2024), treating Indian culture as monolithic. No benchmark examines whether cultural commonsense in India is nationally shared or regionally specific.

Is cultural commonsense in India actually uniform, or does it vary by region? We introduce **INDICA**, the first benchmark designed to answer this question. We collect human-annotated answers from five Indian regions (North, South, East, West, and Central) across 515 questions spanning 8 domains of everyday life, yielding 1,630 region-specific question-answer pairs. Our findings reveal that only 39.4% of questions achieve consensus across all regions (Figure 1), confirming that cultural commonsense in India is predominantly regional, not national. This finding carries implications for any culturally diverse nation, and our methodology provides a generalizable framework for examining sub-national cultural variation, from anthropologically-grounded question design to regional data collection to bias measurement.

We evaluate eight state-of-the-art LLMs and find two critical gaps. First, models achieve only 13.4%–20.9% accuracy, capturing broad cultural concepts but lacking region-specific knowledge (§5.1). Second, when geographic context is removed, all models exhibit implicit geographic bias, over-selecting Central and North Indian answers as the “default” (30–40% more often than expected) while under-representing East and West, as illustrated in Figure 1. Cultural commonsense within diverse nations cannot be assumed uniform; it must be modeled and tested regionally.

2 INDIAN Cultural commonsense Inventory with Cross-regional Answers (INDICA)

INDICA is a benchmark for evaluating regional variation in cultural commonsense within India. Its creation involves three phases: (1) question creation grounded in anthropological taxonomy (§2.1), (2) response collection from participants across five Indian regions (§2.2), and (3) gold standard establishment through intra-region consensus, inter-region agreement, and universal agreement analysis (§2.3). Figure 2 illustrates the complete pipeline.

2.1 Question Creation

Question creation involves three stages: domain selection, topic generation, and question writing.

Stage 1: Domain Selection. To ensure principled coverage, we ground our domain selection in the Outline of Cultural Materials (OCM) (Murdock et al., 2008), an established anthropological taxonomy organizing cultural knowledge into 90+ major categories and 700+ subcategories, widely used in cross-cultural research (Wutich et al., 2014; Van de Vijver and Leung, 1997).

We select 8 domains relevant to everyday cultural knowledge—Interpersonal Relations, Education, Clothing and Adornment, Food Processing and Consumption, Communication, Finance, Festivals and Rituals, and Traffic and Transport Behavior, aligning with recent cultural NLP work (Shi et al., 2024; Chiu et al., 2025). Within each domain, we select OCM subcategories based on three criteria: (1) sufficient diversity to support multiple topics, (2) non-overlapping practices, and (3) everyday rather than institutional knowledge, yielding 18 subcategories across 8 domains. Appendix A.1.2 provides a worked example illustrating how these criteria were applied within the Education domain. Complete domain-to-OCM mappings appear in Appendix Table 5.

Stage 2: Topic Generation. For each subcategory, we use GPT-4-0613 to generate 8–10 specific cultural topics using OCM subcategory definitions as context (prompts in Appendix A.1.3). We then manually select 2–4 topics per subcategory based on four criteria: (1) ability to support at least 15 distinct questions, (2) clear answerable scope, (3) minimal overlap with other topics, and (4) focus on everyday rather than institutional knowledge. This process yields 39 final topics across 18 subcate-

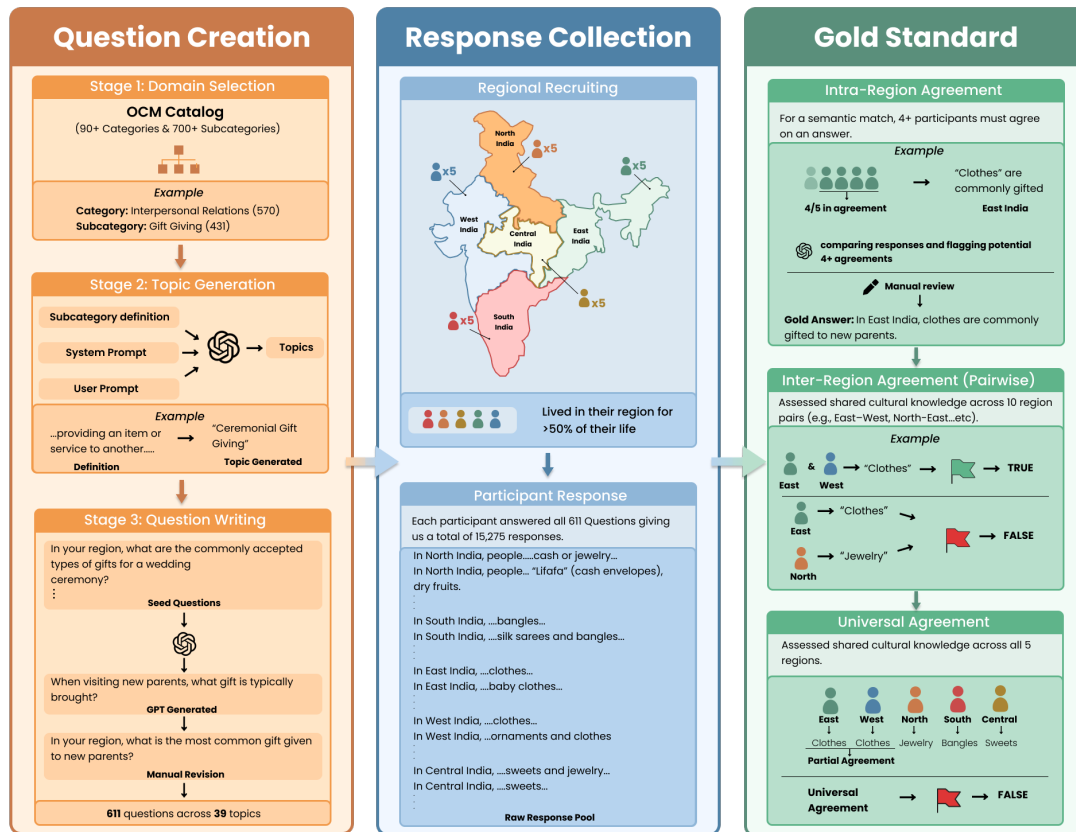


Figure 2: The INDICA creation pipeline: from domain selection to gold standard establishment

gories. Appendix A.1.4 details all generated and selected topics.

Stage 3: Question Writing. For each topic, we manually crafted 3–8 seed questions demonstrating the desired style: open-ended, culturally grounded, and focused on everyday practices. We use GPT-4-0613 with these seeds to generate additional questions, targeting 15+ per topic (Appendix A.1.5). All questions underwent manual review against three criteria: (1) removing ambiguity—e.g., clarifying "When visiting new parents, what gift is typically brought?" to "What is the most common gift given to new parents?"; (2) eliminating semantic redundancy within topics; and (3) ensuring cultural grounding by excluding questions about institutional facts (e.g., government policies). This process yielded 611 unique questions across 39 topics. Appendix A.1.6 contains the seed questions used for generation.

2.2 Response Collection

We collect responses from participants across five Indian regions (North, South, East, West, Central), following regional groupings commonly used in prior large-scale Indian studies (Patidar and Dhi-

man, 2021; Sinha et al., 2025).

We recruited 5 participants per region through Prolific², requiring each to have lived in their region for the majority of their life. Each answered all 611 questions (5 responses per question per region; 15,275 total). The study was IRB-approved with fair-wage compensation. Complete study details (including participant criteria and survey interface) appear in Appendix A.2.

2.3 Gold Standard Establishment

From 15,275 responses, we establish gold standards through automated assistance and complete manual review. GPT-4o provided initial agreement assessments; two independent annotators then manually reviewed every question using a custom annotation tool displaying GPT-4o’s preliminary classification alongside all 5 raw participant responses, verifying each assessment against the raw data. The process involves three levels: intra-region, inter-region, and universal agreement.

²<https://www.prolific.com/>

Domain	Question	North	South	East	West	Central
Interpersonal Relations	What is the most common gift given to new parents?	Jewelry	Bangles	Clothes	Clothes	Jewelry & sweets
Education	Are certain types of extracurricular activities commonly expected or required for university or college admission? If so, which ones?	N/A	N/A	Not expected	N/A	N/A
Clothing & Adornment	What fabrics are seen as auspicious for celebration wear?	Silk	Silk & cotton	N/A	Silk	Silk
Food Processing & Consumption	How is oil traditionally extracted from seeds or nuts for home use?	Crushing/ pressing	N/A	N/A	Crushing/ pressing	N/A
Communication	How do people commonly signal they are full after a meal without speaking?	Touch/pat stomach	N/A	Touch/pat stomach	Touch/pat stomach	N/A
Finance	During which times of year are discounts most commonly offered?	Diwali	Diwali & Pongal	Durga Puja	Diwali	Diwali
Festivals & Rituals	How do pilgrims typically travel to pilgrimage sites, and are there any traditional routes or modes of transport associated with the journey?	Trains & buses	On foot/ buses or cars	Public/ private transport or on foot	Public/ private transport or on foot	Trains & buses
Traffic & Transport Behavior	Are tinted windows allowed on vehicles, and what types of vehicles most commonly have them?	Not allowed	Not allowed	Not allowed	Not allowed	Not allowed

Table 1: Questions across eight domains showing regional variation. Green highlighting (or Green highlighting) indicates regions that agree with each other on that question. N/A entries indicate no consensus was reached in that region (fewer than 4 out of 5 participants agreed).

2.3.1 Intra-Region Agreement

For each question within a region, we require that at least 4 of 5 participants provide semantically equivalent answers. GPT-4o served as initial classifier, then two authors manually verified all cases using a custom annotation tool³. The annotation task was a meta-annotation: reviewing whether 4+ participants provided semantically equivalent answers, rather than making subjective cultural judgments. Inter-annotator agreement was perfect (Fleiss’ $\kappa = 1.0$) between two independent annotators, indicating clear consensus on agreement criteria.⁴ Prompting details appear in Appendix A.4.

Questions with 4+ agreeing participants received gold answers; others were marked “N/A” for that region. Of the 611 original questions, 515 (84.3%) achieved agreement in at least one region and were retained in the final dataset, yielding 1,630 question-answer pairs across all five regions.

2.3.2 Inter-Region Agreement

Beyond individual regions, we analyze whether pairs of regions shared cultural knowledge. For each question and each of the 10 possible region pairs (e.g., North-South, North-East), GPT-4o assessed whether both regions had valid answers expressing similar concepts. We manually reviewed

³Our annotation tool is available for preview [here](#). It displays pre-computed classifications for human review.

⁴ κ measures human-human agreement, not with GPT-4o. Humans overrode GPT-4o in 7.6% of intra-regional, 28.9% of inter-regional, and 24.5% of universal cases (Appendix A.3).

all assessments using the same annotation tool.

We apply strict agreement criteria: two regions were marked as agreeing only if their gold standard answers reflected exactly the same cultural practice. Partial overlaps were not counted. For example, if one region answered “silk” and another answered “silk and cotton” for celebration fabrics, they were not marked as agreeing, as these represent distinct practices despite shared elements. Prompting details appear in Appendix A.5.

2.3.3 Universal Agreement

Finally, we identify questions where all five regions provide valid answers expressing the same cultural concept. GPT-4o assessed all valid answers for universal consensus, and we manually reviewed each assessment. Prompting details appear in Appendix A.6.

2.4 Dataset Characteristics

The final dataset contains 515 questions yielding 1,630 region-specific question-answer pairs across 8 domains, 18 subcategories, and 39 topics. Each question includes: gold standard answers per region (or “N/A” if no consensus was reached), pairwise agreement flags for all 10 region pairs, a universal agreement flag, and metadata (domain, subcategory, topic). Table 1 shows example questions with regional answers and agreement patterns.

2.4.1 Question Distribution

Figure 3 shows question distribution across domains, ranging from Festivals and Rituals (109)

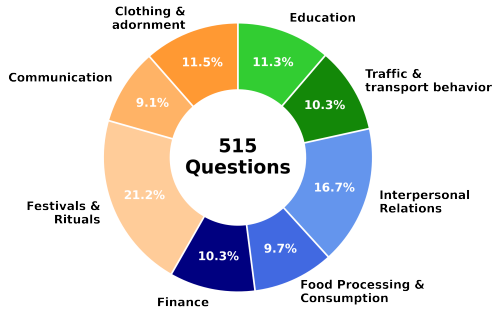


Figure 3: Distribution of 515 questions across 8 domains

to Communication (47). Appendix Table 33 provides the full breakdown by domain, subcategory, and topic.

2.4.2 Regional Coverage

Regional coverage varies across the dataset. Of the 515 questions, West India has intra-region consensus on 354 (68.7%), followed by Central (348, 67.6%), North and South (326 each, 63.3%), and East (276, 53.6%). East India’s lower coverage suggests greater internal diversity within the region.

2.4.3 Cross-Region Agreement Patterns

Pairwise agreement. Figure 4 shows agreement rates for all region pairs, calculated as the percentage of questions where both regions provided valid answers and agreed. North-Central shows the highest pairwise agreement (68.3%), likely reflecting geographic contiguity and linguistic similarities, followed by West-Central (65.0%) and North-West (63.7%). South-East shows the lowest agreement (60.1%), suggesting greater cultural distance between these regions.

Universal agreement. Of 132 questions where all five regions provide valid answers, only 52 (39.4%) have unanimous agreement, confirming cultural commonsense in India is largely regional.

Domain-level variation. Universal agreement varies greatly by domain (Table 2). Traffic & Transport Behavior shows highest agreement (22.6%), likely reflecting nationwide standardization, while Festivals & Rituals (1.8%) and Food Processing & Consumption (6.0%) show lowest, reflecting strong regional traditions. These differences are substantive, not linguistic. For example, harvest festival games yield “Jallikattu” (South India) vs “kite flying” (Central India), fundamentally different practices rather than different names for the same activ-

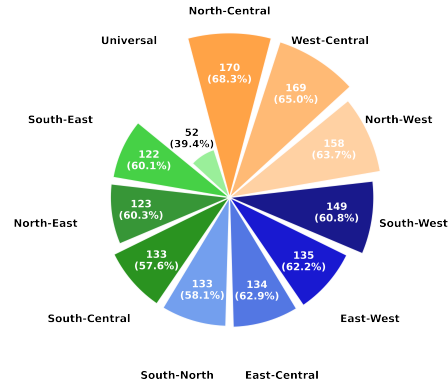


Figure 4: Pairwise and universal agreement rates between all 5 regions. Percentages calculated over questions where both regions provided responses.

Domain	Universal Agreement
Traffic & Transport Behavior	22.6%
Education	13.8%
Clothing & Adornment	13.6%
Communication	12.8%
Interpersonal Relations	10.5%
Finance	7.5%
Food Processing & Consumption	6.0%
Festivals & Rituals	1.8%

Table 2: Universal agreement rates by domain

ity. Even Education achieves only 13.8% despite national curricula, showing regional practices persist in standardized domains.

3 Model Evaluation

We evaluate LLMs on INDICA to answer two questions: (1) Can models generate accurate region-specific cultural knowledge? (2) Do models exhibit implicit geographic bias, favoring certain regions as representative of “Indian culture”? To address these distinct questions, we design two complementary evaluation tasks. Following best practices for MCQ evaluation (Balepur et al., 2025), we use multiple runs with randomized option ordering to ensure robust measurement.

3.1 Region-Anchored Short Answer (RASA)

Purpose. RASA tests whether models can generate accurate region-specific cultural knowledge when given geographic context. Unlike multiple-choice formats, RASA requires free-form generation, testing whether models can produce cultural knowledge rather than merely recognize it.

Construction. For each question where at least one region has a gold standard answer, we create region-specific variants by prepending the region

identifier. For example, “What is the most common gift given to new parents?” becomes “In South India, what is the most common gift given to new parents?” This yields 1,630 region-anchored questions. Appendix Table 34 shows the regional distribution.

Scoring. We use Gemini 3.0 Flash⁵ as an LLM judge to evaluate model responses against gold standard answers.⁶ Each question is run n times to account for response variability, and we compute the average score. Responses are scored as:

- **Correct (1.0):** Response captures the same cultural practice as the gold answer with no significant omissions or additions.
- **Partially Correct (0.5):** Response contains core elements but misses key details or includes extraneous information.
- **Incorrect (0.0):** Response is inconsistent with the gold answer.

We weight partial credit at $w = 0.5$ as a balanced choice. Results are robust to this weighting: varying $w \in \{0.3, 0.5, 0.7\}$ maintains tight model clustering (3–4 percentage points) at each weight (Appendix Table 35).

Appendix Table 36 provides scoring examples and Appendix Section A.8.1 contains LLM judge prompting details.

3.2 Region-Agnostic Multiple Choice Questions (RA-MCQ)

Purpose. RA-MCQ reveals models’ implicit biases by observing which regional practices models select when geographic context is absent. When models must choose between options representing different regions without knowing which region each option corresponds to, their selection patterns reveal which regions’ practices they treat as the “default” for India.

Construction. For questions where three or more regions provided distinct consensus answers, we construct MCQs without regional conditioning. Each option represents one or more regions’ consensus answer:

⁵Gemini 3.0 Flash selected for cost-efficient evaluation at scale (390K+ judgments, 8 models \times 1,630 q’s \times 30 runs)

⁶Validated on 200 responses (100 Qwen, 100 Gemini) by two independent human annotators. Inter-annotator agreement: 94%–100% (97% overall). LLM-human agreement: 87%–88%.

Model sees:

Q: What is the most common gift given to new parents?

Options: A) Jewelry, B) Bangles, C) Clothes, D) Sweets

Example Model answer: A) Jewelry

Our annotation (hidden from model):

A \rightarrow North, B \rightarrow South, C \rightarrow {East, West}, D \rightarrow Central

Credit: North India = 1.0

This yielded 79 RA-MCQ questions. Appendix Table 37 shows the distribution across domains.

Scoring. Each question is evaluated n times with randomized option ordering. We calculate each region’s selection rate as the proportion of times that region’s answer was chosen when available. When an option represents multiple regions, credit is split equally. Under unbiased selection, each region should be selected approximately 20% of the time. We use a chi-square goodness-of-fit test to assess statistical significance, with expected counts accounting for regional availability and varying option counts (details in Appendix A.9.2).

4 Experimental Setup

Models. We evaluate eight state-of-the-art LLMs spanning open and closed-source models across diverse families. Closed-source models include Claude Sonnet 4.5 (Anthropic, 2025), Gemini 3 Flash (DeepMind, 2025), GPT-5.2 (OpenAI, 2025), and Grok-4 Fast (xAI, 2025). Open-source models include DeepSeek-V3.2 (DeepSeek-AI, 2025), Llama 3.3 70B (AI, 2024), Mistral Large 3 (AI, 2025), and Qwen3-VL (Team, 2025).

Evaluation Settings. For both RASA and RA-MCQ, we run each question $n = 30$ times to account for response variability and, for RA-MCQ, to enable randomized option ordering across runs. All models are evaluated with temperature 1.0 to capture the full distribution of model responses. Complete prompts are in Appendix A.10.

Metrics. For RASA, we use three metrics: Fully Correct (response matches the gold answer’s cultural practice with no significant omissions or additions), Partially Correct (response contains core elements but misses details or includes extraneous information), and Overall Accuracy (Fully Correct + 0.5 \times Partially Correct). For RA-MCQ, we report regional selection rates and assess bias using a chi-square goodness-of-fit test against the expected uniform distribution of approx 20% per region.

Model	Full	Partial	Incorr.	Overall
Grok-4 Fast	19.6	65.8	14.5	52.6
GPT-5.2	18.6	67.6	13.8	52.4
Claude Sonnet 4.5	14.1	75.3	10.6	51.7
Qwen3 VL	20.9	61.3	17.8	51.6
Gemini 3.0 Flash	13.4	75.3	11.3	51.1
Mistral Large	16.2	68.4	15.4	50.4
Llama 3.3 70B	18.2	64.2	17.6	50.3
DeepSeek V3.2	14.9	69.3	15.8	49.5

Table 3: Model performance on RASA (%). **Green** = top tercile, **yellow** = middle, **orange** = bottom tercile. **Bold** = best in column. For “Incorrect”, lower is better.

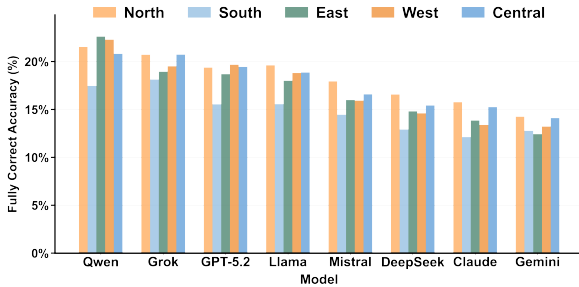


Figure 5: Fully correct accuracy by region on RASA

5 Results

We present results for both evaluation tasks: RASA (§5.1), which measures region-specific cultural knowledge, and RA-MCQ (§5.2), which reveals implicit regional biases.

5.1 Models Capture Broad Cultural Concepts but Lack Regional Specificity

Models achieve overall accuracy between 49.5% and 52.6%, tightly clustered within 3.1 percentage points (Table 3), indicating comparable cultural knowledge across models. However, fully correct rates remain low across all models (13.4%–20.9%), with the majority of responses (61.3%–75.3%) being only partially correct. This pattern indicates that models capture broad cultural concepts but either add extraneous information or omit region-specific details, demonstrating an inability to generate precise cultural knowledge.

An analysis on 100 partially correct responses from Claude Sonnet 4.5 reveals overexplaining (89%) as the dominant pattern over underspecifying (1%) or both (10%), suggesting models elaborate from a generic cultural template rather than drawing on region-specific knowledge. For example, when asked about color avoidance in West India, the gold answer is “avoid black during auspicious holidays,” but models add extraneous information about Amavasya (new moon day), mourning periods, and specific weekdays, burying re-

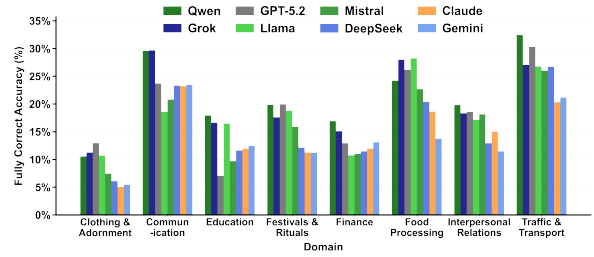


Figure 6: Fully correct accuracy by domain on RASA

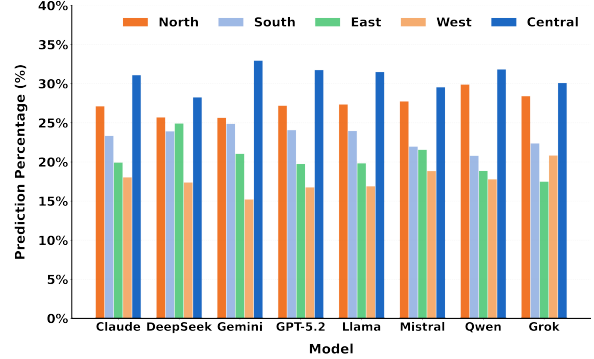


Figure 7: Regional selection rates on RA-MCQs

gional precision under generalized cultural noise (Appendix Table 38).

Minimal regional variation. Model performance remains remarkably uniform across regions (Figure 5). While North (14.3%–21.5%) and Central (13.4%–20.9%) India receive marginally higher fully correct rates, regional differences remain small at 3–5 percentage points. Overall accuracy reveals even tighter clustering at 49–54% across all regions (Figure 8). This uniformity further suggests superficial cultural knowledge everywhere rather than balanced representation.

Domain-level performance. Model accuracy varies across cultural domains (Figure 6). Models achieve highest fully correct rates in Traffic & Transport (20.3–32.4%) and Communication (18.6–29.7%), while struggling with Clothing & Adornment (5%–12.9%) and Finance (10.7%–16.9%). When examining overall accuracy (Appendix Figure 9), domain differences compress (from 17.7 to 8.5 percentage points), with most domains converging to 48–56% accuracy, indicating models possess fragmentary knowledge across all cultural areas.

5.2 Models Default to Central and North Indian Cultural Practices

Figure 7 shows selection patterns on RA-MCQs. Under uniform random selection, each region would be selected approximately 20% of the time. All models deviate significantly from this baseline (chi-square goodness-of-fit, $p < 0.001$, Appendix

Table 39), revealing systematic regional biases.

Over-selection of Central and North. All models consistently over-select Central India (24.7%–28.8%) and North India (22.4%–26.1%), with selection ratios of 1.25–1.46× and 1.14–1.32× expected rates, respectively (Appendix Table 40). Standardized residuals exceed +2.0 in all cases. Central India shows the strongest over-selection: Gemini (28.8%, 1.46× expected), Qwen (27.8%, 1.41× expected), and GPT-5.2 (27.8%, 1.40× expected) select it most frequently. This indicates that when geographic context is absent, models default to Central and North Indian cultural practices as representative of “Indian culture.”

Why do models default to Central and North India? We identify three reinforcing factors. First, Hindi, spoken primarily in North and Central states, dominates Indian-language web content. Kakwani et al. (2020) found Hindi has billions of tokens in Common Crawl, while South and East Indian languages such as Assamese, Odia, and Kannada have 10–100× less data. Second, tokenization compounds this gap: Rust et al. (2021) showed that multilingual tokenizers disproportionately fragment underrepresented languages into less meaningful subword units, degrading downstream performance even on the limited South/East content that exists. Third, Hindi-language media and Bollywood, centered in North/Central India, dominate cultural representation of India both domestically and internationally, reinforcing these regions as the perceived default in training data. These factors combine to produce a training signal that systematically treats North/Central practices as representative of Indian culture.

Conversely, most models under-select West India (12.9%–17.7%, 0.73× expected) and East India (13.3%–18.9%, 0.82× expected), with standardized residuals below -2.0. South India shows variable patterns (16.6%–19.9%, 0.88× expected). See Appendix A.11.1 for detailed analyses.

6 Generalizing Beyond India

While INDICA focuses on India, our framework transfers directly to any culturally heterogeneous nation. We illustrate with China—a nation of 34 provincial-level divisions, 56 recognized ethnic groups, and 130+ languages—where the assumption of cultural uniformity is equally untenable.

Benchmark	Regional Variation	Commonsense (not factual)	Everyday Practices
MILU	✗	✗	✗
SANSKRITI	✗	✗	✓
IndicQuest	✗	✗	✗
DOSA	✓	✗	✓
IndiBias	✓	✓	✗
FairTales	✓	✓	✗
INDICA	✓	✓	✓

Table 4: Comparison with Indian cultural benchmarks.

Question Creation. Researchers can adopt our 8 OCM-grounded domains directly. For China, the same domains yield culturally distinct questions: Food Processing and Consumption (regional staples: Cantonese dim sum vs. Sichuan hotpot vs. Northeastern stews), Festivals and Rituals (Cantonese lion dance traditions vs. Northern temple fairs vs. Southwestern torch festivals), and Clothing and Adornment (traditional cheongsam vs. cotton padded jackets vs. miao batik). The three-stage process (§2.1)—domain selection via OCM taxonomy, topic generation, and question writing—requires only content adaptation, not methodological changes.

Response Collection. China’s established statistical-regions (North, Northeast, East, South Central, Southwest, Northwest) provide natural divisions used in administrative and social research. The protocol transfers directly: (1) recruit 5+ participants per region with majority-of-life residency, (2) collect responses to all questions per participant, (3) ensure fair-wage compensation through platforms supporting Chinese participants.

Gold Standard Establishment. Apply identical consensus thresholds (§2.3): 4/5 intra-regional agreement establishes gold answers, then assess inter-regional agreement across all 15 region pairs, and finally universal agreement across all 6 regions.

Evaluation Tasks. RASA questions become region-anchored: “In Southwest China, what is the traditional gift when visiting someone’s home for the first time?” RA-MCQ reveals bias by presenting options representing different regions’ practices without labels, measuring whether models default to specific regional practices as “Chinese culture.”

Bias Measurement. Chi-square tests against uniform selection ($\approx 16.7\%$ per region) detect geographic bias. Given Eastern China’s higher population density and economical status, models might

over-select eastern practices, parallel to our finding of Central/North India bias.

7 Related Work

Commonsense Reasoning Benchmarks. Commonsense reasoning has been studied extensively through knowledge bases like ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019a), operationalized into benchmarks such as CommonsenseQA (Talmor et al., 2019), SocialIQA (Sap et al., 2019b), PIQA (Bisk et al., 2020), and Winogrande (Sakaguchi et al., 2021). These resources successfully test universally-held reasoning, such as physical commonsense. However, they treat culturally-dependent knowledge as universal truth, encoding answers that vary by culture as singular facts and reflecting the predominantly Western backgrounds of their annotators (Sap et al., 2019b).

Cultural Commonsense. Recent work has expanded commonsense evaluation beyond Western contexts. GeoMLAMA (Yin et al., 2022) probes geo-diverse knowledge across countries, CANDLE (Nguyen et al., 2023) extracts cultural commonsense at scale, CultureBank (Shi et al., 2024) catalogs practices across 120+ cultural groups, CulturalBench (Chiu et al., 2025) evaluates knowledge through questions about customs, FORK (Palta and Rudinger, 2023) tests food-related cultural knowledge, and NORMAD (Rao et al., 2025) measures reasoning about culturally-dependent social norms. Concurrent work by Naous et al. (2025) benchmarks cultural biases across Asian languages. However, these benchmarks treat culture at the *national* level, representing countries as monolithic entities and collapsing within-country diversity into national stereotypes.

Indian Cultural Knowledge. Work on Indian cultural knowledge has focused primarily on factual evaluation. MILU (Verma et al., 2025) and IndicMMLU-Pro (Sankalp et al., 2025) evaluate multi-task knowledge in Indic languages, SAN-SKRITI (Maji et al., 2025) tests knowledge across 16 cultural attributes, IndicQuest (Rohera et al., 2024) evaluates regional knowledge in 19 languages, and DOSA (Seth et al., 2024) tests familiarity with social artifacts from 19 subcultures. On social biases, IndiBias (Sahoo et al., 2024) and FairI Tales (Nawale et al., 2025) measure biases across identity dimensions, while Shankar et al. (2025) show that pan-Asian LLM alignment ob-

scures regional diversity, and Mukhopadhyay et al. (2025) propose mitigation techniques aligned with Indian constitutional values. These works focus on factual knowledge or bias detection and treat India as uniform (Table 4); INDICA addresses both gaps through regionally annotated commonsense.

8 Conclusion

Cultural commonsense is not national—it is *regional*. INDICA provides the first empirical evidence for this claim, demonstrating that only 39.4% of questions achieve consensus across India’s five regions. LLMs fail to capture this diversity: they lack region-specific knowledge and default to Central and North Indian practices when context is absent. Our methodology generalizes beyond India. Any nation with sub-national cultural diversity—Indonesia, Nigeria, Brazil, China—faces this challenge. We release INDICA as both a benchmark and a blueprint: the data to evaluate, and the framework to replicate. Culturally competent AI cannot treat nations as monoliths. It must model diversity where diversity exists. This work is a first step toward greater granularity; finer-grained cultural analysis remains future work.

Limitations

Geographic Scope and Generalizability: INDICA focuses on India as a case study. While the *data* (1,630 region-specific question-answer pairs) reflects Indian contexts, the *framework* generalizes: OCM-grounded question creation, regional response collection, gold standard establishment, dual evaluation tasks (RASA and RA-MCQ), and statistical bias measurement transfer to any culturally heterogeneous nation (Section 6).

Regional Granularity: Our five-region division (North, South, East, West, Central) captures major geographic and cultural boundaries but necessarily aggregates internal diversity. For instance, South India encompasses Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, Telangana, Puducherry, Lakshadweep, Andaman and Nicobar Islands, which have distinct languages (Malayalam, Tamil, Kannada, Telugu, etc.), cuisines, and festival traditions. Analysis of intra-regional annotator agreement reveals varying levels of internal consensus (Sec. 2.4.2), with some regions showing higher unanimity than others, indicating that cultural variation exists at multiple scales. Finer-grained analysis

was constrained by the feasibility of recruiting sufficient annotators per region and ensuring adequate domain coverage within budget.

Importantly, our findings conclusively demonstrate that treating India as culturally uniform is empirically invalid. The question is not whether sub-national variation exists, but at what scales it manifests most strongly.

Participant Demographics: Our participant pool was recruited through Prolific, which may skew toward English-speaking, digitally connected Indians; cultural practices among rural or non-English-speaking populations may differ. Future work should expand sampling to include rural participants.

Temporal Validity: Cultural practices evolve over time. INDICA represents a snapshot of contemporary everyday knowledge as reported by participants in 2025.

Domain Coverage: While our 8 domains provide initial coverage of major cultural dimensions and are grounded in the OCM anthropological taxonomy, culture is multifaceted and additional domains would reveal further regional variation. The OCM includes other relevant categories such as Marriage (OCM 580), Family (OCM 590), Law (OCM 670), and Living Standards and Routines (OCM 510) that could enrich cultural analysis. We prioritized domains that reflect cultural commonsense, shared everyday knowledge learned through participation, rather than personal experiences or institutional facts.

Demographic Stratification: Our analysis examines variation along geographic lines, but cultural practices may also vary along demographic dimensions such as religion, caste, gender, and urban/rural residence within a single region. Future work should incorporate these demographic variables.

Ethical Considerations

This study was approved by our institutional research ethics board. All participants provided informed consent, were compensated at fair wage rates following Prolific guidelines, and could withdraw at any time. We collect cultural practices as reported by participants, not objective truths; regional answers reflect shared knowledge within our sample rather than authoritative claims about entire

populations. We acknowledge that any regional categorization risks oversimplification, and we do not intend our five-region framework to reify or essentialize cultural boundaries. Our goal is to reveal diversity that current benchmarks erase, not to replace one form of stereotyping with another. The dataset will be released for research purposes with documentation encouraging responsible use.

References

1950. [The constitution of India](#). Eighth Schedule.
- Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. [Susu box or piggy bank: Assessing cultural commonsense knowledge between ghana and the US](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9483–9502, Miami, Florida, USA. Association for Computational Linguistics.
- Meta AI. 2024. [Llama 3.3: Advancing state-of-the-art in open foundation models](#). Technical report, Meta.
- Mistral AI. 2025. [Mistral large 3 \(2512\): A new benchmark for open-weight generalists](#). Technical report, Mistral AI.
- Anthropic. 2025. [Claude sonnet 4.5 technical overview](#). Technical report, Anthropic.
- Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. [Which of these best describes multiple choice evaluation with LLMs? a\) forced B\) flawed C\) fixable D\) all of the above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- 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.
- Ernest Davis and Gary Marcus. 2015. [Commonsense reasoning and commonsense knowledge in artificial intelligence](#). *Communications of the ACM*, 58(9):92–103.
- Google DeepMind. 2025. [Gemini 3 flash: Frontier intelligence at scale](#). Technical report, Google.
- DeepSeek-AI. 2025. [Deepseek-v3.2: Integrating sparse attention with enhanced reasoning](#). Technical report, DeepSeek.
- Quyet V. Do, Junze Li, Tung-Duong Vuong, Zhaowei Wang, Yangqiu Song, and Xiaojuan Ma. 2024. [What really is commonsense knowledge?](#) *Preprint*, arXiv:2411.03964.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 15696–15707. PMLR.
- Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. [A systematic investigation of commonsense knowledge in large language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Aniket Maji and 1 others. 2025. [Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture](#). *Preprint*, arXiv:2506.15355.
- Snehasis Mukhopadhyay, Aryan Kasat, Shivam Dubey, Rahul Karthikeyan, Dhruv Sood, Vinija Jain, Aman Chadha, and Amitava Das. 2025. [Ambedkar-a multi-level bias elimination through a decoding approach with knowledge augmentation for robust constitutional alignment of language models](#). *Preprint*, arXiv:2509.02133.
- George P. Murdock, Clellan S. Ford, Alfred E. Hudson, Raymond Kennedy, Leo W. Simmons, and John W. M. Whiting. 2008. *Outline of Cultural Materials*, 6th revised edition with modifications edition. Human Relations Area Files, New Haven.
- Tarek Naous, Anagha Savit, Carlos Rafael Catalan, Geyang Guo, Jaehyeok Lee, Kyungdon Lee, Lheane Marie Dizon, Mengyu Ye, Neel Kothari, Sahajpreet Singh, Sarah Masud, Tanish Patwa, Trung Thanh Tran, Zohaib Khan, Alan Ritter, JinYeong Bak, Keisuke Sakaguchi, Tanmoy Chakraborty, Yuki Arase, and Wei Xu. 2025. [Camel-ia: Benchmarking cultural biases in llms for asian languages](#). *Preprint*, arXiv:2510.05291.
- Janki Atul Nawale, Mohammed Safi Ur Rahman Khan, Janani D, Mansi Gupta, Danish Pruthi, and Mitesh M Khapra. 2025. [FairI tales: Evaluation of fairness in Indian contexts with a focus on bias and stereotypes](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30331–30380, Vienna, Austria. Association for Computational Linguistics.
- Tuan-Phong Nguyen and 1 others. 2023. Extracting cultural commonsense knowledge at scale. In *WWW*.
- OpenAI. 2025. [Gpt-5.2 technical report](#). Technical report, OpenAI.
- Shramay Palta and Rachel Rudinger. 2023. [FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.
- Gopal K. Patidar and Yashaswi Dhiman. 2021. [Distribution of abo and rh \(d\) blood groups in india: A systematic review](#). *ISBT Science Series*, 16(1):37–48.
- 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 Annual Conference of the Nations of the Americas Chapter of the ACL*. Association for Computational Linguistics.
- Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. 2024. [L3cube-indicquest: A benchmark question answering dataset for evaluating knowledge of llms in indic context](#). *Preprint*, arXiv:2409.08706.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. [How good is your tokenizer? on the monolingual performance of multilingual language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.
- Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. [IndiBias: A benchmark dataset to measure social biases in language models for Indian context](#). 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 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [Winogrande: An adversarial Winograd schema challenge at scale](#). In *Communications of the ACM*, volume 64, pages 99–106. ACM New York, NY, USA.
- K. J. Sankalp, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. 2025. [IndicMMLU-Pro: Benchmarking Indic large language models on multi-task language understanding](#). *Preprint*, arXiv:2501.15747.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Socialliqa: Commonsense reasoning about social interactions. In *EMNLP*.
- 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, Italy. ELRA and ICCL.
- Hari Shankar, Vedanta S P, Tejas Cavale, Ponnurangam Kumaraguru, and Abhijnan Chakraborty. 2025. [Sometimes the model doth preach: Quantifying religious bias in open llms through demographic analysis in asian nations](#). *Preprint*, arXiv:2503.07510.
- 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.
- Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério 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.
- Urmila Sinha, Shiv Kumar Mudgal, Ajay Kumar Patel, Vipin Patidar, and Sanjay Kumar. 2025. Mapping the burden prevalence of neural tube defects across indian regions: a systematic review and meta-analysis. *The Pan African Medical Journal*, 52:54.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*.
- Qwen Team. 2025. [Qwen3-vl technical report](#). *arXiv preprint arXiv:2511.21631*.
- Fons J. R. Van de Vijver and Kwok Leung. 1997. *Methods and Data Analysis for Cross-Cultural Research*. SAGE Publications, Thousand Oaks, CA.
- Sshubam Verma and 1 others. 2025. Milu: A multi-task indic language understanding benchmark. In *NAACL*.
- Amber Wutich, Gery Ryan, and H. Russell Bernard. 2014. Text analysis. In H. Russell Bernard, editor, *Handbook of Methods in Cultural Anthropology*, 2nd edition, pages 155–188. Rowman & Littlefield, Lanham, MD.
- xAI. 2025. [Grok-4 fast: Unified reasoning and inference](#). Technical report, xAI.
- Da Yin, Jack Hessel, and 1 others. 2022. Geomlma: Geo-diverse commonsense probing on multilingual pre-trained language models. In *EMNLP*.

A Appendix

A.1 Dataset

A.1.1 Domains, Subcategories, and Topics

Domain	OCM Code	Subcategory	Pattern	Topics
Interpersonal Relations	570	Visiting and hospitality (574)	Cross	Etiquette in the reception of visitors, Occasions for visiting
		Gift giving (431)		Gift Giving Etiquette, Ceremonial gift giving
		Etiquette (576)		Greeting and Salutation Etiquette, Eating, drinking, and smoking etiquette
Festivals and Rituals	—	Rest days and holidays (527)	Cross	Conceptualization of Holidays, Secular Festival Practices, Commemoration of Personal Milestones, Religious Taboos on Holidays
		Ritual (788)		Symbolic Act Performance, Ritual Gestures, Pilgrimage Practices
		Organized ceremonial (796)		Engaging with Religious Music and Dance, Timing of Ceremonies
Traffic and Transport Behavior	—	Streets and traffic (363)	Cross	Understanding Local Traffic Regulations, Adapting to Local Transportation Modes
		Transportation (489)		Carrying Capacity of Transport, Transport during Special Events
Education	870	Education system (871)	Cross	Formal Educational Structure, Attitudes toward Education
		Teachers (875) & Students (877)		Norms for Interacting with Teachers, Student Extracurricular Activities
Clothing and Adornment	290 + 300	Special garments (292)	Merged	Special Occasion Clothing, Headgear and Footwear Norms
		Ornament (301)		Ornamental Attire and Status Indication, Occasions for Wearing Specific Ornaments
Food Processing and Consumption	250 + 260	Food preparation (252)	Merged	Food Preparation Techniques, Cultural Recipes and Ingredients
		Diet (262)		Staple Food Consumption, Seasonal Diet Modifications
Communication	200	Gestures & Signs (201)	Single	Social Expression of Emotions, Non-Verbal Expression of Respect or Disrespect
		Dissemination of News (203)		Navigating the “Grapevine”, Trustworthiness of Information Sources
Finance	450	Credit (452)	Single	Negotiating Credit Advances and Discounts, Navigating Installment Buying
		Saving & Investment (454)		Safekeeping of Valuables, Preferable Investment Forms

Table 5: Complete domain hierarchy showing OCM codes, subcategories, selection patterns, and topics. Pattern indicates: *Cross* = subcategories from multiple OCM categories, *Merged* = combined entire OCM categories, *Single* = subcategories from one OCM category.

A.1.2 Subcategory and Topic Selection

Criteria: Worked Example

We illustrate our inclusion and exclusion criteria using the Education domain.

Subcategory selection. We applied three criteria: (1) sufficient diversity to support multiple topics, (2) non-overlapping practices, and (3) everyday rather than institutional knowledge. We selected “Education System (OCM 871)” because its definition encompasses educational structure, board systems, examination types, and attitudes toward education—each supporting multiple question topics. We excluded “Elementary Education (OCM 872)” as too narrow to support diverse topics, “Vocational Education (OCM 874)” for overlapping substantially with Education System, and “Educational Theory and Methods (OCM 876)” for focusing on institutional rather than everyday knowledge.

Topic selection. GPT-4 generated 8–10 candidate topics per subcategory, which we manually filtered using four criteria: ability to support 15+ questions, clear answerable scope, minimal overlap with other topics, and everyday knowledge focus. For example, “Norms for Interacting with Teachers” easily supports 15+ questions (forms of address, classroom behavior, greeting customs, etc.) and reflects everyday cultural knowledge, so it was retained. “Teachers’ Academic Freedom” was excluded because it yielded only 3–4 potential questions, and “Teacher Training and Proficiency Expectations” was excluded as institutional rather than everyday knowledge.

A.1.3 Topic Generation

All topics in the category hierarchy (Table 5) were extracted using GPT-4 with the following configuration and prompt:

Model Configuration

- **Model:** GPT-4-0613
- **Temperature:** 0.7
- **Top-p:** 1.0

System Prompt The following system prompt was used to establish the extraction framework:

You are a cultural anthropology expert. Your task is to extract concrete, culturally grounded topics

from definitions provided in the Outline of Cultural Materials (OCM), with a focus on common-sense knowledge that reflects everyday norms and expectations within a given society.

These topics will be used to evaluate whether language models possess deep, culturally situated commonsense — the type of knowledge necessary to navigate routine social life in culturally coherent ways.

Goals

Extract topics that:

- *Reflect socially shared knowledge (70%+ agreement within a cultural group)*
- *Are learned through cultural participation, not formal education*
- *Represent normative expectations, not preferences, frequencies, or trivia*
- *Are relevant to practical functioning in society — what people should know to behave appropriately in common social situations*
- *Are stable and generalizable across individuals within a cultural group*
- *Are specific enough to form the basis of a cultural commonsense question*

Output Format (Per Topic)

For each topic, return:

1. *Topic Label (3–7 words): Concise, clear, culturally grounded*
2. *Definition: A 1–2 sentence explanation of the commonsense knowledge it reflects within the society*
3. *Connection to OCM: A sentence showing how the topic derives from specific language or dimensions of the OCM subcategory*

Scope and Standards

- *Focus on cultural norms, interaction expectations, and implicit social logic that people rely on to function in their communities*
- *Avoid abstract academic categories, highly individualized behaviors, or edge cases*
- *Do not include examples or sample questions — your goal is to extract conceptual dimensions, not generate prompts*
- *Prioritize topics that carry social consequences for incorrect behavior (e.g., shame, respect, offense, admiration)*

Cultural Guidance

As you interpret the OCM definition, consider:

- *Hierarchical etiquette systems (e.g., age, gender, ritual authority)*
- *Ritualized or habitual practices around eating, greeting, clothing, interaction*
- *Moral or symbolic underpinnings of routine social expectations*
- *Everyday behavioral norms that guide what is appropriate, respectful, or inappropriate*
- *Local variation, but aim for core practices that are widely shared within the group*

User Prompt Template For each domain and subcategory, the following template was used:

Please analyze the following OCM entry and extract 8–10 culturally grounded cultural common-sense reasoning topics:

Category: [Category Name]

Subcategory: [Subcategory Name]

Definition: [OCM Definition Text]

Focus only on social knowledge that helps people function appropriately in their cultural environment.

Example Application As an illustration, for the Education category with the Students subcategory:

Category: Education

Subcategory: Education System

Definition: Degree of development and elaboration of formal education; prevalent types of educational specialization (e.g., schools, tutors, apprenticeship); source of support of teachers and educational institutions (e.g., fees from students, ecclesiastical aid, private gifts and endowments); systematization of education (e.g., local schools boards, state educational agencies, voluntary organizations of educational administrators); degree of standardization as to levels, policies, language, and curricula; primary objectives of formal education (e.g., piety, morality, citizenship, vocational skills, intellectual leadership); diffusion of education (e.g., educational statistics); attitudes toward education; etc.

A.1.4 Generated and Selected Topics

The following tables present all generated topics organized by domain and subcategory. Selected topics are marked with ✓ .

Subcategory	Generated Topics
Visiting and hospitality (574)	Norms of Reciprocal Visiting · ✓ Etiquette in Reception of Visitors · Informal vs Formal Hospitality · Offering of Food and Drink · Social Etiquette at Feasts and Parties · ✓ Frequency and Occasions for Visiting · Hospitality Duties and Expectations · Rules of Partaking in Hospitality · Significance of Visiting in Social Relationships
Gift giving (431)	✓ Gift Giving Etiquette · Role of Gifts in Social Relationships · Gift Reciprocity Expectations · ✓ Ceremonial Gift Giving · Role of Gifts in Economic Distribution · Recipient's Rights and Privileges · Donor's Rights and Privileges · Gift-Giving on Secular Holidays · Potlatch Practices · Frequency of Gift Giving
Etiquette (576)	✓ Greeting and Salutation Etiquette · Expression of Respect through Obeisances · Complimenting Appropriately · ✓ Eating, Drinking, and Smoking Etiquette · Visiting and Travel Etiquette · Deference to Status Superiors · Noblesse Oblige Responsibilities · Social Sanctions for Breaches of Etiquette

Table 6: Interpersonal Relations (570)

Subcategory	Generated Topics
Rest days and holidays (527)	Observance of Rest Days · ✓ Conceptualization of Holidays · ✓ Secular Festival Practices · ✓ Commemoration of Personal Milestones · Holiday Activity Norms · Patriotic Festival Observance · ✓ Religious Taboos on Holidays · Norms for Economic Activities on Holidays · Festive Home Visit Etiquette · Holiday Feasting Traditions
Ritual (788)	✓ Symbolic Act Performance · ✓ Ritual Gestures · Recital of Religious Formulas · Prescribed Prayer Practices · Mantra and Liturgy Recitation · Participation in Formal Processions · ✓ Pilgrimage Practices · Ritual Etiquette and Decorum · Ritual Preparation and Execution · Recognition and Understanding of Ritual Symbols
Organized ceremonial (796)	Recognizing Religious Holidays and Festivals · Observing Rest Days · Participating in Ritual Ceremonies · Understanding Ceremonial Attire and Paraphernalia · Sequencing of Ritual Rites · Interpreting Ritual Symbolism · ✓ Engaging with Religious Music and Dance · Attending Religious Services · ✓ Timing of Ceremonies · Understanding Religious Dramas and Spectacles

Table 7: Festivals and Rituals

Subcategory	Generated Topics
Streets and traffic (363)	✓ Understanding Local Traffic Regulations · Familiarity with Street Layout and Design · Comprehending Urban Traffic Composition · Awareness of Street Maintenance Practices · Acknowledging Parking Regulations · Respecting Pedestrian Rights · ✓ Adapting to Local Transportation Modes · Navigating Urban Paving and Infrastructure · Recognizing Animal Involvement in Traffic
Transportation (489)	Understanding Public Transportation Systems · Comprehending Transportation Regulations · ✓ Carrying Capacity of Transport · Navigating Traffic Volume · Etiquette on Public Transport · ✓ Transport during Special Events · Transport System Literacy · Dealing with Transport Delays · Ticketing and Fare Procedures · Environmentally Conscious Transport Choices

Table 8: Traffic and Transport Behavior

Subcategory	Generated Topics
Education system (871)	Educational Institutions · Financial Support · ✓ Formal Educational Structure · Education System Standardization · Objectives of Formal Education · ✓ Attitudes Toward Education · Educational Specialization Types · Diffusion of Education · Development of Formal Education
Teachers (875)	Social Status and Respect for Teachers · ✓ Norms for Interacting with Teachers · Understanding Teacher Selection Processes · Teacher Salaries and Livelihood · Teacher Tenure and Advancement · Teachers' Academic Freedom · Teacher Associations and Unions · Teachers' Roles Beyond the Classroom · Understanding Teachers' Qualifications and Specializations · Teacher Training and Proficiency Expectations
Students (877)	Composition of Student Body · Student Organizations and Leadership · Social Status of Students · Student Financial Support · Academic Freedom for Students · ✓ Student Extracurricular Activities · Student Living and Dining Accommodations · Student-Community Relations · Student-Faculty Relationships · Student Group Values and Behaviors

Table 9: Education

Subcategory	Generated Topics
Special garments (292)	✓ Special Occasion Clothing · Weather-Dependent Clothing · Swimwear Etiquette · ✓ Headgear and Footwear Norms · Costume Associated Statuses · Activity-Specific Attire · Methods of Wearing Garments · Clothing and Symbolic Meaning · Dress Code for Rituals · Role-Defined Attire
Ornament (301)	✓ Ornamental Attire and Status Indication · Gender Differences in Ornamental Attire · Cultural Materials for Ornament Manufacture · Age-Based Norms for Ornament Wearing · ✓ Occasions for Wearing Specific Ornaments · Attachment Modes of Different Ornaments · Types of Ornaments and Their Significance · Etiquette of Gifting Ornaments · Cultural Interpretation of Ornaments

Table 10: Clothing and Adornment (290 and 300)

Subcategory	Generated Topics
Food preparation (252)	✓ Food Preparation Techniques · Cultural Cooking Methods · Use of Food Preparation Apparatus · Utilization of Cooking Utensils · ✓ Cultural Recipes and Ingredients · Food Preparation Associated Beliefs · Practices Around Commercialized Food Preparation · Social Norms in Food Processing · Rituals in Food Preparation
Diet (262)	✓ Staple Food Consumption · ✓ Seasonal Diet Modifications · Dietary Proportions · Group-Specific Dietary Practices · Edible and Harmful Food Discrimination · Food Preferences and Avoidances · Cultural Food Taboos · Gender-Based Food Practices · Social Class and Dietary Norms · Age-Appropriate Foods

Table 11: Food Processing and Consumption (250 and 260)

Subcategory	Generated Topics
Gestures & Signs (201)	✓ Social Expression of Emotions · Affirmative and Negative Gestures · Gesture-Based Communication of Size and Shape · Directive Signs in Cultural Communication · Understanding Intertribal Sign Languages · Cultural Nuances in Gesture Interpretation · ✓ Non-Verbal Expression of Respect or Disrespect · Social Consequences of Misinterpreted Gestures · Role of Gestures in Conflict Resolution · Gestures Signifying Community Inclusion
Dissemination of News (203)	Understanding Informal Verbal Transmission · Role of Criers and Heralds · Interpreting Bulletins and Newsletters · ✓ Navigating the "Grapevine" · Etiquette of Sharing News Informally · ✓ Trustworthiness of Information Sources · Reaction to Disseminated Information · Responsibility in Information Sharing · Cultural Significance of News Dissemination

Table 12: Communication (200)

Subcategory	Generated Topics
Credit (452)	Understanding Credit Extension Practices · ✓ Negotiating Credit Advances and Discounts · Credit Ratings and Their Impact · Non-Banking Credit Institutions · ✓ Navigating Installment Buying · Role of Bills of Exchange · Managing Credit Terms and Conditions · Culturally Appropriate Credit Communication
Saving & Investment (454)	Cultural Norms Around Saving · ✓ Safekeeping of Valuables · ✓ Preferable Investment Forms · Norms Around Investment Banking · Role of Specialized Financial Personnel · Financial Regulation Awareness · Cultural Practices Around Underwriting · Norms of Using Savings Accounts · Principles of Hoarding Wealth

Table 13: Finance (450)

A.1.5 Seed Question Prompting Details

This section documents the prompt configuration used for seed question generation.

Model Configuration

- **Model:** GPT-4-0613
- **Temperature:** 0.7
- **Top-p:** 1.0

System Prompt The following system prompt was used:

You are a culturally aware commonsense reasoning assistant.

Your task is to generate culturally grounded, realistic questions that reflect everyday social norms, expectations, or interactions within a specific region.

Each question should:

- *Be relevant to the provided category, subcategory, and topic definition*
- *Reflect what someone in that culture is expected to know or understand*
- *Avoid trivia, preferences, or niche edge cases*
- *Be open ended and lack specific nouns or indicators for the region*
- *Must begin with “In your region”*
- *Take inspiration from the example questions provided for that topic*
- *Be usable in a cultural commonsense benchmark*

Do not include the answer. Focus on what a culturally competent person should ask or understand in social settings.

User Prompt Template For each topic, the following template was instantiated with specific category, subcategory, topic, and region information:

Category: [Category Name]

Subcategory: [Subcategory Name]

Topic Label: [Topic Label]

Topic Definition: [Topic Definition]

Region: [Target Region]

Question Examples for this Topic:

[List of 3-8 example questions demonstrating the style and scope]

Generate 20-25 culturally grounded commonsense questions that conform to this topic definition and [Region], like the example questions.

Only output the questions, no bullet points, no commentary.

Example Application As an illustration, for generating questions about visitor reception etiquette in India:

Category: Interpersonal Relations

Subcategory: Visiting and Hospitality

Topic Label: Etiquette in Reception of Visitors

Topic Definition: The traditional norms and behaviors associated with receiving and entertaining visitors in a culturally appropriate manner.

Region: India

Question Examples for this Topic:

1. In your region, what is the first thing you do when you enter someone’s house? Focus on actions and not greetings.
2. In your region, what is a traditional drink, aside from water, that is offered to a guest when they visit you? Be as specific as possible.
3. In your region, what special food items are made when relatives come from out of town?
4. In your region, how do you traditionally prepare your house for the arrival of guests?
5. In your region, what are the utensils used to serve meals to guests? Are there any changes made from everyday utensils or are different traditional utensils used?
6. In your region, what are the common customs or expectations when an out-of-town guest arrives, such as from an airport or train station, in terms of how they travel to your home?

A.1.6 Seed Questions for each Topic

Topic	Seed Questions
Etiquette in Reception of Visitors	<ol style="list-style-type: none"> 1. In your region, what is the first thing you do when you go to someone's house? Focus on actions and not greetings. 2. In your region, what is a traditional drink, aside from water, that is offered to a guest when they visit you? Be as specific as possible. 3. In your region, what special food items are made when relatives come from out of town? 4. In your region, how do you traditionally prepare your house for the arrival of guests? 5. In your region, what are the utensils used to serve meals to guests? Are there any changes made from everyday utensils or are different traditional utensils used? 6. In your region, what are the common customs or expectations when an out-of-town guest arrives, such as from an airport or train station, in terms of how they travel to your home?
Occasions for Visiting	<ol style="list-style-type: none"> 1. In your region what types of occasions usually prompt people to visit neighbors? 2. In your region, is it common to visit friends or relatives unannounced, or is notice expected? 3. In your region, what is the most common personal occasion for visiting relatives? 4. In your region, what is the most common festival for visiting relatives?

Table 14: Seed questions for the topics in Visiting and Hospitality subcategory

Topic	Seed Questions
Gift Giving Etiquette	<ol style="list-style-type: none"> 1. In your region, what are the most common occasions for giving gifts (e.g., festivals, weddings, birthdays, housewarmings, visits)? 2. In your region, when receiving a gift, is it generally expected to open it immediately in front of the giver, or to open it later? 3. In your region, what types of gifts are considered universally acceptable for common occasions? 4. In your region, how do cultural or religious beliefs influence the choice of gifts or the manner of giving them (e.g., avoiding certain numbers, colors)? 5. In your region, what are considered inappropriate or unlucky gifts to give someone? 6. In your region, what is the social expectation when someone receives a gift, what do they say or do?
Ceremonial Gift Giving	<ol style="list-style-type: none"> 1. In your region, what are the commonly accepted types of gifts for a wedding ceremony? 2. In your region, during a baby naming ceremony or a child's first birthday, what kind of gifts are typically given to the child or the parent? 3. In your region, when attending a religious function at someone's home, is it customary to bring a gift, and if so, what types are appropriate and most common? 4. In your region, is there a custom of giving gifts to priests or religious officiants during ceremonies, and what form do these gifts usually take? 5. In your region, are gifts during religious or spiritual events expected to be new, handmade, or of a particular material? 6. In your region, are there gifts that must not be given during certain ceremonies due to religious or cultural taboos? If so, mention the item and the occasion.

Table 15: Seed questions for the topics in Gift Giving subcategory

Topic	Seed Questions
Greeting and Salutation Etiquette	<ol style="list-style-type: none"> 1. In your region, what are the most common verbal greetings used when meeting someone for the first time, both formally and informally? 2. In your region, what type of physical gestures (like bowing, touching feet, or handshakes) accompany greetings? 3. In your region, what type of salutation is used when addressing someone of high status or authority? 4. In your region, what type of greeting is expected when entering a religious or spiritual place? 5. In your region, what are the customary ways to bid farewell to someone, both in formal and informal situations?
Eating, Drinking, and Smoking Etiquette	<ol style="list-style-type: none"> 1. In your region, what are the customary practices for beginning and ending a meal, especially in a family or formal setting? 2. In your region, where is it generally considered acceptable to smoke (e.g., designated areas, private homes), and where is it strictly prohibited or frowned upon? 3. In your region, is there any type of seating arrangement that is common during formal or family meals? 4. In your region, what type of food is considered inappropriate to refuse? 5. In your region, what type of hand (left or right) is traditionally used for eating, and why?

Table 16: Seed questions for the topics in Etiquette subcategory

Topic	Seed Questions
Conceptualization of Holidays	<ol style="list-style-type: none"> 1. In your region, what is the most anticipated and widely celebrated festival and its significance or purpose? 2. In your region, what is the most anticipated and widely celebrated patriotic holiday and its significance or purpose? 3. In your region, what are some other holidays/festivals that are celebrated and the cultural significance of them?
Secular Festival Practices	<ol style="list-style-type: none"> 1. In your region, what are the typical decor activities and practices associated with a housewarming celebration? 2. In your region, what are the typical decor activities and practices associated with harvest celebrations? 3. In your region, what are the decor typical activities and practices associated with the most widely celebrated festival? 4. In your region, what month is your harvest festival celebrated?
Commemoration of Personal Milestones	<ol style="list-style-type: none"> 1. In your region, what is the first big moment celebrated after your child's birth? 2. In your region, what are the most important birthdays in someone's life? 3. In your region, are there any cultural norms associated with the first anniversary of a couple? If any, what are they? Focus on any celebrations or cultural customs that might be performed. 4. In your region, what other personal milestones of an individual are commonly celebrated apart from anniversaries and birthdays?
Religious Taboos on Holidays	<ol style="list-style-type: none"> 1. In your region, are there any festivals or holidays that have restrictions on activities performed, for example food restrictions or action restrictions? If any, name them and also the restrictions. 2. In your region, are there any days of the week that hold certain constraints, for example food or activity related restrictions? If any, name the day and also the restriction. 3. In your region, are there any hygiene restrictions that hold during certain festivals or holidays? If yes, name the holiday and the restrictions.

Table 17: Seed questions for the topics in Rest Days and Holidays subcategory

Topic	Seed Questions
Symbolic Act Performance	<ol style="list-style-type: none"> 1. In your region, what types of lamps are lit (if any) and what is their significance? 2. In your region, what is a specific ritual that you perform often and what are the specific actions you perform to observe that ritual? 3. In your region, what is often offered to deities? Be as specific as possible. 4. In your region, what are the specific actions you perform when visiting a religious institution? 5. In your region, when receiving spiritual blessings, what physical and verbal responses are customary?
Ritual Gestures	<ol style="list-style-type: none"> 1. In your region, what is the customary gesture for showing reverence to a sacred text, and why is it performed? 2. In your region, what is the most common gesture done during a religious ritual? 3. In your region, what is the customary gesture for offering food to a deity, and how is it performed? 4. In your region, what is a gesture associated with remembrance or honor of a religion? 5. In your region, is there a gesture to show respect after touching someone or something with your feet?
Pilgrimage Practices	<ol style="list-style-type: none"> 1. In your region, what preparatory practices like fasting or special clothing precede important pilgrimages? 2. In your region, what rituals are performed immediately upon reaching a pilgrimage destination? 3. In your region, what sacred items do pilgrims carry back from their destination, and how are these used later? 4. In your region, what is the most popular pilgrimage? 5. In your region, who can undertake pilgrimages? 6. In your region, what changes in movement or attire occur during certain parts of a pilgrimage, and what might these signify?

Table 18: Seed questions for the topics in Ritual subcategory

Topic	Seed Questions
Engaging with Religious Music and Dance	<ol style="list-style-type: none"> 1. In your region, what is the traditional dance form associated with your culture or rituals? 2. In your region, what is the traditional music form associated with your culture or rituals? 3. In your region, are there any specific dances performed during specific rituals, if yes, specify the dance as well as the ritual. 4. In your region, is there specific music that is played during specific rituals, if yes, specify the music as well as the ritual. 5. In your region, who typically participates in these traditional dance forms or who is it performed by?
Timing of Ceremonies	<ol style="list-style-type: none"> 1. In your region, how is an auspicious time for important ceremonies, such as weddings, typically chosen, and who is involved in that decision? 2. In your region, how do spiritual advisors determine the right time to begin a new venture or embark on a journey? 3. In your region, how do agricultural rhythms and seasonal changes influence the timing of festivals, rituals, or community ceremonies related to the land? 4. In your region, are certain days or times avoided for house-related rituals, and what beliefs influence those choices? 5. In your region, are there certain periods during the year when major activities are paused or avoided, and what is the reasoning behind this? 6. In your region, is there a specific calendar followed that is not typical to the calendar in the rest of the world, if yes, specify it. 7. In your region, are there certain timings avoided to take flights or journeys? If yes, please specify. 8. In your region, are there certain times of the year where there are food restrictions, if yes, specify.

Table 19: Seed questions for the topics in Organized Ceremonial subcategory

Topic	Seed Questions
Understanding Local Traffic Regulations	<ol style="list-style-type: none"> 1. In your region, is there a designated spot where pedestrians normally cross the road? If not, specify where and how the pedestrians normally cross the road. 2. In your region, if not in a residential area, where are the cars normally parked? Is there any payment associated with this typical parking norm? 3. In your region, what are the conventions and frequencies of honking? Focus on how often people honk and for what reasons. 4. In your region, what side of the road do people normally drive on?
Adapting to Local Transportation Modes	<ol style="list-style-type: none"> 1. In your region, what is the most widely used form of public/local transportation? 2. In your region, what is the most unique form of local transportation that may not be found in other regions? 3. In your region, are the public/local transportation methods used to transport any other goods or services apart from people? If yes, please specify the mode of transport as well as the goods/services transported. 4. In your region, what are the typical occasions to use public transport? Focus on where people generally commute to using local transportation. 5. In your region, do people use digital tools or local networks to stay informed about delays, route changes, or real-time public transport updates? If so specify the tool as well as the public transport it is used for. 6. In your region, what are the typical apps used to book public transport routes? Specify the app as well as the public transport. 7. In your region, what payment options are commonly used for buses, trains, or other public transport, and which ones are most convenient for daily users?

Table 20: Seed questions for the topics in Streets and Traffic subcategory

Topic	Seed Questions
Carrying Capacity of Transport	<ol style="list-style-type: none"> 1. In your region, for the most used public transport, is there a carrying capacity maintained normally or is the capacity often broken? 2. In your region, what is the most crowded public transport mode? 3. In your region, what accommodations are made to guarantee more space when reaching carrying capacity? For example, sharing seats with the driver in the auto, etc. Be specific.
Transport during Special Events	<ol style="list-style-type: none"> 1. In your region, are there any new routes or transport added to accommodate for special events like festivals, etc? If any, specify the transportation added and the special event it was added for. 2. In your region, what is the influence on public transport during a major festival, is the use increased or decreased? Specify the mode of transportation as well as the change. 3. In your region, is there commonly a change from your regular transportation method to another during a major festival? If so specify the regular transport, the changed transport, and the major festival.

Table 21: Seed questions for the topics in Transportation subcategory

Topic	Seed Questions
Formal Educational Structure	<ol style="list-style-type: none"> 1. In your region, are there any local school boards followed? If yes, please specify them. 2. In your region, what are the common school boards that children go to? Please specify all the common board names. 3. In your region, what are the most important exams that children give in their education up until high school? 4. In your region, what grades are normally present in a school? Be as specific as possible. 5. In your region, do students normally change schools since some grades are absent in the school? If yes, specify when schools are changed. 6. In your region, what are the typical exams that students give post high school, for university/college purposes?
Attitudes Toward Education	<ol style="list-style-type: none"> 1. In your region, how many levels of education are considered necessary for a child to complete? Focus on levels e.g. high school, college, etc. 2. In your region, are there any streams of education considered more beneficial for the child's future over the others? If yes, please specify the streams. 3. In your region, how important is education to get a job? 4. In your region, what educational board is considered the best to provide education to children?

Table 22: Seed questions for the topics in Education System subcategory

Topic	Seed Questions
Norms for Interacting with Teachers	<ol style="list-style-type: none"> 1. In your region, how are teachers addressed when they enter the classroom? Focus on actions and sayings and be descriptive. 2. In your region, what is the typical status of teachers? Focus on status in terms of the amount of respect. 3. In your region, what is considered the appropriate way to address a teacher? Focus on sayings used. 4. In your region, what actions or gestures are considered disrespectful when interacting with teachers?
Student Extracurricular Activities	<ol style="list-style-type: none"> 1. In your region, what are the typical activities that students engage in outside of school? Think about their after school activities. 2. In your region, is it more common for students to engage in extracurricular activities or further academics after school? 3. In your region, when students go for academic help after school, what does that usually look like? (home tutoring, coaching center. etc) 4. In your region, what are the common extracurricular activities carried out within the school?

Table 23: Seed questions for the topics in Teachers and Students subcategory

Topic	Seed Questions
Special Occasion Clothing	<ol style="list-style-type: none"> 1. In your region, what is the traditional outfit of a bride in a wedding ceremony? 2. In your region, what is the traditional color of the bride's outfit? 3. In your region, what is the traditional outfit of a groom in a wedding ceremony? 4. In your region, what is the traditional outfit associated with the harvest festival of your region? Name and describe the outfit as well as the harvest festival. 5. In your region, what is the traditional outfit associated with the most widely celebrated festival of your region? Name and describe the outfit as well as the festival. 6. In your region, what fabrics are seen as auspicious for celebration wear? 7. In your region, what is the traditional attire worn by women if attending a religious ceremony at someone's? Please specify the religious ceremony as well as the attire. 8. In your region, what is the traditional attire worn by men if attending a religious ceremony at someone's? Please specify the religious ceremony as well as the attire.
Headgear and Footwear Norms	<ol style="list-style-type: none"> 1. In your region, what are the occasions of wearing certain head coverings that are not normally worn? For e.g. during religious ceremonies, etc. Be as specific as possible and describe the occasion. 2. In your region, what are the types of head covering, if any, worn during wedding ceremonies by the bride that are not ordinarily worn? 3. In your region, do married woman wear certain head coverings as a signifier of their marriage? If yes, please specify the type of head covering. 4. In your region, what types of head covering, if any, are worn during wedding ceremonies by the groom that are not ordinarily worn? 5. In your region, are there any special types of footwear worn by grooms during wedding ceremonies that are not ordinarily worn? If yes, please specify the type of the footwear. 6. In your region, what are some formal footwear choices for men that are not ordinarily worn? 7. In your region, what are some formal footwear choices for women that are not ordinarily worn? 8. In your region, are there any materials preferred for formal footwear? If yes, please specify the materials.

Table 24: Seed questions for the topics in Special Garments subcategory

Topic	Seed Questions
Ornamental Attire and Status Indication	<ol style="list-style-type: none"> 1. In your region, are there specific ornaments that are seen as signs of wealth and how do people view them? 2. In your region, what are the ornaments worn during a wedding that signify higher social standing? 3. In your region what are ornaments worn by men that showcase a higher status symbol? 4. In your region what are ornaments worn by women that showcase a higher status symbol?
Occasions for Wearing Specific Ornaments	<ol style="list-style-type: none"> 1. In your region, what type of traditional ornaments are usually worn by the bride during weddings? 2. In your region, what type of traditional ornaments are usually worn by the groom during weddings? 3. In your region, what types of ornaments are typically avoided during mourning or funerals? 4. In your region, are there any ornaments that are specially worn only during certain festivals? If yes, specify the ornament as well as the festival. 5. In your region, are there any ornaments that are avoided when traveling or visiting crowded areas? If yes, specify the ornaments.

Table 25: Seed questions for the topics in Ornament subcategory

Topic	Seed Questions
Food Preparation Techniques	<ol style="list-style-type: none"> 1. In your region what are the traditional tools utilized for grinding spices for food preparation? 2. In your region what are typical methods used for preserving food for long periods? 3. In your region what types of stove or cooking setup is used for making traditional dishes? 4. In your region, how are coconut or other hard ingredients broken and processed for cooking? 5. In your region, what types of vessels are more commonly used for everyday cooking? 6. In your region are there traditional condiments made at home? If yes, what are they?
Cultural Recipes and Ingredients	<ol style="list-style-type: none"> 1. In your region, what is the featured spice in most of your traditional food items? 2. In your region, what is the meat (if any) associated with most of your traditional dishes? 3. In your region, what is a traditional vegetable that is featured in most traditional dishes? 4. In your region, what is the traditional dish associated with your most celebrated festival? Name the festival as well as the dish. 5. In your region, is there a food ceremony normally associated with any of your festivals? If yes, please specify the name of the food ceremony as well as the festival. 6. In your region, are any specific condiments used during special occasions? 7. In your region, are any specific condiments avoided during certain occasions? 8. In your region are there condiments people avoid mixing together? If yes, specify.

Table 26: Seed questions for the topics in Food Preparation subcategory

Topic	Seed Questions
Staple Food Consumption	<ol style="list-style-type: none"> 1. In your region, what is considered a staple food source? 2. In your region, what type of grain or cereal is usually served with lunch or dinner? 3. In your region, what type of staple food is usually eaten during fasting periods? 4. In your region, what are the typical breakfast staples? 5. In your region, what is the most common accompaniment served with rice? 6. In your region, what is the most commonly used bread in most dishes? 7. In your region, after a main meal, what is commonly consumed as a digestive aid or mouth freshener?
Seasonal Diet Modifications	<ol style="list-style-type: none"> 1. In your region, what special dishes are prepared during the rainy season? 2. In your region, what changes are made to the diet during winters? 3. In your region, what traditional drinks are preferred during summers? 4. In your region, what traditional drinks are preferred during winters? 5. In your region, are any foods that are avoided in certain seasons? If yes, specify the food as well as the season.

Table 27: Seed questions for the topics in Diet subcategory

Topic	Seed Questions
Social Expression of Emotions	<ol style="list-style-type: none"> 1. In your region how do people express disagreement or disapproval without using any words? 2. In your region what hand gestures are considered rude or insulting? 3. In your region what gestures are used to greet someone warmly versus formally? 4. In your region how do people express gratitude nonverbally? 5. In your region what gestures or signs express anger or frustration?
Non-Verbal Expression of Respect or Disrespect	<ol style="list-style-type: none"> 1. In your region, how do you greet an elder? Focus on gestures. 2. In your region, what gestures signify respect towards teachers? 3. In your region, what gestures are performed to respectfully greet a religious leader? 4. In your region, how do you display respect for the national flag or anthem through gestures? 5. In your region, how do you display respect to a deceased person during a funeral ceremony? Focus on gestures.

Table 28: Seed questions for the topics in Gestures & Signs subcategory

Topic	Seed Questions
Navigating the “Grapevine”	<ol style="list-style-type: none"> 1. In your region, what is the typical informal digital source used to convey news and information? 2. In your region, what role do tea shops, local markets, or hair salons play in spreading informal news? 3. In your region, what are the typical practices done to avoid spreading sensitive information on the “grapevine”? 4. In your region, are digital platforms used to resolve or discuss issues related to the neighborhood or community? If yes specify the platform as well as the issues discussed through it. 5. In your region, do religious institutions use any digital platforms to spread information or keep people aware of religious practices? If yes, specify the digital platform, religion, and the information spread through it.
Trustworthiness of Information Sources	<ol style="list-style-type: none"> 1. In your region, what are the formal digital news platforms that are considered reliable? 2. In your region, what are the informal digital news platforms that are considered reliable? 3. In your region, do people still commonly continue to get their news through newspapers? 4. In your region, what are the most common newspapers that are considered reliable? 5. In your region, are there any magazines that kids normally read? If yes, specify.

Table 29: Seed questions for the topics in Dissemination of News and Information subcategory

Topic	Seed Questions
Negotiating Credit Advances and Discounts	<ol style="list-style-type: none"> 1. In your region, is it common to negotiate prices before buying, or is the price usually fixed? If so specify for what purchases. 2. In your region, are people more likely to offer discounts to familiar customers or strangers? If so specify for what purchases. 3. In your region, what kinds of purchases are most commonly negotiated for better rates? 4. In your region, do certain professions or groups receive more flexibility with credit or discounts? If so specify. 5. In your region, do people expect discounts when buying in bulk? If so specify. 6. In your region, are there particular times of the year or specific events when discounts and promotional offers are more prevalent? If so, specify. 7. In your region, what kinds of services or professional fees are typically open to negotiation for better rates?
Navigating Installment Buying	<ol style="list-style-type: none"> 1. In your region, what kinds of products are most commonly purchased through installment plans? 2. In your region, are people more likely to use formal (bank/store) or informal (person-to-person) installment agreements? 3. In your region, how do family or community opinions influence a person’s decision to opt for installment buying? 4. In your region what are common schemes offered for installment plans (eg: no cost emi, guarantor etc) 5. In your region, how do shopkeepers decide whether a customer is eligible for paying in installments? 6. In your region, are certain professions more likely to use installment buying than others?

Table 30: Seed questions for the topics in Credit subcategory

Topic	Seed Questions
Safekeeping of Valuables	<ol style="list-style-type: none"> 1. In your region, do people commonly keep their important valuables at home? If not, specify where they keep it. 2. In your region, how do people protect valuable things when they travel? 3. In your region, do people prefer keeping cash at home or in a bank, and why? 4. In your region, is there a strong cultural emphasis on keeping precious metals within the household for generations, and how is it typically protected? 5. In your region, how do cultural beliefs or superstitions sometimes influence the placement or display of certain valuables within a home? 6. In your region, what items do people typically hoard as a way of preserving wealth?
Preferable Investment Forms	<ol style="list-style-type: none"> 1. In your region, what is considered a preferred long-term investment for families? (Like gold, land etc) 2. In your region, how do religious or cultural beliefs influence what people invest in? 3. In your region, what kind of property is most commonly purchased as an investment? 4. In your region, what kind of items are bought as “prestige” investments beyond financial value? 5. In your region, do people invest in religious donations or temples as a spiritual investment? If yes, specify. 6. In your region, how do marriage customs influence what type of assets families prioritize?

Table 31: Seed questions for the topics in Saving & Investment subcategory

A.2 Study Details

A.2.1 Participant Criteria

Participants were required to have lived in their target region for more than half their lifetime.

A.2.2 Study Design

Each participant completed 41 forms, with each form containing a maximum of 15 questions (611 total questions divided across forms). To ensure participants provided region-specific responses, each question was prefaced with “In your region.” For example, the question “Are gifts opened in front of the giver or later?” was presented as “In your region, are gifts opened in front of the giver or later?” This framing served as a consistent reminder for participants to draw upon their local cultural knowledge. The form interface is available at: <https://cultural-survey-frontend.vercel.app/>.

Participants responded to questions in their own words, allowing us to capture natural cultural knowledge rather than forcing responses into predetermined categories. Each form took between 20–67 minutes to complete. We implemented various attention checks throughout, and responses were reviewed and scanned for any AI-generated content.

A.2.3 Compensation

Participants were compensated at \$8.00 per hour, aligning with Prolific’s fair payment standards.

A.2.4 Ethics and Consent

This study received approval from our institution’s Research Ethics Board (REB). All participants provided informed consent before beginning the study, acknowledging data usage, anonymization procedures, and their right to withdraw at any time.

A.2.5 Data Collection Period

Responses were collected between October and November 2025.

A.3 Agreement Validation and Override Analysis

GPT-4o provided preliminary classifications for three types of agreement across all question-region combinations: intra-regional consensus (whether 4–5 participants within a region agreed), inter-regional agreement (whether two regions shared the same practice), and universal agreement (whether all five regions agreed).

Two authors independently reviewed all cases using a custom annotation tool displaying: (1) the question, (2) GPT-4o’s preliminary assessment, and (3) all participant responses. For each case, annotators decided whether responses were semantically equivalent and, if so, established the gold standard answer. Inter-annotator agreement between the two human annotators was perfect (Fleiss’ $\kappa = 1.0$) across all judgment types.

A.3.1 Override Rates

Table 32 presents the rates at which human annotators overrode GPT-4o’s preliminary classifications.

Agreement Type	Pairs	Raw LLM	Production	Overrides
Intra-Regional	2,540	1,607 (63.3%)	1,605 (63.2%)	194 (7.6%)
Added (LLM No → Human Yes)				96
Removed (LLM Yes → Human No)				98
Inter-Regional	2,197	1,931 (87.9%)	1,299 (59.1%)	636 (28.9%)
Added (LLM No → Human Yes)				2
Removed (LLM Yes → Human No)				634
Universal	106	75 (70.8%)	49 (46.2%)	26 (24.5%)
Added (LLM No → Human Yes)				0
Removed (LLM Yes → Human No)				26

Table 32: Human override rates of GPT-4o’s preliminary agreement classifications. Intra-regional overrides were balanced between additions and removals (7.6% override rate), while inter-regional (28.9%) and universal (24.5%) cases showed predominantly removals, indicating GPT-4o over-identified cross-regional consensus.

A.4 Intra-Region Agreement Prompting Details

A.4.1 Model Configuration

- **Model:** GPT-4o-2024-08-06
- **Temperature:** 0.1
- **Max Tokens:** 800–3000 (progressive increase across retry attempts)
- **Retry Logic:** Up to 5 retry attempts with increasing token limits (800 → 1200 → 1500 → 2000 → 2500 → 3000)

A.4.2 System Prompt

The following system prompt was used:

You are an expert at analyzing cultural agreement. You must respond with valid, complete JSON only, no additional text.

A.4.3 User Prompt Template

For each question with regional responses, the following user prompt template was used:

Analyze agreement among [N] responses from [REGION] about regional practices.

QUESTION: [Question Text]

NUMBERED RESPONSES:

Response 1: [Answer 1]

Response 2: [Answer 2]

...

CORE INSTRUCTION: Only look for concepts that directly answer the question asked. This is imperative above all else.

ANALYSIS RULES:

1. Look for the SAME underlying concept across responses (semantic similarity counts)
2. Spelling variations, spacing differences, and synonyms count as the SAME concept
3. You must quote exact text but recognize when different words mean the same thing
4. [THRESHOLD]+ different responses must mention the same underlying concept

where [THRESHOLD] = 4 if $N \geq 5$, otherwise $\max(2, N-1)$

CONCEPT IDENTIFICATION EXAMPLES:

Question: “What foods are eaten during festivals?”

- Response mentions “sweets” → Answers the question
- Response mentions “celebration” → Doesn’t answer what food

Question: “What nonverbal actions are disrespectful to elders?”

- Response mentions “pointing” → Answers the question (specific action)
- Response mentions “being rude” → Doesn’t answer what action

Question: “In your region, what is the customary gesture for offering food to a deity, and how is it performed?”

- Response: “We offer food to deity’s by putting it on a clean plate” → Concept extracted should be “putting it on a clean plate” since that is what answers the question
- Response: “We offer food to deity’s by putting it on a clean plate” → Concept extracted should NOT be “food is offered”, that is not the answer to the question

SEMANTIC MATCHING EXAMPLES:

- “Raksha Bandhan” = “Rakshabandhan” = “rakhi” (same festival)
- “clean house” = “cleaning home” = “tidy up house” (same activity)
- “Diwali” = “Deepawali” (same festival)
- “new clothes” = “fresh clothing” = “new garments” (same concept)

STEP-BY-STEP ANALYSIS:

1. Analyse if the response is even answering the question before moving on to getting the concepts
2. Extract concepts from each response with exact quotes that answer the question
3. Group semantically similar concepts together (consider spelling, synonyms, variants)
4. Count how many different responses mention each concept group
5. Agreement exists if any concept group appears in [THRESHOLD]+ responses

VERIFICATION FORMAT:

For each concept group, show the evidence and reasoning.

A.4.4 Required JSON Output Format

The model was required to return structured JSON with the following schema:

```
{
  "step_by_step_extraction": {
    "response_1_concepts": ["concept from response 1"],
    "response_2_concepts": ["concept from response 2"],
    ...
  },
  "semantic_grouping": {
    "concept_group_name": {
      "responses_and_quotes": {
        "1": "exact quote from response 1",
        "2": "exact quote from response 2",
        ...
      }
    },
    "semantic_explanation": "Why these quotes represent the same concept",
  }
}
```

```

    "count": N
  }
},
"agreement_found": true/false,
"threshold_met": "X out of Y responses
mention the same
    underlying concept",
"common_concepts": [
  {
    "concept": "unified concept name",
    "responses_mentioning": [1, 2, 3],
    "exact_quotes_proof": ["quote 1",
"quote 2", "quote 3"],
    "semantic_note": "Explanation of any
spelling/synonym variations"
  }
],
"summary": "Brief explanation recognizing
semantic similarity while showing
evidence"
}

```

A.4.5 Retry Mechanism

To handle truncated or incomplete responses, an automatic retry system was implemented:

- **Trigger Conditions:** Empty responses, un-parseable JSON, missing required fields, or agreement found with empty concept lists
- **Progressive Token Increase:** Each retry attempt increased `max_tokens` (800 → 1200 → 1500 → 2000 → 2500 → 3000)
- **Maximum Attempts:** 5 retries per question

A.4.6 Agreement Threshold Logic

The agreement threshold was dynamically calculated based on the number of responses:

$$\text{Threshold} = \begin{cases} 4 & \text{if } N \geq 5 \\ \max(2, N - 1) & \text{otherwise} \end{cases} \quad (1)$$

where N is the total number of responses for a given question in a region. In our dataset, all questions had exactly $N = 5$ responses per region, resulting in a consistent agreement threshold of 4 responses. This means that for agreement to be found, at least 4 out of 5 responses (80%) needed to mention the same underlying concept.

A.5 Inter-Region Agreement Prompting Details

A.5.1 Model Configuration

- **Model:** GPT-4o-2024-08-06
- **Temperature:** 0.1
- **Max Tokens:** 800

A.5.2 Question Matching Methodology

Questions were matched between regions using normalized question text rather than question numbers. Text normalization involved:

- Converting to lowercase
- Removing extra whitespace
- Stripping leading/trailing spaces

A.5.3 System Prompt

The following system prompt was used:

You are an expert at comparing cultural concepts for semantic similarity. You must respond with valid JSON only.

A.5.4 User Prompt Template

For each question where both regions had intra-region agreement, the following prompt template was used:

Compare already-identified concepts from two regions to find inter-regional agreement.

QUESTION: *[Question Text]*

[REGION1] AGREED-UPON CONCEPTS (from intra-region analysis):

1. Concept: '[Concept Name]'

Evidence: '[Quote 1]', '[Quote 2]', '[Quote 3]'

...

[REGION2] AGREED-UPON CONCEPTS (from intra-region analysis):

1. Concept: '[Concept Name]'

Evidence: '[Quote 1]', '[Quote 2]', '[Quote 3]'

...

TASK: *Determine if any concept from [REGION1] matches any concept from [REGION2].*

INTER-REGIONAL AGREEMENT CRITERIA:

- Agreement exists if ANY concept from [REGION1] is semantically similar to ANY concept from [REGION2]
- For specific festivals and traditions, the festival or tradition names have to be an exact match for agreement
- Both concepts must answer the same question

- Semantic similarity includes synonyms, variations, and different ways of expressing the same idea

SEMANTIC MATCHING EXAMPLES:

- “emergency situations” matches “urgent circumstances” (same underlying concept)
- “cleaning house” matches “home tidying” (same activity)
- “touching feet” matches “feet touching” (same gesture)
- “festival sweets” matches “celebratory desserts” (same food category)
- “August” matches “august” matches “month of August” (same month)

NO SEMANTIC MATCHING EXAMPLES:

- “pongal” and “lohari” are both names of a harvest festival with the first for south and second for north but they are not semantically similar
- “godh bharaai” and “valaikappu” are both names of a pregnancy ceremony with the first for north and second for south but they are not semantically similar

ANALYSIS PROCESS:

1. Compare each [REGION1] concept with each [REGION2] concept
2. Look for semantic similarity in the concept names and evidence quotes
3. If any pair matches, inter-regional agreement exists
4. If no concepts match, no inter-regional agreement

A.5.5 Required JSON Output Format

The model was required to return structured JSON with the following schema:

```
{
  "concept_comparisons": [
    {
      "region1_concept": "concept name from Region 1",
      "region2_concept": "concept name from Region 2",
      "semantic_match": true,
      "matching_explanation": "Why these concepts represent the same underlying idea"
    }
  ],
  "inter_regional_agreement": true/false,
  "matched_concepts": [
    {
      "unified_concept_name": "shared concept name",
      "region1_concept": "original concept from Region 1",
      "region2_concept": "original concept from Region 2",
      "semantic_explanation": "How these concepts are semantically similar"
    }
  ]
}
```

```
"agreement_summary": "Brief explanation of  
whether and why inter-regional agreement  
was found"  
}
```

A.5.6 Example Application

As an illustration, for a question about respectful gestures toward elders:

Question: What is a common respectful gesture shown to elders in your region?

South Region Agreed Concept: "Touching feet" (4/5 responses mentioned variants: "touch feet", "feet touching", "touching their feet")

North Region Agreed Concept: "Touching the feet" (4/5 responses mentioned variants: "we touch feet", "touching feet of elders", "feet touching")

Inter-Regional Analysis Result: Agreement found

Unified Concept: "Touching feet as a gesture of respect"

Semantic Explanation: Both regions independently converged on the same physical gesture (touching feet) as a sign of respect toward elders, with only minor variations in phrasing.

A.6 Universal Agreement Prompting Details

A.6.1 Model Configuration

- **Model:** GPT-4o-2024-08-06
- **Temperature:** 0.1
- **Max Tokens:** 2000

A.6.2 Question Matching Methodology

Questions were matched between regions using normalized question text rather than question numbers. Text normalization involved:

- Converting to lowercase
- Removing extra whitespace
- Stripping leading/trailing spaces

A.6.3 System Prompt

The following system prompt was used:

You are an expert at comparing cultural concepts across multiple regions for semantic similarity. You must respond with valid JSON only.

A.6.4 User Prompt Template

For each question where all five regions had intra-region agreement, the following user prompt template was used:

Compare already-identified concepts from 5 regions to find inter-regional agreement.

QUESTION: [Question Text]

REGIONAL CONCEPTS:

=====

NORTH - AGREED-UPON CONCEPTS

=====

Concept 1: '[Concept Name]'

Supporting Evidence:

1. "[Quote 1]"

2. "[Quote 2]"

...

=====

SOUTH - AGREED-UPON CONCEPTS

=====

[Similar format for South region]

[Similar sections for EAST, WEST, and CENTRAL]

TASK: *Determine if there is agreement across ANY or ALL of these regions: North, South, East, West, Central*

Systematically compare concepts across all 5 regions to determine:

1. Whether there is **UNIVERSAL** agreement (all 5 regions share the same concept)

2. Whether there is **PARTIAL** agreement (some but not all regions share concepts)
3. Whether there is **NO** agreement (each region has completely different concepts)

INTER-REGIONAL AGREEMENT CRITERIA:

- Universal agreement exists if at least one concept from ALL regions (North, South, East, West, and Central) is semantically similar
- Partial agreement exists if at least one concept is semantically similar for some and not ALL 5 regions
- For specific festivals and traditions, names must be exact matches
- All concepts must answer the same question
- Semantic similarity includes synonyms, variations, and different expressions of the same idea

SEMANTIC MATCHING EXAMPLES:

- "emergency situations" matches "urgent circumstances" (same underlying concept)
- "cleaning house" matches "home tidying" (same activity)
- "touching feet" matches "feet touching" (same gesture)
- "festival sweets" matches "celebratory desserts" (same food category)

NO SEMANTIC MATCHING EXAMPLES:

- "pongal" and "lohari" are both harvest festivals but they are not semantically similar (different regional festivals)
- "godh bharai" and "valaikappu" are both pregnancy ceremonies but they are not semantically similar (different regional ceremonies)

NO UNIVERSAL AGREEMENT EXAMPLE:

- If North, South, West, and Central regions mention Diwali as the most popular festival but East mentions Durga Puja, that is not counted as universal agreement, instead it is partial agreement

ANALYSIS PROCESS:

STEP 1: CREATE A COMPARISON MATRIX

- List all concepts from all 5 regions
- For each unique concept group, identify which regions mention it
- Example format:
Concept Group 1: "Fasting/Observing Fast"
→ Present in: North, South, West
Concept Group 2: "Pongal"
→ Present in: South only

STEP 2: APPLY SEMANTIC MATCHING RULES

- Compare each concept from Region A with each concept from Regions B, C, D, E
- Use the matching rules above to determine if concepts are semantically similar

- *Remember: Similar category \neq Semantic match (e.g., both are festivals, but different festivals)*

STEP 3: IDENTIFY AGREEMENT PATTERNS

- **Universal Agreement:** At least ONE concept is shared by ALL 5 regions
- **Partial Agreement:** At least ONE concept is shared by SOME regions (2 or more, but not all)
- **No Agreement:** Each region has completely different concepts OR no semantic matches found

STEP 4: DOCUMENT MATCHES

For each matched concept group, clearly state:

- The unified concept name (the general term that encompasses all variations)
- Which regions share this concept
- What each region calls it (regional variations)
- Why they are semantically similar (evidence from quotes)

A.6.5 Required JSON Output Format

The model was required to return structured JSON with the following schema:

```
{
  "universal_agreement": true/false,
  "agreement_type":
  "universal|partial|none",
  "regions_in_agreement": ["region1",
  "region2", ...],
  "matched_concepts": [
    {
      "unified_concept_name": "shared
      concept name",
      "regions_sharing": ["region1",
      "region2", ...],
      "regional_variations": {
        "region1": "concept name in region1",
        "region2": "concept name in region2"
      },
      "semantic_explanation": "How these
      concepts are semantically similar"
    }
  ],
  "concept_matrix": [
    {
      "region1": "concept_name",
      "region2": "concept_name",
      "region3": "concept_name",
      "semantic_match": true/false,
      "explanation": "why they match or
      don't match"
    }
  ],
  "agreement_summary": "Brief explanation of
  inter-regional agreement patterns"
}
```

A.6.6 Agreement Classification

Universal agreement was classified into three mutually exclusive categories:

- **Universal Agreement:** At least one concept is semantically similar across all 5 regions (North, South, East, West, Central)
- **Partial Agreement:** At least one concept is semantically similar across 2–4 regions, but not all 5
- **No Agreement:** No concepts are semantically similar across any subset of regions

A.6.7 Example Application

As an illustration, for a question about harvest festivals:

Question: What is the main festival celebrated in your region?

Regional Agreed Concepts:

- **North:** “Diwali” (4/5 responses)
- **South:** “Diwali” (4/5 responses)
- **East:** “Durga Puja” (4/5 responses)
- **West:** “Ganesh Chaturti” (4/5 responses)
- **Central:** “Diwali” (4/5 responses)

Universal Agreement Analysis Result: Partial agreement (not universal)

Agreement Pattern:

- **Concept Group 1:** “Diwali” shared by North, South, and Central (3 regions)

Classification: Partial agreement (North, South, and Central share a concept, but not all 5 regions)

Explanation: While all regions celebrate festivals, the specific festival names are region-specific and not semantically similar according to the “exact match for festival names” rule. However, North, South, and Central share the same festival name (Diwali), resulting in partial agreement.

A.7 Dataset Structure

Domain	Subcategories	Topics	Questions
Clothing and adornment	2	4	59
Communication	2	4	47
Education	2	4	58
Festivals and Rituals	3	9	109
Finance	2	4	53
Food processing and consumption	2	4	50
Interpersonal Relations	3	6	86
Traffic and transport behavior	2	4	53
Total	18	39	515

Table 33: Question distribution across domains, subcategories, and topics

A.8 Model Evaluation RASA

A.8.1 LLM-as-Judge Evaluation Details

Model Configuration

- **Judge Model:** Google Gemini Flash 3.0 (via OpenRouter API⁷)
- **Temperature:** 0.0

System Prompt The following system prompt was used to establish the evaluation framework:

You are an expert evaluator for cultural knowledge questions about India.

Your task: Determine if a model's answer is factually correct compared to the gold standard answer.

Guidelines:

1. **CORRECT:** The model's answer conveys **ONLY** the same core information as the gold answer

- Paraphrasing is acceptable
- ALL details in the model's answer must appear in the gold answer
- Equivalent terms are acceptable (e.g., "clay stove" = "chulha")
- The answer should not contain significant additional facts, examples, or details beyond what the gold answer provides

2. **PARTIALLY_CORRECT:** The model's answer has some correct information from the gold answer but:

- Misses details from the gold answer
- Contains additional information, facts, or examples that are not present in the gold answer (even if factually correct)
- Provides extra context or details that go beyond the scope of the gold answer
- Is too vague or incomplete

3. **INCORRECT:** The model's answer:

- Contradicts the gold answer with wrong facts
- Provides completely different information
- Misses the main point entirely

Important:

- Focus on factual accuracy, not writing style
- Consider cultural context and regional variations
- Be strict about factual contradictions (e.g., "jewelry" \neq "cash")
- If the model adds information not in the gold answer (like additional examples, regional variations, or extra details), mark as **PARTIALLY_CORRECT** even if the added information is accurate
- The gold answer defines the scope - answers should not exceed that scope

User Prompt Template For each question-answer pair, the following template was used:

Question: [Question Text]

Gold Standard Answer: [Gold Answer]

Model's Answer: [Predicted Answer]

Evaluate the model's answer.

Required JSON Output Format The judge model was required to return structured JSON with the following schema:

```
{
  "label": "CORRECT" | "PARTIALLY_CORRECT" | "INCORRECT",
  "reasoning": "brief explanation",
  "key_discrepancies": ["list any factual errors or significant additions"]
}
```

A.8.2 Question Distribution

Region	Number of Questions
North	326
South	326
East	276
West	354
Central	348
Total	1630

Table 34: Distribution of Region-Anchored Short Answer questions by region

A.8.3 RASA Sensitivity Analysis

Model	w=0.3	w=0.5	w=0.7
Grok-4 Fast	39.4%	52.6%	65.7%
GPT-5.2	38.9%	52.4%	65.9%
Qwen3 VL	39.3%	51.6%	63.8%
Llama 3.3 70B	37.5%	50.3%	63.2%
Claude Sonnet 4.5	36.7%	51.7%	66.8%
Mistral Large	36.7%	50.4%	64.1%
Gemini 3.0	36.0%	51.1%	66.1%
DeepSeek V3.2	35.7%	49.5%	63.4%
Range	3.7pp	3.1pp	3.6pp

Table 35: Sensitivity analysis: RASA overall accuracy under different partial credit weights. Model rankings remain stable across weights, with all models clustering within 3–4 percentage points at each weight setting. This confirms our findings are robust to the choice of partial credit weighting. Models are sorted by w=0.5 performance (our primary metric).

⁷<https://openrouter.ai/>

A.8.4 Scoring Criteria

Score	Question	Gold	Model Response	Rationale
1.0	In North India, during annual school functions or inter-school competitions, what transportation is typically arranged?	Buses	Schools typically arrange a bus for students' transportation	Correctly identifies the transportation arrangement
0.5	In North India, do students usually remain in the same classroom while teachers rotate, or do they move between classes?	Students remain in the classroom and teachers rotate	Students usually remain in the same classroom while teachers rotate, especially in primary and middle grades, though some secondary schools have students move for specialized subjects	Correct core practice but adds content not in gold answer
0.0	In South India, how do drivers typically behave toward pedestrians waiting at zebra crossings?	Drivers wait for pedestrians to cross	Drivers often do not stop for pedestrians waiting at zebra crossings, so pedestrians typically wait for a safe gap to cross	Contradicts the gold answer by describing opposite behavior

Table 36: Examples of the three-tier scoring system for RASA questions. Green text indicates correct information matching the gold answer; red text indicates incorrect or additional information.

A.9 Model Evaluation RA-MCQ

A.9.1 Question Distribution

Domain	Number of Questions
Interpersonal Relations	16
Education	12
Clothing & Adornment	8
Food Processing	9
Communication	5
Finance	9
Festivals & Rituals	18
Traffic & Transport	2
Total	79

Table 37: Distribution of Region-Agnostic Multiple Choice Questions by domain.

A.9.2 Chi-Square Test for Regional Selection Bias

We used a chi-square goodness-of-fit test to assess whether models exhibit regional selection bias in RA-MCQ questions.

Null Hypothesis The model selects uniformly at random from available options, with no regional preference.

Observed and Expected Counts Observed count O_r for region r is calculated by aggregating the model’s actual selections:

$$O_r = \sum_{q \in Q} \begin{cases} \frac{1}{|R_{\text{selected}}(q)|} & \text{if } r \in R_{\text{selected}}(q) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where Q is the set of all question instances (30 runs per unique question) and $R_{\text{selected}}(q)$ is the set of regions represented by the option the model selected in question instance q . If an option represents multiple regions, credit is split equally.

Expected count E_r under uniform random selection:

For each question instance q with n_q options, if option i represents region set R_i , each region receives:

$$\text{ExpectedCredit}_r = \frac{1}{n_q} \times \frac{1}{|R_i|} \quad (3)$$

Total expected count for region r across all instances:

$$E_r = \sum_{q \in Q} \sum_{i: r \in R_i} \frac{1}{n_q} \times \frac{1}{|R_i|} \quad (4)$$

This accounts for: (1) varying numbers of options per question (3–5) and (2) multiple regions sharing the same option.

Example Two question instances:

Instance 1 (3 options): A→North, B→South, C→{East,West}

Model selects Option A (North)

Observed: North=1, South=0, East=0, West=0, Central=0

Expected: North=1/3, South=1/3, East=1/6, West=1/6, Central=0

Instance 2 (5 options): A→North, B→South, C→East, D→West, E→Central

Model selects Option C (East)

Observed: North=0, South=0, East=1, West=0, Central=0

Expected: Each region=1/5

Aggregated across both instances:

North: $O = 1.0$, $E = 0.533$; South: $O = 0.0$, $E = 0.533$; East: $O = 1.0$, $E = 0.367$; West: $O = 0.0$, $E = 0.367$; Central: $O = 0.0$, $E = 0.200$

Test Statistic The chi-square goodness-of-fit statistic is:

$$\chi^2 = \sum_r \frac{(O_r - E_r)^2}{E_r}, \quad df = 4 \quad (5)$$

where df is the degrees of freedom (number of regions minus 1). Statistical significance was assessed at $\alpha = 0.05$.

Standardized Residuals To identify which specific regions deviate significantly from expectation, we calculate standardized residuals:

$$z_r = \frac{O_r - E_r}{\sqrt{E_r}} \quad (6)$$

Values $|z_r| > 1.96$ indicate significant deviation at $\alpha = 0.05$; $|z_r| > 2.58$ at $\alpha = 0.01$; $|z_r| > 3.29$ at $\alpha = 0.001$. Positive residuals indicate over-selection; negative indicate under-selection.

A.10 Model Evaluation Prompts and Configuration

A.10.1 Model Configuration

- **Temperature:** 1.0
- **Number of runs per question:** 30
- **API providers:** OpenAI API⁸ (for GPT models), OpenRouter⁹ API (for all other models)

A.10.2 Region-Anchored Short Answer

Prompt Template:

You are answering a question about Indian cultural practices. Please provide a concise answer that directly answers the question.

Question: [Question Text]

IMPORTANT: Provide a direct answer in 1 sentence. Do not use conversational filler or any justifications. Simply answer the question in the most brief way possible.

Answer:

A.10.3 Region-Agnostic Multiple Choice (RA-MCQ)

User Prompt:

You are answering a question about Indian cultural practices. Please select the most appropriate answer from the given options.

Question: [Question Text]

Options:

A. [Option 1]

B. [Option 2]

C. [Option 3]

...

IMPORTANT: Respond with ONLY the letter of your chosen answer (A, B, C, D, or E). Do not provide any explanation or additional text.

Your answer:

⁸<https://platform.openai.com/docs/models>

⁹<https://openrouter.ai/>

A.11 Additional Tables, Figures, and Analyses

Question	Model	Error Type	Gold Answer	Predicted Answer
In West India, what behaviors are commonly considered disrespectful while a teacher is speaking in class?	Llama	Over-explaining	Speaking to other students while a teacher is speaking is considered disrespectful.	In West India, behaviors such as interrupting , making eye contact , or crossing arms and legs are commonly considered disrespectful while a teacher is speaking in class.
In West India, what religious holidays require people to avoid wearing or displaying certain colors?	Claude	Over-explaining	People typically avoid wearing black during auspicious religious holidays.	During Amavasya (new moon day) and certain mourning periods , wearing black is traditionally avoided, while white is avoided during festivals, and during inauspicious days like Tuesdays and Saturdays some avoid black in certain communities.
In West India, how is a friend's parent usually greeted? Focus on the titles used to greet them.	Llama	Both	The standard greeting is Namaskar , followed by "Uncle" or "Aunty," sometimes regional titles like "Kaka/Kaku" are used as well.	In West India, a friend's parent is usually greeted with titles such as "Kaka" or "Kaki" for the father and "Aaji" or "Aai" for the mother.
In Central India, what ornaments are commonly worn by men on a daily basis?	Qwen	Over-explaining	Rings are most commonly worn by men on a daily basis.	Men in Central India commonly wear a simple silver or brass chain , a wristwatch , and sometimes a ring or a thread (mangalsutra-style) on a daily basis.

Table 38: Examples of partially correct model responses on RASA with error type annotations. Over-explaining: correct core information but extraneous additions; Both: extraneous additions and key omissions; Underspecifying: lacks core information. **Green** = correct, **red** = erroneous/extraneous, **orange** = omitted.

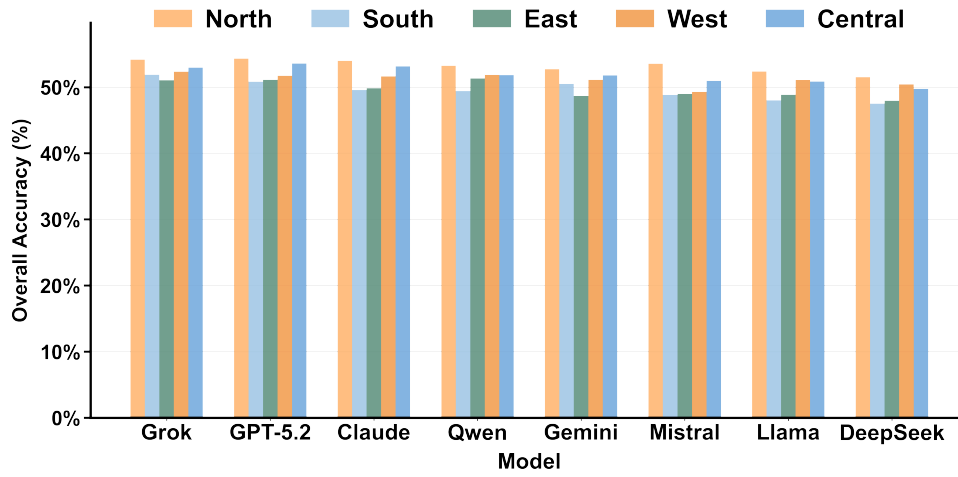


Figure 8: Overall Accuracy by regions of all models on RASA

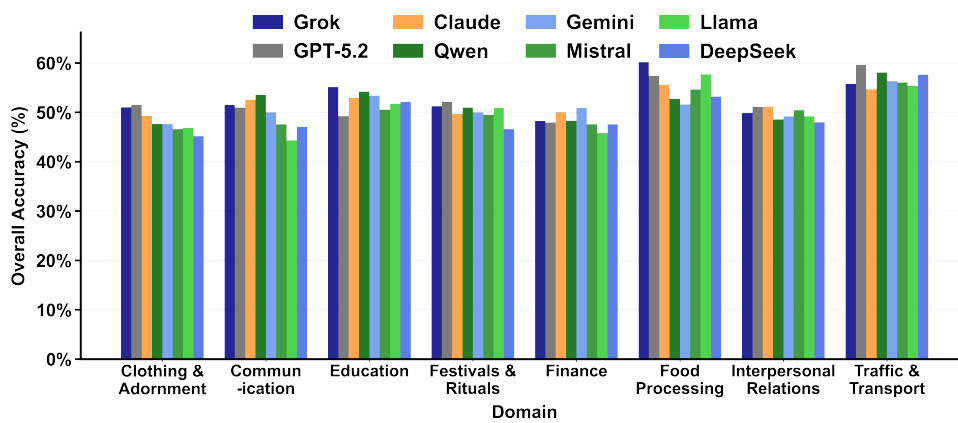


Figure 9: Overall Accuracy by domain of all models on RASA

Model	χ^2	df	p-value
Qwen3 VL	209.46	4	<0.001***
Gemini 3.0 Flash	186.41	4	<0.001***
GPT-5.2	165.04	4	<0.001***
Llama 3.3 70B	160.63	4	<0.001***
Grok-4 Fast	143.32	4	<0.001***
Claude Sonnet 4.5	141.03	4	<0.001***
Mistral Large	114.15	4	<0.001***
DeepSeek V3.2	80.70	4	<0.001***

Table 39: Chi-square goodness-of-fit tests for regional selection bias in RA-MCQ. All models deviate significantly from uniform random selection (expected approx 20% per region).

Model	Observed (% of Total)					Selection Ratio (Obs/Exp)				
	North	South	East	West	Central	North	South	East	West	Central
Qwen3 VL	619 (26.1%)	394 (16.6%)	340 (14.3%)	358 (15.1%)	659 (27.8%)	1.32×	0.79×	0.76×	0.73×	1.41×
Gemini 3.0	531 (22.4%)	471 (19.9%)	379 (16.0%)	306 (12.9%)	683 (28.8%)	1.14×	0.95×	0.84×	0.63×	1.46×
GPT-5.2	563 (23.8%)	456 (19.2%)	356 (15.0%)	337 (14.2%)	658 (27.8%)	1.20×	0.92×	0.79×	0.69×	1.40×
Llama 3.3	566 (23.9%)	453 (19.1%)	357 (15.1%)	340 (14.3%)	652 (27.5%)	1.21×	0.91×	0.80×	0.70×	1.39×
Grok-4 Fast	589 (24.8%)	423 (17.9%)	315 (13.3%)	419 (17.7%)	624 (26.3%)	1.26×	0.85×	0.70×	0.86×	1.33×
Claude Sonnet	562 (23.7%)	442 (18.7%)	359 (15.1%)	363 (15.3%)	644 (27.2%)	1.20×	0.89×	0.80×	0.75×	1.38×
Mistral Large	575 (24.2%)	416 (17.5%)	389 (16.4%)	379 (16.0%)	612 (25.8%)	1.23×	0.84×	0.86×	0.78×	1.31×
DeepSeek V3.2	526 (22.4%)	450 (19.2%)	444 (18.9%)	346 (14.7%)	579 (24.7%)	1.14×	0.91×	1.00×	0.72×	1.25×
Average	561 (23.7%)	438 (18.5%)	360 (15.2%)	356 (15.0%)	639 (26.9%)	1.21×	0.88×	0.82×	0.73×	1.37×

Table 40: Regional selection in adversarial MCQs. Left: Observed counts and percentage of total selections (approx 2,370 per model). Right: Selection ratio (observed/expected), where expected counts derived from chi-square methodology. Ratio >1.0 = over-selection; <1.0 = under-selection. Bold indicates systematic over-selection across all models. Central India selected 1.37× expected rate (37% over-selection); North India 1.21× (21% over-selection); West India 0.73× (27% under-selection).

A.11.1 Regional Selection Details.

West India. West India experiences the most severe under-selection across all models (12.9%–17.7%, 0.63–0.86× expected). Standardized residuals range from -3.1 (Grok) to -8.2 (Gemini), all significantly below expected rates.

East India. East India shows similarly strong under-selection (13.3%–18.9%, 0.70–1.00× expected). Standardized residuals range from -1.9 (Mistral) to -5.7 (Grok). DeepSeek is a notable exception, selecting East India at near-expected rates (18.9%, 1.00×, residual = 0.00), though it still under-selects West India.

South India. South India shows heterogeneous patterns across models (16.6%–19.9%, 0.79–0.95× expected). Three models significantly under-select: Qwen (16.6%, 0.79×, residual = -4.7), Mistral (17.5%, 0.84×, residual = -3.7), and Grok (17.9%, 0.85×, residual = -3.3). Five models approach expected frequencies: Gemini (19.9%, 0.95×, residual = -1.2), GPT-5.2 (19.2%, 0.92×, residual = -1.9), DeepSeek (19.2%, 0.91×, residual = -2.0), Llama (19.1%, 0.91×, residual = -2.0), and Claude (18.7%, 0.89×, residual = -2.5). Unlike the consistent over-selection of Central/North or under-selection of West/East, South India's treatment varies by model.