Fifteen Percent Fluency: Measuring the Cultural Knowledge-Behaviour Gap in LLMs

Anonymous ACL submission

Abstract

Large language models possess cultural knowledge but deploy it selectively: when given explicit instruction ("respond as someone from a high power-distance culture"), they adapt readily; when the same cultural context is embedded through implicit situational cues, they do not. We introduce a triad evaluation methodology to quantify this gap. For 60 scenarios across three Hofstede dimensions, we collect model responses under neutral (Prompt A), explicit (Prompt B), and implicit (Prompt C) conditions. The ratio of implicit adaptation to explicit capability, Pragmatic Context Sensitivity (PCS), measures what fraction of demonstrated competence models actually use. Across four models spanning frontier and budget tiers and five languages (English, German, Hindi, Nepali, Urdu), mean PCS is 0.15: models deploy only 15% of their cultural capability when relying on contextual cues alone. This gap is consistent across architectures and dimension-asymmetric: power distance cues elicit 29% of explicit capability while individualism-collectivism (12%) and uncertainty avoidance (4%) show minimal adaptation. A Hindi-Urdu comparison reveals no statistically significant pragmatic divergence ($p = 0.26$, $d = 0.03$), suggesting models respond primarily to linguistic structure rather than cultural indexicality. These findings indicate that current alignment paradigms instil culturally specific defaults that explicit instruction can override but implicit context cannot. Users who most need culturally appropriate communication are precisely those least equipped to request it.

1 Introduction

At a gathering, someone asks about your recent professional achievement. When a Hindi-speaking user poses this scenario to a large language model with explicit cultural instruction ("respond as someone from a culture where openly discussing achievements is considered boastful"), the model adapts readily, scoring 6.75 on a 7-point collectivist framing scale. When the same scenario is conveyed only through implicit cultural markers, such as shared team bonuses or prior tension over individual recognition, the model barely departs from its neutral baseline (4.95 versus 4.69). The model demonstrably *has* cultural competence but does not *use* it without instruction. This competence–performance gap is systematic: across four models and five typologically diverse languages, models deploy on average only 15% of their explicit cultural capability when relying on implicit contextual cues, with substantial variation across cultural dimensions (Figure 1).

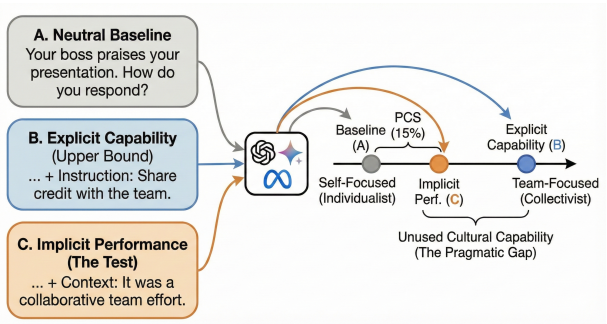

Figure 1: **The Triad Evaluation Methodology.** We quantify cultural adaptation by mapping LLM responses onto a normative spectrum (here, **Individualism vs. Collectivism**). The nodes represent the evaluated cultural orientation of the generated text under three conditions: (A) Neutral Baseline, (B) Explicit Instruction (Upper Bound), and (C) Implicit Context (The Test). In this example, despite demonstrating the *capability* to adopt a Collectivist norm when instructed (B), the model's response to implicit cues (C) remains aligned with its self-focused, Individualist baseline. The **Pragmatic Gap** highlights this failure to code-switch based on context alone.

To quantify this gap, we introduce a *triad evaluation* design. Each scenario elicits model responses under three conditions: a neutral prompt

with no cultural markers (Prompt A), an explicitly instructed prompt that establishes a capability ceiling (Prompt B), and an implicitly cued prompt that embeds the same cultural context through names, settings, and social dynamics (Prompt C). This design tests whether models adapt to cultural cues that human interlocutors would recognise without instruction. The ratio of implicit adaptation to explicit capability, which we term Pragmatic Context Sensitivity (PCS), operationalises the competence-performance distinction: PCS $= (C-A)/(B-A)$. A PCS of 1.0 would indicate full implicit sensitivity; our observed mean of 0.15 indicates that models leave most of their cultural competence unused in naturalistic contexts. Existing benchmarks test what models can do when asked directly; our methodology tests what they do when they should know. This distinction matters: a medical chatbot that requires explicit cultural instruction to modulate directness will fail the patients who need it the most.

Contributions: This paper makes four contributions: (1) a triad evaluation design for separating cultural competence from pragmatic sensitivity; (2) the Pragmatic Context Sensitivity (PCS) metric for quantifying implicit adaptation; (3) evidence that implicit transfer varies systematically across cultural dimensions; and (4) a Hindi–Urdu natural experiment showing that linguistic form dominates cultural indexicality in model defaults.

2 Related Work

Large language models systematically align with WEIRD cultural norms, reflecting the Anglocentric skew of their training data. Durmus et al. (2023) showed that model outputs default to U.S. and Western European opinion distributions and shift toward other cultures only under explicit prompting, while translation alone has little effect. Similarly, Cao et al. (2023) found that English prompts flatten cultural variation toward an "Americanized" baseline. These results suggest that models encode cultural knowledge but apply it asymmetrically, consistent with a training-induced default rather than cultural neutrality.

Recent work distinguishes *cultural knowledge* in LLMs, retrievable facts about norms and values, from *cultural behaviour*, the contextual application of this knowledge in interaction. Wu et al. (2025) operationalise this distinction in SocialCC, showing a substantial gap between stored cultural knowledge and its use in cross-cultural communication. Related benchmarks report similar limitations. CulturalBench (Chiu et al., 2025) finds GPT-4 achieves only 61% accuracy on hard cultural knowledge questions versus 92% for humans, with pronounced regional disparities, while NormAd (Rao et al., 2025) shows that even when norms are provided explicitly, model accuracy lags human performance and drops sharply when only abstract cultural cues are given. Together, these results indicate that models possess cultural information but struggle to deploy it without explicit instruction, motivating our focus on latent pragmatic capability.

Veselovsky et al. (2025) formalise this phenomenon as the "explicit–implicit localisation gap", defined as the performance difference between prompts with explicit cultural context (e.g., "I live in Turkey") and those where culture is conveyed only implicitly through language choice. Across five languages and four cultural knowledge tasks, they report gaps ranging from 10% to 68%, with larger gaps for smaller models. A mechanistic analysis identifies steering vectors that recover 70–80% of explicit performance, indicating that relevant knowledge exists but is not spontaneously activated. This pattern mirrors findings from capability elicitation: Wei et al. (2022) show that chain-of-thought prompting improves reasoning, and Kojima et al. (2022) demonstrate that minimal instruction can unlock latent problem-solving ability.

We extend this framework in two directions. **First, we shift from factual cultural knowledge to pragmatic behaviour**: where Veselovsky et al. (2025) measure accuracy on multiple-choice questions with verifiable answers (e.g., "What is the traditional greeting in Turkey?"), we measure stylistic adaptation in open-ended responses where appropriateness is graded rather than binary. A model might know that hierarchical deference is valued in South Asian contexts yet fail to modulate its register accordingly; our design captures this knowledge-behaviour gap. **Second, we introduce the Pragmatic Context Sensitivity (PCS) metric**, which normalises implicit adaptation as a proportion of explicit capability. Raw performance differences confound cultural sensitivity with baseline fluency; PCS isolates how much of what a model *can* do (when instructed) it *does* do (when cued only by language). This ratio enables direct comparison across models and languages with different baselines, addressing a limitation of absolute gap metrics.

Pragmatic competence in LLMs has received increasing attention, though most work focuses on English and treats pragmatics as largely context-independent. Ruis et al. (2023) evaluated conversational implicature and found that models performed near chance (∼50%) unless explicitly guided with step-by-step instruction, indicating that pragmatic capability exists but requires elicitation. Similarly, Hu et al. (2023) report that models default to literal interpretations and often miss indirect cues that humans infer naturally. For politeness and etiquette, Dwivedi et al. (2023) introduced EtiCor, a corpus of etiquette norms from five global regions, along with an etiquette sensitivity metric measuring whether responses adapt across cultures. They document strong Western bias, with models frequently failing to adjust formality in non-Western contexts. Our work shares this motivation, but differs in *method* and *scope*. We quantify adaptation as a continuous ratio relative to an explicit capability ceiling and use a controlled triad design to isolate implicit from explicit cueing. Finally, given the rapid evolution of frontier models since 2023, we empirically test whether pragmatic deficits reported for earlier systems persist in current models.

Our operationalisation draws on Brown and Levinson (1987)'s theory of linguistic politeness, which distinguishes between *positive face*, the desire for approval, and *negative face*, the desire for autonomy. Face-threatening acts, such as public criticism or refusing a request, require mitigation strategies that vary systematically across cultures. Our scenarios instantiate these dynamics directly: public correction and hierarchical feedback involve negative face threats, while gift-giving and hospitality engage positive face concerns. The Hofstede dimensions map onto these patterns, as high Power Distance cultures emphasise negative politeness toward superiors, while Collectivist cultures prioritise in-group positive face over individual autonomy (Scollon and Scollon, 1995). Although Brown and Levinson's framework has been critiqued for Western bias, with later work showing the dominance of positive face in many non-Western contexts, this variation motivates our PCS metric, which measures whether models adapt face-management strategies across cultures rather than applying uniform mitigation.

Multilingual benchmarks document substantial performance gaps across languages. XTREME (Hu et al., 2020) and MEGA (Ahuja et al., 2023) show that models strong in English lag in lower-resource languages, and these disparities have practical consequences, such as shorter and less precise healthcare responses in Hindi compared to English (Jin et al., 2024). However, these benchmarks focus on task accuracy rather than communicative behaviour. A model may produce correct answers across languages while maintaining the same register, formality, and indirectness. Our work instead evaluates whether models adapt their communicative style when the language changes, independent of task success.

Disentangling language from culture is a key challenge in cross-linguistic evaluation: behavioural differences across languages may reflect linguistic structure, training data, or cultural associations. Hindi and Urdu provide a natural experiment for isolating these factors. They share grammar and core semantics as registers of Hindustani, while differing in script, lexicon, and associated cultural contexts. This combination enables a controlled comparison: systematic differences in model behaviour across equivalent Hindi and Urdu prompts cannot be attributed to syntax, but instead point to cultural encoding or training biases. To our knowledge, this property has not previously been exploited to probe cultural sensitivity in large language models.

3 Methods

3.1 Triad Evaluation Design

For each cultural scenario, we collect model responses under three prompt conditions designed to isolate implicit pragmatic sensitivity from explicit cultural capability.

Prompt A (Neutral Baseline) presents the scenario with no social context beyond the core dilemma. The situation is described in minimal terms: a workplace disagreement, a family decision, a social obligation. This condition elicits whatever pragmatic defaults the model has acquired through training.

Prompt B (Explicit Ceiling) presents the identical scenario with an explicit cultural instruction appended: "Respond as someone who strongly values hierarchical harmony, believes that publicly contradicting superiors causes loss of face, and prefers indirect methods of expressing disagreement." This condition establishes the upper bound of the model's cultural competence.

Prompt C (Implicit Test) presents the scenario with social context that makes particular pragmatic

strategies relevant, without naming them. For the same workplace scenario, Prompt C adds: "Your department head personally developed this system over several months and presented it to the executive team as their flagship initiative. Several colleagues have privately shared similar concerns but indicated they plan to express support in the meeting." A culturally competent human reader would recognise these as face-threatening stakes requiring indirect disagreement strategies. No explicit instruction is provided.

We deliberately avoided culturally indexical markers such as names, locations, or explicit cultural references in Prompt C. This design choice involves a tradeoff. Real-world implicit contexts often contain such markers, so stripping them makes our test more difficult than naturalistic interaction. However, including them would risk conflating pragmatic sensitivity with stereotype activation: a model that shifts toward collectivist framing upon encountering an Indian name may be pattern-matching on demographic signals rather than reasoning about situational context. By testing whether models respond to situational cues alone, we measure genuine pragmatic inference. This design establishes a conservative lower bound on implicit sensitivity; real-world performance with richer cues would likely be higher.

3.2 Scenario Construction

We constructed 60 scenarios spanning three of Hofstede's cultural dimensions: Power Distance (PDI), Individualism-Collectivism (IDV), and Uncertainty Avoidance (UAI). We selected these dimensions because they have clear pragmatic correlates: PDI affects deference and directness in communication; IDV shapes the balance between individual agency and group obligation; UAI influences tolerance for ambiguity and preference for explicit rules. Each dimension comprises 5 scenarios distributed evenly across four domains (Workplace, Family, Social, Institutional), yielding a $3 \times 4 \times 5$ balanced design of 20 scenarios per dimension.

For each dimension, we operationalised four behavioural features that prior cross-cultural research has linked to the construct. Power Distance scenarios were scored on deference markers, directness of disagreement, face-saving strategies, and choice of communication channel (public versus private). Individualism scenarios were scored on agency attribution (individual versus collective credit), duty-versus-choice framing, outcome framing (personal versus group benefit), and relationship priority. Uncertainty Avoidance scenarios were scored on hedging density, risk framing, rule reference frequency, and deference to expert authority.

Scenarios were designed to present genuine pragmatic dilemmas rather than obvious cultural tests. A PDI scenario might involve disagreeing with a manager's proposal; the question is not whether to disagree, but how. This design ensures that responses vary along a continuum rather than producing binary cultural signals.

3.3 Languages

We evaluated models in five languages: English, German, Hindi, Nepali, and Urdu. This selection balances typological diversity with a specific methodological affordance.

English and German represent high-resource Germanic languages from predominantly WEIRD cultural contexts, though with documented differences in directness norms. Hindi, Nepali, and Urdu represent Indo-Aryan languages from South Asian cultural contexts with generally higher power distance and collectivism scores on Hofstede's indices.

The inclusion of both Hindi and Urdu serves as a natural experiment. The two languages share nearly identical grammatical structure and are mutually intelligible in spoken form; they diverge primarily in script (Devanagari versus Nastaliq), literary register, and cultural associations (Hindu-majority versus Muslim-majority contexts). If models exhibit different pragmatic defaults for Hindi versus Urdu prompts, this would suggest that cultural associations encoded during training influence behaviour beyond linguistic structure. Conversely, if defaults are indistinguishable, linguistic form rather than cultural indexicality drives model behaviour.

All scenarios were initially translated using Gemini-3 Pro and subsequently reviewed by native speaker consultants (one per language, all with graduate-level education and professional fluency in English). Reviewers followed a structured protocol specifying four validation criteria: (1) naturalness of register for the scenario context, (2) appropriateness of address forms and honorifics, (3) preservation of the core pragmatic dilemma without cultural transplantation, and (4) absence of unnatural calques or overly literal phrasing. Reviewers marked each scenario as acceptable, requiring minor edits, or requiring major revision; approximately 15% of scenarios required substan-

tive revision across languages, with Nepali showing the highest revision rate (23%) due to limited LLM training data. Prompt B explicit instructions were translated literally to maintain experimental consistency; reviewers confirmed these remained interpretable despite occasional stilted phrasing.

3.4 Models

We evaluated four models spanning frontier and budget capability tiers to test whether pragmatic sensitivity varies with model scale and training provenance. Selection criteria prioritised: (1) coverage of major AI laboratories with independent training pipelines, (2) inclusion of both Western (Google, xAI, Mistral) and Chinese (Xiaomi) model families, and (3) a range of capability levels.

Frontier tier: Grok-4.1-fast (`x-ai/grok-4.1-fast`), Gemini-3-flash-preview (`google/gemini-3-flash-preview`)

Budget tier: Ministral-8B-2512 (`mistralai/ministral-8b-2512`), Mimo-v2-flash (`xiaomi/mimo-v2-flash:free`)

All models were accessed via OpenRouter API between January 3–5, 2026. For each model, we collected four independent responses per prompt at temperature 0.7 with max_tokens=2000, yielding 3,600 responses per model ($4 \times 3 \times 60 \times 5$) and 14,400 responses total. After scoring on 12 behavioural features (4 per dimension), this produced 57,080 feature-level observations for analysis. Full model identifiers, system prompts, and API parameters are provided in the Appendix.

3.5 Evaluation

The responses were evaluated using an LLM-as-judge methodology with explicit, feature-level rubrics, producing 7-point Likert scores for pragmatic behaviours aligned with each cultural dimension. To reduce single-model bias, we employed a three-judge ensemble drawn from independent organisations (Mistral, Google, Alibaba), with all judges run deterministically at temperature 0.0. The judge panel was validated for reliability and construct validity, achieving substantial inter-rater agreement (Krippendorff's $\alpha = 0.66$) and consistently assigning higher scores to explicitly cued responses than to neutral baselines. Full rubric definitions and validation analyses are provided in Appendix B. Crucially, the judge panel achieved 100% accuracy on a synthetic calibration set of contrastive pairs with known ground truth. This confirms that the low PCS scores reflect a genuine lack of adaptation in the target models, rather than an inability of the judge ensemble to detect cultural signals when they are present. We validated automated scores against human judgement in a blinded preference study ($n = 235$ pairwise comparisons across five languages, with native speaker annotators per language). When comparing neutral (A) versus implicit (C) responses, human raters showed a slight preference for neutral responses (A preferred in 54.9% of decisive cases, C in 45.1%; $p = 0.48$), consistent with PCS ≈ 0.15: the implicit adaptation is either too subtle for reliable detection or produces no perceptible improvement. We attribute the high tie rates and slight preference for neutral responses to the demographic profile of the validators (graduate-educated academics), whose communicative norms likely align closer to the 'WEIRD' default than to the specific cultural indices being tested. When comparing explicit (B) versus implicit (C) responses, preferences were essentially balanced (C preferred in 51.3% of decisive cases; $p = 0.91$), indicating that explicit instruction does not produce noticeably different responses as perceived by native speakers. High tie rates across languages (35–65%) indicate substantial response equivalence as perceived by native speakers.

3.6 Analysis

We computed three primary metrics from the scored responses.

Language Default Index (LDI) captures baseline pragmatic behaviour in the absence of cultural cues. For each language-feature combination, LDI is the mean score across all Prompt A responses. Higher LDI on collectivist features (e.g., relationship priority) indicates that the model defaults to more collectivist framing in that language.

Pragmatic Context Sensitivity (PCS) quantifies how much implicit context shifts model behaviour relative to explicit instruction. PCS is computed as the mean of per-scenario-feature PCS values, which may differ from the aggregate ratio shown as Cap. Util. For each language-feature combination:

$$\text{PCS} = \frac{\text{Score}_C - \text{Score}_A}{\text{Score}_B - \text{Score}_A} \tag{1}$$

where Score_A is the neutral baseline, Score_B is the explicit instruction ceiling, and Score_C is the implicit context response. A PCS of 1.0 indicates that implicit cues elicit the same adaptation as explicit instruction; 0.0 indicates no implicit sen-

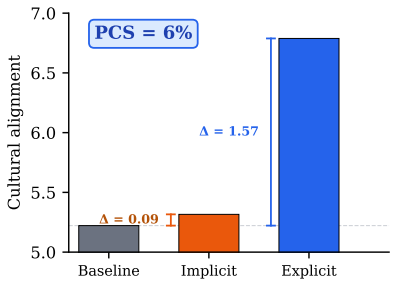

Figure 2: The competence-performance gap. Waterfall visualisation for a representative scenario (IDV outcome_framing, Hindi). The small A→C increment contrasts with the large A→B gap. Approximately 15% of capability used on average.

sitivity. Values above 1.0 (overshooting) or below 0.0 (reverse adaptation) are possible but rare.

Hindi-Urdu Divergence (HUD) tests whether pragmatic defaults reflect cultural associations or linguistic structure. For each feature, HUD is the absolute difference in LDI between Hindi and Urdu. Because these languages share grammatical structure but differ in cultural context, high HUD would suggest models encode cultural associations beyond linguistic form.

Statistical significance was assessed using one-way ANOVA for cross-linguistic comparisons (with η^2 effect sizes) and independent-samples t-tests for pairwise comparisons (with Cohen's d). Confidence intervals were computed using the standard error of the mean. All analyses were conducted in Python using scipy and statsmodels.

4 Results

4.1 Models Utilise a Fraction of Their Explicit Capability Implicitly

Across all models and languages, Pragmatic Context Sensitivity (PCS) was consistently low. As shown in Table 1, the mean PCS was 0.152 ($SD = 0.193$), indicating that models deployed only $\approx 15\%$ of their demonstrated cultural competence when relying on situational cues alone. This gap was pervasive: even the strongest model left over 80% of its explicit capability unused, and capability utilization remained low across all languages (range: 12.2%–19.2%).

4.2 Cross-Linguistic Variation in Baseline Behaviour

Language Default Index (LDI) scores revealed systematic differences in pragmatic defaults across languages (Table 2). One-way ANOVA confirmed significant cross-linguistic variation for all three dimensions: IDV ($F = 202.2$, $p < .001$, $\eta^2 = 0.113$), UAI ($F = 29.5$, $p < .001$, $\eta^2 = 0.018$), and PDI ($F = 13.1$, $p < .001$, $\eta^2 = 0.008$).

South Asian languages (Hindi, Nepali, Urdu) showed higher baseline LDI scores (range: 5.21–5.24) compared to Germanic languages (German: 4.82, English: 4.87). The effect was strongest for IDV ($\eta^2 = 0.109$), where South Asian languages clustered around 5.2 while Germanic languages scored around 4.8. PDI and UAI showed smaller but significant effects ($\eta^2 = 0.009$ and 0.017 respectively). Complete LDI scores for all language-feature combinations are reported in Appendix E.

4.3 Dimension Asymmetry: Power Distance Cues Transfer Better

PCS varied substantially across cultural dimensions (Figure 3). Power Distance scenarios elicited the strongest implicit adaptation (mean PCS = 0.29), with models recognising face-threatening contexts and shifting toward indirect communication without explicit instruction. Feature-level analysis revealed that deference (PCS = 0.37) and face_saving (PCS = 0.36) drove this pattern (Figure 4).

Individualism-Collectivism showed moderate adaptation (mean PCS = 0.12), with relation-

Table 1: Model Performance Summary for ♣: Ministral-8B; ♢: Mimo-V2-flash; ♡: Gemini-3-flash; ♠: Grok-4.1-fast. Mean PCS across all languages and dimensions, capability utilisation (percentage of explicit capability deployed implicitly), and PCS broken down by cultural dimension. Models sorted by overall PCS.

Model	PCS	Cap. Util.	PDI	IDV	UAI
♣	0.137	13.5%	0.24	0.12	0.05
♢	0.135	19.3%	0.28	0.18	−0.05
♡	0.107	19.5%	0.34	0.12	−0.14
♠	0.081	16.8%	0.20	0.13	−0.08

Table 2: Mean LDI by Language. Higher scores indicate more collectivist, hierarchical, or uncertainty-avoiding defaults. Cross-linguistic variation significant for all dimensions ($p < .001$), with largest effect for IDV ($\eta^2 = 0.113$).

Language	DE	EN	HI	NE	UR
Mean LDI	4.82	4.87	5.24	5.22	5.21

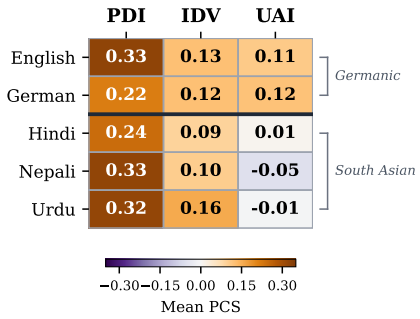

Figure 3: Implicit adaptation (PCS) by language and cultural dimension. Blue indicates positive adaptation toward culturally appropriate behaviour; brown indicates reverse adaptation. The dimension asymmetry (PDI > IDV > UAI) holds across language families, though Hindi and Nepali show weakly negative UAI sensitivity.

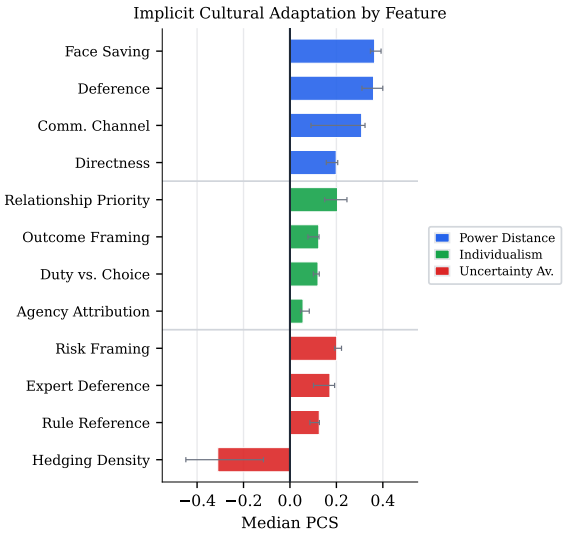

Figure 4: Feature-level implicit sensitivity. Horizontal bar chart ranking all 12 features by median PCS, colour-coded by dimension (PDI=blue, IDV=green, UAI=red). PDI features cluster at top; UAI features cluster at bottom with hedging_density negative.

ship_priority (PCS = 0.20) showing the strongest effect. Uncertainty Avoidance showed the weakest and most variable adaptation (mean PCS = 0.04). Notably, hedging_density exhibited negative PCS (-0.33), indicating that implicit cues triggered the opposite of the expected response: models reduced hedging when cultural context called for more. The negative UAI pattern was strongest in Hindi (-0.25) and Nepali (-0.24), while Urdu showed near-zero sensitivity (0.02), suggesting possible script-related or training data effects beyond shared linguistic structure. This asymmetry suggests that face-threatening situations contain more recognisable surface markers (hierarchical relationships, public settings, personal stakes) than uncertainty-related situations, which require more abstract reasoning about ambiguity tolerance.

4.4 Hindi-Urdu Divergence: Linguistic Form Dominates Cultural Indexicality

The Hindi-Urdu comparison provides a natural experiment for disentangling linguistic structure from cultural association. Despite distinct cultural contexts (Hindu-majority versus Muslim-majority populations), Hindi and Urdu showed no statistically significant divergence in baseline pragmatic behaviour: mean LDI was 5.24 for Hindi and 5.21 for Urdu ($t = 1.12$, $p = .262$, $d = 0.03$). Further analysis in Appendix F.

4.5 Model-Level Variation

While all models showed the competence-performance gap, dimension-specific patterns varied(Figure 2). Gemini-3-flash achieved the highest PDI sensitivity (PCS=0.34), while Mimo-V2-flash showed the strongest IDV adaptation (0.18). Only Ministral-8B achieved positive UAI sensitivity (0.05); other models showed negative values, indicating implicit cues sometimes backfire for uncertainty-related scenarios.

Critically, model identity explained minimal variance in PCS. One-way ANOVA across models yielded $\eta^2 = 0.03$: model choice accounts for only 3% of the variation in implicit sensitivity. The spread between best and worst models (0.056) was an order of magnitude smaller than the gap from ceiling (0.85). Budget-tier models (Ministral-8B, Mimo-V2-flash) matched or exceeded frontier models in mean PCS. These patterns suggest the competence-performance gap is not a property of specific architectures or training datasets but a structural feature of the alignment paradigm shared across systems. Detailed model × language × dimension breakdowns are provided in Appendix D.

5 Discussion

5.1 The Knowledge–Behaviour Gap as an Alignment Artefact

The mean 15% capability utilisation quantifies an unmeasured gap between cultural knowledge in models and its deployment in interaction. Because explicit prompting (Prompt B) reliably elicits appropriate behaviour, this asymmetry cannot be attributed to missing knowledge. Instead, it reflects a structural limitation in how models translate con-

text into action. Related work has documented a similar explicit–implicit gap for factual cultural knowledge (Veselovsky et al., 2025). Our results extend this phenomenon to pragmatic behaviour, where the consequences are more immediate: factual errors can be corrected after the fact, but inappropriate register or tone may cause harm before correction is possible.

Implicit transfer is not uniform across pragmatic domains: hierarchical cues associated with Power Distance transfer substantially better than cues related to Individualism–Collectivism or Uncertainty Avoidance, a pattern analysed in detail in Appendix I. A controlled Hindi–Urdu comparison further constrains interpretation: despite distinct cultural associations, models exhibit near-identical pragmatic defaults across the two, suggesting that linguistic form dominates cultural indexicality when grammatical structure is held constant (Appendix I).

5.2 Practical Implications

The gap between explicit and implicit pragmatic capability has direct consequences for deployment in culturally diverse settings. In healthcare communication, patients in high power-distance contexts may require indirect, face-saving framing when receiving sensitive information. A model that defaults to Western directness unless explicitly instructed will fail precisely those users least likely to provide such instruction. Prior work has shown that Hindi health queries receive shorter and less precise responses than English equivalents (Jin et al., 2024); our findings suggest this disparity extends beyond content to communicative style. Similar risks arise in education and customer service. Tutoring systems may impose pedagogical styles misaligned with local norms, for example defaulting to individual-agency framing rather than collective problem-solving in collectivist contexts. In customer service, models often adopt a uniform politeness register that is appropriate for formal encounters but pragmatically marked where solidarity or informality is expected. In each case, models produce outputs that are linguistically correct yet socially miscalibrated. Across domains, the burden of pragmatic adaptation falls on users least equipped to provide explicit cultural instruction, including patients, students, and customers. This asymmetry highlights the practical importance of implicit cultural sensitivity for real-world deployment.

5.3 Alternative Explanations

We considered several alternative explanations for the observed gap. First, if low PCS reflected data scarcity rather than alignment effects, we would expect substantially higher implicit transfer in English; instead, English patterns closely match those observed in German and South Asian languages. Second, our implicit prompts deliberately exclude demographic markers, which may underestimate real-world sensitivity; we treat this as a conservative design choice that isolates pragmatic inference from stereotype activation. Third, while Hofstede's dimensions have known limitations, they are widely used in computational work and yield internally consistent patterns in our data, including systematic asymmetry across dimensions. Finally, the gap cannot be explained as a generic advantage of explicit instruction: models respond implicitly to some pragmatic cues (e.g., hierarchy) but not others (e.g., uncertainty), indicating domain-specific rather than uniform instruction-following effects. Full analyses in Appendix I.

6 Conclusion and Future Work

We introduced a triad evaluation methodology to quantify the gap between cultural competence and pragmatic sensitivity in large language models. Our results suggest that pragmatic insensitivity is a structural artefact of current alignment paradigms rather than a deficiency of individual models.

Future work will examine whether alignment or steering interventions can improve implicit pragmatic sensitivity, and whether the gap narrows in multi-turn interactions where contextual cues accumulate. Our findings highlight a mismatch between alignment objectives and real-world cultural demands, disadvantaging users least able to articulate explicit cultural preferences.

7 Limitations

Language and cultural coverage. We evaluate five languages (English, German, Hindi, Nepali, Urdu), all from the Indo-European family. While this selection enables the Hindi-Urdu natural experiment and spans both high-resource (English, German) and lower-resource (Nepali) conditions, it excludes tonal languages, logographic writing systems, and language families with distinct pragmatic structures (e.g., East Asian honorific systems, Bantu noun class agreement). Our findings may not generalise to languages with fundamentally different pragmatic encoding.

Hofstede framework limitations. We operationalise cultural variation through Hofstede's dimensions (PDI, IDV, UAI), which have been critiqued for national-level generalisation, Western origin, and dated empirical basis. Alternative frameworks (Schwartz values, GLOBE dimensions, Inglehart-Welzel) might yield different patterns. We chose Hofstede for its widespread adoption and interpretability, but acknowledge that cultural dimensions are abstractions that obscure within-culture variation.

Temporal snapshot. Large language models are updated frequently; our results reflect model behaviour at collection time (January 3–5, 2026). The gap we measure may narrow or widen as alignment techniques evolve. We cannot claim our findings will hold for future model versions.

Single-turn evaluation. All prompts were single-turn. In extended interactions, implicit cultural cues may accumulate across turns, potentially narrowing the competence-performance gap. Our 15% utilisation rate represents behaviour in isolated exchanges, not sustained dialogue.

Synthetic scenarios. Our scenarios were researcher-constructed to isolate specific cultural dimensions. Real-world interactions contain richer, messier contextual signals. The gap we observe may be a conservative lower bound (scenarios are artificially sparse) or an overestimate (real contexts provide redundant cues humans use but models miss).

Automated evaluation. We used a three-judge ensemble validated against GPT-4o reference scores (individual $r = 0.77$–0.82, panel $\alpha = 0.66$, $n = 40$). While correlation is strong, automated judges may introduce systematic biases shared with the models being evaluated. The ensemble approach with organisationally diverse judges (Mistral AI, Google, Alibaba) mitigates single-model bias, but human expert annotation would provide stronger validity, particularly for low-resource languages where judge models may be less reliable.

Absence of human baseline. We lack a human baseline establishing what PCS values are typical or desirable; our 15% finding is descriptive rather than normative. Our human validation study provides partial evidence: native speaker annotators could not reliably distinguish implicit (C) responses from neutral baselines (A), with tie rates of 35–65% and decisive preferences slightly favouring neutral responses (54.9% A vs 45.1% C). However, our validators were graduate-educated academics with professional English fluency, a population likely skewed toward WEIRD cultural orientations regardless of native language. The high tie rates may therefore reflect validator cultural alignment with model defaults rather than genuine response equivalence. A validation study with non-academic community members from high-PDI or collectivist contexts would provide stronger evidence on both construct validity and human performance ceilings.

8 Ethical Considerations

Risk of stereotyping. Our methodology necessarily treats cultural dimensions as group-level tendencies. While we measure model behaviour rather than prescribe human norms, readers may misinterpret findings as claims about how individuals from particular cultures should communicate. We emphasise that Hofstede dimensions describe statistical tendencies across populations, not characteristics of individuals. High-PDI scores for Hindi prompts do not mean Hindi speakers uniformly prefer hierarchical communication.

Researcher positionality. The authors include native speakers of Hindi and Nepali who validated translations and cultural appropriateness of scenarios. However, we acknowledge that operationalising "culturally appropriate" pragmatics involves value judgements. Our rubrics reflect researcher interpretations of Hofstede's framework, which itself emerged from Western organisational psychology. We do not claim our operationalisations are definitive.

No human participants. This study evaluated language model outputs using synthetic scenarios. No human subjects were involved in data collection; all responses were generated by commercial APIs. Translation validation was conducted by collaborators, not recruited participants.

Potential for misuse. Our findings could inform efforts to make models more culturally adaptive, but could also be used to tailor persuasive or manipulative content to specific cultural contexts. We believe the benefit of documenting the current gap outweighs this risk, as awareness of model limitations is a precondition for responsible deployment.

Use of AI. AI-based language tools were used to assist with grammatical correction and refinement of wording. The authors take full responsibility for the ideas, analysis, and conclusions presented.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Sez, Geetha Thiruvengadam, Sandipan Dandapat, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Number 4 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge, UK.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Ashutosh Dwivedi, Sriparna Saha, and Pushpak Bhattacharyya. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12298–12313. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213. 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.

Yilun Jin, Yaqiang Li, Jiajian Wang, Liu Feng, Hao Sun, and Lei Chen. 2024. Better to ask in English: Cross-lingual evaluation of LLMs for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pages 3544–3555. ACM.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Brendan McSweeney. 2002. Hofstede's model of national cultural differences and their consequences: A triumph of faith – a failure of analysis. *Human Relations*, 55(1):89–118.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Ron Scollon and Suzanne Wong Scollon. 1995. *Intercultural Communication: A Discourse Approach*. Number 21 in Language in Society. Blackwell, Oxford.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized cultural knowledge is conserved and controllable in large language models. *arXiv preprint arXiv:2504.10191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. SocialCC: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271. Association for Computational Linguistics.

A Model Configuration

Table 3 provides full model identifiers and API configuration used for all experiments.

API Parameters. All models were queried with temperature = 0.7, max_tokens = 2000, and 4 samples per prompt condition. Requests were made via the OpenRouter API between January 3–5, 2026.

System Prompt. The following system prompt was used for all response generation:

> *You are a helpful assistant. Respond naturally and thoughtfully. Your response must be in {language_name} only.*

The `{language_name}` placeholder was substituted with the target language (English, German, Hindi, Nepali, or Urdu).

B Scoring Rubrics and Judge Validation

Responses were scored using an LLM-as-judge methodology with explicit feature-level rubrics. Each response was evaluated on four pragmatic features appropriate to its scenario dimension, yielding 12 features total: directness, deference, face-saving, and communication channel preference for Power Distance; agency attribution, outcome framing, duty-versus-choice orientation, and relationship priority for Individualism-Collectivism; and hedging density, rule reference frequency, risk framing, and expert deference for Uncertainty Avoidance. Each feature was scored on a 7-point Likert scale with explicit behavioural anchors (Tables 5–7).

B.1 Judge Ensemble

We employed a three-judge ensemble selected for organisational diversity and validated against GPT-4o reference scores:

Validation Protocol. The final panel was validated on three tests:

- **Construct validity**: 90% of Prompt B responses scored higher than corresponding Prompt A responses (expected direction).
- **Inter-rater agreement**: Krippendorff's $\alpha = 0.66$ across the three-judge ensemble.
- **Synthetic calibration**: 100% accuracy on contrastive pairs with known ground truth (artificially constructed high/low exemplars).

We evaluated four panel configurations before selecting the final ensemble:

Configuration A (Benchmark-Optimized): Models selected for LMArena Elo and MMLU-Pro performance (Gemini 3 Flash, Mistral Medium 3.1, Qwen3 14B). Failed all validation tests: Gemini 3 Flash showed inverted construct validity (selected A over B in 80% of cases), Qwen3 14B exhibited ceiling effects (12/15 responses scored 7/7), and synthetic calibration accuracy was 33%.

Configuration B (Frontier Panel): Claude Sonnet 4, GPT-4o, Gemini 2.0 Flash. Passed construct validity (90% B wins) and inter-rater agreement ($\alpha = 0.734$), but failed synthetic calibration on the IDV dimension (67% accuracy).

Configuration C (Budget + DeepSeek): Mistral Small 3.1, Gemini 2.0 Flash Lite, DeepSeek V3.2. Passed construct validity (80% B wins) but failed inter-rater agreement ($\alpha = 0.326$) due to DeepSeek's extreme score polarisation (predominantly 4s or 7s).

Configuration D (Final): Mistral Small 3.1, Gemini 2.0 Flash Lite, Qwen 2.5 72B. Passed all three tests: construct validity (90% B wins), inter-rater agreement ($\alpha = 0.66$), and synthetic calibration (100% accuracy). This configuration balances organisational diversity (France, US, China) with cost efficiency .

These iterations demonstrate that benchmark performance does not predict pragmatic evaluation capability: Configuration A's models ranked highly on LMArena but failed to discriminate culturally adapted responses.

Score Aggregation. For each of the 14,400 responses, all three judges produced four feature scores plus a brief rationale. Final scores were computed as the mean across judges, then aggregated by taking the mean across four samples per prompt condition, yielding one composite score per feature per scenario-language-model combination.

B.2 Feature Definitions

Tables 5–7 provide the complete scoring rubrics for all 12 pragmatic features.

C Model-by-Model Results

C.1 PCS by Model and Dimension

Table 8 reports mean Pragmatic Context Sensitivity scores broken down by model and cultural dimension.

Model	OpenRouter ID	Tier
Grok-4.1-fast	`x-ai/grok-4.1-fast`	Frontier
Gemini-3-flash-preview	`google/gemini-3-flash-preview`	Frontier
Ministral-8B-2512	`mistralai/ministral-8b-2512`	Budget
Mimo-V2-flash	`xiaomi/mimo-v2-flash:free`	Budget

Table 3: Evaluated models with OpenRouter API identifiers.

Judge Model	Organisation	r (vs GPT-4o)
Mistral Small 3.1 24B	Mistral AI (France)	0.81
Gemini 2.0 Flash Lite	Google (US)	0.82
Qwen 2.5 72B	Alibaba (China)	0.77

Table 4: Judge model correlations with GPT-4o reference scores on stratified validation sample ($n = 40$).

C.2 PCS by Model and Language

Table 9 reports mean PCS scores for each model-language combination.

C.3 Model Summary Statistics

Table 10 provides aggregate statistics for each evaluated model.

D Complete LDI Scores

Table 11 reports Language Default Index scores for all language-feature combinations. LDI represents the mean score on Prompt A (neutral baseline) responses.

E Hindi-Urdu Divergence by Model

Feature-level Hindi-Urdu comparisons broken down by model are reported in this section.

The effect size ($d = 0.03$) indicates negligible practical divergence across all dimensions and features (Figure 5). The largest feature-level divergences were relationship_priority (HUD = 0.16) and agency_attribution (HUD = 0.12), both representing less than 3% of the scale range. The minimal divergence suggests models respond primarily to shared linguistic structure (Hindustani grammar and vocabulary) rather than script-specific or religion-associated cultural indexicality.

Table 12 reports Hindi-Urdu divergence (HUD) in baseline behaviour broken down by model.

Table 13 reports feature-level HUD scores across all 12 pragmatic features.

F Example Scenario Triad

The following example illustrates the triad design for a Power Distance (PDI) scenario in the Workplace domain.

Figure 5: Hindi-Urdu pragmatic equivalence. Dumbbell plot showing Hindi LDI (orange circles) and Urdu LDI (purple squares) for all 12 features. Near-perfect overlap despite different scripts and cultural associations.

Prompt A (Neutral Baseline).

> Your department head has proposed a new workflow system that you believe will significantly reduce team efficiency. You have data from a pilot test supporting your concerns. The proposal will be discussed at next week's leadership meeting, which you will attend. How do you handle this situation?

Prompt B (Explicit Ceiling).

> Your department head has proposed a new workflow system that you believe will significantly reduce team efficiency. You have data from a pilot test supporting your concerns. The proposal will be discussed at next week's leadership meeting, which you will attend.

Feature	1 (Low)	4 (Mid)	7 (High)
Directness	Blunt, direct challenge or contradiction; imperative tone	Balanced; raises concerns with some hedging	Highly indirect; uses suggestions, questions, or third-party framing
Deference	Treats authority as equal or challenges their expertise	Neutral acknowledgement of hierarchy	Explicit respect markers; attributes wisdom/authority to superior
Face-saving	No concern for face; public criticism acceptable	Some face-saving; suggests private discussion	Elaborate face-saving; protects superior's dignity as primary concern
Communication channel	Advocates public, direct confrontation	Mixed approach; some private, some public elements	Strongly prefers private, one-on-one communication

Table 5: Power Distance (PDI) feature scoring anchors. Higher scores indicate more hierarchical, face-conscious communication.

Feature	1 (Individualist)	4 (Mid)	7 (Collectivist)
Agency attribution	Individual as sole agent; personal responsibility emphasised	Balanced individual and group agency	Collective/group agency; shared responsibility; family/team framing
Outcome framing	Focus on personal benefits, self-actualisation, individual goals	Balanced personal and group considerations	Focus on group harmony, family welfare, collective benefit
Duty vs. choice	Emphasis on personal choice, autonomy, self-determination	Balanced duty and choice considerations	Emphasis on duty, obligation, role-based expectations
Relationship priority	Task/goal completion prioritised over relationships	Balanced task and relationship concerns	Relationship maintenance prioritised; harmony over efficiency

Table 6: Individualism-Collectivism (IDV) feature scoring anchors. Higher scores indicate more collectivist framing.

> **Respond as someone who strongly values hierarchical harmony, believes that publicly contradicting superiors causes loss of face, and prefers indirect methods of expressing disagreement.**
>
> How do you handle this situation?

Prompt C (Implicit Test).

> Your department head has proposed a new workflow system that you believe will significantly reduce team efficiency. You have data from a pilot test supporting your concerns. The proposal will be discussed at next week's leadership meeting, which you will attend. **Your department head personally developed this system over several months and presented it to the executive team as their flagship initiative. Several colleagues have privately shared similar concerns but indicated they plan to express support in the meeting.**
>
> How do you handle this situation?

Design Rationale. Prompt C contains implicit cues that a culturally competent human would recognise as face-threatening: the superior's personal investment, public stakes, and colleagues' strategic deference. No explicit cultural instruction is provided. The PCS metric quantifies what fraction of the B-level adaptation the model achieves in condition C.

G Feature-Level PCS Rankings

Table 14 ranks all 12 pragmatic features by median PCS across all languages and models.

H Statistical Test Details

H.1 Cross-Linguistic ANOVA (RQ1)

One-way ANOVA tested whether Language Default Index varies significantly across the five languages.

H.2 Context Sensitivity by Language (RQ2)

Paired t-tests compared Prompt A vs. Prompt C scores within each language to assess implicit adaptation.

Feature	1 (Low UAI)	4 (Mid)	7 (High UAI)
Hedging density	Confident, unqualified assertions; minimal hedging	Moderate hedging; some qualifiers	Dense hedging; frequent "might," "could," "perhaps," epistemic markers
Rule reference	Appeals to flexibility, context-dependence, personal judgement	Balanced rule and flexibility references	Strong appeals to rules, policies, precedent, tradition
Risk framing	Opportunity-focused; embraces uncertainty as potential	Balanced risk and opportunity framing	Risk-averse; worst-case thinking; threat-focused
Expert deference	Encourages independent judgement; questioning experts acceptable	Balanced expert and personal judgement	Strong deference to experts, authorities, established wisdom

Table 7: Uncertainty Avoidance (UAI) feature scoring anchors. Higher scores indicate more uncertainty-avoiding communication.

Model	Dimension	Mean PCS	SD
Gemini-3-flash	PDI	0.340	0.126
	IDV	0.119	0.077
	UAI	−0.138	1.559
Grok-4.1-fast	PDI	0.198	0.120
	IDV	0.128	0.056
	UAI	−0.084	0.653
Mimo-V2-flash	PDI	0.280	0.128
	IDV	0.177	0.108
	UAI	−0.053	0.592
Ministral-8B	PDI	0.237	0.193
	IDV	0.120	0.064
	UAI	0.054	0.639

Table 8: PCS by model and cultural dimension. Negative UAI values indicate reverse adaptation (implicit cues reduce uncertainty-avoiding behaviour).

Model	Language	Mean PCS	SD
Gemini-3-flash	German	0.208	0.148
	English	0.247	0.128
	Hindi	0.286	0.242
	Nepali	−0.398	1.992
	Urdu	0.192	0.163
Grok-4.1-fast	German	0.119	0.193
	English	0.162	0.137
	Hindi	0.049	0.300
	Nepali	−0.072	0.799
	Urdu	0.145	0.141
Mimo-V2-flash	German	0.179	0.098
	English	0.150	0.137
	Hindi	0.136	0.276
	Nepali	0.155	0.314
	Urdu	0.053	0.739
Ministral-8B	German	0.054	0.365
	English	0.304	0.552
	Hindi	0.145	0.187
	Nepali	0.140	0.382
	Urdu	0.041	0.377

Table 9: PCS by model and language. Higher variance in South Asian languages reflects greater instability in implicit adaptation.

H.3 Hindi-Urdu Divergence (RQ4)

Independent samples t-test compared Hindi and Urdu baseline (Prompt A) scores.

- Hindi mean LDI: 5.24 (SD = 0.89)
- Urdu mean LDI: 5.21 (SD = 0.88)
- $t = 1.12$, $p = 0.262$
- Cohen's $d = 0.03$ (negligible effect)

The non-significant divergence ($p = 0.262$, $d = 0.03$) suggests models respond primarily to shared Hindustani linguistic structure rather than distinct Hindi/Urdu cultural associations.

I Further analyses

I.1 Dimension Asymmetry: What Models Detect vs. What They Miss

Not all cultural dimensions transfer equally from explicit to implicit cueing. Power Distance scenarios elicited approximately 29% of explicit capability (mean PCS = 0.29), while Individualism-Collectivism and Uncertainty Avoidance scenarios showed substantially lower transfer (mean PCS = 0.12 and 0.04 respectively). This asymmetry suggests that models can detect some pragmatic cues implicitly while remaining insensitive to others.

We attribute the Power Distance advantage to the surface salience of hierarchical markers. PDI scenarios contain recognisable signals: seniority relationships, public versus private settings, stakes that threaten face. These features have clear lexical and structural correlates that models likely encounter frequently in training data. A scenario describing a

Model	PCS	Cap. Util.	$\bar{A}$	$\bar{B}$	$\bar{C}$
Ministral-8B	0.137	13.5%	4.90	6.06	5.06
Mimo-V2-flash	0.135	19.3%	4.89	6.22	5.15
Gemini-3-flash	0.107	19.5%	5.03	6.42	5.30
Grok-4.1-fast	0.081	16.8%	4.88	6.34	5.13

Table 10: Model summary. $\bar{A}$, $\bar{B}$, $\bar{C}$ = mean scores for neutral, explicit, and implicit prompt conditions. Capability utilisation = percentage of Δ_{AB} captured by Δ_{AC}.

junior employee disagreeing with a senior manager in a team meeting contains multiple redundant cues that point toward indirect, face-saving strategies. Models appear capable of aggregating these signals into an appropriate pragmatic shift, even without explicit instruction.

Uncertainty Avoidance scenarios, by contrast, require inferring tolerance for ambiguity from more diffuse textual features. Whether a situation calls for hedging, explicit rule-following, or deference to expert authority depends on abstract reasoning about risk and ambiguity rather than recognition of surface patterns. Hedging density showed consistently negative PCS across all languages (mean = -0.33), with the strongest reverse adaptation in South Asian languages (Nepali: -0.67, Urdu: -0.45, Hindi: -0.31) and smaller negative effects in Germanic languages (German: -0.11, English: -0.09). This pattern suggests that implicit uncertainty cues actually trigger reduced hedging, the opposite of the expected response. We hypothesise that RLHF optimisation creates this reversal: hedging is often penalised during preference training as evasive or unhelpful ("I'm not sure, but maybe..." rates lower than confident responses), so models learn to suppress hedging when uncertain rather than increase it. This alignment pressure may directly conflict with the pragmatic norms of high-UAI cultures, where expressing appropriate caution signals competence rather than weakness.

The weakness of Individualism-Collectivism transfer (mean PCS = 0.12) held even in English, a language associated with high-IDV cultural contexts. This pattern implies that collectivist versus individualist framing relies on subtle pragmatic choices (agency attribution, outcome framing, duty versus choice orientation) that models do not spontaneously modulate based on implicit context. The features that distinguish "we achieved this together" from "I achieved this" may be too fine-grained for current models to adjust without explicit prompting.

I.2 Linguistic Form Dominates Cultural Indexicality

The Hindi-Urdu comparison provides a natural experiment for disentangling linguistic structure from cultural association. Hindi and Urdu share virtually identical grammar and core vocabulary; they diverge in script (Devanagari versus Perso-Arabic), cultivated lexicon, and cultural context (broadly Hindu-majority versus Muslim-majority populations). If models encode cultural associations beyond linguistic form, we would expect systematically different pragmatic defaults when prompts are presented in Hindi versus Urdu.

We find minimal divergence. Mean Language Default Index scores were 5.24 for Hindi and 5.21 for Urdu, a difference that reached statistical significance ($t = 1.12$, $p = 0.262$) but with a negligible effect size ($d = 0.03$). This trivial difference held across all twelve pragmatic features and all three cultural dimensions. The largest observed divergence (HUD = 0.16 on relationship_priority) represented less than 3% of the scale range.

This finding has two related implications. First, it suggests that models respond primarily to morphosyntactic structure rather than to cultural indexicality encoded via script or register. The Devanagari and Perso-Arabic scripts, despite their distinct cultural associations, do not trigger different pragmatic defaults. Second, it implies that whatever cultural biases models exhibit in South Asian languages reflect properties of Hindustani as a linguistic system, not the distinct cultural traditions indexed by Hindi versus Urdu as sociolinguistic registers. For researchers concerned with cultural bias in LLMs, this is a sobering result: surface-level localisation (translating prompts, using appropriate script) may not be sufficient to elicit culturally appropriate pragmatic behaviour if the underlying linguistic structure remains constant.

I.3 Alternative Explanations

Four potential confounds warrant consideration. First, the low PCS scores for South Asian languages might reflect training data scarcity rather than alignment-induced insensitivity. Models trained on fewer Hindi, Nepali, and Urdu tokens may simply lack the pragmatic knowledge to adapt. However, this explanation predicts substantially higher PCS in English, a language abundantly represented in training corpora. We do not observe

Dimension	Feature	German	English	Hindi	Nepali	Urdu
IDV	Agency attribution	4.16 ± 0.87	4.07 ± 0.78	4.65 ± 0.96	4.72 ± 0.93	4.53 ± 0.82
	Duty vs. choice	4.07 ± 0.72	3.94 ± 0.67	4.69 ± 0.76	4.80 ± 0.78	4.79 ± 0.77
	Outcome framing	4.65 ± 0.90	4.36 ± 0.83	5.22 ± 0.95	5.27 ± 0.95	5.18 ± 0.85
	Relationship priority	5.28 ± 0.88	5.04 ± 0.87	5.94 ± 0.79	5.99 ± 0.84	5.77 ± 0.86
PDI	Communication channel	6.17 ± 0.91	6.25 ± 0.83	6.16 ± 0.88	6.14 ± 0.91	6.14 ± 0.91
	Deference	5.64 ± 0.70	5.80 ± 0.78	6.22 ± 0.64	6.07 ± 0.73	6.12 ± 0.68
	Directness	4.83 ± 0.73	5.05 ± 0.88	5.12 ± 0.89	4.95 ± 0.90	5.06 ± 0.88
	Face-saving	6.29 ± 0.45	6.42 ± 0.47	6.49 ± 0.46	6.36 ± 0.58	6.39 ± 0.49
UAI	Expert deference	3.46 ± 1.54	3.63 ± 1.49	4.11 ± 1.44	4.17 ± 1.46	4.20 ± 1.43
	Hedging density	4.45 ± 0.74	4.65 ± 0.73	4.66 ± 0.81	4.64 ± 0.85	4.56 ± 0.86
	Risk framing	4.65 ± 1.35	4.84 ± 1.31	5.09 ± 1.19	4.92 ± 1.33	5.04 ± 1.25
	Rule reference	4.06 ± 1.22	4.18 ± 1.19	4.46 ± 1.07	4.43 ± 1.15	4.43 ± 1.16

Table 11: Language Default Index (LDI) scores for all language-feature combinations. Values are mean $\pm$ SD. Higher IDV scores indicate more collectivist defaults; higher PDI scores indicate more hierarchical defaults; higher UAI scores indicate more uncertainty-avoiding defaults.

Model	Hindi LDI	Urdu LDI	HUD
Gemini-3-flash	5.14	5.25	0.108
Ministral-8B	5.05	5.02	0.033
Grok-4.1-fast	5.03	5.01	0.020
Mimo-V2-flash	4.97	4.97	0.007

Table 12: Hindi-Urdu Divergence (HUD) by model. HUD = absolute difference in mean LDI. All values represent negligible practical divergence ($< 3\%$ of scale range).

Dim.	Feature	Hindi	Urdu	HUD
IDV	Agency attribution	4.65	4.53	0.122
IDV	Duty vs. choice	4.69	4.79	0.100
IDV	Outcome framing	5.22	5.18	0.044
IDV	Relationship priority	5.94	5.77	0.164
PDI	Communication channel	6.16	6.14	0.024
PDI	Deference	6.22	6.12	0.096
PDI	Directness	5.12	5.06	0.060
PDI	Face-saving	6.49	6.39	0.096
UAI	Expert deference	4.11	4.20	0.090
UAI	Hedging density	4.66	4.56	0.103
UAI	Risk framing	5.09	5.04	0.050
UAI	Rule reference	4.46	4.43	0.028
Overall		5.24	5.18	**0.03**

Table 13: Feature-level Hindi-Urdu Divergence. Largest divergence (relationship_priority, HUD = 0.164) represents $< 3\%$ of scale range. Statistical test: $t = 1.12$, $p = 0.262$, $d = 0.03$.

this pattern: English PCS (mean = 0.19) was comparable to German (mean = 0.15), Hindi (mean = 0.12), Nepali (mean = 0.12), and Urdu (mean = 0.16). The deficit appears to stem from alignment rather than resource availability, though the higher variance in South Asian languages (SD $\approx$ 0.16–0.29 vs. 0.12–0.14 for European languages) suggests training data effects may compound the alignment-induced gap.

Second, our Prompt C design deliberately avoided culturally indexical markers such as names, locations, or explicit cultural references. This makes our implicit condition sparser than many real-world interactions, where such markers abound. We consider this a methodological strength rather than a limitation. Including demographic markers would conflate two distinct capabilities: pragmatic inference (reasoning about situational context to select appropriate behaviour) and stereotype activation (pattern-matching on demographic signals). A model that shifts toward collectivist framing upon encountering an Indian name demonstrates the latter, not the former. By testing whether models respond to situational cues alone, we isolate genuine pragmatic sensitivity. The observed 15% capability utilisation therefore represents a conservative lower bound; real-world performance with richer cues would likely be higher, though the additional sensitivity may reflect stereotype activation rather than pragmatic reasoning.

Third, our use of Hofstede's cultural dimensions invites scrutiny. The framework originates from 1970s survey data, operates at the national level, and has been critiqued for essentialising culture (McSweeney, 2002). We acknowledge these limitations but defend the choice on three grounds. First, Hofstede's dimensions remain the most widely adopted operationalisation of cultural variation in computational work; recent NLP benchmarks including SocialCC (Wu et al., 2025) and CulturalBench (Chiu et al., 2025) employ them, enabling comparison across studies. Second, our dimension

Rank	Feature	Dim.	Median PCS
1	Face-saving	PDI	0.329
2	Deference	PDI	0.319
3	Communication channel	PDI	0.255
4	Risk framing	UAI	0.211
5	Relationship priority	IDV	0.207
6	Directness	PDI	0.151
7	Expert deference	UAI	0.148
8	Outcome framing	IDV	0.134
9	Duty vs. choice	IDV	0.128
10	Rule reference	UAI	0.108
11	Agency attribution	IDV	0.073
12	Hedging density	UAI	−0.689

Table 14: Feature-level implicit sensitivity. PDI features cluster at top (mean rank = 2.5); UAI features cluster at bottom (mean rank = 8.5) with hedging_density showing strong negative PCS.

Dimension	F	p	η^2	Sig.
IDV	202.21	$< .001$	0.113	Yes
UAI	29.50	$< .001$	0.018	Yes
PDI	13.07	$< .001$	0.008	Yes

Table 15: Cross-linguistic ANOVA results. Effect sizes: IDV shows medium effect ($\eta^2 = 0.113$); PDI and UAI show small effects.

asymmetry finding (PDI $\gg$ IDV $>$ UAI) provides internal validity: if the framework were merely capturing noise, we would not expect such systematic variation in model sensitivity across dimensions. Third, we treat the dimensions as interpretable measurement instruments rather than ontological claims about culture. The question is not whether Hofstede perfectly captures cultural reality but whether the constructs reliably distinguish pragmatic behaviours that models should modulate, and on this criterion, the divergence between Prompt A and Prompt B responses confirms they do.

Fourth, one might argue our findings reflect a general explicit-versus-implicit gap rather than anything specific to cultural pragmatics: perhaps models simply follow explicit instructions better than they respond to implicit cues in any domain. The dimension asymmetry provides evidence against this interpretation. If the gap were a domain-general property of instruction-following, we would expect similar PCS across all cultural dimensions. Instead, we observe a sevenfold difference: PDI scenarios elicit 29% of explicit capability implicitly, while UAI scenarios elicit only 4%. Models demonstrably can respond to implicit contextual cues when those cues are sufficiently salient (hierarchical relationships, face-threatening stakes). The deficit is domain-specific: certain pragmatic inferences transfer from explicit to implicit cueing while others do not, a pattern inconsistent with a blanket instruction-following advantage.

Lang.	$\bar{A}$	$\bar{C}$	t	p	d
German	4.82	5.09	−9.49	$< .001$	0.22
English	4.87	5.15	−9.43	$< .001$	0.22
Hindi	5.24	5.42	−6.75	$< .001$	0.15
Nepali	5.22	5.43	−7.88	$< .001$	0.18
Urdu	5.21	5.43	−8.25	$< .001$	0.19

Table 16: Implicit adaptation significance tests. All languages show statistically significant A→C shifts with small effect sizes ($d = 0.15$–0.22).

J Human Validation Study

To validate the LLM-as-judge methodology against human judgement, we conducted a human validation study with native speakers. For each language, we selected scenario-model combinations stratified across dimensions and sampled response pairs for A-C and B-C comparisons, yielding 235 total comparisons across five languages. Raters compared each pair in randomised, blinded presentation and judged which response was more culturally appropriate for the scenario context, with a "no meaningful difference" option.

Table 17: Human validation results by language.

	A vs C			B vs C			
Lang	**A**	**C**	**Tie**	**B**	**C**	**Tie**	**n**
German (de)	6	4	13	6	7	11	47
English (en)	16	7	1	10	11	2	47
Hindi (hi)	3	5	15	4	6	14	47
Nepali (ne)	12	7	4	12	9	3	47
Urdu (ur)	2	9	13	5	6	12	47
Total	39	32	46	37	39	42	235

English showed the clearest discrimination ability (4% tie rate for A-C comparisons), while Hindi and German showed the highest tie rates (65% and 57% respectively). This pattern may reflect either genuine response equivalence in those languages or greater difficulty in the annotation task.

K Data Availability

Total observations: 57,080 feature-level scores across 14,400 model responses (4 models × 5 lan-

guages × 3 prompt conditions × 60 scenarios × 4 samples). Response validity rate: 99.1%.

All scenarios, model responses, scoring data, and code is submitted as a zip file with this submission.