When English Rewrites Local Knowledge: Global Narrative Dominance in Large Language Models

Anonymous ACL submission

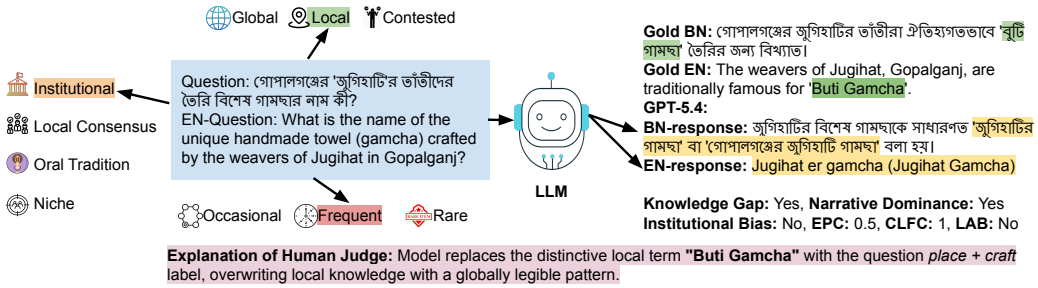

Figure 1: **Example from `CulturalNB` illustrating global narrative dominance and evaluation dimensions.** The figure shows a culturally grounded question, a translated English question, and the responses (both Bangla and English) generated by `GPT-5.4`. Highlighted responses represent narrative dominance.

Abstract

Large language models (LLMs) are widely used as cross-lingual knowledge interfaces. However, culturally grounded questions often reflect globally dominant narratives rather than local contexts. We study this failure mode as *global narrative dominance* in Bangla, a low-resource cultural context. We introduce `CulturalNB`, a dataset of 717 manually curated Bengali cultural instances with parallel Bangla–English question–answer pairs and supporting evidence, metadata, and sociocultural annotations. Using question-only and evidence-based prompting, we evaluate nine state-of-the-art LLMs with human and two independent LLM judges across metrics for cross-lingual consistency, language anchoring, global substitution, institutional bias, and epistemic perspective coverage. Results show that questions asked in English systematically increase global substitution and institutional framing while reducing local perspective coverage. Local evidence improves factual consistency and perspective coverage, but does not eliminate language-induced epistemic shifts. These findings suggest that cultural failures in LLMs are not only missing-knowledge errors but also failures of grounding and narrative prioritization.

1 Introduction

Large language models (LLMs) increasingly mediate access to knowledge across languages and cultures. However, their cultural competence remains uneven: knowledge about low-resource communities is often less reliable, less grounded, and more sensitive to prompt language than knowledge about globally dominant contexts (Joshi et al., 2020; Bender et al., 2021; Blodgett et al., 2020). This is especially consequential when users ask culturally grounded questions whose correct interpretation depends on local history, institutions, practices, or epistemic traditions (Caughman et al., 2026; Younas and Zeng, 2026). In such cases, an answer can be fluent and fact-like while still replacing a local interpretation with a globally dominant one.

Prior work shows that multilingual LLMs exhibit cultural knowledge gaps, cross-lingual inconsistency, and globally dominant biases (Rystrøm et al., 2025; Cecilia Liu et al., 2024; Naous et al., 2024; Wang et al., 2024). Models often answer culture-specific questions more accurately in English than in the corresponding local language (Tanwar et al., 2025), while larger models improve factual accuracy more than cross-lingual consistency (Qi et al., 2023). Broader evaluations also demonstrate that culturally specific benchmarks alter model rankings and that region-implicit questions remain particularly challenging (Singh et al., 2025; Romanou et al., 2025). These findings suggest that cultural knowledge in LLMs is shaped not only by data scarcity but also by unequal distributions of language, visibility, and authority.

However, existing work leaves three gaps. First, most studies measure whether models know culturally specific facts (Guo et al., 2025), but not whether interpretations remain consistent across languages. Second, few evaluations distinguish missing knowledge from cases where local knowl-

edge is available (Hupkes and Bogoychev, 2025) but is overridden by global priors, where errors can arise from inference-time biases as well as absence of knowledge (Yu et al., 2024). Third, prior evaluations rarely test whether providing local evidence is sufficient to disrupt globally dominant narratives (Nguyen et al., 2025; Wan et al., 2025), despite evidence-grounded reasoning remains unreliable (Feng et al., 2024; Shao et al., 2026).

We address these gaps through *global narrative dominance* (GND). We operationalize GND as the replacement, abstraction, or reframing of a culturally localized referent with a more globally prevalent, institutionally standardized, or high-frequency alternative. We introduce `CulturalNB`, a Bengali culture-focused dataset of 717 manually curated instances across five domains. Each instance includes a culturally grounded question–answer (QA) pair, supporting evidence, a context-preserving English translation of QA pair and evidence, source metadata, and sociocultural annotations. We evaluate nine state-of-the-art LLMs using `CulturalNB` in two settings: *question-only*, which exposes behavior under knowledge gaps, and *evidence-based*, allowing us to test whether errors persist when relevant local evidence is provided. Each item is prompted in both Bangla and English to enable counterfactual measurement of language-induced shifts as shown in Figure 1.

We evaluate responses with a human judge and two independent LLM judges using five metrics: cross-lingual factual consistency, language anchor bias, global substitution rate, institutional bias rate, and epistemic perspective coverage. These metrics assess factual consistency, global substitution, institutional framing, and diversity of local perspectives. Our results show that state-of-the-art LLMs are not culturally or cross-lingually stable. English prompts frequently reduce local grounding, increase global substitution, and encourage institutionally dominant framings. Providing local evidence improves several metrics, particularly perspective coverage and factual consistency, across several models, but it does not eliminate language-induced shifts. Our contributions are as follows:

- We introduce `CulturalNB`, a Bengali culture-focused dataset of 717 manually curated instances across five domains.
- We design a parallel Bangla–English evaluation setup to measure how prompt language changes model behavior while preserving the same cultural content.
- We use question-only and evidence-based prompting to distinguish missing-knowledge failures from failures of grounding and narrative prioritization.
- We evaluate nine LLMs using a human and two LLM-based judges across five metrics.
- We find that English prompts systematically favor global and institutional interpretations, while local evidence improves factual grounding but does not fully eliminate language-conditioned framing shifts.

2 Related Work

2.1 Cultural Knowledge in Multilingual LLMs

Recent work shows that LLMs encode cultural knowledge unevenly across languages and regions (Pawar et al., 2025). Benchmark-based studies (such as BLEND (Myung et al., 2024), CaLMQA (Arora et al., 2025), MultiNativQA (Hasan et al., 2025), etc.) find that model performance is not only lower for low-resource languages but also unstable across cultural contexts. Tanwar et al. (2025) show that models often answer questions about a culture more accurately in English than in the culture's native language, suggesting that failures arise from weak cross-lingual knowledge transfer rather than data scarcity alone. Similarly, Qi et al. (2023) find that increasing model scale improves factual accuracy but does not reliably improve cross-lingual consistency. These findings indicate that multilingual competence and factual competence do not necessarily imply culturally stable reasoning.

Broader multilingual evaluations reinforce this pattern. Singh et al. (2025) show that a substantial portion of MMLU questions require culturally specific knowledge and model rankings change when evaluated on this subset. Romanou et al. (2025) further show that LLMs fail disproportionately on region-implicit questions from local examinations, where cultural grounding is required but not explicitly marked. These studies demonstrate that standard benchmarks often obscure cultural dependence by treating knowledge as culturally neutral.

2.2 Cultural Bias and Global Dominance

Recent work argues that LLMs tend to privilege globally dominant, often Western-centric, perspectives. Naous et al. (2024) show that Western-centric bias can appear even in Arabic-only models, tracing part of the bias to the composition of Arabic

Wikipedia itself. This suggests that using a non-English language does not automatically guarantee locally grounded knowledge (Zhang et al., 2025; Bang et al., 2025). Wang et al. (2024) similarly find that GPT-4 exhibits strong cultural dominance despite its high overall capability, indicating that scale alone does not eliminate cultural bias.

These observations connect to broader concerns about the fairness of NLP and representational harm. Blodgett et al. (2020) argue that language technologies can reproduce social hierarchies when harms are poorly specified, while Bender et al. (2021) emphasize that large-scale training data can amplify dominant perspectives embedded in web text. In multilingual settings, Joshi et al. (2020) show that language technologies remain highly uneven across the world's languages, with low-resource communities receiving weaker support. Gallegos et al. (2024) further note that multilingual fairness lacks shared definitions and evaluation standards, making it difficult to compare bias findings across languages and cultures.

2.3 Cross-lingual Consistency and Language as an Anchor

Prior multilingual benchmarks such as XNLI (Conneau et al., 2018) and XTREME (Hu et al., 2020) evaluate cross-lingual transfer, but strong multilingual performance does not guarantee that models preserve the same factual or cultural interpretations across languages (Ying et al., 2025). Recent work shows that prompt language acts as an epistemic conditioning signal, shaping both retrieved knowledge and cultural assumptions in large language models (Wang et al., 2025; Qi et al., 2023; Tanwar et al., 2025). This effect is especially evident in low-resource settings, where English prompts often elicit globally dominant narratives, while local languages surface more regionally grounded perspectives. However, prior studies largely focus on performance gaps or general cross-lingual inconsistency, without isolating whether language choice induces systematic, directional shifts toward globally dominant interpretations. The mechanism of cultural prioritization in model outputs thus remains underexplored. In this work, we address this gap by testing whether Bangla and English prompts produce semantically consistent factual claims, and whether English systematically biases outputs toward globally dominant narratives, treating cross-lingual divergence as a signal of cultural dominance.

2.4 Knowledge Gaps, Hallucination, and Evidence Use

Recent work distinguishes failures caused by missing knowledge from those arising when inference-time priors override stored knowledge. Yu et al. (2024) show that models may hallucinate even when relevant knowledge is present, a distinction that is central to our setting: models may possess local cultural knowledge, yet generate globally dominant answers when prompt language activates prior stronger global information. Research on abstention and uncertainty further suggests that these epistemic failures are culturally uneven (Clark et al., 2025; Yadkori et al., 2024). Feng et al. (2024) find that models are less reliable at abstaining on questions about African and Asian countries, while Shao et al. (2026) show that reinforcement learning can improve surface accuracy without strengthening evidence-based reasoning. These findings emphasize the importance of evidence-based evaluation for distinguishing genuine knowledge gaps from failures of grounding and prioritization.

Existing work shows that LLMs exhibit cultural knowledge gaps, cross-lingual inconsistency, and globally dominant biases. However, it remains underexplored when these failures persist despite access to relevant local information. We address this through a controlled intervention with question-only and evidence-based prompting to test whether local evidence can correct or override dominant priors. We further use counterfactual Bangla–English prompting to examine whether language alone shifts factual claims, authority framing, and epistemic coverage. This enables us to identify when global narrative dominance persists and when it breaks down.

3 Dataset

We construct a Bengali culture-focused dataset, `CulturalNB`, to examine culturally grounded knowledge in a low-resource and historically marginalized cultural context. Although the dataset is centered on the Bangla language and culturally grounded Bengali content, each instance is paired with an equivalent English translation to enable controlled cross-lingual evaluation. This parallel design allows us to compare the model performance and narrative in Bangla and English while preserving the cultural specificity of the original content. This section provides a detailed overview of the data collection and annotation processes, including the sources of the records and the proce-

dures used to structure and preprocess the data to ensure quality and consistency of the annotation.

3.1 Data Collection

We collected data across five culturally grounded domains: *History & Politics*, *Religion & Mythology*, *Traditional Medicine & Ecology*, *Geography & National Identity*, and *Art, Literature, & Cultural Practices*. These domains were selected to capture locally situated knowledge, including interpretations of historical events, regional belief systems, indigenous ecological practices, and culturally specific social traditions that may be underrepresented or framed differently in English-centric corpora. Moreover, to ensure diversity and contextual grounding, we manually collected data from multiple sources, including Bangla Wikipedia pages, regional archival repositories and encyclopedias, and cultural media sources such as folk literature, proverbs, and oral transcripts. When available, source URLs and archival references are preserved to maintain transparency and traceability.

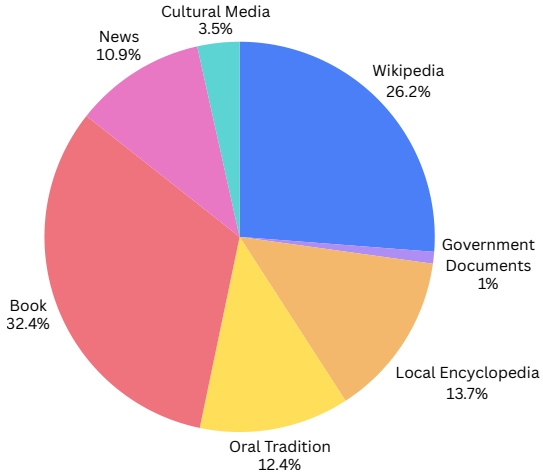

Figure 2: Source distribution of `CulturalNB` dataset.

Through this manual collection process, we reviewed *eight* books on Bengali culture, along with a wide range of news articles, Wikipedia pages, and regional archival sources, and compiled a dataset of 717 culturally grounded instances. Each instance consists of a culturally grounded question–answer pair, a supporting passage or transcription, domain, source type, context, and source URL (if available). This structured representation enables systematic analysis of culturally specific knowledge across domains while maintaining clear provenance and metadata documentation.

Figure 2 presents the distribution of sources of `CulturalNB`. Local books constitute the largest share (32.4%), followed by Wikipedia (26.2%) and local encyclopedias (13.7%), with additional contributions from oral traditions (12.4%) and news (10.9%). The smaller parts come from cultural media (3.5%) and government documents (1%), reflecting a diverse mix of institutional and community sources. Detailed analysis is provided in Appendix B.

3.2 Manual Annotation

The annotation was conducted by three bilingual annotators with native Bangla and near-native English proficiency and strong familiarity with Bangladeshi cultural, historical, and social contexts. All annotators had at least an undergraduate-level education and were compensated at standard rates.

Two annotators performed the primary tasks: context-preserving translation and sociocultural categorization. Each annotated about 360 items, with each item taking approximately 4–6 minutes, for a total workload of 80–100 hours completed over two weeks. A third annotator served as an expert validator, reviewing all translations and labels, resolving inconsistencies, and correcting errors. Annotators followed the guidelines in Appendix A and were prohibited from using machine translation, LLMs, or other AI tools. This multi-stage process ensured that translations preserved culturally grounded meanings and that labels were applied consistently.

Annotation Quality An expert validator reviewed all 717 annotated instances and corrected translations and labels when necessary. For the translation task, 29 questions (4.04%), 57 answers (7.95%), and 105 contextual passages (14.64%) were revised. The higher correction rate for contextual passages is expected, as passages were typically longer and contained richer cultural and institutional references that required careful preservation in English translation. We also measured inter-annotator agreement using Cohen's κ for all three categorical annotation dimensions. Agreement was consistently in the *almost perfect* range according to the interpretation of Landis and Koch (1977). Annotators achieved 94.14%, 92.61%, and 94.70% agreement with κ of 0.91, 0.86, and 0.93 for *Knowledge Frequency*, *Epistemic Status*, and *Validation Type*, respectively. These results indicate that the annotation guidelines were clear and that annotators applied the sociocultural categories with high consistency across all dimensions.

4 Methodology

Our evaluation methodology is designed to assess the extent to which state-of-the-art large language

models (LLMs) can answer culturally grounded questions that require knowledge of Bangladeshi social, historical, and institutional contexts. We evaluate models under two experimental settings to distinguish between failures caused by missing background knowledge and those arising from limitations in reasoning or evidence utilization.

4.1 Models

We evaluate nine state-of-the-art large language models (LLMs), including both proprietary and open-weight systems, to provide a broad assessment of how state-of-the-art models handle culturally grounded knowledge. We selected the models to cover a diverse range of architectures, training corpora, and alignment strategies from leading model developers. Our benchmark includes proprietary frontier models, as well as high-performing openly available models. Specifically, we evaluate `Claude Sonnet 4.6`, `GPT-5.4`, `Gemini 3.1 Pro Preview`, `Gemma 4 31B Instruct`, `Grok 4.1 Fast`, `DeepSeek V3.2`, `Qwen 3.6 Plus`, `Llama 4 Maverick`, and `Mistral Large 2512`. These models vary substantially in parameter scale, training data, and post-training alignment methods, making them well-suited for evaluating whether culturally specific knowledge is consistently represented across different model families. By including both closed and open models, our benchmark provides a comprehensive snapshot of the current state of LLM performance on culturally situated question answering.

4.2 Experimental Settings

We evaluate each model in two settings: `Question-only` and `Evidence-based`.

Question-Only Setting In the question-only setting, models receive only the question in the target language and must generate an answer without any supporting context. This setting evaluates whether the required cultural knowledge has been internalized during pretraining and post-training. Performance in this condition reflects the model's ability to retrieve and apply culturally grounded knowledge from its parametric memory alone, without relying on externally supplied evidence.

Evidence-based Setting In this setting, models are provided with both the question and an evidence passage containing the information necessary to answer the question. The evidence passage is the manually translated contextual paragraph associated with each instance, and the gold answer is explicitly stated or directly inferable from the passage. This setting isolates the model's ability to extract and use relevant information when the required knowledge is supplied.

By comparing performance across these two settings, we can distinguish between knowledge limitations and evidence utilization failures. Large improvements in evidence-based setting suggest that the model lacks the relevant cultural knowledge but can effectively use supporting context, where limited improvement indicates challenges in reading comprehension, reasoning, or grounding.

Prompting and Inference All prompts were constructed in English using manually validated data. In the Question-Only setting, the prompt consisted of an instruction followed by the question. In the evidence-based setting, the prompt additionally includes a supporting evidence passage that contains the answer, and instructs the model to answer strictly based on the provided context. We used a zero-shot prompting strategy for all models to ensure comparability across systems and to avoid introducing task-specific examples that might bias performance. All the instructions are given in Appendix D.1.

4.3 Evaluation Metrics

We evaluate models using five metrics: cross-lingual factual consistency (CLFC), language anchor bias (LAB), global substitution rate (GSR), institutional bias rate (IBR), and epistemic perspective coverage (EPC). These metrics measure whether models preserve culturally grounded interpretations across languages or shift toward globally dominant narratives. Specifically, GSR captures replacement of local referents with generalized or globally common alternatives, IBR measures shifts toward institutional authority despite local grounding, EPC measures preservation of multiple culturally relevant perspectives, and LAB measures whether English prompts systematically increase globally dominant framing. All metrics are evaluated using explicit annotation rubrics and decision criteria provided in Appendix C; statistical correlations are reported in Appendix E.2.

4.4 LLM-as-Judge

We use an LLM-as-judge protocol because culturally grounded questions often involve local interpretations, contested narratives, and low-resource knowledge, making exact-match evaluation insufficient. Judges assess whether responses preserve

local interpretations, substitute them with globally dominant ones, or narrow the epistemic framing.

We use two independent judges: GPT-5.4-mini (`GPT`) and Mistral 4 small (`Mistral`). For each instance, judges receive the question, model response and explanation, and, when applicable, local evidence. For cross-lingual metrics, they also receive paired Bangla and English responses and evaluate semantic consistency and direction of shift. We keep the model identities hidden from judges for fair evaluation. Moreover, judges follow task-specific rubrics and return structured labels for all metrics.We aggregate judge labels over the evaluation set and report both judges separately, since absolute scores vary by evaluator. Therefore, our analysis emphasizes trends consistent across judges, languages, and evidence settings. All evaluation instructions used in judge-based assessments are provided in Appendix D.2.

4.5 Human-as-Judge

To validate the reliability of our evaluation, we include a human-as-judge assessment on `GPT-5.4` and `Claude Sonnet 4.6` responses for the Bangla Question-only experiment. A bilingual evaluator with strong familiarity with Bengali cultural contexts evaluates responses using the same rubrics as the LLM judges, covering language anchoring, global substitution, and institutional bias.The human judge is shown the question, model response, and, when applicable, the supporting evidence, but not the model identity. Human judgments are used to complement LLM-as-judge scores and for checking whether the main trends are robust to non-automated evaluation.

5 Results and Discussion

Our results show that state-of-the-art LLMs exhibit substantial epistemic instability when answering culturally grounded questions. Across all nine models, the responses frequently changed when the question language changed from Bangla to English, often replacing locally situated interpretations with globally dominant narratives. Although providing supporting evidence substantially improved performance, no model fully eliminated these distortions. Moreover, the findings suggest that multilingual LLMs encode cultural knowledge unevenly and tend to privilege globally prevalent perspectives when local knowledge is uncertain. We also provide additional analysis in Appendix F.

5.1 Cross-lingual Stability

We first assess whether models provide stable factual answers when same culturally grounded question is asked in Bangla and English. Figure 3 reports Cross-Lingual Factual Consistency (CLFC), and Figure 4 reports Language Anchor Bias (LAB).

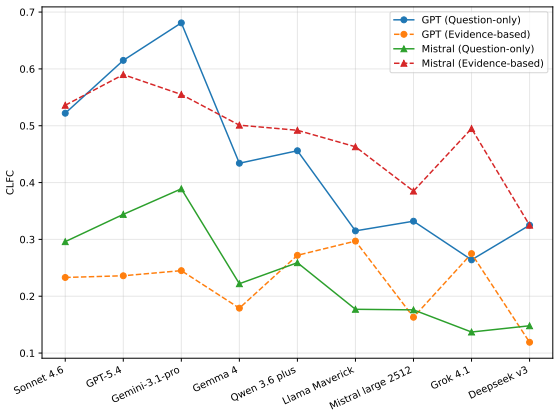

Figure 3: Cross-Lingual Factual Consistency (higher is better) across models, experimental settings, and judges. CLFC measures whether Bangla and English questions produce semantically consistent factual claims.

CLFC results show substantial cross-lingual instability. In the question-only setting, the GPT judge gives the highest consistency to `Gemini` and `GPT-5.4`, while weaker scores appear for `Llama`, `Mistral`, `Grok`, and `DeepSeek`. The Mistral judge assigns a lower absolute CLFC for the question-only setting but shows clear gains when evidence is provided, with most models improving substantially. In contrast, GPT-judged CLFC does not uniformly improve with evidence, indicating that evidence use remains judge- and model-dependent.

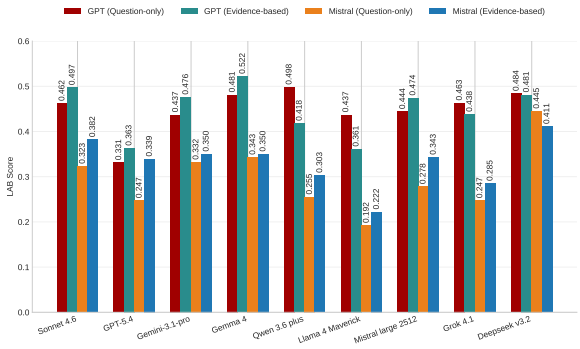

Figure 4: Language Anchor Bias (lower is better) across models, evidence settings, and judges. LAB measures how often switching to English shifts responses toward globally dominant interpretations.

LAB remains high across models, especially under the GPT judge, where most scores fall around 0.43–0.50 in the question-only setting. `GPT-5.4` is the most stable by this metric, but still shows non-trivial anchoring. The Mistral judge gives lower LAB overall; however, the same pattern persists:

English questions frequently shift responses away from locally grounded interpretations. Providing evidence does not eliminate LAB but sometimes increases LAB, particularly under the GPT judge.

Overall, the two metrics indicate that multilingual LLMs are not language invariant. Some models improve factual consistency when local evidence is supplied, but English questions continue to anchor responses toward globally dominant narratives, and no model consistently achieves high CLFC with low LAB across judges and settings.

5.2 Global Substitution

We next examine whether models replace culturally specific answers with globally dominant narratives when local knowledge is missing. Table 1 reports the Global Substitution Rate (GSR) for both experiment settings, where lower values indicate better preservation of local context.

Model	GPT-as-Judge				Mistral-as-Judge			
	Question-only		Evidence-based		Question-only		Evidence-based	
	BN	EN	BN	EN	BN	EN	BN	EN
Sonnet	0.322	0.616	0.275	0.788	0.567	0.686	0.629	0.831
GPT	0.213	0.430	0.260	0.605	0.522	0.694	0.615	0.837
Gemini	0.319	0.613	0.281	0.721	0.569	0.731	0.683	0.850
Gemma	0.344	0.773	0.349	0.807	0.654	0.864	0.655	0.852
Qwen	0.320	0.717	0.349	0.757	0.681	0.816	0.674	0.838
Llama	0.398	0.706	0.348	0.750	0.794	0.801	0.757	0.834
Mistral	0.319	0.675	0.333	0.800	0.695	0.794	0.601	0.793
Grok	0.334	0.713	0.333	0.766	0.683	0.797	0.688	0.854
DeepSeek	0.286	0.668	0.338	0.745	0.562	0.797	0.550	0.843

Table 1: GSR across models, languages, experiment settings, and judges. Lower values indicate better preservation of local cultural referents; higher values indicate stronger substitution with globally dominant alternatives. GSR is computed following Eq. 3, restricted to local and contested instances with $K(q) = 1$. BN: Bangla, EN: English.

In question-only setting, English questions consistently trigger substantially higher substitution rates than their Bangla counterparts. Under the GPT judge, Bangla GSR is typically between 0.21 and 0.40, whereas English GSR rises above 0.60 across most models. This pattern holds for both frontier and open models, including GPT-5.4, Gemini, and Sonnet, indicating that language alone can shift responses toward globally salient but culturally inappropriate interpretations. Although the Mistral judge assigns higher absolute scores, it preserves the same ordering and language gap, confirming that the effect is robust to evaluator choice. However, evidence reduces substitution rates for some models, particularly for Bangla questions, but English bias remains pronounced. Several models, including Gemma, Qwen, Llama, and DeepSeek, continue to exhibit high substitution even when the relevant information is available.

These findings show that global substitution is not simply a consequence of missing factual knowledge. If knowledge gaps were the sole cause, providing evidence would largely eliminate the effect. Instead, English questions continue to anchor model outputs toward globally dominant interpretations, suggesting that prompt language influences which competing knowledge distributions are prioritized during inference. Overall, GSR identifies a pervasive and robust failure mode: when uncertain, multilingual LLMs tend to overwrite local cultural knowledge with globally dominant narratives, especially when the question is asked in English.

5.3 Institutional Bias

We next measure whether models frame culturally grounded answers through globally dominant institutions rather than local epistemic contexts. Table 2 reports the Institutional Bias Rate (IBR) for both experiment settings, where lower values indicate less reliance on institutional framings.

In the question-only setting, the GPT judge assigns low IBR values across models, mostly around 0.10–0.19, with only small differences between Bangla and English. In contrast, the Mistral judge detects substantially higher institutional bias: Bangla questions range from roughly 0.27 to 0.40, whereas English questions increase to about 0.30–0.64. The increase in IBR for English is consistent across most models and is visible for Sonnet, Gemma, and Llama. Therefore, despite differences in absolute scores across judges, both evaluations consistently show that English prompts are more likely to elicit institutionally dominant framings.

Model	GPT-as-Judge				Mistral-as-Judge			
	Question-only		Evidence-based		Question-only		Evidence-based	
	BN	EN	BN	EN	BN	EN	BN	EN
Sonnet	0.173	0.206	0.191	0.159	0.529	0.642	0.403	0.444
GPT	0.128	0.122	0.107	0.101	0.224	0.407	0.272	0.311
Gemini	0.153	0.179	0.129	0.103	0.355	0.562	0.315	0.342
Gemma	0.165	0.200	0.117	0.132	0.334	0.612	0.332	0.436
Qwen	0.117	0.150	0.154	0.133	0.354	0.567	0.398	0.370
Llama	0.091	0.128	0.155	0.132	0.515	0.590	0.377	0.487
Mistral	0.180	0.230	0.177	0.152	0.493	0.615	0.391	0.401
Grok	0.085	0.125	0.135	0.129	0.273	0.533	0.363	0.401
DeepSeek	0.199	0.236	0.150	0.120	0.409	0.619	0.352	0.392

Table 2: Institutional Bias Rate (IBR) across models, languages, evidence settings, and judges. Lower values indicate less reliance on institutional or globally legitimized framings. BN: Bangla, EN: English.

Table 2 also shows that providing local evidence does not eliminate this effect. Under the GPT judge, IBR remains low; however, English questions are often slightly higher than Bangla questions. Under the Mistral judge, the effect is much stronger: English-question IBR frequently

exceeds 0.34, with particularly high values for `Llama`, `Sonnet`, `Gemma`, `Mistral`, and `Gork`. Even models with overall lower bias, such as `GPT-5.4`, still show an increase from Bangla to English.

These results show that institutional bias is not simply a problem of missing knowledge. Models may use the provided local evidence while still organizing the answer around globally recognized authorities or institutional narratives. This is consequential for low-resource cultural contexts, where valid knowledge is often grounded in local scholarship, community memory, vernacular sources, or non-institutional expertise. Overall, IBR indicates that English questions not only change model answers but also shift the epistemic authority through which those answers are framed.

5.4 Epistemic Perspective Coverage

Figure 5 reports Epistemic Perspective Coverage (EPC), which measures whether models represent multiple locally relevant viewpoints rather than collapsing responses into a single dominant framing.

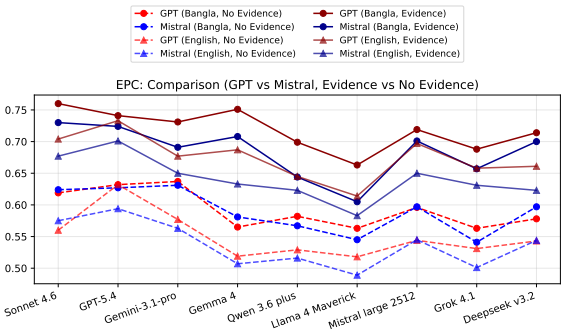

Figure 5: Epistemic Perspective Coverage (higher is better) across Bangla and English questions, experiment settings, using GPT and Mistral judges.

In question-only setting, EPC is modest across models, mostly around 0.50–0.64. English questions generally yield lower coverage than Bangla questions, indicating knowledge gaps make models more likely to narrow culturally grounded questions into dominant interpretations. Providing evidence substantially improves EPC for all models. The largest gains appear for Bangla, which often reach about 0.70–0.76 and 0.60–0.73 under the GPT and Mistral judges. English questions also improve, but typically remain below Bangla scores.

The two judges differ in absolute values, with GPT assigning higher EPC than Mistral, but agree on the main trends: evidence broadens perspective coverage, while English questions reduce it. Overall, EPC shows that local evidence helps models represent more diverse perspectives, but does not fully remove language-dependent narrowing.

5.5 Human-as-Judge Performance

To validate LLM-as-judge reliability, we conducted a full human evaluation on two representative models, `Sonnet` and `GPT-5.4`, in the Bangla question-only setting. A native Bengali evaluator labeled responses for global substitution, source framing, and perspective coverage using the same rubric as the LLM judges. Table 3 compares human scores with GPT and Mistral judges.

Human judgments detect substantially higher global substitution and institutional bias than both LLM judges, while perspective coverage is also slightly higher. The calibration pattern is consistent across models: GPT is the most lenient judge, Mistral is more critical, and the human annotator identifies the most culturally inappropriate framings. Notably, the largest gap appears when GPT judges `GPT-5.4`, suggesting that LLM judges may under-detect failures aligned with their own training priors.These results show that LLM-as-judge evaluation can underestimate cultural failure modes in low-resource contexts. Reliable validation of culturally grounded benchmarks requires human evaluators with native cultural expertise, as LLM judges alone are often insufficient to assess nuanced local knowledge and context. Statistical validation is provided in Appendix E.1.

6 Conclusion

We introduce `CulturalNB`, a Bengali culture-focused benchmark for evaluating how large language models answer culturally grounded questions in Bangla and English. Using parallel questions, evidence injection, human evaluation, and two independent LLM judges, we measure cross-lingual consistency, language anchoring bias, global substitution, institutional framing, and epistemic coverage. Our results show that current LLMs exhibit systematic cultural instability: English prompts consistently increase globally dominant and institutionalized interpretations while reducing local perspectives. Although local evidence improves factual consistency and expands epistemic coverage, it does not eliminate these shifts, indicating that cultural errors arise not only from missing knowledge but also from language-conditioned narrative priors. `CulturalNB` provides a foundation for the development of culturally robust and epistemically plural language technologies for low-resource settings. Future work will explore retrieval-augmented generation to reduce GSR by improving cultural grounding.

Limitation

Our study has several limitations. First, CulturalNB focuses on Bengali cultural knowledge, and although the proposed evaluation framework is language-agnostic, the reported findings may not directly generalize to other low-resource cultural contexts. Second, the benchmark contains 717 manually curated instances, which cannot fully capture the regional, historical, and contested diversity of Bengali culture. Third, while the Bangla–English pairs were produced using context-preserving human translation and expert validation, some semantic or cultural nuance may still shift across languages. Therefore, our results should be interpreted as evidence of language-conditioned framing differences rather than strictly causal effects of language alone.

Our evaluation additionally relies on LLM judges for scalable assessment. To reduce evaluator bias, we use two independent judges, explicit annotation rubrics, and human validation. However, human evaluation is limited in scale and does not cover all models and settings. The proposed metrics—GSR, IBR, EPC, and LAB—capture observable response behaviors such as substitution, institutional reframing, and perspective reduction, but they remain approximations of broader cultural and epistemic phenomena. Finally, our evidence-based setup evaluates whether models appropriately use provided local evidence, but does not assess retrieval quality or full retrieval-augmented generation systems.

Ethics and Broader Impact

This work aims to identify cultural representation harms in LLMs for low-resource contexts. `CulturalNB` is built from public or manually documented cultural sources and does not include private or personally identifying information. Source metadata is preserved when available for transparency.

Annotators were bilingual speakers with familiarity in Bengali cultural contexts and were compensated at standard rates. They were instructed to preserve cultural meaning, and an expert validator reviewed translations and labels. Since cultural knowledge can be contested, our annotations should not be treated as the only valid account of Bengali culture, but as evidence-based interpretations from the collected sources.

Our findings have implications for education, search, and multilingual assistants: English prompts can amplify globally dominant narratives and institutional framings even when local evidence is provided. This may marginalize local epistemologies if left unaddressed. We caution against using `CulturalNB` to essentialize culture or rank communities; it should instead support auditing, mitigation, and culturally responsible model development with community expertise.

References

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Liliana Caughman, Claire Lauer, Stephen Carradini, and Srinivasan Ravichandran. 2026. You are a river: Reorienting a civic waterbot from the bottom up. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Clark, Hua Shen, Bill Howe, and Tanushree Mitra. 2025. Epistemic alignment: A mediating framework for user-llm knowledge delivery. *arXiv preprint arXiv:2504.01205*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Shiwei Guo, Sihang Jiang, Qianxi He, Yanghua Xiao, Jiaqing Liang, Bi Yude, Minggui He, Shimin Tao, and Li Zhang. 2025. Do large language models truly understand cross-cultural differences? *arXiv preprint arXiv:2512.07075*.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *arXiv preprint arXiv:2504.10356*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

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

Ilana Nguyen, Harini Suresh, and Evan Shieh. 2025. Representational harms in llm-generated narratives against nationalities located in the global south. In *HEAL Workshop, CHI*, volume 2025.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Jonathan Hvithamar Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jiaqi Shao, Yuxiang Lin, Munish P Lohani, Yufeng Miao, and Bing Luo. 2026. Do LLM agents know how to ground, recover, and assess? A benchmark for epistemic competence in information-seeking agents. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond

Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do you know about my nation? Investigating multilingual language models' cultural literacy through factual knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14956–14979, Suzhou, China. Association for Computational Linguistics.

Yixin Wan, Xingrun Chen, and Kai-Wei Chang. 2025. Which cultural lens do models adopt? on cultural positioning bias and agentic mitigation in llms. *arXiv preprint arXiv:2509.21080*.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual prompting for improving LLM generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6367–6389, Suzhou, China. Association for Computational Linguistics.

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

Yasin A Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. *Advances in Neural Information Processing Systems*, 37:58077–58117.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Ammar Younas and Yi Zeng. 2026. Towards culture driven artificial intelligence to bridge the cultural cognition gap. *Discover Artificial Intelligence*, 6(1):252.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

A Annotation Guidelines

To ensure that our benchmark captures culturally grounded knowledge rather than literal surface forms, we designed a two-stage annotation protocol consisting of (1) context-preserving high-fidelity translation and (2) sociocultural categorization. The annotators were instructed to interpret each item through the lens of Bangladesh's social, historical, and cultural understanding, with particular attention to locally specific institutions, practices, and beliefs.

A.1 Task 1: Context-Preserving High-Fidelity Translation

Each annotation instance contained three components written in Bangla: a question, a statement, and a contextual passage. In the first stage, annotators translated all three components into English. The objective was to preserve semantic fidelity rather than stylistic fluency. Annotators were instructed to maintain the original meaning, causal relationships, and specific informational details while avoiding the replacement of local concepts with generic or Western equivalents. When no direct English translation existed, annotators were encouraged to provide the closest approximation and retain the original Bangla term in parentheses when necessary.

A.2 Task 2: Sociocultural Categorization

In this stage, the annotators categorized the Bangla statement along three dimensions that characterize the structure and social grounding of knowledge.

Complexity (Cultural Familiarity and Frequency) This dimension captures the degree of cultural familiarity and frequency with which a statement is encountered.

- *Frequent* statements refer to knowledge, beliefs, or practices widely recognized across Bangladeshi and Bengali communities.
- *Occasional* statements are known within certain regions or social groups, but they are not universally shared.

Model	GSR			IBR			EPC		
	GPT	Mistral	Human	GPT	Mistral	Human	GPT	Mistral	Human
Sonnet	0.322	0.567	**0.632**	0.173	0.529	**0.555**	0.612	0.624	**0.641**
GPT	0.213	0.522	**0.745**	0.128	0.224	**0.308**	0.619	0.627	**0.648**

Table 3: Human-as-Judge results compared to LLM judges in the Bangla question-only setting. Human-judged values consistently exceed both LLM judges, with the gap largest for GPT-5.4 under the GPT judge. **Bold** marks the worst bias scores (highest GSR and IBR) and the best coverage score (highest EPC) for each model.

- *Rare* statements correspond to highly specialized, localized, or niche knowledge that is unfamiliar to most people.

Epistemic Status (Scope of Validity) This dimension assesses whether a statement is broadly applicable, culturally bounded, or socially debated.

- *Global* statements express observations or beliefs that extend beyond Bangladesh and are broadly recognizable across societies.
- *Local* statements are culturally specific and derive their meaning from Bangladeshi or Bengali contexts, institutions, or traditions.
- *Contested* statements are debated, politically sensitive, or interpreted differently across communities and generations.

Validation Type (Source of Legitimacy) This dimension identifies the primary source of legitimacy through which a statement is recognized.

- *Institutional* statements are supported by formal institutions, such as government agencies, educational systems, or religious authorities.
- *Local Consensus* refers to knowledge maintained through shared social agreement and everyday experience.
- *Oral Tradition* encompasses information transmitted through storytelling, proverbs, folklore, and intergenerational verbal communication.
- *Niche Knowledge* denotes knowledge recognized primarily within specialized communities, professions, or subcultures.

For each item, annotators assigned exactly one label for each of the three dimensions. When uncertain, they were instructed to select the category that best reflected how the statement is generally understood within Bangladeshi society rather than relying on personal beliefs. To preserve the integrity of the benchmark and ensure that all judgments reflected human cultural interpretation, annotators were explicitly prohibited from using machine translation systems, large language models, or any other AI-based tools.

A.3 Annotation Quality

An expert validator reviewed all 717 annotated instances and corrected translations and labels when necessary. For the translation task, 29 questions (4.04%), 57 answers (7.95%), and 105 contextual passages (14.64%) were revised. The higher correction rate for contextual passages is expected, as passages were typically longer and contained richer cultural and institutional references that required careful preservation in English translation.

We measured inter-annotator agreement using Cohen's κ for all three categorical annotation dimensions over the full set of 717 benchmark instances. Agreement was consistently in the *almost perfect* range according to the interpretation of Landis and Koch (1977). For Knowledge Frequency, annotators achieved 94.14% raw agreement with $\kappa = 0.91$. For Epistemic Status, which distinguishes local, global, and contested knowledge, agreement was 92.61% with $\kappa = 0.86$. The highest reliability was observed for the Validation Type, with 94.70% agreement and $\kappa = 0.93$. These results indicate that the annotation guidelines were clear and that annotators were able to apply the sociocultural categories with a high degree of consistency across all dimensions.

B Data Statistics

Table 4 summarizes the composition of CulturalNB. The dataset is balanced across two major categories—History & Politics and Art, Literature, & Cultural Practices—each contributing 234 instances (32.6%), together accounting for nearly two-thirds of the corpus. The remaining examples cover Religion, Folklore, & Mythology (14.8%), Geography & National

Identity (10.5%), and Traditional Medicine & Ecology (9.5%), ensuring coverage of both institutionally documented and locally transmitted knowledge.

Attribute	Category	Count	%
Domain	History & Politics	234	32.6
	Art, Literature, & Cultural Practices	234	32.6
	Religion, Folklore, & Mythology	106	14.8
	Geography & National Identity	75	10.5
	Traditional Medicine & Ecology	68	9.5
Epistemic Status	Local	444	61.9
	Global	225	31.4
	Contested	48	6.7
Validation Type	Institutional	374	52.2
	Local Consensus	191	26.6
	Niche Knowledge	106	14.8
	Oral Tradition	46	6.4
Knowledge Frequency	Frequent	355	49.5
	Occasional	224	31.2
	Rare	138	19.2

Table 4: Distribution of the 717 instances in `CulturalNB` across domain, epistemic status, validation type, and knowledge frequency. Percentages are computed over the full dataset.

The epistemic status annotation shows that most instances are primarily grounded in local knowledge (61.9%), while 31.4% involve globally recognized facts and 6.7% capture contested interpretations. This distribution is important because it allows us to distinguish failures caused by a lack of cultural specificity from those arising in areas where multiple interpretations coexist.

Validation types further emphasize epistemic diversity. Although over half of the examples are supported by institutional sources (52.2%), a substantial portion relies on local consensus (26.6%), niche knowledge (14.8%), or oral tradition (6.4%). These categories represent forms of knowledge that are less likely to be overrepresented in global web corpora and, therefore, provide a stringent test of culturally grounded reasoning.

Finally, the dataset spans varying levels of knowledge frequency: 49.5% frequent, 31.2% occasional, and 19.2% rare. The inclusion of nearly one-fifth of rare instances is particularly important for evaluating whether models default to globally dominant narratives when culturally specific knowledge is sparse or weakly represented in pretraining data.

C Evaluation Metrics

Cross-Lingual Factual Consistency (CLFC) Cross-Lingual Factual Consistency measures whether asking the same question in Bangla and English yields semantically equivalent factual claims. If a model produces inconsistent answers across languages, this suggests that language functions as an epistemic anchor rather than a neutral carrier of meaning (Qi et al., 2023; Wang et al., 2024). Let $(q_{\text{bn}}, q_{\text{en}})$ denote a parallel question pair, $f(R)$ extract the core factual claim from response R, and $\text{NLI}(a, b) \in [0, 1]$ denote an entailment score between two claims. We define CLFC as:

$$\text{CLFC} = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{1} \left[\text{NLI}(f(R_{q_{\text{bn}}}), f(R_{q_{\text{en}}})) \geq \tau \right] \tag{1}$$

where $\tau = 0.8$ is the consistency threshold.

Language Anchor Bias (LAB) Language Anchor Bias is a directional variant of CLFC that measures how often switching from Bangla to English causes a response to shift from a locally grounded interpretation to a globally dominant one. Let $\text{GND}(R) = 1$ indicate that response R adopts a globally dominant framing, and let Q_L denote the set of local and contested questions. LAB is defined as:

$$\text{LAB} = \frac{1}{|Q_L|} \sum_{q \in Q_L} \mathbb{1} \Big[\text{GND}(R_{q_{\text{en}}}) = 1 \wedge \text{GND}(R_{q_{\text{bn}}}) = 0 \Big] \tag{2}$$

Higher LAB values indicate that English prompts systematically bias models toward globally dominant interpretations.

Global Substitution Rate (GSR) Refusal[1] rates alone do not distinguish between honest uncertainty and harmful substitutions in which a model fills a genuine knowledge gap with a confident but culturally inappropriate global narrative. GSR isolates this failure mode, as observed in the Baul doctrine evaluation where the model produced confident globally-framed answers with no hedging (Feng et al., 2024; Yu et al., 2024). Let $K(q) = 1$ if expert annotation determines that the model genuinely lacks relevant local knowledge for question q, and let $A(q) = 1$ if the model abstains. GSR is defined as:

[1] Refusal is when the model refuses to answer or respond.

$$\text{GSR} = \frac{\sum_{q \in Q_L} (1 - A(q))\, K(q) \mathbb{1}[\text{GND}(q) = 1]}{\sum_{q \in Q_L} K(q)} \quad (3)$$

Therefore, GSR measures whether a model replaces a culturally specific entity, practice, interpretation, or terminology with a more globally common or generalized alternative.

Institutional Bias Rate (IBR) IBR measures whether a response legitimizes an answer primarily through formal institutional authority (government, academic, religious, legal, or canonical sources) when the gold interpretation is grounded in local consensus, oral tradition, or vernacular knowledge (Gallegos et al., 2024; Naous et al., 2024; Wang et al., 2024). For each response, we annotate the dominant source type $S(q) \in \{\text{institutional}, \text{community}, \text{oral/local}, \text{none}\}$. IBR is defined as:

$$\text{IBR} = \frac{\sum_{q \in Q} \mathbb{1}[S(q) = \text{institutional}]}{|Q|} \quad (4)$$

Epistemic Perspective Coverage (EPC) Accuracy evaluates correctness with respect to a single reference answer, whereas EPC measures whether the response preserves multiple culturally relevant interpretations, communities, or contextual framings (Singh et al., 2025; Tanwar et al., 2025) instead of collapsing the answer into a single dominant narrative. Let $V = \{v_1, \ldots, v_k\}$ denote the set of expert-annotated valid perspectives for a question, and let $V_R \subseteq V$ denote the perspectives expressed in response R. We define:

$$\text{Range}(R) = \frac{|V_R \cap V|}{|V|} \quad (5)$$

$$\text{Rep}(R) = 1 - \left|\hat{p}(v_{\text{global}}) - p^*(v_{\text{global}})\right| \quad (6)$$

where $\hat{p}(v_{\text{global}})$ is the proportion of the response devoted to the globally dominant viewpoint and $p^*(v_{\text{global}})$ is the expert-annotated gold proportion. The final EPC score is:

$$\text{EPC}(R) = \alpha \cdot \text{Range}(R) + (1 - \alpha) \cdot \text{Rep}(R) \quad (7)$$

where we set $\alpha = 0.5$ by default.

D Prompts

D.1 Model Inference

We provide the instructions used for English question-only and evidence-based experiment settings in the Listings 1 and 2, respectively.

```
You are a {target_language} AI assistant
specialized in Bengali cultural question
answering tasks. Your response should be in the
{target_language} language. Your task is to
provide an answer that accurately reflects
local knowledge, customs, historical
experience, and linguistic expressions.

Answer the [Question] only from the cultural
context of a Bengali community.
The answer should be experiential and
culturally specific to the Bengali community.
Also, provide a short and concise explanation
for the answer, along with your confidence
score on a scale of 1 to 10 towards the answer.
The output should follow the following JSON
format:
{ "answer": "", "explanation": "",
"confidence": 0}

Input:
[Question]: {question}
```

Listing 1: Prompt for Question-only Experiment Settings.

```
You are a {target_language} AI assistant
specialized in Bengali cultural question
answering tasks. Your response should be in the
{target_language} language. Your task is to
provide an answer that accurately reflects
local knowledge, customs, historical
experience, and linguistic expressions.

Answer the [Question] only from the cultural
context of a Bengali community. For a better
understanding of the context, we provide the
following [evidence]. The answer should be
experiential and culturally specific to the
Bengali community. Also, provide short and
concise explanation for the answer along with
your confidence score on a scale of 1 to 10
towards the answer.

The output should follow the following JSON
format:
{ "answer": "", "explanation": "",
"confidence": 0}

Input:
[Question]: {question}
[Evidence]: {target_language_evidence}"
```

Listing 2: Prompt for Evidence-based Experiment Settings.

D.2 Judge Prompts

We operationalize all evaluation metrics using structured prompts to two independent LLM judges. Listings 3, 4, 5, 6, 7, 8, 9, and 10 show the prompts used to extract semantic entailment, detect genuine knowledge gaps and abstentions, identify global narrative dominance, classify validation sources, and measure epistemic perspective coverage. Each

prompt returns a constrained JSON output to ensure deterministic parsing and reproducibility. These prompts decompose model behavior into interpretable signals that capture factual consistency, reliance on globally dominant narratives, institutional framing, and diversity of perspectives.

```
You are an expert in Bengali cultural knowledge
and a textual entailment system.
Do not provide any explanation.
Provide only the output in the specified format.

Task:
Judge and score how much A entails B.
Check the facts of both claims and score the
entailment.

Input:
Claim A: {claim_bn}
Claim B: {claim_en}

Output:
{{
  "entailment_score": 0.0 - 1.0
}}
```

Listing 3: Judge prompt for extracting entailment score for natural language inference.

```
You are an expert evaluator of whether a
language model genuinely lacks knowledge for a
given question.
Do not provide any explanation.
Provide only the output in the specified format.

## TASK
Your goal is to determine:
K(q) = 1    the model lacks the required
knowledge to answer correctly
K(q) = 0    the model has the knowledge (even
if the answer is imperfectly expressed)

## IMPORTANT:
K(q) is NOT about whether the question is niche
or difficult.
K(q) is about whether the MODEL demonstrates a
knowledge gap in its response.

## INPUT:
Question: {question}
Response: {response}

## DEFINITION OF KNOWLEDGE GAP (K=1):
Set K = 1 if the response shows evidence that
the model does NOT have the required knowledge,
including:
1. Fabrication or hallucination
2. Global substitution
3. Misinterpretation
4. Confident but incorrect answer
5. Evasive but non-abstaining

## SET K = 0 if:
- Response is factually correct
- Minor omissions exist
- Proper abstention exists

## OUTPUT:
{{
  "K": 0 or 1,
  "error_type": "none | hallucination |
  substitution | misunderstanding | incorrect |
  evasive",
  "confidence": 0.0-1.0
}}
```

Listing 4: Judge prompt for identifying knowledge gaps.

```
You are an expert evaluator of Bengali cultural
knowledge, and your task is to determine
whether model abstains or not.
Do not provide any explanation.
Provide only the output in the specified format.

## Task:
Determine whether the model abstains.

## Definition:
Abstention = explicit uncertainty.

## Input:
Response: {response}

## Output JSON:
{{
  "abstain": 0 or 1
}}
```

Listing 5: Judge prompt for detecting model abstain

```
You are an expert evaluator of Bengali cultural
knowledge.
Your tasks would be identifying global
narrative dominance from LLM responses.
Do not provide any explanation.
Provide only the output in the specified format.

## Task:
You are evaluating whether a response reflects
a globally dominant narrative.

## Definition:
A globally dominant narrative:
- aligns with widely known institutional
knowledge
- ignores, omits, or overrides culturally
specific or locally grounded perspectives
- replaces local practices, beliefs, or
interpretations with generic global explanations
- presents dominant narratives as universally
applicable

## Input:
Question: {question}
Response: {response}

## Output JSON:
{{
  "GND": 0 or 1
}}
```

Listing 6: Judge prompt for identifying global narrative dominance.

```
You are an expert evaluator of Bengali cultural
knowledge.
Your task is to detect the validation source of
the LLM explanation.
Do not provide any explanation.
Provide only the output in the specified format.

## TASK
Classify the PRIMARY validation source used in
the model explanation.

## Categories:
- institutional (Wikipedia, research papers,
formal history)
- local consensus (shared practices, everyday
experience, or social agreement, even without
formal recognition)
- oral tradition (storytelling, folklore,
proverbs, or intergenerational verbal
communication, rather than written sources)
- niche knowledge: recognized mainly within a
specialized group (e.g., particular profession,
regional community, activist group, or cultural
subcommunity).
- none (no justification given)

## INPUT:
Explanation: {model_explanation}

## Output JSON:
{{
  "source": "institutional | local_consensus |
  oral | niche | none"
}}
```

Listing 7: Judge prompt for classifying primary validation source.

```
You are an expert evaluator of Bengali cultural
knowledge.
Your task is to extract the viewpoints of the
LLM response.
Do not provide any explanation.
Provide only the output in the specified format.

## Task
Extract distinct viewpoints expressed in the
response.
## Definition:
A viewpoint = a culturally or epistemically
distinct interpretation.

## Input:
Question: {question}
Response: {response}

## Output JSON:
{{
  "perspectives": ["", "", ...]
}}
```

Listing 8: Judge prompt for extracting viewpoints expressed in the LLM response

```
You are an information extraction system. Your
task to break the response into atomic meaning
units (claims or sentences), then classify the
epistemic perspective.

## Rules for breaking into atomic meaning units:
- Each unit should express one idea
- Keep them short and self-contained

## Definitions for epistemic perspective:
1. Global: generic, widely documented,
institutional, Western/global narrative
2. Local: culturally specific, community-based,
oral, contested, non-global

Response: {response}

## Output JSON format:
{{
  "units": [
      {{"unit": "atomic unit 1", "label":
      "global | local | neutral"}},
      {{"unit": "atomic unit 2", "label":
      "global | local | neutral"}},
      ....
  ]
}}
```

Listing 9: Judge prompt for extracting atomic units along with epistemic perspective classification.

```
You are an expert evaluator of Bengali cultural
knowledge.
Your task is to count how many perspectives
match.
Do not provide any explanation.
Provide only the output in the specified format.

## TASK
You are given two lists of perspectives
expressed in the response and statement.
Your task is to count how many viewpoints
expressed in response match the viewpoints in
the statement.

## Input:
statement viewpoints: {statement_perspectives}
response viewpoints: {response_perspectives}

## Output JSON:
{{
  "match": integer
}}
```

Listing 10: Judge prompt for matching perspectives between LLM response and gold statement.

E Statistical Analysis on Judges and Metrics

E.1 Validation of Human and LLM Judges

Statistical analysis evaluates the agreement between LLM-based judges (GPT and Mistral) and human evaluation across the proposed cultural metrics. Pearson correlation analysis shows generally strong positive relationships between automated and human judgments, although most correlations

Model	Comparison	Pearson r	p-value	Interpretation
Sonnet	GPT vs Human	0.8187	0.3894	Very strong positive correlation (not significant)
Sonnet	Mistral vs Human	0.8561	0.3458	Very strong positive correlation (not significant)
Sonnet	GPT vs Mistral	0.9977	0.0435	Very strong positive correlation (significant)
GPT	GPT vs Human	0.4559	0.6987	Moderate positive correlation (not significant)
GPT	Mistral vs Human	0.8930	0.2972	Very strong positive correlation (not significant)
GPT	GPT vs Mistral	0.8076	0.4015	Very strong positive correlation (not significant)

Table 5: Pearson correlation analysis between LLM judges and human evaluation across GSR, IBR, and EPC metrics.

Model	Comparison	p-value	Conclusion
Sonnet	GPT vs Human	0.5000	No significant difference
Sonnet	Mistral vs Human	1.0000	No significant difference
GPT	GPT vs Human	1.0000	No significant difference
GPT	Mistral vs Human	1.0000	No significant difference

Table 6: McNemar test results comparing binary agreement patterns between LLM judges and human evaluation.

do not reach statistical significance due to the limited sample size ($n = 3$ metrics per comparison).

For the Sonnet model, both GPT and Mistral judges exhibit very strong correlations with human judgments ($r = 0.8187$ and $r = 0.8561$, respectively), indicating that both judges broadly capture similar evaluation trends as humans. The strongest agreement is observed between GPT and Mistral judges themselves ($r = 0.9977$, $p < 0.05$), suggesting substantial consistency between the two automated evaluators for Sonnet outputs.

For the GPT model, Mistral demonstrates stronger alignment with human judgments ($r = 0.8930$) than GPT self-evaluation ($r = 0.4559$). This finding supports the observation that GPT-based judging may under-detect culturally inappropriate framings when evaluating outputs generated by models with similar training priors. Although the correlations remain positive, the lack of statistical significance indicates that these findings should be interpreted cautiously and validated with larger-scale human evaluation.

The McNemar tests further show no statistically significant disagreement between LLM judges and human annotations across all evaluated settings ($p > 0.05$). This suggests that, at a binary decision level, the automated judges do not systematically diverge from human evaluators. However, the perfect or near-perfect agreement patterns also reflect the limited sample size and coarse binary aggregation.

Overall, the statistical results indicate that LLM judges capture similar directional trends as human evaluators, particularly for detecting global substitution and institutional framing biases. Nevertheless, the stronger correlations observed for Mistral compared to GPT self-evaluation reinforce the paper's broader argument that LLM-as-judge frameworks may inherit model-specific epistemic priors. These findings motivate the continued inclusion of culturally grounded human evaluation when auditing multilingual cultural reasoning in LLMs.

E.2 Correlation Among Evaluation Metrics

The Pearson correlation analysis reveals several important relationships among the evaluation metrics. The strongest positive correlation is observed between GSR and IBR ($r = 0.556$, $p < 0.001$), indicating that these two measures capture related aspects of the underlying phenomenon. This suggests that increases in global substitution are associated with higher institutional bias representation scores. However, GSR captures narrative dominance toward globally prevalent interpretations, whereas IBR captures reliance on institutional authority as the primary source of legitimacy.

In contrast, GSR shows significant negative correlations with EPC ($r = -0.316$, $p = 0.0068$) and LAB ($r = -0.499$, $p < 0.001$). These results imply that higher GSR values tend to correspond with lower EPC and LAB scores, indicating potential conceptual divergence between these metrics.

Similarly, IBR is negatively correlated with EPC ($r = -0.366$, $p = 0.0016$), suggesting that the constructs measured by EPC may capture behavior distinct from both GSR and IBR. Meanwhile, EPC and LAB exhibit a moderate positive correlation ($r = 0.395$, $p < 0.001$), implying that these metrics may reflect partially overlapping dimensions.

Notably, CLFC demonstrates weak and statistically non-significant correlations with all other metrics, with coefficients ranging from -0.143 to 0.153. This indicates that CLFC captures a rela-

Metric Pair	Pearson's r	p-value	Interpretation
GSR vs IBR	0.5560	0.0000	Moderate positive correlation (significant)
GSR vs EPC	-0.3161	0.0068	Weak-to-moderate negative correlation (significant)
GSR vs CLFC	0.0674	0.5737	Negligible correlation (not significant)
GSR vs LAB	-0.4991	0.0000	Moderate negative correlation (significant)
IBR vs EPC	-0.3660	0.0016	Moderate negative correlation (significant)
IBR vs CLFC	-0.1432	0.2300	Weak negative correlation (not significant)
IBR vs LAB	-0.1212	0.3106	Weak negative correlation (not significant)
EPC vs CLFC	0.1532	0.1988	Weak positive correlation (not significant)
EPC vs LAB	0.3948	0.0006	Slight positive correlation (significant)
CLFC vs LAB	-0.0977	0.4144	Negligible negative correlation (not significant)

Table 7: Pairwise Pearson Correlations Among Evaluation Metrics

tively independent characteristic and provides complementary information rather than redundant measurement.

Overall, the correlation patterns suggest that while certain metrics share moderate relationships, none of the correlations are excessively high (e.g., $r > 0.80$). This supports the discriminant validity of the evaluation framework, indicating that the metrics measure related but distinct dimensions rather than redundant constructs.

F Additional Result Analysis

F.1 Question-only Setting

F.1.1 Domain-Level Analysis

Global Substitution Table 8 reveals strong domain and language effects in global substitution. Across both judges, Bangla prompts generally yield lower GSR than English prompts, indicating that asking questions in English more often shifts responses toward globally salient interpretations. The effect is most pronounced in History & Politics and Geography & National Identity, where culturally specific entities are frequently replaced by internationally dominant narratives. Traditional Medicine & Ecology exhibits the lowest GSR under the GPT judge, suggesting that some locally grounded knowledge remains relatively stable in Bangla, though this advantage largely disappears under English prompting. Domain rankings are broadly consistent across the two judges despite differences in absolute scores, indicating robust agreement on which cultural areas are most vulnerable to substitution. Overall, the results show that global narrative dominance is systematic rather than uniform: it is amplified in domains tied to national identity and historical interpretation, where local meanings directly compete with globally dominant frames.

Institutional Bias Table 9 shows that institutional bias varies substantially across domains, languages, and judges. Across both judges, History & Politics consistently exhibits the highest IBR, indicating that responses in this domain are most likely to privilege globally recognized institutional narratives over locally grounded interpretations. Geography & National Identity also shows elevated bias, particularly under the Mistral judge, suggesting that questions involving national identity are especially susceptible to institutional framing. In contrast, Traditional Medicine & Ecology yields the lowest IBR across nearly all models, reflecting greater reliance on community-based or practice-oriented knowledge. English prompts generally increase IBR relative to Bangla prompts, especially under the Mistral judge, indicating that language choice amplifies dependence on globally legitimized sources. Despite differences in absolute scores, both judges agree that institutional bias is concentrated in domains where historical authority and national narratives compete with local epistemic traditions.

Epistemic Perspective Coverage Table 10 shows that providing evidence substantially improves epistemic perspective coverage across all models and domains, with most scores falling between 0.55 and 0.70. The highest coverage consistently appears in *Traditional Medicine & Ecology* and *Religion, Folklore & Mythology*, suggesting that explicit local evidence helps models preserve plural and culturally grounded interpretations in domains where knowledge is often community-specific or orally transmitted. In contrast, *His-*

Language	Domain	Sonnet	GPT	Gemini	Gemma	Qwen	Llama	Mistral	Grok	DeepSeek
					Judge: GPT-5.4-mini					
Bangla	Geography & National Identity	0.333	0.143	0.143	0.387	0.250	0.294	0.179	0.233	0.242
	Traditional Medicine & Ecology	0.105	0.059	0.182	0.042	0.222	0.200	0.200	0.074	0.167
	Art, Literature & Cultural Practices	0.204	0.119	0.277	0.309	0.226	0.338	0.206	0.208	0.299
	Religion, Folklore & Mythology	0.040	0.208	0.261	0.158	0.156	0.222	0.172	0.000	0.158
	History & Politics	0.415	0.217	0.425	0.536	0.417	0.545	0.433	0.392	0.366
					Judge: Mistral 4 Small					
Bangla	Geography & National Identity	0.846	0.609	0.613	0.811	0.645	0.837	0.650	0.705	0.565
	Traditional Medicine & Ecology	0.625	0.600	0.750	0.565	0.704	0.659	0.762	0.725	0.524
	Art, Literature & Cultural Practices	0.582	0.531	0.586	0.644	0.735	0.779	0.680	0.674	0.627
	Religion, Folklore & Mythology	0.667	0.480	0.611	0.643	0.585	0.759	0.842	0.659	0.526
	History & Politics	0.466	0.495	0.500	0.626	0.688	0.845	0.667	0.676	0.542
					Judge: GPT-5.4-mini					
English	Geography & National Identity	0.600	0.375	0.625	0.789	0.606	0.838	0.719	0.781	0.657
	Traditional Medicine & Ecology	0.375	0.462	0.600	0.711	0.618	0.714	0.576	0.737	0.658
	Art, Literature & Cultural Practices	0.636	0.505	0.582	0.754	0.703	0.705	0.616	0.648	0.659
	Religion, Folklore & Mythology	0.595	0.273	0.538	0.643	0.647	0.560	0.542	0.681	0.518
	History & Politics	0.675	0.433	0.662	0.852	0.806	0.723	0.788	0.754	0.741
					Judge: Mistral 4 Small					
English	Geography & National Identity	0.767	0.826	0.784	0.872	0.689	0.804	0.731	0.816	0.784
	Traditional Medicine & Ecology	0.722	0.833	0.862	0.879	0.750	0.778	0.875	0.784	0.926
	Art, Literature & Cultural Practices	0.701	0.730	0.745	0.843	0.903	0.809	0.727	0.820	0.852
	Religion, Folklore & Mythology	0.893	0.708	0.667	0.865	0.711	0.825	0.909	0.875	0.674
	History & Politics	0.598	0.606	0.700	0.870	0.842	0.792	0.816	0.748	0.780

Table 8: Domain-wise Global Substitution Rate (GSR) in the question-only setting. Lower values indicate better preservation of culturally grounded answers, while higher values indicate more frequent replacement of local referents with globally dominant ones. Computed following Eq. 3.

tory & Politics remains the most difficult domain, exhibiting the lowest EPC scores for nearly all models, which indicates a continued tendency to collapse contested narratives into narrower institutional accounts. Across models, `Gemini-3.1 Pro` achieves the strongest and most consistent perspective coverage under both judges, while `Claude Sonnet 4.6` and `GPT-5.4` also perform well. Open-weight models such as `Llama 4 Maverick` and `Qwen 3.6` generally produce lower EPC, particularly in politically sensitive domains. The two judges show strong agreement on overall trends, reinforcing the robustness of these findings. Together, the results demonstrate that evidence improves cultural grounding but does not fully eliminate domain-specific limitations, especially where multiple competing historical interpretations coexist.

Cross-lingual Factual Consistency Table 11 presents cross-lingual factual consistency, which is uniformly low, indicating that models often produce substantially different claims when the same culturally grounded question is asked in Bangla and English. Under the stricter GPT-5.4-mini judge, most scores remain below 0.15, suggesting near-complete semantic divergence across languages. The Mistral judge yields higher absolute values but preserves the same ranking patterns, confirming that cross-lingual instability is robust to judge choice. Google, Anthropic, and OpenAI frontier models perform best overall, with Google Gemini-3.1 Pro achieving the highest consistency across most domains (up to 0.453 in Religion, Folklore & Mythology), followed by Anthropic Sonnet 4.6 and OpenAI GPT-5.4. However, even these strongest systems remain far from stable cross-lingual grounding. Domain effects are pronounced: History & Politics and Geography & National Identity generally exhibit higher consistency, whereas Art, Literature & Cultural Practices and Traditional Medicine & Ecology show the lowest scores, reflecting greater susceptibility to culturally specific reinterpretation. These results demonstrate that multilingual competence does not guarantee semantic equivalence across languages; instead, changing the language of interaction can substantially alter the factual content of model responses.

Language Anchor Bias Table 12 presents language anchor bias, which is consistently positive across all models and domains, demonstrating that the language of interaction systematically affects cultural interpretation. Asking the same question in English reliably shifts responses toward globally dominant narratives, even when the underlying cultural content is unchanged. This pattern is robust across both judges, with `GPT-5.4-mini` assigning

Language	Domain	Sonnet	GPT	Gemini	Gemma	Qwen	Llama	Mistral	Grok	DeepSeek
					Judge: GPT-5.4-mini					
Bangla	Geography & National Identity	0.173	0.120	0.120	0.173	0.080	0.107	0.213	0.080	0.160
	Traditional Medicine & Ecology	0.090	0.029	0.015	0.059	0.000	0.029	0.029	0.015	0.044
	Art, Literature & Cultural Practices	0.077	0.064	0.056	0.068	0.060	0.026	0.068	0.051	0.103
	Religion, Folklore & Mythology	0.057	0.057	0.075	0.066	0.075	0.028	0.057	0.057	0.113
	History & Politics	0.312	0.239	0.325	0.333	0.239	0.197	0.389	0.150	0.385
					Judge: Mistral 4 Small					
Bangla	Geography & National Identity	0.520	0.187	0.453	0.400	0.293	0.653	0.520	0.293	0.387
	Traditional Medicine & Ecology	0.179	0.015	0.074	0.074	0.059	0.176	0.103	0.029	0.103
	Art, Literature & Cultural Practices	0.444	0.120	0.205	0.179	0.235	0.355	0.342	0.154	0.291
	Religion, Folklore & Mythology	0.387	0.142	0.189	0.132	0.283	0.377	0.330	0.142	0.283
	History & Politics	0.735	0.415	0.611	0.624	0.585	0.782	0.816	0.513	0.679
					Judge: GPT-5.4-mini					
English	Geography & National Identity	0.213	0.147	0.120	0.173	0.133	0.120	0.280	0.120	0.347
	Traditional Medicine & Ecology	0.044	0.029	0.059	0.029	0.029	0.059	0.044	0.029	0.059
	Art, Literature & Cultural Practices	0.085	0.060	0.077	0.103	0.073	0.038	0.124	0.051	0.120
	Religion, Folklore & Mythology	0.132	0.047	0.057	0.123	0.104	0.075	0.113	0.057	0.094
	History & Politics	0.359	0.226	0.376	0.392	0.308	0.244	0.423	0.256	0.432
					Judge: Mistral 4 Small					
English	Geography & National Identity	0.733	0.427	0.600	0.720	0.520	0.680	0.747	0.627	0.747
	Traditional Medicine & Ecology	0.265	0.206	0.324	0.368	0.338	0.441	0.324	0.338	0.309
	Art, Literature & Cultural Practices	0.573	0.363	0.453	0.534	0.513	0.427	0.517	0.410	0.526
	Religion, Folklore & Mythology	0.538	0.330	0.472	0.491	0.500	0.481	0.481	0.434	0.519
	History & Politics	0.705	0.509	0.684	0.772	0.718	0.722	0.786	0.688	0.774

Table 9: Domain-wise Institutional Bias Rate (IBR) in the question-only setting. Lower values indicate greater reliance on locally grounded explanations, while higher values indicate stronger dependence on globally legitimized institutional framing.

Language	Judge	Domain	Sonnet	GPT	Gemini	Gemma	Qwen	Llama	Mistral	Grok	DeepSeek
Bangla	GPT	History & Politics	0.583	0.605	0.596	0.499	0.546	0.540	0.567	0.512	0.526
		Art, Literature & Cultural Practices	0.633	0.655	0.653	0.601	0.592	0.565	0.604	0.591	0.598
		Religion, Folklore & Mythology	0.658	0.626	0.671	0.585	0.579	0.573	0.626	0.571	0.612
		Geography & National Identity	0.612	0.646	0.671	0.608	0.630	0.623	0.615	0.633	0.580
		Traditional Medicine & Ecology	0.638	0.646	0.636	0.597	0.627	0.550	0.602	0.554	0.638
	Mistral	History & Politics	0.545	0.566	0.540	0.518	0.507	0.507	0.536	0.470	0.555
		Art, Literature & Cultural Practices	0.660	0.677	0.684	0.624	0.615	0.588	0.627	0.590	0.625
		Religion, Folklore & Mythology	0.668	0.618	0.671	0.580	0.556	0.541	0.621	0.586	0.620
		Geography & National Identity	0.640	0.622	0.631	0.583	0.564	0.552	0.610	0.516	0.583
		Traditional Medicine & Ecology	0.685	0.680	0.697	0.645	0.626	0.528	0.648	0.577	0.622
English	GPT	History & Politics	0.520	0.598	0.539	0.450	0.500	0.482	0.483	0.489	0.488
		Art, Literature & Cultural Practices	0.583	0.659	0.607	0.568	0.540	0.543	0.575	0.558	0.589
		Religion, Folklore & Mythology	0.612	0.638	0.620	0.585	0.567	0.546	0.572	0.531	0.574
		Geography & National Identity	0.574	0.621	0.601	0.514	0.549	0.531	0.559	0.555	0.523
		Traditional Medicine & Ecology	0.520	0.656	0.511	0.486	0.504	0.495	0.583	0.549	0.551
	Mistral	History & Politics	0.499	0.507	0.484	0.432	0.453	0.423	0.462	0.435	0.488
		Art, Literature & Cultural Practices	0.621	0.665	0.617	0.562	0.555	0.522	0.575	0.537	0.566
		Religion, Folklore & Mythology	0.628	0.625	0.623	0.512	0.581	0.534	0.588	0.540	0.554
		Geography & National Identity	0.573	0.539	0.541	0.490	0.475	0.505	0.543	0.492	0.593
		Traditional Medicine & Ecology	0.601	0.661	0.577	0.586	0.545	0.515	0.664	0.551	0.594

Table 10: Domain-wise Epistemic Perspective Coverage (EPC) in question-only setting across Bangla and English language, evaluated by GPT and Mistral judges. Higher values indicate broader inclusion of culturally relevant perspectives.

higher absolute scores and `Mistral 4 Small` preserving the same relative ordering. Traditional Medicine & Ecology and Art, Literature & Cultural Practices exhibit the strongest anchoring effects, with several models exceeding 0.50, indicating that culturally specific practices are particularly vulnerable to reinterpretation through globally familiar frames. Among the models, GPT-5.4 shows the lowest overall LAB, while Gemma 4, Qwen 3.6 Plus, and DeepSeek v3.2 display the largest shifts. Frontier systems such as Sonnet 4.6 and Gemini-3.1 Pro reduce but do not eliminate this effect. In particular, no model approaches zero across domains, confirming that multilingual capability does not guarantee language-invariant reasoning. Instead, English functions as a strong contextual prior that systematically alters how LLMs retrieve and prioritize cultural knowledge.

Judge	Domain	Sonnet	GPT	Gemini	Gemma	Qwen	Llama	Mistral	Grok	DeepSeek
GPT	History & Politics	0.107	0.094	0.124	0.082	0.158	0.115	0.085	0.090	0.056
	Art, Literature & Cultural Practices	0.077	0.060	0.081	0.056	0.073	0.090	0.038	0.030	0.013
	Religion, Folklore & Mythology	0.057	0.047	0.132	0.019	0.075	0.094	0.047	0.075	0.009
	Geography & National Identity	0.120	0.067	0.147	0.080	0.093	0.107	0.040	0.053	0.053
	Traditional Medicine & Ecology	0.045	0.044	0.147	0.059	0.132	0.088	0.103	0.132	0.029
Mistral	History & Politics	0.338	0.338	0.427	0.241	0.308	0.188	0.209	0.158	0.175
	Art, Literature & Cultural Practices	0.295	0.338	0.346	0.179	0.256	0.145	0.141	0.107	0.107
	Religion, Folklore & Mythology	0.217	0.340	0.453	0.198	0.236	0.179	0.170	0.132	0.123
	Geography & National Identity	0.373	0.387	0.333	0.347	0.200	0.253	0.200	0.187	0.280
	Traditional Medicine & Ecology	0.194	0.353	0.368	0.206	0.206	0.162	0.162	0.118	0.088

Table 11: Domain-wise Cross-Lingual Factual Consistency (CLFC) for nine LLMs, evaluated by two independent judges. Higher scores indicate stronger semantic consistency between Bangla and English responses to the same culturally grounded question.

Domain	Sonnet	GPT	Gemini	Gemma	Qwen	Llama	Mistral	Grok	DeepSeek
GPT-as-Judge									
History & Politics	0.470	0.261	0.432	0.375	0.466	0.329	0.427	0.410	0.432
Art, Literature & Cultural Practices	0.479	0.406	0.440	0.551	0.538	0.470	0.457	0.470	0.521
Religion, Folklore & Mythology	0.509	0.321	0.425	0.509	0.538	0.481	0.406	0.491	0.491
Geography & National Identity	0.413	0.267	0.400	0.453	0.467	0.467	0.533	0.493	0.533
Traditional Medicine & Ecology	0.358	0.397	0.500	0.588	0.441	0.588	0.412	0.544	0.471
Mistral-as-Judge									
History & Politics	0.303	0.184	0.350	0.302	0.252	0.141	0.248	0.214	0.385
Art, Literature & Cultural Practices	0.333	0.265	0.363	0.380	0.252	0.218	0.278	0.282	0.496
Religion, Folklore & Mythology	0.406	0.302	0.311	0.406	0.302	0.302	0.358	0.321	0.472
Geography & National Identity	0.293	0.293	0.280	0.240	0.267	0.147	0.307	0.213	0.440
Traditional Medicine & Ecology	0.254	0.265	0.250	0.368	0.191	0.162	0.221	0.162	0.441

Table 12: Domain-wise Language Anchor Bias (LAB) for nine LLMs, evaluated by two independent judges. Higher scores indicate a stronger tendency for English prompts to shift responses toward globally dominant interpretations relative to equivalent Bangla prompts.

F.1.2 Analysis on Knowledge Frequency, Epistemic Status, and Validation Type

Global Substitution Table 13 reports Global Substitution Rate (GSR) by language, model, judge, and sociocultural category. Since GSR measures whether a locally grounded answer is replaced by a globally dominant one, it is computed only for local and contested instances; the global epistemic-status column is therefore omitted.

The dominant pattern is a large Bangla–English gap. Under the GPT judge, Bangla prompts generally yield lower GSR, often below 0.35 for models such as `GPT`, `Grok`, `Sonnet`, and `DeepSeek`. In contrast, English prompts sharply increase global substitution, with many models reaching 0.60–0.80 across knowledge-frequency groups. This effect is especially strong for `Gemma`, `Qwen`, `Llama`, `Grok`, `Mistral`, and `DeepSeek`, indicating that the same Bengali cultural content is much more likely to be resolved through globally dominant referents when asked in English.

The Mistral judge assigns higher absolute GSR values, but confirms the same trend. Even in Bangla, several models show moderate substitution, while English prompts further increase GSR across nearly all models and categories. Under this judge, `Gemma`, `Qwen`, `Llama`, and `Grok` often exceed 0.80 in English, showing strong reliance on globally salient interpretations. Therefore, judge calibration differs, but both judges agree that English substantially amplifies global substitution.

The category-level breakdown shows that this effect is not limited to rare knowledge. English-side GSR remains high across occasional, frequent, and rare items, suggesting that global substitution is not simply a data-sparsity problem. Similarly, both local and contested items are vulnerable, with contested items often showing particularly high substitution. Across validation types, institutional, niche, local-consensus, and oral-tradition items all show increased GSR in English, indicating that models struggle to preserve local sources of legitimacy as well as local factual referents.

Overall, Table 13 provides strong evidence that global substitution is systematic and language-conditioned. English prompts consistently push models away from Bengali cultural grounding and toward globally dominant interpretations, and this pattern persists across judges, models, knowledge-frequency levels, epistemic-status categories, and validation types.

Institutional Bias Table 14 reports the Institutional Bias Rate (IBR) across languages, models, judges, and sociocultural categories. The main pattern is that English prompts generally increase institutional framing. Under the GPT judge,

Language	Model	Knowledge Frequency			Epistemic Status			Validation Type			
		Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
					GPT-as-Judge						
Bangla	GPT	0.135	0.169	0.196	0.138	0.250	–	0.125	0.222	0.065	0.154
	Qwen	0.233	0.317	0.309	0.246	0.545	–	0.341	0.294	0.250	0.267
	Gemini	0.289	0.365	0.269	0.274	0.364	–	0.282	0.371	0.225	0.333
	Gemma	0.393	0.357	0.338	0.318	0.714	–	0.296	0.464	0.290	0.095
	DeepSeek	0.294	0.298	0.257	0.250	0.520	–	0.255	0.311	0.269	0.267
	Mistral	0.315	0.297	0.232	0.203	0.500	–	0.217	0.351	0.255	0.133
	Llama	0.477	0.376	0.310	0.346	0.619	–	0.267	0.468	0.349	0.333
	Grok	0.231	0.261	0.200	0.173	0.345	–	0.186	0.338	0.143	0.000
	Sonnet	0.273	0.205	0.300	0.225	0.200	–	0.250	0.338	0.116	0.154
English	GPT	0.526	0.408	0.353	0.429	0.346	–	0.449	0.462	0.403	0.167
	Qwen	0.739	0.694	0.728	0.695	0.733	–	0.648	0.759	0.727	0.545
	Gemini	0.651	0.646	0.520	0.602	0.548	–	0.478	0.695	0.558	0.542
	Gemma	0.806	0.777	0.723	0.740	0.857	–	0.797	0.794	0.733	0.692
	DeepSeek	0.710	0.680	0.592	0.617	0.781	–	0.617	0.739	0.598	0.577
	Mistral	0.711	0.654	0.667	0.635	0.765	–	0.641	0.743	0.598	0.545
	Llama	0.661	0.740	0.707	0.699	0.647	–	0.671	0.738	0.694	0.600
	Grok	0.644	0.756	0.729	0.685	0.704	–	0.705	0.697	0.747	0.727
	Sonnet	0.667	0.598	0.591	0.585	0.381	–	0.542	0.687	0.562	0.455
					Mistral-as-Judge						
Bangla	GPT	0.563	0.496	0.532	0.518	0.273	–	0.611	0.504	0.556	0.300
	Qwen	0.761	0.656	0.642	0.699	0.619	–	0.714	0.689	0.636	0.643
	Gemini	0.544	0.588	0.551	0.579	0.273	–	0.564	0.547	0.633	0.556
	Gemma	0.732	0.622	0.607	0.648	0.615	–	0.688	0.681	0.585	0.533
	DeepSeek	0.579	0.573	0.520	0.556	0.462	–	0.537	0.586	0.524	0.545
	Mistral	0.691	0.670	0.756	0.694	0.636	–	0.756	0.748	0.568	0.444
	Llama	0.838	0.758	0.811	0.785	0.781	–	0.787	0.805	0.779	0.789
	Grok	0.728	0.676	0.634	0.679	0.545	–	0.632	0.720	0.689	0.474
	Sonnet	0.545	0.545	0.640	0.594	0.174	–	0.595	0.587	0.514	0.417
English	GPT	0.712	0.710	0.643	0.714	0.571	–	0.762	0.661	0.756	0.500
	Qwen	0.865	0.806	0.771	0.815	0.719	–	0.746	0.824	0.871	0.722
	Gemini	0.716	0.768	0.682	0.730	0.563	–	0.641	0.725	0.843	0.684
	Gemma	0.839	0.892	0.831	0.854	0.938	–	0.877	0.847	0.889	0.889
	DeepSeek	0.766	0.857	0.732	0.787	0.750	–	0.696	0.799	0.896	0.643
	Mistral	0.711	0.748	0.667	0.796	0.769	–	0.850	0.810	0.720	0.667
	Llama	0.819	0.812	0.755	0.784	0.714	–	0.727	0.811	0.812	0.905
	Grok	0.788	0.782	0.840	0.797	0.741	–	0.826	0.751	0.851	0.900
	Sonnet	0.747	0.646	0.673	0.698	0.310	–	0.727	0.680	0.647	0.800

Table 13: GSR scores by model, knowledge frequency, epistemic status, and validation type. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition. Following Eq. 3; Global epistemic-status column omitted.

IBR remains relatively low overall, especially for Bangla prompts, where many models stay below 0.20 across knowledge-frequency categories. However, English prompts raise IBR for most models, particularly `DeepSeek`, `Mistral`, `Gemma`, and `Sonnet`, indicating a stronger tendency to frame answers through institutional or globally legitimized sources.

The Mistral judge assigns substantially higher IBR values, but preserves the same directional trend. Bangla responses already show moderate institutional bias for several models, while English responses increase sharply, often exceeding 0.50 and sometimes reaching above 0.70 for `Gemma`, `DeepSeek`, `Mistral`, and `Sonnet`. This suggests that absolute IBR values are judge-dependent, but both judges agree that English prompts make institutional framing more likely.

The category-level results show that institutional bias is strongest for contested and globally framed items, as well as items whose validation type is institutional. This is expected, since such items are more likely to activate formal or dominant knowledge sources. However, English also increases IBR for niche, local-consensus, and oral-tradition items, showing that even community-grounded knowledge can be reframed through institutional authority when prompted in English.

Language	Model	Knowledge Frequency			Epistemic Status			Validation Type			
		Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
					GPT-as-Judge						
Bangla	GPT	0.103	0.149	0.087	0.063	0.083	0.249	0.047	0.198	0.047	0.000
	Qwen	0.107	0.127	0.109	0.090	0.125	0.169	0.057	0.166	0.068	0.065
	Gemini	0.152	0.175	0.080	0.077	0.208	0.280	0.047	0.246	0.052	0.000
	Gemma	0.134	0.220	0.072	0.088	0.271	0.293	0.075	0.257	0.073	0.000
	DeepSeek	0.179	0.231	0.138	0.117	0.375	0.316	0.057	0.291	0.120	0.065
	Mistral	0.165	0.220	0.116	0.106	0.333	0.302	0.075	0.302	0.052	0.000
	Llama	0.076	0.107	0.072	0.063	0.125	0.138	0.038	0.142	0.037	0.022
	Grok	0.089	0.096	0.043	0.050	0.083	0.151	0.019	0.131	0.047	0.000
	Sonnet	0.152	0.181	0.130	0.106	0.271	0.250	0.113	0.241	0.068	0.022
English	GPT	0.098	0.152	0.065	0.063	0.104	0.231	0.028	0.206	0.026	0.000
	Qwen	0.116	0.197	0.116	0.088	0.271	0.267	0.075	0.219	0.105	0.043
	Gemini	0.147	0.211	0.123	0.104	0.250	0.298	0.085	0.281	0.052	0.022
	Gemma	0.179	0.243	0.123	0.111	0.229	0.369	0.094	0.316	0.063	0.067
	DeepSeek	0.205	0.285	0.159	0.140	0.312	0.409	0.123	0.366	0.089	0.043
	Mistral	0.183	0.287	0.152	0.151	0.354	0.356	0.170	0.342	0.079	0.065
	Llama	0.080	0.163	0.080	0.070	0.250	0.196	0.019	0.198	0.058	0.000
	Grok	0.107	0.158	0.065	0.072	0.229	0.204	0.038	0.198	0.058	0.000
	Sonnet	0.156	0.251	0.094	0.115	0.229	0.333	0.057	0.299	0.089	0.043
					Mistral-as-Judge						
Bangla	GPT	0.223	0.242	0.138	0.137	0.375	0.338	0.151	0.316	0.105	0.022
	Qwen	0.299	0.377	0.341	0.275	0.521	0.449	0.321	0.471	0.178	0.087
	Gemini	0.371	0.369	0.261	0.239	0.521	0.529	0.255	0.508	0.162	0.043
	Gemma	0.321	0.363	0.261	0.212	0.479	0.533	0.236	0.495	0.136	0.022
	DeepSeek	0.433	0.431	0.312	0.295	0.583	0.596	0.292	0.561	0.241	0.130
	Mistral	0.464	0.544	0.399	0.358	0.729	0.702	0.358	0.684	0.277	0.109
	Llama	0.527	0.541	0.413	0.383	0.688	0.729	0.377	0.698	0.319	0.109
	Grok	0.263	0.307	0.196	0.180	0.479	0.409	0.179	0.401	0.120	0.065
	Sonnet	0.531	0.540	0.420	0.390	0.667	0.728	0.415	0.665	0.361	0.152
English	GPT	0.362	0.428	0.377	0.320	0.438	0.542	0.358	0.497	0.288	0.130
	Qwen	0.531	0.575	0.580	0.498	0.771	0.644	0.500	0.663	0.455	0.326
	Gemini	0.518	0.555	0.507	0.432	0.667	0.707	0.453	0.639	0.414	0.370
	Gemma	0.619	0.616	0.572	0.532	0.667	0.747	0.547	0.716	0.503	0.311
	DeepSeek	0.594	0.645	0.536	0.507	0.729	0.782	0.491	0.733	0.476	0.413
	Mistral	0.576	0.617	0.623	0.516	0.750	0.751	0.575	0.727	0.445	0.348
	Llama	0.536	0.577	0.551	0.491	0.646	0.676	0.481	0.647	0.466	0.413
	Grok	0.504	0.538	0.500	0.439	0.646	0.653	0.425	0.631	0.403	0.326
	Sonnet	0.589	0.631	0.529	0.527	0.562	0.747	0.434	0.714	0.518	0.370

Table 14: Institutional Bias Rate (IBR) by model, knowledge frequency, epistemic status, and validation type. Higher values indicate stronger reliance on institutional or globally dominant sources over local cultural perspectives. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition.

Overall, Table 14 supports the claim that institutional bias is a systematic language-conditioned effect. Models not only change what they answer across languages; they also change which sources of authority they implicitly privilege, with English prompts pushing responses toward globally recognized and institutionally dominant framings.

Epistemic Perspective Coverage Table 15 reports epistemic perspective coverage (EPC) across languages, models, judges, and sociocultural categories. Since higher EPC indicates broader coverage of locally relevant viewpoints, the table shows how well models preserve cultural plurality rather than collapsing answers into a single dominant framing.

The main pattern is that Bangla prompts generally produce higher EPC than English prompts. Under both GPT- and Mistral-as-judge settings, most models lose perspective coverage when the same cultural content is asked in English. This drop is especially visible for `Qwen`, `Gemma`, `Llama`, `Grok`, and `Sonnet`, suggesting that English prompts not only increase global substitution and institutional bias, but also narrow the range of local interpretations represented in the response.

Language	Model	Knowledge Frequency			Epistemic Status			Validation Type			
		Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
					GPT-as-Judge						
Bangla	GPT	0.645	0.642	0.588	0.636	0.587	0.636	0.593	0.633	0.644	0.668
	Qwen	0.565	0.611	0.536	0.571	0.519	0.618	0.513	0.584	0.617	0.586
	Gemini	0.641	0.649	0.600	0.643	0.539	0.646	0.601	0.639	0.654	0.632
	Gemma	0.559	0.590	0.512	0.572	0.514	0.564	0.540	0.558	0.584	0.606
	DeepSeek	0.574	0.592	0.550	0.579	0.473	0.599	0.559	0.568	0.592	0.654
	Mistral	0.587	0.613	0.566	0.597	0.549	0.603	0.567	0.594	0.606	0.640
	Llama	0.549	0.583	0.532	0.559	0.515	0.580	0.520	0.572	0.557	0.611
	Grok	0.547	0.589	0.522	0.562	0.477	0.585	0.527	0.559	0.590	0.570
	Sonnet	0.623	0.632	0.576	0.619	0.603	0.622	0.563	0.612	0.649	0.676
English	GPT	0.624	0.655	0.585	0.634	0.562	0.641	0.589	0.619	0.658	0.722
	Qwen	0.521	0.549	0.489	0.527	0.420	0.554	0.484	0.523	0.553	0.572
	Gemini	0.581	0.588	0.541	0.587	0.471	0.579	0.550	0.566	0.611	0.582
	Gemma	0.523	0.540	0.457	0.526	0.416	0.526	0.469	0.507	0.552	0.596
	DeepSeek	0.547	0.559	0.497	0.555	0.431	0.543	0.520	0.519	0.602	0.549
	Mistral	0.543	0.552	0.524	0.559	0.384	0.547	0.537	0.521	0.574	0.620
	Llama	0.490	0.532	0.526	0.527	0.459	0.512	0.495	0.500	0.549	0.590
	Grok	0.538	0.537	0.501	0.535	0.441	0.541	0.514	0.523	0.547	0.562
	Sonnet	0.565	0.575	0.512	0.571	0.448	0.560	0.523	0.543	0.598	0.623
					Mistral-as-Judge						
Bangla	GPT	0.636	0.627	0.612	0.639	0.512	0.626	0.620	0.597	0.673	0.688
	Qwen	0.574	0.574	0.535	0.580	0.431	0.570	0.530	0.553	0.605	0.601
	Gemini	0.631	0.644	0.598	0.656	0.468	0.616	0.582	0.606	0.692	0.691
	Gemma	0.595	0.585	0.546	0.591	0.440	0.590	0.564	0.552	0.639	0.614
	DeepSeek	0.584	0.607	0.591	0.603	0.502	0.605	0.573	0.584	0.617	0.670
	Mistral	0.608	0.601	0.569	0.608	0.506	0.595	0.588	0.568	0.636	0.691
	Llama	0.547	0.552	0.524	0.556	0.504	0.532	0.550	0.535	0.558	0.560
	Grok	0.537	0.544	0.540	0.557	0.456	0.529	0.532	0.520	0.569	0.618
	Sonnet	0.632	0.633	0.587	0.645	0.493	0.610	0.559	0.597	0.695	0.691
English	GPT	0.604	0.612	0.530	0.608	0.477	0.592	0.560	0.552	0.666	0.711
	Qwen	0.542	0.530	0.437	0.539	0.394	0.496	0.467	0.486	0.580	0.602
	Gemini	0.567	0.580	0.511	0.578	0.455	0.555	0.556	0.526	0.625	0.621
	Gemma	0.529	0.520	0.440	0.517	0.354	0.521	0.479	0.480	0.571	0.528
	DeepSeek	0.557	0.563	0.475	0.549	0.467	0.551	0.474	0.526	0.591	0.663
	Mistral	0.527	0.565	0.525	0.561	0.353	0.555	0.549	0.508	0.584	0.679
	Llama	0.477	0.490	0.508	0.520	0.382	0.451	0.500	0.441	0.548	0.612
	Grok	0.499	0.511	0.477	0.512	0.312	0.519	0.457	0.487	0.532	0.582
	Sonnet	0.585	0.595	0.509	0.601	0.464	0.549	0.502	0.542	0.657	0.677

Table 15: Epistemic perspective coverage (EPC) scores by by model, knowledge frequency, epistemic status, and validation type. Higher scores indicate broader EPC. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition.

Model-level differences are also clear. `GPT`, `Gemini`, and `Sonnet` generally achieve the strongest EPC, especially in Bangla, while `Gemma`, `Qwen`, `Llama`, and `Grok` tend to provide narrower coverage. The two judges differ slightly in calibration, but they agree on the overall ranking and the language effect: Bangla responses preserve more epistemic diversity than English responses.

The category-level breakdown shows that rare and contested knowledge are the most difficult. EPC is often lower for rare items than for frequent or occasional items, indicating that models cover fewer perspectives when the cultural knowledge is less visible. Contested items also receive consistently lower EPC than local or global-status items, which suggests that models struggle to represent disagreement or plural interpretations. By validation type, local-consensus and oral-tradition items often receive relatively high EPC, while niche and institutional categories are more variable, indicating that models are better at broad cultural descriptions than at preserving specialized or authority-sensitive epistemic contexts.

Overall, Table 15 supports the claim that language affects not only factual content but also epistemic breadth. English prompts systematically reduce perspective coverage across models and cultural categories, reinforcing that global narrative

dominance operates by narrowing the range of locally grounded viewpoints available in model responses.

Cross-Lingual Factual Consistency Table 16 reports cross-lingual factual consistency (CLFC) across models and sociocultural categories. Since higher CLFC indicates stronger semantic consistency between Bangla and English responses, lower values reflect greater language-induced factual instability.

The table shows that cross-lingual consistency is generally weak and highly category-dependent. Under the GPT judge, CLFC values are low across most models and categories, often below 0.15. `Gemini`, `Qwen`, `Llama`, and `Sonnet` show relatively better consistency, while `DeepSeek` is consistently among the weakest. Under the Mistral judge, absolute CLFC values are higher, but the model-level pattern is similar: `Gemini`, `GPT`, and `Sonnet` are more stable, whereas `DeepSeek`, `Grok`, and `Mistral` remain less consistent.

The category breakdown further shows that rare and niche knowledge tend to have lower CLFC than frequent or institutional knowledge. This suggests that models are less able to preserve factual meaning across languages when the knowledge is less visible or less standardized. Global and institutional categories often receive higher CLFC, indicating that models are more stable when the content aligns with widely circulated or formally documented knowledge. In contrast, local, contested, oral tradition, and niche categories are more fragile, reflecting the difficulty of preserving culturally specific interpretations across Bangla and English.

Overall, Table 16 supports the central claim that cross-lingual instability is not uniform: it is strongest for culturally local, rare, contested, and community-grounded knowledge. Although judges differ in calibration, both show that even stronger models fail to maintain consistently high factual equivalence across languages.

Language Anchor Bias Table 17 reports Language Anchor Bias (LAB) across models and sociocultural categories. Higher LAB indicates that model responses are more strongly influenced by the language of the prompt, meaning that the same culturally grounded question elicits different framings when asked in Bangla versus English.

The results show that language anchoring is pervasive across all evaluated models. Under the GPT judge, most systems exhibit moderate to high LAB, with values typically ranging from 0.40 to 0.55 across categories. `GPT` is the most stable model, maintaining relatively low LAB (approximately 0.33 across most knowledge-frequency groups) and showing particularly weak anchoring oral-tradition items. In contrast, `Qwen`, `Gemma`, `DeepSeek`, `Grok`, and `Sonnet` display substantially higher LAB, indicating greater sensitivity to whether questions are asked in Bangla or English.

The Mistral judge assigns lower absolute LAB values for some models, but the overall ranking remains consistent. `Llama` and `GPT` appear comparatively stable, whereas `DeepSeek`, `Gemini`, `Gemma`, and `Sonnet` show stronger language dependence. `DeepSeek` stands out under this judge, with particularly high LAB for frequent knowledge, contested items, institutional validation, and local-consensus categories. This suggests that language anchoring persists even for culturally prominent and socially validated knowledge.

The category-level analysis further shows that LAB is not limited to rare or obscure items. Frequent and occasional knowledge often exhibit anchoring comparable to, or greater than, rare knowledge, indicating that these differences cannot be explained solely by knowledge scarcity. Contested items generally show higher LAB than local items, suggesting that models are especially sensitive to prompt language when cultural interpretations are disputed. Across validation types, institutional and local-consensus categories tend to exhibit elevated LAB, while oral-tradition items show greater variation across models and judges.

Overall, Table 17 demonstrates that prompt language serves as a powerful epistemic anchor. Even when the underlying question remains identical, models often shift their framing depending on whether the question is written in Bangla or English. These findings support our central claim that global narrative dominance arises not only from factual substitution, but also from language-conditioned differences in how models prioritize and present cultural knowledge.

F.2 Evidence-based Setting

F.2.1 Domain-Level Analysis

Global Substitution Table 18 shows that global substitution persists even when local supporting evidence is provided. Across both judges, English prompts yield higher GSR than Bangla prompts in

Model	Knowledge Frequency			Epistemic Status			Validation Type			
	Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
GPT-as-Judge										
GPT	0.073	0.062	0.065	0.102	0.052	0.062	0.088	0.057	0.047	0.022
Qwen	0.127	0.107	0.065	0.138	0.095	0.104	0.136	0.057	0.099	0.043
Gemini	0.124	0.121	0.087	0.147	0.104	0.083	0.126	0.075	0.120	0.109
Gemma	0.073	0.040	0.065	0.067	0.054	0.104	0.075	0.085	0.031	0.022
DeepSeek	0.051	0.018	0.007	0.044	0.020	0.083	0.048	0.009	0.021	0.000
Mistral	0.085	0.045	0.029	0.084	0.045	0.104	0.075	0.047	0.052	0.022
Llama	0.115	0.103	0.058	0.129	0.083	0.125	0.115	0.085	0.089	0.065
Grok	0.079	0.067	0.043	0.093	0.059	0.042	0.080	0.038	0.063	0.065
Sonnet	0.102	0.085	0.043	0.125	0.070	0.042	0.113	0.028	0.068	0.065
Mistral-as-Judge										
GPT	0.392	0.321	0.261	0.382	0.333	0.271	0.364	0.292	0.346	0.304
Qwen	0.285	0.250	0.210	0.320	0.236	0.188	0.291	0.170	0.257	0.217
Gemini	0.389	0.415	0.348	0.413	0.381	0.354	0.414	0.321	0.366	0.435
Gemma	0.263	0.193	0.167	0.258	0.204	0.229	0.244	0.208	0.209	0.133
DeepSeek	0.163	0.138	0.123	0.160	0.131	0.250	0.179	0.113	0.136	0.022
Mistral	0.220	0.143	0.116	0.204	0.155	0.229	0.195	0.123	0.168	0.174
Llama	0.223	0.170	0.072	0.249	0.146	0.125	0.217	0.085	0.162	0.130
Grok	0.152	0.165	0.051	0.164	0.119	0.167	0.155	0.066	0.162	0.043
Sonnet	0.350	0.281	0.181	0.388	0.236	0.417	0.340	0.198	0.272	0.261

Table 16: Cross-Lingual Factual Consistency (CLFC) by model, knowledge frequency, epistemic status, and validation type under two independent judges. Lower values indicate less consistency between Bangla and English responses.

Model	Knowledge Frequency			Epistemic Status			Validation Type			
	Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
GPT-as-Judge										
GPT	0.327	0.339	0.326	0.373	0.327	–	0.356	0.330	0.314	0.196
Qwen	0.470	0.518	0.536	0.476	0.520	–	0.500	0.472	0.513	0.478
Gemini	0.428	0.509	0.341	0.484	0.432	–	0.460	0.330	0.476	0.326
Gemma	0.463	0.516	0.471	0.444	0.516	–	0.432	0.519	0.539	0.556
DeepSeek	0.451	0.545	0.471	0.476	0.502	–	0.487	0.519	0.471	0.435
Mistral	0.431	0.451	0.464	0.440	0.450	–	0.452	0.453	0.429	0.413
Llama	0.454	0.384	0.478	0.382	0.473	–	0.393	0.453	0.508	0.457
Grok	0.465	0.433	0.507	0.476	0.464	–	0.422	0.453	0.529	0.543
Sonnet	0.452	0.473	0.471	0.446	0.471	–	0.450	0.481	0.492	0.391
Mistral-as-Judge										
GPT	0.239	0.223	0.304	0.196	0.273	–	0.225	0.274	0.257	0.326
Qwen	0.259	0.232	0.283	0.227	0.264	–	0.246	0.236	0.283	0.261
Gemini	0.344	0.299	0.355	0.289	0.340	–	0.313	0.302	0.356	0.457
Gemma	0.333	0.354	0.348	0.276	0.373	–	0.308	0.358	0.372	0.467
DeepSeek	0.420	0.482	0.449	0.364	0.500	–	0.396	0.443	0.560	0.370
Mistral	0.256	0.290	0.312	0.231	0.306	–	0.233	0.245	0.351	0.413
Llama	0.197	0.188	0.188	0.164	0.214	–	0.152	0.160	0.251	0.348
Grok	0.228	0.246	0.297	0.200	0.257	–	0.222	0.255	0.283	0.283
Sonnet	0.319	0.330	0.319	0.277	0.336	–	0.298	0.311	0.372	0.348

Table 17: Language Anchor Bias (LAB) by model across knowledge frequency, epistemic status, and validation type. Higher values indicate a stronger tendency for model responses to differ depending on whether the question is asked in Bangla or English, reflecting greater language-dependent framing bias.

nearly all thematic categories, indicating that evidence does not fully prevent models from replacing Bengali cultural referents with globally dominant alternatives. Under the GPT judge, Bangla GSR is comparatively lower, often below 0.35 for several models and domains, while English GSR frequently rises above 0.70, especially in traditional medicine and ecology, religion and mythology, and geography and national identity. The Mistral judge assigns higher absolute GSR values, but pre-

Model	History & Politics		Art, Literature, & Cultural Practices		Traditional Medicine & Ecology		Religion, Folklore, & Mythology		Geography & National Identity	
	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN
GPT-as-Judge										
GPT	0.295	0.629	0.203	0.577	0.053	0.657	0.226	0.551	0.286	0.618
Qwen	0.367	0.771	0.247	0.732	0.190	0.857	0.400	0.746	0.409	0.707
Gemini	0.287	0.765	0.176	0.687	0.222	0.833	0.273	0.603	0.231	0.714
Gemma	0.308	0.831	0.202	0.796	0.143	0.824	0.317	0.776	0.241	0.765
DeepSeek	0.347	0.705	0.176	0.791	0.167	0.824	0.333	0.764	0.290	0.667
Mistral	0.444	0.844	0.128	0.791	0.292	0.757	0.344	0.732	0.233	0.811
Llama	0.447	0.734	0.313	0.717	0.240	0.784	0.429	0.836	0.417	0.778
Grok	0.350	0.802	0.235	0.724	0.130	0.833	0.316	0.712	0.519	0.775
Sonnet	0.365	0.808	0.197	0.759	0.083	0.867	0.414	0.783	0.370	0.735
Mistral-as-Judge										
GPT	0.568	0.829	0.727	0.854	0.579	0.828	0.667	0.893	0.478	0.781
Qwen	0.680	0.878	0.680	0.878	0.679	0.892	0.708	0.741	0.613	0.667
Gemini	0.711	0.874	0.679	0.872	0.640	0.919	0.650	0.760	0.655	0.756
Gemma	0.726	0.838	0.540	0.910	0.650	0.844	0.636	0.816	0.692	0.765
DeepSeek	0.586	0.855	0.563	0.868	0.526	0.806	0.429	0.810	0.520	0.800
Mistral	0.658	0.838	0.558	0.853	0.667	0.636	0.591	0.710	0.455	0.697
Llama	0.748	0.845	0.736	0.843	0.778	0.744	0.778	0.924	0.800	0.707
Grok	0.726	0.907	0.653	0.860	0.688	0.789	0.688	0.826	0.656	0.725
Sonnet	0.655	0.858	0.615	0.866	0.533	0.840	0.824	0.800	0.500	0.676

Table 18: Global Substitution Rate (GSR) in the evidence-based setting across categories. Lower values indicate better preservation of culturally appropriate answers, while higher values indicate greater substitution of local references with globally dominant alternatives even when supporting evidence is provided. See Eq. 3. BN: Bangla, EN: English.

serves the same trend: English responses remain highly substitutional across domains, often exceeding 0.80. The effect is broad rather than domain-specific, appearing in history and politics, cultural practices, ecological knowledge, religious folklore, and national identity. A few model–domain exceptions occur, but they do not alter the overall pattern. These results strengthen our central claim that global narrative dominance is not only caused by missing knowledge; even when local evidence is available, English prompts continue to anchor models toward globally dominant interpretations.

Institutional Bias Table 19 shows that institutional framing persists even when models are given local evidence. Under the GPT judge, IBR is generally low in most domains, especially Traditional Medicine and Ecology and Religion, Folklore, and Mythology. However, History and Politics remains consistently higher across models, and Geography and National Identity also shows elevated bias for several systems, indicating that evidence does not fully prevent models from relying on formal or dominant institutional narratives in politically and nationally salient domains. The Mistral judge assigns higher absolute IBR values, but reveals the same domain structure more strongly: History and Politics is the most institutionally biased category, often exceeding 0.60, followed by Geography and National Identity. English prompts often further increase IBR under the Mistral judge, particularly for Gemma, Llama, Grok, and Sonnet. Overall, the table shows that institutional bias is both domain-sensitive and language-conditioned: local evidence reduces some errors, but models continue to privilege institutional authority in domains where cultural knowledge is closely tied to history, statehood, and national identity.

Epistemic Perspective Coverage Table 20 shows that local evidence substantially improves perspective coverage, but coverage remains language-dependent. Across both judges, Bangla prompts generally achieve higher EPC than English prompts in most model–domain combinations, indicating that English still narrows the range of locally relevant perspectives even when supporting evidence is available. Under the GPT judge, EPC is high for Art, Literature, and Cultural Practices, Religion, Folklore, and Mythology, and Geography and National Identity, often exceeding 0.70 in Bangla. Traditional Medicine and Ecology is consistently lower, especially in English, suggesting that ecological and vernacular knowledge remains

Model	History & Politics		Art, Literature, & Cultural Practices		Traditional Medicine & Ecology		Religion, Folklore, & Mythology		Geography & National Identity	
	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN
GPT-as-Judge										
GPT	0.209	0.214	0.060	0.047	0.015	0.000	0.047	0.028	0.093	0.107
Qwen	0.299	0.265	0.094	0.068	0.029	0.015	0.028	0.038	0.200	0.173
Gemini	0.252	0.214	0.077	0.056	0.000	0.015	0.038	0.028	0.147	0.107
Gemma	0.209	0.274	0.085	0.068	0.015	0.029	0.028	0.047	0.133	0.107
DeepSeek	0.291	0.248	0.085	0.073	0.044	0.000	0.028	0.019	0.147	0.107
Mistral	0.350	0.312	0.094	0.085	0.059	0.029	0.019	0.019	0.227	0.147
Llama	0.312	0.278	0.073	0.056	0.059	0.044	0.066	0.038	0.133	0.120
Grok	0.278	0.256	0.085	0.077	0.015	0.015	0.028	0.028	0.120	0.160
Sonnet	0.338	0.308	0.128	0.090	0.103	0.088	0.028	0.057	0.227	0.160
Mistral-as-Judge										
GPT	0.526	0.568	0.145	0.179	0.044	0.059	0.132	0.170	0.267	0.360
Qwen	0.744	0.667	0.244	0.231	0.059	0.088	0.179	0.179	0.427	0.427
Gemini	0.607	0.585	0.175	0.248	0.044	0.088	0.132	0.123	0.387	0.387
Gemma	0.594	0.701	0.205	0.308	0.044	0.176	0.151	0.208	0.440	0.573
DeepSeek	0.594	0.615	0.252	0.333	0.118	0.044	0.179	0.236	0.373	0.400
Mistral	0.701	0.688	0.214	0.282	0.132	0.118	0.142	0.151	0.547	0.493
Llama	0.620	0.726	0.209	0.376	0.191	0.235	0.255	0.292	0.493	0.533
Grok	0.645	0.692	0.231	0.278	0.103	0.118	0.142	0.160	0.387	0.453
Sonnet	0.692	0.722	0.265	0.303	0.118	0.147	0.160	0.217	0.533	0.613

Table 19: Institutional Bias Rate (IBR) in the evidence-based setting across thematic categories. Lower values indicate less reliance on institutional or globally legitimized framings when local supporting evidence is provided. BN: Bangla, EN: English.

Model	History & Politics		Art, Literature, & Cultural Practices		Traditional Medicine & Ecology		Religion, Folklore, & Mythology		Geography & National Identity	
	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN
GPT-as-Judge										
GPT	0.739	0.723	0.769	0.786	0.596	0.570	0.748	0.741	0.784	0.735
Qwen	0.698	0.610	0.697	0.683	0.572	0.520	0.737	0.695	0.766	0.680
Gemini	0.703	0.673	0.766	0.726	0.594	0.487	0.756	0.702	0.799	0.671
Gemma	0.749	0.692	0.782	0.727	0.646	0.558	0.749	0.681	0.759	0.671
DeepSeek	0.690	0.611	0.731	0.711	0.624	0.553	0.763	0.679	0.750	0.730
Mistral	0.699	0.675	0.750	0.733	0.641	0.559	0.720	0.718	0.751	0.748
Llama	0.632	0.577	0.711	0.661	0.525	0.493	0.675	0.643	0.712	0.653
Grok	0.672	0.640	0.731	0.687	0.516	0.476	0.715	0.689	0.726	0.750
Sonnet	0.745	0.680	0.799	0.748	0.628	0.528	0.768	0.715	0.795	0.779
Mistral-as-Judge										
GPT	0.665	0.662	0.784	0.739	0.689	0.672	0.744	0.749	0.719	0.658
Qwen	0.591	0.566	0.697	0.667	0.616	0.570	0.692	0.681	0.603	0.627
Gemini	0.625	0.570	0.742	0.722	0.681	0.631	0.743	0.691	0.675	0.637
Gemma	0.666	0.550	0.740	0.691	0.712	0.619	0.721	0.694	0.719	0.636
DeepSeek	0.621	0.554	0.748	0.680	0.732	0.631	0.751	0.664	0.701	0.590
Mistral	0.612	0.577	0.755	0.705	0.735	0.643	0.741	0.683	0.722	0.659
Llama	0.541	0.507	0.669	0.660	0.569	0.512	0.641	0.627	0.586	0.583
Grok	0.600	0.584	0.707	0.684	0.618	0.579	0.686	0.651	0.677	0.636
Sonnet	0.656	0.626	0.790	0.734	0.741	0.606	0.758	0.717	0.721	0.663

Table 20: Epistemic Perspective Coverage (EPC) in the evidence-based setting across categories. Higher values indicate broader coverage of locally relevant perspectives when supporting evidence is provided. BN: Bangla, EN: English.

difficult for models to represent broadly. The Mistral judge gives slightly lower absolute scores but preserves the same pattern: Bangla responses usually retain broader epistemic coverage, while English responses decline most clearly in History and Politics, Geography and National Identity, and Tra-

Model	History & Politics	Art, Literature, & Cultural Practices	Traditional Medicine & Ecology	Religion, Folklore, & Mythology	Geography & National Identity
GPT-as-Judge					
GPT	0.214	0.201	0.176	0.151	0.240
Qwen	0.329	0.248	0.309	0.189	0.400
Gemini	0.286	0.218	0.265	0.179	0.267
Gemma	0.179	0.162	0.103	0.132	0.200
DeepSeek	0.171	0.120	0.088	0.075	0.187
Mistral	0.192	0.124	0.074	0.104	0.213
Llama	0.329	0.239	0.368	0.283	0.400
Grok	0.363	0.235	0.382	0.236	0.400
Sonnet	0.256	0.188	0.147	0.142	0.267
Mistral-as-Judge					
GPT	0.560	0.573	0.529	0.660	0.693
Qwen	0.517	0.483	0.397	0.481	0.547
Gemini	0.517	0.551	0.485	0.623	0.653
Gemma	0.470	0.504	0.544	0.538	0.493
DeepSeek	0.338	0.312	0.338	0.340	0.293
Mistral	0.419	0.368	0.279	0.368	0.453
Llama	0.487	0.423	0.412	0.472	0.547
Grok	0.547	0.449	0.441	0.472	0.560
Sonnet	0.547	0.534	0.471	0.557	0.533

Table 21: Cross-Lingual Factual Consistency (CLFC) in the evidence-based setting across thematic categories. Higher values indicate stronger semantic consistency between Bangla and English responses to the same culturally grounded question.

Model	History & Politics	Art, Literature, & Cultural Practices	Traditional Medicine & Ecology	Religion, Folklore, & Mythology	Geography & National Identity
GPT-as-Judge					
GPT	0.316	0.406	0.382	0.358	0.360
Qwen	0.359	0.453	0.544	0.377	0.440
Gemini	0.427	0.538	0.471	0.481	0.427
Gemma	0.466	0.568	0.559	0.547	0.480
DeepSeek	0.372	0.607	0.515	0.481	0.400
Mistral	0.419	0.573	0.382	0.434	0.480
Llama	0.286	0.419	0.412	0.406	0.307
Grok	0.397	0.487	0.500	0.443	0.347
Sonnet	0.444	0.564	0.500	0.481	0.467
Mistral-as-Judge					
GPT	0.365	0.312	0.388	0.333	0.307
Qwen	0.278	0.333	0.265	0.330	0.280
Gemini	0.325	0.346	0.338	0.396	0.387
Gemma	0.282	0.385	0.382	0.415	0.333
DeepSeek	0.402	0.419	0.353	0.453	0.413
Mistral	0.325	0.321	0.324	0.396	0.413
Llama	0.252	0.218	0.162	0.255	0.147
Grok	0.299	0.286	0.206	0.321	0.253
Sonnet	0.372	0.372	0.397	0.406	0.400

Table 22: Language Anchor Bias (LAB) across categories. Higher values indicate stronger language-dependent shifts between Bangla and English prompts, reflecting greater anchoring of responses to the prompt language.

ditional Medicine and Ecology. Model-wise, GPT, Gemini, Gemma, and Sonnet tend to show stronger coverage, whereas Llama is consistently lower. Overall, the table supports our central claim that evidence helps models include more local perspectives, but does not fully prevent English-induced epistemic narrowing.

Cross-lingual Factual Consistency Table 21 shows that cross-lingual factual consistency remains limited even when local evidence is provided. Under the GPT judge, CLFC is low across most domains, often below 0.30, with the weakest consistency appearing in Religion, Folklore, and Mythology and Traditional Medicine and Ecology. Llama, Grok, and Qwen achieve the highest GPT-judged scores in some domains, especially Geography and National Identity, but no model is consistently stable across all categories. The Mistral judge assigns higher absolute CLFC values, yet preserves the same broad pattern: GPT, Gemini, Grok, and Sonnet are comparatively stronger, while DeepSeek and Mistral remain weaker. Geography and National Identity is often the most stable domain, likely because such knowledge is more standardized, whereas ecological, religious, and folklore-based knowledge remains more cross-lingually fragile. Overall, the table supports our claim that evidence improves grounding but does not guarantee language-invariant factual answers.

Language Anchor Bias Table 22 shows that language anchoring is substantial across domains. Under the GPT judge, most models show moderate to high LAB, often between 0.40 and 0.60. The effect is strongest in Art, Literature, and Cultural Practices and Traditional Medicine and Ecology, where models such as Gemma, DeepSeek, Mistral, and Sonnet show particularly large Bangla–English shifts. GPT and Llama are comparatively more stable, but still exhibit non-trivial anchoring across all domains. The Mistral judge assigns lower absolute LAB values for several models, yet preserves the same conclusion: prompt language continues to alter model framing, especially for DeepSeek, Gemma, Sonnet, and Mistral. Domain-wise, culturally interpretive domains such as folklore, cultural practices, and ecological knowledge remain especially sensitive to language. Overall, the table supports our claim that language is not a neutral interface: asking the same Bengali cultural question in Bangla or English can shift the epistemic framing of the response, even when the underlying content is unchanged.

F.2.2 Analysis on Knowledge Frequency, Epistemic Status, and Validation Type

Global Substitution Table 23 reports GSR in the evidence-based setting across sociocultural annotations. Because GSR measures substitution away from a locally grounded interpretation, it is computed only for local and contested items; the global

Language	Model	Knowledge Frequency			Epistemic Status			Validation Type			
		Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
					GPT-as-Judge						
Bangla	GPT	0.216	0.269	0.219	0.206	0.154	–	0.263	0.299	0.107	0.133
	Qwen	0.375	0.269	0.365	0.288	0.310	–	0.294	0.362	0.286	0.188
	Gemini	0.221	0.269	0.208	0.216	0.103	–	0.164	0.326	0.150	0.150
	Gemma	0.266	0.302	0.174	0.219	0.167	–	0.175	0.315	0.236	0.150
	DeepSeek	0.258	0.302	0.235	0.234	0.267	–	0.182	0.338	0.261	0.130
	Mistral	0.309	0.364	0.205	0.272	0.261	–	0.250	0.400	0.233	0.130
	Llama	0.308	0.414	0.412	0.357	0.400	–	0.370	0.453	0.292	0.222
	Grok	0.270	0.312	0.342	0.312	0.172	–	0.300	0.345	0.260	0.227
	Sonnet	0.279	0.344	0.229	0.267	0.250	–	0.164	0.392	0.246	0.200
English	GPT	0.589	0.621	0.600	0.589	0.471	–	0.625	0.627	0.626	0.292
	Qwen	0.741	0.775	0.745	0.759	0.714	–	0.753	0.726	0.800	0.846
	Gemini	0.674	0.744	0.736	0.713	0.605	–	0.707	0.708	0.783	0.615
	Gemma	0.808	0.847	0.738	0.785	0.769	–	0.744	0.845	0.802	0.708
	DeepSeek	0.746	0.768	0.704	0.760	0.556	–	0.768	0.736	0.769	0.636
	Mistral	0.763	0.858	0.745	0.777	0.829	–	0.714	0.832	0.840	0.667
	Llama	0.715	0.793	0.722	0.749	0.722	–	0.671	0.742	0.791	0.885
	Grok	0.785	0.813	0.657	0.759	0.684	–	0.722	0.774	0.811	0.640
	Sonnet	0.757	0.856	0.720	0.795	0.743	–	0.716	0.800	0.853	0.632
					Mistral-as-Judge						
Bangla	GPT	0.650	0.535	0.717	0.644	0.667	–	0.590	0.535	0.778	0.857
	Qwen	0.686	0.653	0.701	0.677	0.480	–	0.500	0.716	0.726	0.700
	Gemini	0.683	0.702	0.650	0.672	0.696	–	0.638	0.678	0.778	0.375
	Gemma	0.571	0.726	0.614	0.626	0.633	–	0.568	0.709	0.643	0.444
	DeepSeek	0.525	0.621	0.462	0.560	0.667	–	0.476	0.577	0.605	0.364
	Mistral	0.529	0.607	0.674	0.581	0.500	–	0.700	0.598	0.564	0.286
	Llama	0.692	0.782	0.788	0.758	0.750	–	0.742	0.753	0.772	0.765
	Grok	0.708	0.685	0.667	0.686	0.692	–	0.550	0.706	0.723	0.833
	Sonnet	0.632	0.675	0.556	0.607	0.864	–	0.512	0.663	0.641	0.700
English	GPT	0.905	0.765	0.894	0.860	0.625	–	0.864	0.818	0.881	0.625
	Qwen	0.842	0.855	0.804	0.831	0.917	–	0.808	0.828	0.882	0.842
	Gemini	0.870	0.811	0.899	0.864	0.778	–	0.879	0.820	0.902	0.800
	Gemma	0.864	0.846	0.849	0.859	0.781	–	0.815	0.865	0.857	0.824
	DeepSeek	0.820	0.855	0.847	0.848	0.733	–	0.797	0.868	0.846	0.737
	Mistral	0.782	0.815	0.768	0.764	0.789	–	0.754	0.801	0.839	0.667
	Llama	0.847	0.818	0.848	0.833	0.794	–	0.782	0.835	0.864	0.840
	Grok	0.855	0.846	0.870	0.847	0.903	–	0.822	0.847	0.910	0.800
	Sonnet	0.826	0.815	0.865	0.834	0.708	–	0.833	0.838	0.875	0.571

Table 23: Global Substitution Rate (GSR) in the evidence-based setting across sociocultural annotations. Lower values indicate better preservation of local cultural referents; higher values indicate stronger substitution with globally dominant alternatives. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition. Global epistemic-status column omitted; see Eq. 3.

epistemic-status column is therefore omitted. The central pattern is a large and consistent Bangla–English gap. Under the GPT judge, Bangla prompts generally produce low-to-moderate substitution rates, mostly around 0.20–0.40, whereas English prompts increase GSR sharply, often to 0.70–0.85. This increase appears for every model and across all annotation types, including frequent knowledge, local-consensus items, institutional knowledge, and oral traditions. Therefore, global substitution is not confined to rare or obscure cultural content.

The Mistral judge assigns higher absolute GSR values, especially for Bangla responses, but confirms the same directional effect. English prompts remain substantially more substitutional, with many scores above 0.80 across knowledge-frequency, epistemic-status, and validation-type categories. The effect is particularly strong for Qwen, Gemini, Gemma, Llama, and Grok, while GPT is relatively more robust under the GPT judge but still shows large English-side increases. Validation-type results are especially important: even local-consensus and oral-tradition items, whose legitimacy is grounded in community knowledge rather than global institutions, show high English GSR. Overall, the table provides strong evidence that

global narrative dominance persists despite evidence injection. Models are not merely missing local knowledge; when prompted in English, they continue to prioritize globally salient referents over the local cultural grounding provided in the context.

Institutional Bias Table 24 shows that institutional bias persists even when models are given local evidence, but its strength depends strongly on judge calibration and sociocultural category. Under the GPT judge, IBR remains low overall, especially for local-consensus, niche, and oral-tradition items. However, institutional and global-status items consistently show higher IBR, and contested items often show elevated values for models such as `Mistral`, `DeepSeek`, `Grok`, and `Sonnet`. This indicates that even with evidence, models are more likely to invoke institutional authority when the item already involves formal, global, or disputed knowledge.

The Mistral judge assigns substantially higher IBR values, but reveals the same structure more clearly. Institutional validation is the most biased category across almost all models, often exceeding 0.50 and reaching 0.655 for `Llama` in English. Global and contested items also show high IBR, frequently above 0.55, whereas local items remain lower. This gap suggests that models distinguish, implicitly or explicitly, between locally grounded knowledge and knowledge that is more easily framed through formal authority. Importantly, English prompts often increase IBR under the Mistral judge, especially for `Gemma`, `DeepSeek`, `Llama`, `Grok`, and `Sonnet`. Therefore, English not only changes the content of responses, but also shifts the source of epistemic authority toward globally legitimized institutions.

Overall, the table strengthens our claim that evidence availability is insufficient for cultural robustness. Models may incorporate local evidence while still organizing answers through institutional frames, particularly for contested, global, and institutionally validated knowledge. This indicates that global narrative dominance operates not only through factual substitution, but also through authority substitution: local evidence is reinterpreted through dominant institutional epistemologies rather than preserved on its own cultural terms.

Epistemic Perspective Coverage Table 25 shows that evidence substantially improves epistemic perspective coverage, but does not make coverage language-invariant. Across both judges, Bangla prompts usually produce higher EPC than English prompts, especially for contested items and for responses judged by Mistral. This is important because contested knowledge requires representing plural local interpretations; the consistent drop in English indicates that English prompts still narrow culturally situated disagreement into less diverse framings. The pattern is strongest for `Qwen`, `Gemma`, `DeepSeek`, `Llama`, and `Sonnet`, while `GPT` is comparatively more stable.

The sociocultural breakdown shows that EPC degradation is not simply a function of rarity. Rare items often receive coverage comparable to, or higher than, frequent items once evidence is provided, suggesting that local evidence can help models recover less common knowledge. In contrast, contested and global-status items are more fragile, particularly in English, indicating that models struggle less with the availability of facts than with preserving epistemic plurality and culturally specific framing. Validation-type results further support this interpretation: local-consensus and oral-tradition items often receive relatively high EPC, but English prompts still reduce coverage for many models. Overall, the table reinforces the central finding that evidence helps models include more local perspectives, yet English continues to induce epistemic narrowing rather than fully preserving the plurality present in Bengali cultural knowledge.

Cross-lingual Factual Consistency Table 26 shows that providing local evidence improves cross-lingual factual consistency, but does not make model responses language-invariant. Under the GPT judge, CLFC remains low for most models, often below 0.30 across knowledge-frequency, epistemic-status, and validation-type groups. `Llama`, `Grok`, and `Qwen` are comparatively stronger under this judge, while `DeepSeek`, `Mistral`, and `Gemma` show consistently weak alignment between Bangla and English responses. Under the Mistral judge, absolute CLFC scores are higher, with `GPT`, `Gemini`, `Sonnet`, and `Grok` showing the strongest consistency, but several models still remain far from stable.

The sociocultural breakdown reveals where cross-lingual instability is most pronounced. Local-consensus and oral-tradition items often receive lower CLFC than institutional or globally framed items, especially under the GPT judge. This suggests that models preserve factual meaning more

Language	Model	Knowledge Frequency			Epistemic Status			Validation Type			
		Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
					GPT-as-Judge						
Bangla	GPT	0.089	0.146	0.029	0.054	0.188	0.191	0.019	0.190	0.016	0.000
	Qwen	0.143	0.192	0.087	0.090	0.250	0.267	0.075	0.265	0.026	0.000
	Gemini	0.107	0.166	0.065	0.070	0.188	0.231	0.057	0.214	0.031	0.000
	Gemma	0.098	0.158	0.036	0.056	0.104	0.236	0.028	0.206	0.016	0.000
	DeepSeek	0.125	0.189	0.072	0.079	0.271	0.253	0.066	0.243	0.037	0.000
	Mistral	0.152	0.228	0.087	0.095	0.396	0.293	0.066	0.294	0.052	0.000
	Llama	0.121	0.200	0.094	0.081	0.250	0.280	0.066	0.259	0.037	0.000
	Grok	0.134	0.163	0.072	0.063	0.292	0.249	0.075	0.227	0.026	0.000
	Sonnet	0.174	0.228	0.116	0.126	0.333	0.284	0.123	0.294	0.063	0.022
English	GPT	0.071	0.130	0.072	0.059	0.146	0.173	0.066	0.166	0.016	0.000
	Qwen	0.107	0.158	0.116	0.074	0.271	0.222	0.094	0.214	0.031	0.000
	Gemini	0.080	0.138	0.058	0.059	0.104	0.196	0.057	0.174	0.021	0.000
	Gemma	0.098	0.177	0.072	0.081	0.229	0.213	0.047	0.222	0.031	0.022
	DeepSeek	0.125	0.118	0.109	0.070	0.229	0.191	0.085	0.190	0.026	0.000
	Mistral	0.156	0.169	0.094	0.079	0.271	0.267	0.075	0.251	0.031	0.000
	Llama	0.125	0.172	0.036	0.054	0.229	0.262	0.028	0.225	0.037	0.000
	Grok	0.121	0.161	0.072	0.070	0.188	0.240	0.066	0.219	0.026	0.000
	Sonnet	0.134	0.208	0.094	0.097	0.271	0.271	0.094	0.257	0.058	0.000
					Mistral-as-Judge						
Bangla	GPT	0.290	0.304	0.152	0.169	0.417	0.440	0.160	0.422	0.099	0.000
	Qwen	0.348	0.454	0.341	0.270	0.729	0.582	0.292	0.594	0.162	0.043
	Gemini	0.277	0.377	0.239	0.189	0.542	0.529	0.217	0.489	0.115	0.022
	Gemma	0.330	0.372	0.239	0.221	0.500	0.520	0.226	0.497	0.147	0.022
	DeepSeek	0.299	0.408	0.297	0.245	0.583	0.516	0.274	0.503	0.168	0.087
	Mistral	0.339	0.454	0.304	0.257	0.667	0.591	0.255	0.583	0.162	0.065
	Llama	0.366	0.431	0.261	0.268	0.542	0.560	0.264	0.559	0.173	0.022
	Grok	0.330	0.403	0.283	0.241	0.604	0.533	0.236	0.521	0.168	0.087
	Sonnet	0.366	0.470	0.290	0.275	0.583	0.618	0.283	0.588	0.178	0.109
English	GPT	0.277	0.358	0.254	0.191	0.438	0.524	0.208	0.473	0.110	0.087
	Qwen	0.326	0.414	0.341	0.252	0.583	0.564	0.274	0.543	0.168	0.065
	Gemini	0.339	0.369	0.261	0.257	0.333	0.502	0.264	0.487	0.152	0.087
	Gemma	0.415	0.482	0.355	0.309	0.667	0.640	0.340	0.620	0.209	0.109
	DeepSeek	0.420	0.386	0.355	0.311	0.604	0.502	0.377	0.521	0.215	0.087
	Mistral	0.384	0.439	0.333	0.293	0.562	0.582	0.274	0.575	0.215	0.065
	Llama	0.491	0.524	0.355	0.374	0.688	0.649	0.340	0.655	0.293	0.174
	Grok	0.379	0.439	0.326	0.286	0.583	0.582	0.292	0.572	0.188	0.109
	Sonnet	0.406	0.513	0.333	0.340	0.625	0.613	0.321	0.623	0.262	0.043

Table 24: Institutional Bias Rate (IBR) in the evidence-based setting across sociocultural annotations. Lower values indicate less reliance on institutional or globally legitimized framings when local evidence is provided. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition.

reliably when knowledge is formalized or globally legible, but struggle when validity depends on community memory, oral transmission, or local consensus. Contested items are also fragile for several models, indicating that cross-lingual inconsistency is amplified when culturally grounded questions involve multiple possible interpretations.

Overall, the table supports a central claim of our study: evidence availability is not sufficient for cultural robustness. Even when the same local evidence is provided, models frequently produce semantically different claims across Bangla and English. This indicates that cross-lingual failures are not merely retrieval errors, but reflect language-conditioned differences in how models select, organize, and prioritize culturally situated knowledge.

Language Anchor Bias Table 27 shows that language anchoring persists even in the evidence-based setting, where models are explicitly given local supporting context. Under the GPT judge, LAB remains moderate to high across most models and sociocultural categories, often exceeding 0.45 for Gemma, Sonnet, DeepSeek, and Mistral. GPT and Llama are comparatively more stable, but still show non-trivial anchoring, indicating that evi-

Language	Model	Knowledge Frequency			Epistemic Status			Validation Type			
		Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
GPT-as-Judge											
Bangla	GPT	0.753	0.720	0.775	0.740	0.743	0.743	0.733	0.759	0.715	0.722
	Qwen	0.719	0.680	0.715	0.696	0.735	0.696	0.671	0.701	0.710	0.697
	Gemini	0.771	0.699	0.747	0.747	0.710	0.704	0.732	0.738	0.720	0.721
	Gemma	0.786	0.721	0.772	0.747	0.757	0.758	0.749	0.773	0.704	0.772
	DeepSeek	0.723	0.693	0.755	0.726	0.717	0.691	0.733	0.721	0.670	0.796
	Mistral	0.761	0.681	0.748	0.729	0.706	0.702	0.691	0.730	0.693	0.793
	Llama	0.701	0.626	0.695	0.685	0.604	0.631	0.658	0.663	0.656	0.697
	Grok	0.735	0.648	0.715	0.697	0.702	0.668	0.693	0.691	0.675	0.716
	Sonnet	0.798	0.725	0.788	0.769	0.727	0.750	0.780	0.776	0.721	0.745
English	GPT	0.790	0.685	0.765	0.753	0.673	0.707	0.767	0.738	0.699	0.754
	Qwen	0.692	0.610	0.661	0.663	0.564	0.628	0.659	0.642	0.634	0.685
	Gemini	0.712	0.646	0.698	0.688	0.689	0.652	0.700	0.675	0.666	0.677
	Gemma	0.719	0.635	0.767	0.705	0.647	0.659	0.746	0.683	0.664	0.678
	DeepSeek	0.689	0.627	0.703	0.681	0.560	0.642	0.683	0.649	0.653	0.742
	Mistral	0.737	0.654	0.741	0.719	0.642	0.664	0.754	0.689	0.673	0.723
	Llama	0.647	0.586	0.632	0.626	0.574	0.600	0.617	0.620	0.592	0.649
	Grok	0.691	0.629	0.680	0.671	0.643	0.636	0.672	0.651	0.661	0.673
	Sonnet	0.732	0.671	0.740	0.720	0.605	0.691	0.726	0.705	0.685	0.714
Mistral-as-Judge											
Bangla	GPT	0.734	0.707	0.749	0.742	0.667	0.700	0.764	0.695	0.739	0.796
	Qwen	0.643	0.631	0.681	0.668	0.596	0.608	0.654	0.616	0.681	0.696
	Gemini	0.714	0.673	0.702	0.714	0.662	0.652	0.700	0.667	0.727	0.723
	Gemma	0.718	0.692	0.735	0.734	0.686	0.663	0.743	0.686	0.716	0.771
	DeepSeek	0.712	0.686	0.719	0.723	0.624	0.671	0.719	0.666	0.741	0.772
	Mistral	0.721	0.678	0.727	0.732	0.672	0.646	0.741	0.663	0.737	0.764
	Llama	0.618	0.589	0.624	0.633	0.513	0.570	0.620	0.574	0.643	0.661
	Grok	0.669	0.637	0.690	0.685	0.627	0.609	0.656	0.627	0.705	0.712
	Sonnet	0.756	0.704	0.753	0.761	0.688	0.677	0.726	0.691	0.791	0.791
English	GPT	0.693	0.697	0.721	0.729	0.670	0.652	0.707	0.675	0.724	0.795
	Qwen	0.619	0.613	0.655	0.650	0.581	0.579	0.639	0.588	0.668	0.687
	Gemini	0.655	0.642	0.664	0.691	0.569	0.587	0.674	0.619	0.692	0.675
	Gemma	0.652	0.607	0.667	0.669	0.496	0.592	0.663	0.584	0.697	0.693
	DeepSeek	0.616	0.623	0.631	0.660	0.473	0.580	0.603	0.593	0.671	0.707
	Mistral	0.632	0.638	0.706	0.685	0.527	0.606	0.674	0.611	0.690	0.736
	Llama	0.595	0.573	0.590	0.616	0.477	0.541	0.556	0.557	0.627	0.673
	Grok	0.637	0.618	0.656	0.656	0.607	0.588	0.614	0.603	0.690	0.664
	Sonnet	0.682	0.666	0.695	0.701	0.540	0.657	0.664	0.641	0.740	0.727

Table 25: Epistemic Perspective Coverage (EPC) in the evidence-based setting across sociocultural annotations. Higher values indicate broader coverage of locally relevant perspectives. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition.

dence reduces neither the role of prompt language nor the tendency to reframe answers differently in Bangla and English.

The Mistral judge assigns lower absolute LAB values, but confirms the same pattern: language-dependent shifts remain visible despite evidence injection. This is especially important because the relevant local information is available in both language conditions. Therefore, the observed shifts cannot be explained solely by missing knowledge or retrieval failure. Instead, they suggest that models use the provided evidence differently depending on the prompt language.

The sociocultural breakdown further shows that evidence does not uniformly stabilize all types of knowledge. LAB remains substantial for frequent and occasional items, demonstrating that anchoring is not restricted to rare knowledge. Local-consensus, niche, and oral-tradition items also show elevated LAB for many models, suggesting that community-grounded forms of legitimacy are especially vulnerable to language-conditioned reframing. Overall, Table 27 strengthens our central claim: even when local evidence is supplied, prompt language continues to act as an epistemic anchor that changes how models interpret and present Bengali cultural knowledge.

Model	Knowledge Frequency			Epistemic Status			Validation Type			
	Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
GPT-as-Judge										
GPT	0.237	0.175	0.203	0.180	0.167	0.244	0.198	0.243	0.136	0.109
Qwen	0.304	0.273	0.297	0.282	0.396	0.276	0.311	0.326	0.220	0.196
Gemini	0.281	0.231	0.217	0.241	0.229	0.253	0.208	0.262	0.246	0.174
Gemma	0.183	0.144	0.174	0.151	0.188	0.178	0.142	0.222	0.084	0.043
DeepSeek	0.174	0.101	0.152	0.146	0.104	0.116	0.094	0.182	0.073	0.087
Mistral	0.170	0.132	0.152	0.146	0.083	0.164	0.123	0.187	0.115	0.022
Llama	0.339	0.290	0.283	0.297	0.333	0.311	0.292	0.356	0.246	0.152
Grok	0.326	0.304	0.290	0.291	0.333	0.338	0.292	0.366	0.230	0.196
Sonnet	0.201	0.211	0.210	0.198	0.250	0.218	0.179	0.257	0.126	0.217
Mistral-as-Judge										
GPT	0.603	0.586	0.580	0.588	0.583	0.596	0.509	0.596	0.618	0.609
Qwen	0.522	0.468	0.507	0.493	0.500	0.489	0.481	0.532	0.429	0.457
Gemini	0.540	0.569	0.543	0.570	0.438	0.551	0.472	0.567	0.565	0.609
Gemma	0.518	0.487	0.507	0.509	0.521	0.480	0.585	0.527	0.408	0.478
DeepSeek	0.330	0.296	0.391	0.360	0.208	0.280	0.321	0.332	0.325	0.283
Mistral	0.424	0.361	0.384	0.387	0.312	0.396	0.358	0.425	0.356	0.239
Llama	0.460	0.482	0.420	0.464	0.396	0.476	0.377	0.505	0.424	0.478
Grok	0.504	0.487	0.500	0.502	0.458	0.489	0.415	0.543	0.466	0.413
Sonnet	0.531	0.510	0.609	0.545	0.604	0.502	0.557	0.559	0.487	0.500

Table 26: Cross-Lingual Factual Consistency (CLFC) in the evidence-based setting across sociocultural annotations. Higher values indicate stronger semantic consistency between Bangla and English responses to the same culturally grounded question. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition.

Model	Knowledge Frequency			Epistemic Status			Validation Type			
	Occ.	Fre.	Rare	Local	Cont.	Global	Niche	Inst.	Local C.	Oral
GPT-as-Judge										
GPT	0.362	0.352	0.391	0.394	0.250	–	0.377	0.326	0.435	0.326
Qwen	0.388	0.451	0.384	0.453	0.417	–	0.415	0.369	0.487	0.543
Gemini	0.460	0.459	0.543	0.489	0.479	–	0.557	0.422	0.508	0.587
Gemma	0.567	0.493	0.522	0.545	0.521	–	0.547	0.479	0.560	0.652
DeepSeek	0.513	0.476	0.442	0.529	0.417	–	0.585	0.422	0.508	0.609
Mistral	0.464	0.465	0.514	0.502	0.521	–	0.519	0.425	0.524	0.565
Llama	0.344	0.361	0.391	0.403	0.375	–	0.349	0.289	0.471	0.522
Grok	0.473	0.423	0.420	0.455	0.438	–	0.462	0.404	0.471	0.522
Sonnet	0.522	0.479	0.500	0.534	0.521	–	0.594	0.433	0.560	0.522
Mistral-as-Judge										
GPT	0.335	0.350	0.316	0.346	0.319	–	0.371	0.349	0.289	0.391
Qwen	0.290	0.313	0.297	0.311	0.396	–	0.349	0.302	0.267	0.348
Gemini	0.366	0.315	0.413	0.376	0.292	–	0.415	0.318	0.361	0.413
Gemma	0.384	0.321	0.370	0.367	0.271	–	0.377	0.334	0.361	0.370
DeepSeek	0.397	0.397	0.471	0.435	0.333	–	0.406	0.409	0.398	0.500
Mistral	0.357	0.332	0.348	0.347	0.354	–	0.349	0.334	0.346	0.391
Llama	0.259	0.217	0.174	0.230	0.271	–	0.151	0.235	0.230	0.239
Grok	0.254	0.296	0.304	0.277	0.292	–	0.321	0.305	0.241	0.217
Sonnet	0.379	0.369	0.420	0.394	0.229	–	0.377	0.393	0.356	0.413

Table 27: Language Anchor Bias (LAB) in the evidence-based setting across sociocultural annotations. Higher values indicate stronger language-dependent shifts between Bangla and English prompts. Occ.=Occasional, Fre.=Frequent, Cont.=Contested, Inst.=Institutional, Local C.=Local Consensus, Oral=Oral Tradition. The Global column is omitted for this analysis.

F.3 Abstention, Knowledge Gap, and Global Narrative Dominance

Table 28 shows that global narrative dominance is strongly conditioned by prompt language and only

Lang.	Model	GPT-as-Judge			Mistral-as-Judge		
		Abst.	KGAP	GND	Abst.	KGAP	GND
Question-only Experiment Setting							
BN	Sonnet	0.080	0.338	0.402	0.084	0.300	0.490
	GPT	0.046	0.308	0.308	0.071	0.315	0.457
	Gemini	0.043	0.304	0.344	0.045	0.314	0.449
	Gemma	0.032	0.434	0.414	0.024	0.354	0.547
	Qwen	0.021	0.336	0.414	0.050	0.363	0.584
	Llama	0.008	0.396	0.492	0.039	0.548	0.766
	Mistral	0.031	0.367	0.400	0.021	0.275	0.537
	Grok	0.014	0.463	0.425	0.035	0.501	0.622
	DeepSeek	0.050	0.460	0.347	0.047	0.293	0.418
EN	Sonnet	0.073	0.446	0.703	0.084	0.356	0.763
	GPT	0.029	0.435	0.497	0.031	0.310	0.601
	Gemini	0.052	0.512	0.642	0.070	0.477	0.736
	Gemma	0.021	0.577	0.819	0.022	0.460	0.848
	Qwen	0.017	0.513	0.734	0.036	0.494	0.743
	Llama	0.050	0.551	0.782	0.073	0.632	0.880
	Mistral	0.025	0.523	0.715	0.032	0.325	0.748
	Grok	0.029	0.501	0.725	0.059	0.515	0.791
	DeepSeek	0.040	0.554	0.750	0.040	0.434	0.806
Evidence-based Experiment Setting							
BN	Sonnet	0.008	0.356	0.307	0.006	0.271	0.438
	GPT	0.010	0.349	0.262	0.006	0.268	0.435
	Gemini	0.022	0.382	0.303	0.014	0.317	0.516
	Gemma	0.014	0.416	0.335	0.015	0.315	0.536
	Qwen	0.008	0.347	0.346	0.018	0.360	0.552
	Llama	0.008	0.377	0.368	0.013	0.481	0.702
	Mistral	0.007	0.364	0.349	0.010	0.248	0.442
	Grok	0.007	0.406	0.346	0.035	0.438	0.593
	DeepSeek	0.018	0.438	0.339	0.013	0.279	0.403
EN	Sonnet	0.011	0.513	0.805	0.018	0.396	0.741
	GPT	0.007	0.519	0.630	0.013	0.393	0.699
	Gemini	0.013	0.604	0.771	0.021	0.519	0.795
	Gemma	0.015	0.586	0.828	0.021	0.490	0.842
	Qwen	0.007	0.584	0.780	0.021	0.509	0.768
	Llama	0.010	0.591	0.805	0.049	0.639	0.876
	Mistral	0.011	0.558	0.820	0.015	0.371	0.720
	Grok	0.004	0.601	0.782	0.032	0.563	0.821
	DeepSeek	0.033	0.597	0.801	0.025	0.470	0.760

Table 28: Abstention, knowledge gap (KGAP), and global narrative dominance (GND) rates across languages, experiment settings, and LLM judges. Higher KGAP and GND indicate more culturally grounded failure modes. BN: Bangla, EN: English.

partially mitigated by evidence. In the question-only setting, English prompts consistently produce higher knowledge gap (KGAP) and global narrative dominance (GND) rates than Bangla prompts across both LLM judges. Under the GPT judge, GND rises from roughly 0.31–0.49 in Bangla to 0.50–0.82 in English; under the Mistral judge, the same shift is even stronger, with several English-prompt GND scores above 0.75. This indicates that when local knowledge is uncertain, models are more likely to resolve culturally grounded questions through globally dominant narratives in English.

Evidence reduces abstention substantially across nearly all models, showing that models use the provided context to produce answers rather than refuse. However, evidence does not eliminate narrative dominance. In Bangla, GND often decreases or remains moderate, but in English it remains high and in some cases increases, with GPT-judge English GND frequently around 0.75–0.83 and Mistral-judge GND often above 0.75. This pattern is central to our claim: the failure is not only absence of local knowledge, since the relevant evidence is available, but also language-conditioned prioritization of dominant narratives. Model differences are visible, with GPT-5.4 generally more robust and Llama, Gemma, Grok, and DeepSeek showing higher GND, but no model is stable across language and evidence conditions.

Model	GPT-as-Judge			Mistral-as-Judge			Human-as-Judge		
	Abst.	KGAP	GND	Abst.	KGAP	GND	Abst.	KGAP	GND
Sonnet	0.080	0.338	0.402	0.084	0.300	0.490	0.069	0.338	0.214
GPT	0.046	0.308	0.308	0.071	0.315	0.457	0.058	0.371	0.276

Table 29: Human validation subset for the Bangla question-only setting. Human judgments are reported for two representative models and compared against both LLM judges. Abst.: Abstention, KGAP: knowledge-gap, GND: global narrative dominance.

Table 29 provides a validity check for the LLM-judge results on the Bangla question-only subset. Human and LLM judgments are broadly comparable for abstention and KGAP, but differ for GND calibration. The human annotator assigns lower GND than the LLM judges for Sonnet, while assigning a similar or slightly lower value for GPT-5.4 relative to GPT-as-judge and a substantially lower value relative to Mistral-as-judge. This suggests that LLM judges, especially Mistral in this setting, may over-detect global narrative dominance in some Bangla responses, while still preserving the comparative trend that GPT-5.4 is more robust than Sonnet. The human validation supports using LLM judges as scalable diagnostic tools, while motivating judge-specific reporting rather than treating any single judge as definitive.

G Data Release

The CulturalNB dataset will be publicly released under the CC BY-NC-SA 4.0 license.[2]

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/