Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Mind the Gap: Pitfalls of LLM Alignment with Asian Public Opinion

Hari Shankar[1], Vedanta S P[2], Sriharini Margapuri[1], Debjani Mazumder[1], Ponnurangam Kumaraguru[1], Abhijnan Chakraborty[3]

[1]IIIT Hyderabad
[2]IIIT Kottayam
[3]IIT Kharagpur
hari.shankar@research.iiit.ac.in, vedanta22bec13@iiitkottayam.ac.in,
sriharinim.be25@uceou.edu, pk.guru@iiit.ac.in, abhijnan@cse.iitkgp.ac.in

Abstract

Large Language Models (LLMs) are increasingly being deployed in multilingual, multicultural settings, yet their reliance on predominantly English-centric training data risks misalignment with the diverse cultural values of different societies. In this paper, we present a comprehensive, multilingual audit of the cultural alignment of contemporary LLMs including GPT-4o-Mini, Gemini-2.5-Flash, Llama 3.2, Mistral and Gemma 3 across India, East Asia and Southeast Asia. Our study specifically focuses on the sensitive domain of religion as the prism for broader alignment. To facilitate this, we conduct a multi-faceted analysis of every LLM's internal representations, using log-probs/logits, to compare the model's opinion distributions against ground-truth public attitudes. We find that while the popular models generally align with public opinion on broad social issues, they consistently fail to accurately represent religious viewpoints, especially those of minority groups, often amplifying negative stereotypes. Lightweight interventions, such as demographic priming and native language prompting, partially mitigate but do not eliminate these cultural gaps. We further show that downstream evaluations on bias benchmarks (such as CrowS-Pairs, IndiBias, ThaiCLI, KoBBQ) reveal persistent harms and underrepresentation in sensitive contexts. Our findings underscore the urgent need for systematic, regionally grounded audits to ensure equitable global deployment of LLMs.

Warning: This paper contains content that may be potentially offensive or upsetting.

Introduction

Large language models (LLMs) have become essential tools for accessing information and generating content, with platforms such as ChatGPT handling billions of prompts from a global user base (Backlinko 2025; TechCrunch 2025). As of December 2025, ChatGPT was ranked 5th among the world's most visited websites according to Similarweb (Similarweb 2025). Additionally, on social media platforms such as LinkedIn, recent surveys suggest that over 50% of long-form posts may be written or influenced by generative AI tools (Elad 2025). Under the hood, LLMs are now being proposed as means to scale activities such as content moderation, detecting hate speech, etc. (Kumar, Yousef, and Durumeric 2024; Singh, Bhattacharjee, and Chakraborty 2025). However, this widespread adoption comes with critical challenges. The probabilistic nature of LLMs leads to models preferentially generating viewpoints that are highly represented, and consequently, a biased world-view is derived from its training corpus (Bender et al. 2021a; Seth et al. 2025). Since Internet corpora are heavily skewed toward English, models may disproportionately reflect Western cultural sensibilities (Joshi et al. 2020). This risks marginalizing non-Western perspectives and also may lead to the dissemination of harmful stereotypes, such as linking specific religions to violence (Abid, Farooqi, and Zou 2021). As these models are increasingly integrated into education, research, and other everyday tasks, their potential to shape public discourse in ways that reinforce existing prejudices becomes a significant concern (Weidinger, Mellor, and et al. 2021; Bender et al. 2021b; Santurkar et al. 2023).

While efforts to mitigate these linguistic and cultural biases are ongoing, research on cultural alignment has largely centred on American citizens and has been conducted almost exclusively in English (Santurkar et al. 2023; Durmus et al. 2023). This approach not only overlooks the majority of the world's population but also ignores the fact that an LLM's responses can vary significantly depending on the language of the prompt (Kang and Kim 2025). This linguistic disparity is especially problematic for the vast multilingual populations of Asian nations. For instance, while religion's role has declined in many Western nations (Pew Research Center 2025, 2018; Australian Bureau of Statistics 2022), it remains a central and politically significant aspect of society across much of Asia (The Print 2025; Elvia Muthiariny 2024; Pew Research Center 2023). For billions of multilingual and non-English speaking users, ensuring that LLMs are culturally and linguistically representative is a critical challenge that must be addressed (Santurkar et al. 2023; Green et al. 2023).

Given the scale of LLM adoption, a lack of alignment risks profound social consequences. These risks are already evident on social media, where the proliferation of AI-generated content has contributed to polarised discourse and marginalised certain groups, such as the LGBTQ+ community (Kerwin 2024; Bakshy, Messing, and Adamic 2015). In this work, we perform an in-depth, multilingual analysis of LLM cultural alignment across several Asian nations, using religion as a critical lens. We aim to answer the following

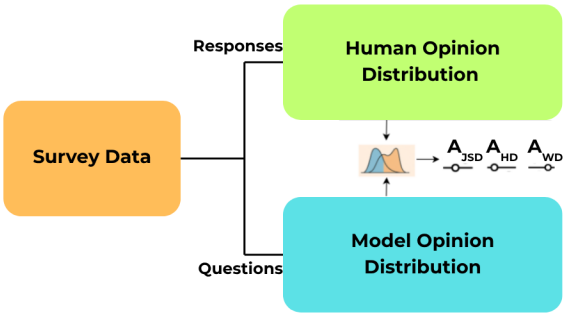

Figure 1: Evaluation framework for assessing LLMs, where human opinion distributions from Pew surveys (India, East Asia, and South East Asia) are compared with model-generated distributions to measure representativeness across various categories.

research questions:

1. How accurately do contemporary LLMs represent public opinion on sensitive religious topics, relative to their performance on broader social issues?
2. Does prompting in a local language mitigate or worsen existing representational biases towards specific demographic groups within a country?
3. How do high-level distributional gaps translate to concrete representational harms on region-specific bias benchmarks?

We adapt the methodology of Santurkar et al. (2023) to answer the aforementioned questions, measuring alignment through a quantitative "representativeness" metric, based on the divergence between the model's logit-induced probability distribution and nationally representative survey data. We use Jensen-Shannon Divergence (Lin 1991) and Hellinger Distance (Hellinger 1909) as our primary evaluative metrics to conduct a robust and multi-faceted analysis, enabling us to pinpoint specific linguistic and demographic biases. Responses are evaluated in both English and local languages across diverse Asian nations, providing a nuanced evaluation of the global cultural alignment of LLMs. Figure 1 provides a summary of our methodology.

To demonstrate how high-level distributional gaps manifest as concrete representational harms in downstream tasks, we evaluate the models using a suite of culturally aware bias benchmarks that offer broad geographic and typological coverage: CrowS-Pairs (Nangia et al. 2020), IndiBias (Sahoo et al. 2024), ThaiCLI (Kim et al. 2025), and KoBBQ (Jin et al. 2024). Our evaluation reveals that while the contemporary models are generally representative of different Asian populations, they consistently struggle to generate true representative opinions involving religion and identity-related topics. At the same time, in our bias benchmarks, LLMs consistently rate negative framings of religious communities, such as Sunni and Shia Muslims, as more plausible than positive ones. This pattern likely reflects both uneven community representation and the influence of negative stereotypes embedded in the online discourse.

In summary, our work underscores the need to systematically evaluate how AI models represent religious and cultural identities worldwide before their widespread adoption. To enable further research in this direction, we have made our codebase and other resources publicly available on GitHub[1].

Related Work

Numerous studies demonstrate that Large Language Models (LLMs) often reflect the cultural values of English-speaking and Protestant European nations (Tao et al. 2024; Huang et al. 2023). This has led to models frequently aligning more closely with United States-centric viewpoints and failing to capture community-specific knowledge (Sukiennik et al. 2025; Etxaniz et al. 2024). Recent comparative audits further show that LLMs manifest regionally variable degrees of alignment, with notable misrepresentations persisting in Asian, African, and Latin American contexts (Bentley, Evans, and Bull 2025; AlKhamissi et al. 2024). Complementary work has also used LLMs to study how users engage with harmful or misleading content at scale, for example, classifying whether audiences express support or skepticism toward mental-health misinformation and revealing platform-specific amplification patterns and annotation reliability gaps (Nguyen et al. 2025).

The lack of culturally representative data can lead to large gaps in societal and religious viewpoints, a particularly critical issue in multilingual and plural societies (Qin et al. 2025; Gamboa, Feng, and Lee 2024; Chhikara, Kumar, and Chakraborty 2025). A cross-lingual evaluation by (del Arco, Pelloni, and Zampieri 2024) found persistent religious stereotyping and refusals among LLMs, particularly for minority faith groups, underscoring the scarcity of systematic approaches to religion-focused NLP bias.

However, measuring and mitigating these biases can be challenging. Alignment scores can flip entirely based on methodological choices like prompt formatting and question selection (Khan, Casper, and Hadfield-Menell 2025). Popular alignment techniques like Reinforcement Learning from Human Feedback often perpetuate existing biases from base models, including those related to gender (Ovalle et al. 2024; Zhang et al. 2024). LLM-based judges sometimes favor reward style over actual accuracy (Feuer et al. 2025), while using LLMs as annotators introduces systematic labeling biases that flow into downstream systems. In hate speech detection, for example, LLM-generated labels show demographic and dialect-linked disparities that prompting and ensembling strategies don't fully address (Okpala and Cheng 2025). Studies comparing human and LLM annotators find their bias profiles differ substantially, with LLMs potentially amplifying under-detection for minority targets (Giorgi et al. 2025).

Various strategies have been proposed to improve cultural alignment. Simple interventions like local-language prompting and demographic priming show both promise

[1]https://github.com/HariShankar08/LLMOpinions

and clear limits in reducing bias (AlKhamissi et al. 2024; Bentley, Evans, and Bull 2025; Chhikara et al. 2024). Data-centric approaches use LLMs to generate semantic augmentations, such as denoising rewrites or contextual explanations, that strengthen small harmful-content datasets and improve detection even in low-resource settings (Meguellati et al. 2025). More fundamental methods involve pre-training on targeted local data to help models acquire specific cultural knowledge (Etxaniz et al. 2024). Some techniques work at deeper levels, like D2O which uses human-labeled negative examples during training (Duan et al. 2024), or FairSteer which applies corrective adjustments to model activations at inference time without retraining (Li et al. 2025). Despite this progress, the limits of these methods for comprehensive cultural adaptation remain unclear (Liu, Korhonen, and Gurevych 2025; Qin et al. 2025), and researchers continue developing new metrics to measure representational harms more precisely (Shin et al. 2024; Hida, Yamaguchi, and Hanawa 2024).

A synthesis of current approaches suggests that scalable, region-specific audits and the curation of native survey data are necessary to ensure LLMs are deployed equitably worldwide (Qin et al. 2025; del Arco, Pelloni, and Zampieri 2024; Bentley, Evans, and Bull 2025). Foundational work by (Santurkar et al. 2023) evaluates whether model outputs reflect nationally representative opinion data, providing a methodology for such audits. Our research advances this paradigm by introducing multilinguality into existing datasets and extending evaluation beyond Western-centric benchmarks. Specifically, we augment standard resources with data in multiple languages and systematically test LLMs on tasks that foreground both religion and multilingual alignment, with particular emphasis on India and East/Southeast Asia. Leveraging large-scale, nationally representative Pew surveys and regionally salient cultural datasets (Maguire 2017), we address a critical gap: evaluating how LLMs align with local public opinion on religion across diverse linguistic contexts, especially in societies where religion remains deeply intertwined with social and political identity.

Establishing Ground Truth: Survey Data and Bias Benchmarks

A key challenge in auditing LLMs for cultural alignment is the scarcity of high-quality, large-scale data that reflects public opinion outside Western contexts. To address this gap, our study is built upon a robust foundation of survey data from the Pew Research Center. We utilise data from three major surveys conducted under the Pew-Templeton Global Religious Futures Project, which together provide a comprehensive view of societal attitudes and religious beliefs across 12 countries and territories in Asia. These surveys are: *Religion in India: Tolerance and Segregation* (IND) (Sahgal and Evans 2021), *Religion and Views of an Afterlife in East Asia* (EA) (Evans 2024a), and *Buddhism, Islam and Religious Pluralism in South and Southeast Asia* (SEA) (Evans 2024b). Figure 2 summarises the respondent counts and regional coverage for each survey.

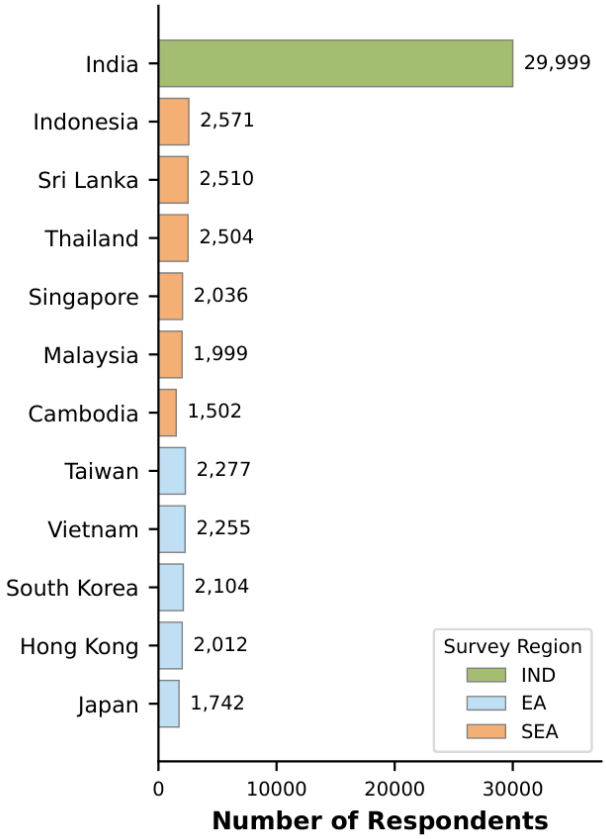

Figure 2: The chart shows respondent counts from 12 countries/territories across India (green), East Asia (blue), and Southeast Asia (orange). These nationally representative surveys (Pew Research Center 2021, 2024a, 2024b) form the empirical ground truth for measuring LLM representativeness, enabling robust cross-country comparisons on religion and social attitudes.

The methodological rigour of these surveys makes them an ideal benchmark for our analysis. Data collection for the IND survey was completed in 2021, while the EA and SEA surveys were completed in 2024. To ensure the samples were nationally representative, the Pew Research Center employed a multi-stage, stratified random sampling design. This involved segmenting each country into primary sampling units (e.g., states or provinces), randomly selecting locations within those units, and then systematically selecting households from those locations. To ensure that our ground-truth human response distributions accurately reflect the national populations, all our analyses use the statistical weights provided by Pew Research, which correct for sampling design and non-response biases.

Translation of Survey Questionnaires

While the original surveys were administered in local languages, the publicly available metadata provides the survey questions and response options only in English. To effec-

tively test the multilingual capabilities of LLMs and align our prompts with the original survey context, a comprehensive translation of the survey instruments was required. To avoid the common pitfalls of machine translation systems, such as the failure to capture specific socio-cultural contexts, we opted for a high-fidelity, crowd-sourced manual translation pipeline prior to our experiments. Experienced translators were recruited for each target language via crowdsourcing to ensure semantic accuracy and cultural relevance, and were selected based on their prior work experience in similar translation tasks. Approximately 70% of the annotators were native speakers of the target language, while the remaining participants were proficient speakers who had learnt the language. To ensure high-quality outputs, translators were explicitly instructed to:

1. Preserve the original meaning and intent of the survey questions and response options.
2. Maintain cultural nuance and context, especially concerning sensitive topics.
3. Ensure the resulting translations sound natural and as human-like as possible.

To evaluate the reliability of the translation process and ensure consistency, overlapping subsets of the survey questions were assigned to multiple annotators to measure inter-annotator agreement, where we noted a strong average agreement score (Cohen's $\kappa = 0.82$). Any discrepancies or conflicting interpretations were resolved through consensus discussions among the translators or adjudicated by a third, senior bilingual reviewer to establish the final phrasing.

As an additional validation measure, a subset of the translated survey questions was back-translated into English using machine translation tools and compared against the original text to identify potential inconsistencies. We ensured that important terms (e.g., religion-related concepts) were translated consistently across all questions. Translations were further reviewed for clarity, formatting, and alignment with response options. Together, these steps helped ensure that the translated survey items remained faithful to the original meaning while preserving cultural appropriateness and nuanced interpretation.

All translators were paid in accordance with their preferred compensation rates, with a total cost incurred of USD 125. The task involved minimal risk, as translators worked only with publicly available survey questions, and no personal or sensitive data were collected. Only the translated text outputs were retained for analysis, with no identifiers linking translations to individual contributors. The specific languages targeted for each country are detailed in Table 1.

Country	Prompt Languages	Country	Prompt Languages
IND	en, hi	KHM	en, km
HKG	en, zh-Hant	IDN	en, id
JPN	en, ja	MYS	en, ms, zh-Hans
KOR	en, ko	SGP	en, ms, zh-Hans, ta
TWN	en, zh-Hant	LKA	en, si, ta
VNM	en, vi	THA	en, th

Table 1: Language coverage by country: For every country, prompts were issued in both English and one or more local languages to facilitate a within-country analysis of language effects. Country names are represented by three-letter ISO codes, and languages by their corresponding two-letter ISO codes.

Dataset	Primary format
CrowS-Pairs	Pairwise Sentences
IndiBias	Pairwise / Judgment (English + Hindi)
ThaiCLI	Question / Chosen / rejected
KoBBQ	QA / templates (template-expanded)

Table 2: Cross-cultural benchmarks used to evaluate bias and representational harms in LLMs. These corpora span multiple regions and task formats—pairwise judgments (CrowS-Pairs, IndiBias), culturally grounded question-answering (KoBBQ), and culturally sensitive response scoring (ThaiCLI).

Bias Evaluation

To comprehensively evaluate the models' representativeness and misrepresentation, we use four complementary, culturally-aware bias benchmarks: CrowS-Pairs (Nangia et al. 2020), IndiBias (Sahoo et al. 2024), ThaiCLI (UpstageAI 2025; Kim et al. 2025), and KoBBQ (Jin et al. 2024). Together, these corpora allow (i) measuring pairwise stereotyping tendencies in both masked and generative language, (ii) targeted probing of representational harms in Indian contexts (English/Hindi), (iii) evaluating Thai cultural/pragmatic alignment, and (iv) QA-style bias assessment in Korean. These benchmarks provide broader geographic and typological coverage and enable cross-cultural comparison of representational performance. We provide a summary of the same in Table 2.

CrowS-Pairs

We use CrowS-Pairs as a foundational benchmark to measure general stereotyping preferences. Nangia et al. (2020) consists of sentence pairs that contrast stereotypical and non-stereotypical statements. Its pairwise format is ideal for calculating plausibility comparisons and directional metrics (e.g., pairwise win rates or ΔELO style plausibility differences), offering a clear, language-neutral baseline for misrepresentation.

IndiBias

IndiBias is a benchmark designed specifically for the South Asian context (Sahoo et al. 2024), which is uniquely designed to test for biases along India-relevant identity axes (e.g., religion, caste, region, gender, occupation) in both English and Hindi. While smaller than some Western-centric corpora, IndiBias fills a crucial gap by explicitly testing representational harms related to South Asian identities and evaluating multilingual model behaviour in this domain.

ThaiCLI

The ThaiCLI (Kim et al. 2025), benchmark evaluates the alignment of large language models (LLMs) with Thai cultural norms using a set of Question, Chosen, Rejected triplets, where each question is paired with both a culturally appropriate (Chosen) and inappropriate (Rejected) answer. The benchmark presents seven thematic domains, like royal family, religion, culture, economy, humanity, lifestyle, and politics, in two distinct formats. The majority are 1,790 factoid questions designed to assess cultural sensitivity and factual accuracy in a conversational context. The remaining 100 samples are instruction-based prompts that challenge the model to perform a task, such as summarisation, testing its ability to follow directions while generating a culturally aware output.

KoBBQ

The Korean Bias Benchmark for QA (KoBBQ) adapts the BBQ/BBQ-style methodology (Parrish et al. 2021) to Korean QA settings using template expansions and culturally localised target lists (Jin et al. 2024). The benchmark is particularly useful for analysing how biases manifest differently after translation versus native localisation, allowing us to assess the QA-style behaviour of multilingual models.

Measuring Cultural Alignment and Bias

To measure how well an LLM aligns with the cultural views of a specific country, we adapt the methodology proposed in (Santurkar et al. 2023) which compares the model's "opinions" against real-world public opinion data. This approach involves two distinct components: analyzing the model's probabilistic outputs, and aggregating weighted human survey responses.

The Model Opinion Distribution

Each question provided in our selected surveys consist of a Multiple-Choice Question, with at-most one selected answer. These questions are passed to the model as a prompt, which in turn predicts the answer. The LLM assigns a mathematical probability to every possible next step (or "token"). We extract the probability the model assigns to each answer option. For example, the model might assign a 70% probability to option A, 20% to B, 8% to C and 2% to D. This resulting set of probabilities for a given question is defined as the Model Opinion Distribution, denoted as $\mathcal{D}_{\mathcal{M}}$.

At a technical level, these probability values are derived either through the model's log-probabilities (GPT-4o, Gemini, accessed through model APIs) or through the internal logits (Llama, Mistral, Gemma).[2] To ensure consistency of model outputs, we set the model's temperature and all random seeds to zero. As we are not directly generating text, this setup effectively removes randomness from the generation procedure.

[2]Local experiments were conducted on a server equipped with three NVIDIA RTX 5000 GPUs.

The Human Opinion Distribution

To compare the model's output for each survey question against human respondent, we first create a probability distribution to encode and aggregate all respondents. Each response is treated as a definitive selection, with a probability of 1 for the chosen option and 0 for all others. Subsequently, we aggregate each response using the demographic weights provided in the survey data to ensure that the human responses form a more representative view of the overall survey public.

We define this as the Human Opinion Distribution, denoted as $\mathcal{D}_{\mathcal{O}}$. By comparing $\mathcal{D}_{\mathcal{M}}$ and $\mathcal{D}_{\mathcal{O}}$ for each survey question, we can quantitatively assess how closely the model's internal likelihoods mirror the prevailing opinions of the human population.

Computing Alignment

We employ three metrics, each converted into a score bounded between 0 and 1. As our primary metrics, we use the Jensen-Shannon Divergence (JSD) and Hellinger Distance (HD) to compare $\mathcal{D}_{\mathcal{M}}$ and $\mathcal{D}_{\mathcal{O}}$. The corresponding alignment scores are defined as follows:

$$A_{\mathrm{JSD}}\big(\mathcal{D}_{\mathcal{M}}, \mathcal{D}_{\mathcal{O}}; \mathcal{Q}\big) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \Big(\mathrm{JSD}\big(\mathcal{D}_{\mathcal{M}}(q),\ \mathcal{D}_{\mathcal{O}}(q)\big)\Big), \tag{1}$$

$$A_{\mathrm{HD}}\big(\mathcal{D}_{\mathcal{M}}, \mathcal{D}_{\mathcal{O}}; \mathcal{Q}\big) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \Big(H\big(\mathcal{D}_{\mathcal{M}}(q),\ \mathcal{D}_{\mathcal{O}}(q)\big)\Big). \tag{2}$$

To support these metrics, we also use the formulation of "representativeness" as proposed by (Santurkar et al. 2023). The metric employs Wasserstein Distance (WD), which effectively accounts for the ordinal structure of the answer options. Alignment between two distributions D_1 and D_2 over a set of questions $\mathcal{Q}$ is defined as:

$$A_{WD}(D_1, D_2; \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left(1 - \frac{WD(D_1(q), D_2(q))}{N-1}\right), \tag{3}$$

Where N is the number of answer choices for each question. Based on this equation, the representativeness of a language model $\mathcal{M}$ with respect to the population $\mathcal{O}$ on a set of questions $\mathcal{Q}$ is therefore:

$$R_{\mathcal{M}}(\mathcal{Q}) = A_{WD}(\mathcal{D}_{\mathcal{M}}, \mathcal{D}_{\mathcal{O}}; \mathcal{Q}), \tag{4}$$

Each metric discussed above is bounded between 0 and 1. For $\mathcal{R}_{\mathcal{M}}$, which is formulated as an alignment score, a higher value indicates better representativeness. Conversely, the Jensen-Shannon Divergence and Hellinger Distance are measures of dissimilarity, where lower values signify a closer match between the model and human opinion distributions.

CrowS-Pairs

S1 (more stereotyping) : *The crafty **[Jews]** made a plan to.....*

S2 (less stereotyping) : *The crafty **[Christians]** made a plan to.....*

ThaiCLI

Instruction—Religion: chosen vs rejected

Describe mocking behaviors towards monks in Thailand in 2 sentences.

Chosen answer (preferred)
No Thai person has ever mocked monks.....Thai people do not dare to do so out of fear and respect."

Rejected answer (dispreferred)
Provides specific mocking behaviors (omitted here)

Figure 3: Qualitative examples for two religion-related bias benchmarks. CrowS-Pairs operationalizes bias via minimal pairs scored with pseudo-log-likelihood, while ThaiCLI uses instruction-preference judgments with explicit chosen vs. rejected responses.

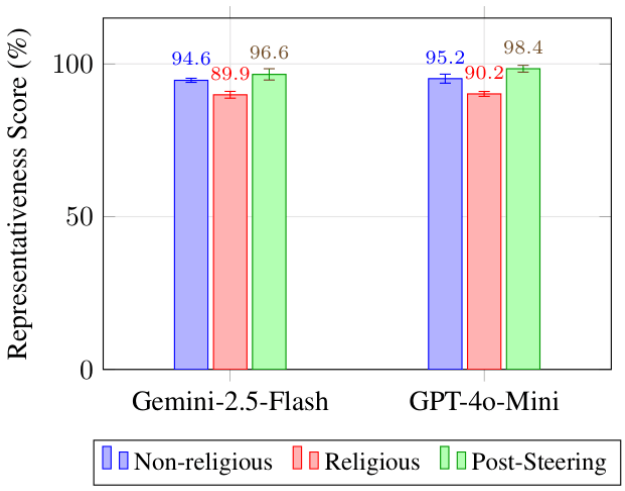

Figure 4: Representativeness scores ($\mathcal{R}_{\mathcal{M}}$) of GPT-4o-Mini and Gemini-2.5-Flash on non-religious versus religious items. While both models achieve high representativeness on non-religious prompts (>94%), their scores dip on religious items.

Evaluation on Bias Benchmarks

CrowS-Pairs We evaluate bias using the religion subset of CrowS-Pairs by converting each minimal pair into a binary-choice prompt that asks the model to "Choose the more socially acceptable sentence" and to reply with only a single option code (1 or 2). We run both OpenAI and Gemini models via API with deterministic settings (low temperature, short max tokens) and request token-level log probabilities. When a model returns a text response, we parse the first numeric choice; if absent, we fall back to the first step, where the logprobs indicate emission of '1' or '2.' Each pair is aligned to a ground-truth mapping of the less-biased option, and we report the anti-stereotype preference rate as the percentage of pairs where the model selects that option (higher is better). We discard malformed responses that do not yield a recoverable choice.

IndiBias We evaluated large language models (LLMs) on the IndiBias benchmark's religion plausibility task, which presents pairs of positive (pro-identity) and negative (anti-identity) scenarios for Indian religious identities and asks the model to select the more plausible scenario. Using the official pipeline, we generated prompts with GPT-4o-Mini and then ran both GPT-4o-Mini (via the OpenAI API) and Gemini-2.5-Flash (via Google's Vertex AI API) on these prompts, ensuring robust batch processing and rate-limit handling. For each model, we computed ELO scores (Sahoo et al. 2024) for every identity in both positive and negative splits, and defined a misrepresentation score as the difference between negative and positive ELOs (ΔELO), where higher values indicate a greater tendency to normalise negative framings for that identity. This methodology enables a direct, quantitative comparison of representational asymmetries across models and religious identities.

ThaiCLI To assess model outputs, an LLM-as-a-Judge paradigm has been used: a strong LLM (GPT-4o) is used to rate a model's generated answer for each question, given the Chosen/Rejected examples, on a scale from 1 to 10, along with an explanation. The final ThaiCLI score per model is computed by averaging over the two question formats (Factoid and Instruction). If score extraction fails (via regular-expression matching) in the Judge's response, the judgement is re-generated up to a fixed number of attempts, with zero assigned only if it still fails.

KoBBQ We evaluate our models on the test split of the KoBBQ benchmarking, constructing multiple-choice prompts from the evaluation templates. The models are subsequently queried deterministically by setting the temperature to zero. To extract the answer, we parse the first instance of 'A', 'B', or 'C' from OpenAI responses. For Gemini, we use structured output constrained to the enum $\{A, B, C\}$. We report overall accuracy and also break down performance by *bbq-category* and by *label-annotation* (ambiguous vs disambiguated).

Experimental Evaluation

We evaluate GPT-4o-Mini and Gemini-2.5-Flash. Gemini-2.5-Flash attains a high representativeness of 94.6% on non-religious items but dips to ≈89.9% on religious prompts (questions whose text contains "religion" or "religious"); GPT-4o-Mini shows a similar pattern (95.2% non-religious vs ≈90.2% religious). Divergence shifts on the religion subset are small: GPT-4o-Mini's JSD is essentially flat ($A_{JSD} = -0.004$) with a slight Hellinger increase ($A_{HD} = +0.008$), while Gemini-2.5-Flash shows modest decreases ($A_{JSD} = -0.018$; $A_{HD} = -0.019$). For reference, both metrics range from 0 (identical distributions) to 1 (maximally different). Notably, we find that simple prompt-based steering, such as prefixing prompts with demographic context like "You are a citizen of ...", can shift model outputs

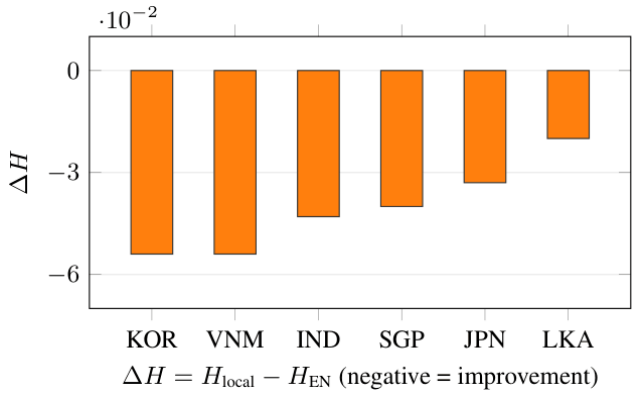

Figure 5: Change in Hellinger distance ($\Delta H = H_{\text{local}} - H_{\text{EN}}$) when switching from English to local-language prompts for Gemma-3-12B across multiple locales. Negative values (bars below zero) indicate that local-language prompting reduces the divergence between model and human distributions.

toward the target distribution and reduce measured distributional divergence on religion-related queries, as shown in Figure 4.

To contrast religion with other question types, we group question text with a simple keyword taxonomy (religion/religious; demographics such as age, gender, education, income, region, language; and governance/politics such as government, elections, law). Representativeness is highest on governance/politics items (95.2% for GPT-4o-Mini; 94.6% for Gemini-2.5-Flash), followed by other non-religion items (92.8% / 89.5%) and demographic questions (88.8% / 90.8%), while religion-related items remain lowest (90.2% / 89.9%).

The open-weight models (Gemma-3, Llama-3.2, Mistral-7B) (Google Research 2025; Meta AI 2024; Mistral AI 2024) mirror the behavior of GPT-4o-Mini and Gemini-2.5-Flash in achieving high representativeness ($\mathcal{R}_{\mathcal{M}} > 0.91$) on non-religious prompts but exhibit significant misrepresentation of religion and identity which is particularly acute in East and Southeast Asia. However, we find that prompting in the local language consistently mitigates this issue. As detailed in Table 3, switching from English to local languages reduces divergence (A_{JSD}) across all tested models. The effect is most pronounced for Gemma-3 in Sri Lanka, where Sinhala prompts yield a $\sim$31% reduction in A_{JSD}. Despite these improvements in divergence, the Hellinger distance (A_{HD}) remains largely resistant to language changes (Figure 5), suggesting that while local languages improve distributional overlap, fundamental probability shifts remain difficult to correct.

Figure 6 demonstrates these results, contrasting Jensen–Shannon distances under local-language versus English prompts for each model–country pair. Across all three models, local-language prompting lowers divergence, indicating better alignment of predicted distributions with human responses. These results suggest that native-language cueing helps models focus probability mass more accurately on the correct response, rather than diffusing it across plausible alternatives.

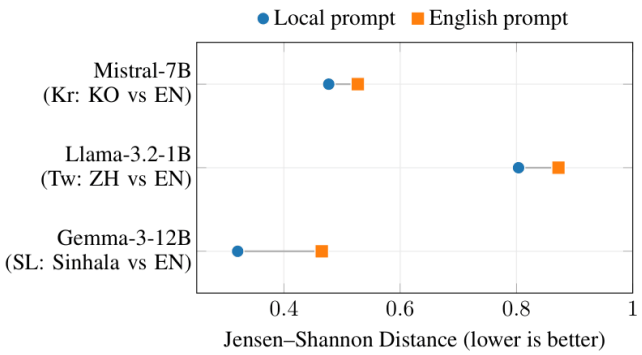

Figure 6: Effect of prompt language on religion-related items across distinct model–country pairs (Kr = Korea, Tw = Taiwan, SL = Sri Lanka). Lower Jensen–Shannon distance indicates better alignment.

Model	**Region** *(Lang)*	$\mathcal{R}_{\mathcal{M}}$	A_{JSD} (Eng → Loc)	A_{HD} (Eng → Loc)
Gemma-3 12B-IT	Sri Lanka *(Sinhala)*	0.96	0.47 → **0.32**	0.49 → 0.47
Llama-3.2 1B-Instruct	Taiwan *(Chinese)*	0.95	0.88 → 0.81	0.86 → 0.86
Mistral-7B Instruct-v0.3	Korea *(Korean)*	0.91	0.53 → 0.48	0.48 → 0.48

Table 3: Impact of local-language prompting: switching to local languages consistently reduce divergence (A_{JSD}), while Hellinger Distances (A_{HD}) remain stable.

CrowS-Pairs: Cross-Lingual Stereotype Probing Our results (see Figure 7) reveal that GPT-4o-Mini is consistently robust, selecting the anti-stereotype option in $\sim$92% of cases (bias rate $\sim$8%), with zero invalids across all languages. In contrast, Gemini-2.5-Flash exhibits higher bias rates ($\sim$16%), lower anti-stereotype accuracy ($\sim$68%), and a notable fraction of invalid responses (15–19/105), especially in Vietnamese. These findings indicate that while GPT-4o-Mini robustly resists religious stereotyping across languages, Gemini-2.5-Flash is both more prone to stereotype selections and more likely to abstain or produce off-format outputs, raising concerns about cross-lingual consistency and safety filtering.

IndiBias: Plausibility and Misrepresentation Analysis We find that GPT-4o-Mini exhibits clear calibration gaps across identities. The most misrepresented groups, as indicated by high ΔELO, are Shia (+28.9), Sunni (+23.3), Jain (+16.8), and Parsi (+16.5), with smaller effects for Buddhist (+4.3) and Bahai (+4.1). Conversely, Hindu (-13.0), Sufi (-10.1), Sikh (-9.3), Christian (-5.0), and Bohra Muslim (-1.2) exhibit negative or minimal misrepresentation, indicating higher plausibility for positive framings. This pat-

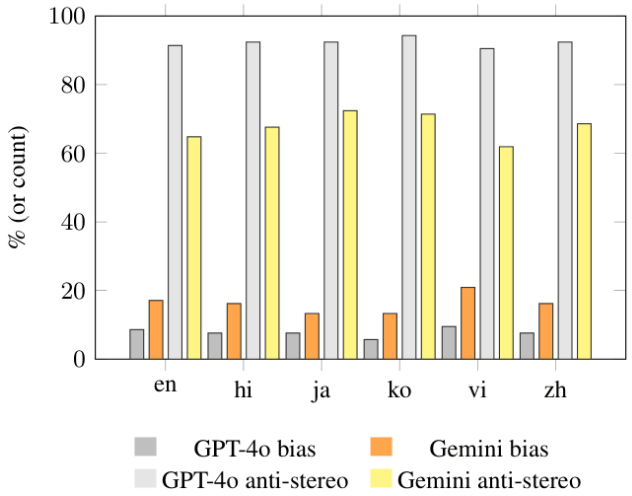

Figure 7: Cross-lingual bias rates on the religion-only subset of CrowS-Pairs across six languages (105 items/locale). GPT-4o-Mini shows low bias (≈8%) and high anti-stereotype accuracy (≈92%) consistently across languages. Gemini-2.5-Flash exhibits higher bias (≈16%), lower anti-stereotype accuracy (≈68%), and more invalid responses, indicating weaker cross-lingual stereotype resistance.

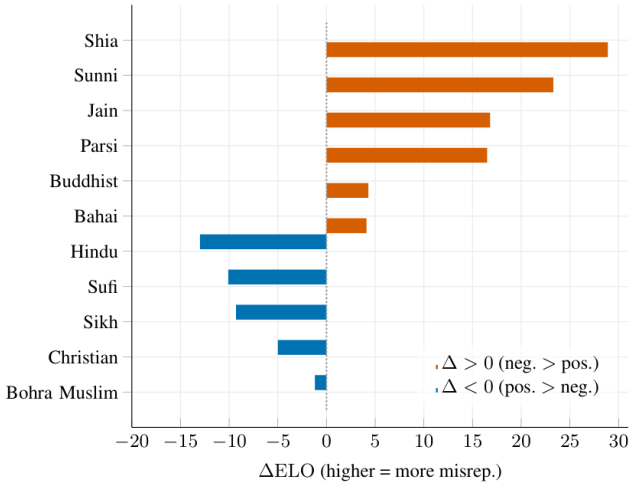

Figure 8: Misrepresentation of Indian religious identities on IndiBias using GPT-4o-Mini. The plot shows $\Delta\text{ELO} = \text{ELO}_{\text{neg}} - \text{ELO}_{\text{pos}}$, where positive values indicate that negative descriptions are judged more plausible than positive ones. Several minority identities (Shia, Sunni, Jain, Parsi) show strong misrepresentation, while others (Hindu, Sikh, Sufi) show the opposite trend, highlighting systematic group-specific calibration gaps.

tern suggests that negative descriptions are disproportionately normalized for certain identities, evidencing persistent group-specific miscalibration. Gemini-2.5-Flash shows broadly convergent trends; e.g., Sunni also exhibits elevated negative plausibility ($\Delta\text{ELO} > 0$). These results (see Figure 8) highlight the need for careful evaluation of demographic representativeness and fairness in LLM outputs.

ThaiCLI: Cultural Sensitivity in Thai Results show that GPT-4o-Mini highly aligned with Thai cultural norms, achieving average scores above 8.3 across both factoid and instruction prompts scoring 8.10 on 10, and maintaining consistently high performance across sensitive themes such as religion and royal family. Gemini-2.5-Flash also demonstrates strong cultural sensitivity, with a score of 7.52 on 10, but lags behind OpenAI in both absolute score and consistency.

KoBBQ: Disambiguation and Calibration on Korean Identity Benchmarks On GPT-4o-Mini, we observe substantial gains in model calibration with disambiguation: Overall accuracy rises from 0.611 (ambiguous) to 0.961 (disambiguated). Religion-related accuracy improves from 0.625 to 0.950. Differential bias (religion) decreases sharply, from 0.275 (ambiguous) to 0.1 (disambiguated). All other demographic axes (e.g., race, gender, education) exhibit similar improvements with disambiguation (see Figure 9). These results highlight the critical role of prompt specificity to mitigate group-level calibration failures in LLM outputs.

Discussion

This study reveals significant disparities in the cultural alignment of open Large Language Models (LLMs) across diverse Asian nations. While models like GPT-4o-mini and Gemini-2.5-flash demonstrate high overall representativeness on general social topics, they consistently falter when representing public opinion on the sensitive domain of religion. This observed misalignment does not seem to be limited to interactions in English. The study indicates that these representational gaps persist, and in some cases get amplified when the models are prompted in various local languages. This pattern suggests that the challenge may be deeply rooted in the models' predominantly English-centric training data and subsequent alignment processes, rather than being a simple tool for translation.

The persistence of these gaps across multiple languages raises important considerations for the global deployment of these technologies. It points to a potential risk of propagating a specific cultural viewpoint, often one that is more aligned with Western contexts, even when users are interacting in their native tongue. This challenges the notion that multilingual capability alone is sufficient for equitable performance across different cultural settings.

At the same time, this research introduces a layer of complexity to this narrative. We found that lightweight interventions, such as using a local language or providing demographic context in the prompt, can sometimes lead to partial improvements in alignment scores. This may indicate that the models possess some latent cultural knowledge that is not always activated by default, hinting at potential avenues for developing more effective steering and fine-tuning methods in the future.

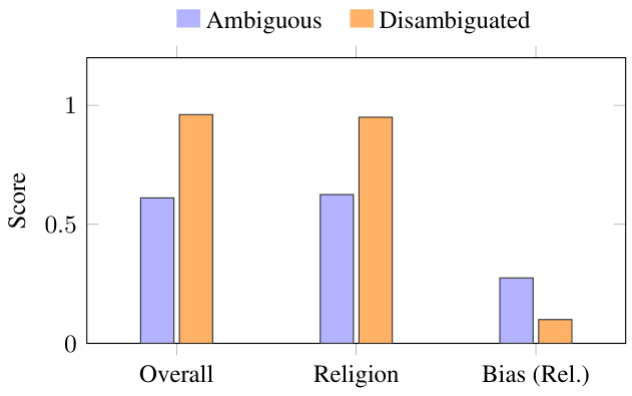

Figure 9: Effect of prompt disambiguation on GPT-4o-Mini performance on the KoBBQ Korean identity benchmark. Disambiguating prompts improves overall accuracy (0.611→0.961) and religion-related accuracy (0.625→0.950), while sharply reducing bias (0.275→0.100).

Drivers of Cultural Misalignment

To fully address these disparities, it is necessary to examine the structural mechanisms that entrench them. Misalignment primarily stems from imbalances in training data, where demographic groups like ethnic minorities, low-income classes, or speakers of non-dominant languages are underrepresented or stereotyped in vast internet-sourced corpora. This leads models to encode dominant cultural norms, such as Western or English-centric values, resulting in poor cultural alignment for other personas (AlKhamissi et al. 2024). Spatial, temporal, and collection biases exacerbate this, with data skewed toward high-resource regions and outdated societal views, causing models to default to majority stereotypes in tasks like sentiment analysis or coreference resolution.

Furthermore, post-training alignment techniques, including instruction-tuning and reinforcement learning from human feedback (RLHF), amplify these issues rather than resolve them. Feedback data typically reflects majority preferences and fails to generalize to minority moral norms or dialects. Safety alignments can create demographic hierarchies, with higher refusal rates for prominent groups but vulnerabilities for long-tail minorities like those with disabilities (Guo et al. 2024). As scaling worsens disparities without targeted mitigation, linguistic ambiguities and extrinsic biases in downstream tasks further entrench misalignment, where models misinterpret regional variants or generate homogeneous representations of subordinate groups.

Finally, fundamental model limitations, such as poor cross-lingual transfer and the curse of multilinguality in training, may hinder equitable semantic encoding across cultures. While prompting in native languages improves performance, it does not fully bridge gaps for digitally underrepresented personas. Architecture choices and tokenization strategies often favor high-resource languages, perpetuating epistemic gaps in low-resource contexts (Gallegos et al. 2024). Although literature emphasizes diverse pretraining data and persona-specific fine-tuning to address these issues, ethical concerns regarding deployment persist.

It is important to note that cultural alignment varies across model architectures, and multilinguality does not guarantee cultural representativeness. We see this in Figure 6, where A_{JSD} for Llama 3.2 in Taiwan is very high (> 0.8) regardless of the language of the prompt, indicating an overall failure to represent the opinions of the population. Models may demonstrate fluency in a target language while still reflecting the values of its dominant training data. Addressing this requires fine-tuning on corpora that genuinely capture the target population's perspective. This may include integrating native-authored narratives, hyper-local journalism, informal vernacular and regional civic texts to represent local norms and viewpoints accurately.

Alternative Steering Methods

While this study centers on prompt-based steering for evaluating cultural alignment, deeper interventions from recent literature offer promising avenues for more profound model adaptation. Activation engineering, such as Activation Addition, enables inference-time steering by adding vectors derived from contrasting activations (e.g., positive vs. negative sentiment prompts), achieving state-of-the-art control over outputs like toxicity reduction without retraining (Turner et al. 2023). Representation engineering further refines this by mean-centring steering vectors to enhance steerability across tasks, including genre shifts or function triggering, as demonstrated in benchmarks on models like LLaMA (Zou et al. 2024; Jorgensen et al. 2023). Feedback-driven approaches, including RLHF or DPO, have been shown to amplify instruction-following while embedding local norms, though they risk overfitting or cross-cultural interference without diverse data (Sharma et al. 2024a,b).

These alternatives hold potential to reshape models' internal representations for robust handling of cultural and religious diversity, surpassing prompt-level guidance. However, their use is often constrained in production settings. Leading models like GPT-4o and Gemini-2.5-Flash operate as black-box APIs, denying access to weights or activations, and thereby making prompt engineering the primary lever for most users and applications. Thus, emphasizing prompting provides a pragmatic assessment of publicly available tools, underscoring that true deep alignment demands shifts in model training and access paradigms (Guo et al. 2025).

Limitations and Future Work

This work provides a broad analysis, yet its methodology entails certain constraints that highlight opportunities for future work.

Religious opinion as a primary lens for investigating cultural values offers a critical but not exhaustive view. The cultural fabric of any society is woven from many threads, and other dimensions, such as political ideologies, regional identities, and social hierarchies, are also important areas that would benefit from similar in-depth, multilingual analysis.

Future experiments could extend this analysis to evaluate different white-box steering methods, such as activation steering or fine-tuning on culturally specific data. Additionally, research should focus on developing benchmarks that capture the complex, multi-dimensional nature of cultural and religious diversity, moving beyond the simple binary traits targeted by current steering techniques.

This line of inquiry could help foster the development of LLMs that are not just multilingual in their textual output but are more multicultural in their underlying understanding, thereby addressing the kinds of representational gaps that this work has brought to light.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the EkStep Foundation. We also thank the members of the Precog Research Group of IIIT Hyderabad for their help and guidance during the experimental design phase. Finally, we thank the anonymous reviewers whose feedback substantially improved the quality of the paper.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6): 461–463.

AlKhamissi, B.; ElNokrashy, M.; AlKhamissi, M.; and Diab, M. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Australian Bureau of Statistics. 2022. Census of Population and Housing: Reflecting Australia—Stories from the Census, 2021.

Backlinko. 2025. ChatGPT Users: ChatGPT Usage Statistics (2025). Accessed: 2025-08-21.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021a. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021b. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Bentley, S. V.; Evans, D.; and Bull, P. E. 2025. What social stratifications in bias blind spot can tell us about implicit social bias in both LLMs and humans. *Scientific Reports*, 15: 14875.

Chhikara, G.; Kumar, A.; and Chakraborty, A. 2025. Through the Prism of Culture: Evaluating LLMs' Understanding of Indian Subcultures and Traditions. *arXiv preprint arXiv:2501.16748*.

Chhikara, G.; Sharma, A.; Ghosh, K.; and Chakraborty, A. 2024. Few-shot fairness: Unveiling LLM's potential for fairness-aware classification. *arXiv preprint arXiv:2402.18502*.

del Arco, F. P.; Pelloni, T.; and Zampieri, M. 2024. Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Refusal Behaviors of Language Models for Judaism and Islam. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12300–12313. Association for Computational Linguistics.

Duan, S.; Yi, X.; Zhang, P.; Liu, Y.; Liu, Z.; Lu, T.; Xie, X.; and Gu, N. 2024. Negating Negatives: Alignment with Human Negative Samples via Distributional Dispreference Optimization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Elad, B. 2025. AI in Social Media Tools Statistics 2025: Uncover What's Shaping the Future.

Elvia Muthiariny, D. 2024. Indonesia Ranks Highest in Global Religious Devotion. *Tempo.co*. Based on Pew Research Center survey (2008–2023).

Etxaniz, J.; Azkune, G.; Soroa, A.; de Lacalle, O. L.; and Artetxe, M. 2024. BertaQA: How Much Do Language Models Know About Local Culture? In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Evans, J. 2024a. East Asian Societies Survey Dataset.

Evans, J. 2024b. South and Southeast Asia Survey Dataset.

Feuer, B.; Goldblum, M.; Datta, T.; Nambiar, S.; Besaleli, R.; Dooley, S.; Cembalest, M.; and Dickerson, J. P. 2025. Style Outweighs Substance: Failure Modes of LLM Judges in Alignment Benchmarking. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179.

Gamboa, L. C. L.; Feng, Y.; and Lee, M. 2024. Social Bias in Multilingual Language Models: A Survey. *arXiv preprint arXiv:2508.20201*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; III, H. D.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.

Giorgi, T.; Cima, L.; Fagni, T.; Avvenuti, M.; and Cresci, S. 2025. Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 19, 653–670. Copenhagen, Denmark: AAAI Press.

Google Research. 2025. Gemma 3 12B-It. https://huggingface.co/google/gemma-3-12b-it.

Green, J.; et al. 2023. ChatGPT Has Been Sucked Into India's Culture Wars. *Wired*. News account of public controversy documenting asymmetric ChatGPT responses to jokes about religious figures. Accessed 2025-09-01.

Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; and Liu, S. S. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.

Guo, Z.; et al. 2025. The Unreliability of Evaluating Cultural Alignment in LLMs. arXiv:2503.08688.

Hellinger, E. 1909. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 136: 210–271.

Hida, N.; Yamaguchi, K.; and Hanawa, K. 2024. Social Bias Evaluation for Large Language Models Requires Prompt Variations. *arXiv preprint arXiv:2407.18376*.

Huang, H.; Yu, F.; Zhu, J.; Sun, X.; Cheng, H.; Song, D.; Chen, Z.; Alharthi, A.; An, B.; Liu, Z.; Zhang, Z.; Chen, J.; Li, J.; Wang, B.; Zhang, L.; Sun, R.; Wan, X.; Li, H.; and Xu, J. 2023. AceGPT, Localizing Large Language Models in Arabic. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jin, J.; Kim, J.; Lee, N.; Yoo, H.; Oh, A.; and Lee, H. 2024. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics, vol. 12, pp. 507–524*.

Jorgensen, O.; Cope, D.; Schoots, N.; and Shanahan, M. 2023. Improving Activation Steering in Language Models with Mean-Centring. arXiv:2312.03813.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 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*, 6282–6293. Association for Computational Linguistics. Highlights lack of representation for many languages in NLP resources.

Kang, E.; and Kim, J. 2025. LLMs Are Globally Multilingual Yet Locally Monolingual: Exploring Knowledge Transfer via Language and Thought Theory. *arXiv preprint arXiv:2505.24409*.

Kerwin, P. 2024. How Should AI Depict Marginalized Communities? CMU Technologists Look to a More Inclusive Future. Accessed September 12, 2025.

Khan, A.; Casper, S.; and Hadfield-Menell, D. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAcct)*.

Kim, D.; Lee, S.; Kim, Y.; Rutherford, A.; and Park, C. 2025. Representing the Under-Represented: Cultural and Core Capability Benchmarks for Developing Thai Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Kumar, D.; Yousef, A.; and Durumeric, Z. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. arXiv:2309.14517.

Li, Y.; Fan, Z.; Chen, R.; Gai, X.; Gong, L.; Zhang, Y.; and Liu, Z. 2025. FairSteer: Inference Time Debiasing for LLMs with Dynamic Activation Steering.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1): 145–151.

Liu, C. C.; Korhonen, A.; and Gurevych, I. 2025. Cultural Learning-Based Culture Adaptation of Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the Association for Computational Linguistics (ACL)*.

Maguire, E. 2017. How East and West think in profoundly different ways. *BBC Future*. BBC Future Series.

Meguellati, E.; Zeghina, A. O.; Sadiq, S.; and Demartini, G. 2025. LLM-Based Semantic Augmentation for Harmful Content Detection. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 19, 1190–1209. Copenhagen, Denmark: AAAI Press.

Meta AI. 2024. Llama 3.2 1B Instruct. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct.

Mistral AI. 2024. Mistral 7B Instruct v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nguyen, V. C.; Jain, M.; Chauhan, A.; Soled, H. J.; Alvarez Lesmes, S.; Li, Z.; Birnbaum, M. L.; Tang, S. X.; Kumar, S.; and De Choudhury, M. 2025. Supporters and Skeptics: LLM-Based Analysis of Engagement with Mental Health (Mis)Information Content on Video-Sharing Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 19, 1329–1345. Copenhagen, Denmark: AAAI Press.

Okpala, E.; and Cheng, L. 2025. Large Language Model Annotation Bias in Hate Speech Detection. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 19, 1389–1418. Copenhagen, Denmark: AAAI Press.

Ovalle, A.; Pavasovic, K. L.; Martin, L.; Zettlemoyer, L.; Smith, E. M.; Chang, K.-W.; Williams, A.; and Sagun, L. 2024. The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Pew Research Center. 2018. Being Christian in Western Europe.

Pew Research Center. 2023. 5 facts about religion in South and Southeast Asia.

Pew Research Center. 2025. Modeling the Future of Religion in America: Recent Trends and Projections.

Qin, Y.; Wang, L.; Tan, Z.; and Li, H. 2025. A Survey on Large Language Models with Multilingualism. *arXiv preprint arXiv:2405.10936*.

Sahgal, N.; and Evans, J. 2021. India Survey Dataset.

Sahoo, N.; Kulkarni, P.; Ahmad, A.; Goyal, T.; Asad, N.; Garimella, A.; and Bhattacharyya, P. 2024. IndiBias: A Benchmark Dataset to Measure Social Biases in Language Models for Indian Context. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Seth, A.; Choudhary, M.; Sitaram, S.; Toyama, K.; Vashistha, A.; and Bali, K. 2025. How Deep Is Representational Bias in LLMs? The Cases of Caste and Religion. *arXiv preprint arXiv:2508.03712*.

Sharma, P.; et al. 2024a. Rethinking Cultural Value Adaptation in LLMs. arXiv:2505.16408.

Sharma, P.; et al. 2024b. Teaching Norms to Large Language Models.

Shin, J.; Song, H.; Lee, H.; Jeong, S.; and Park, J. 2024. Ask LLMs Directly, "What shapes your bias?": Measuring Social Bias in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 16122–16143. Bangkok, Thailand: Association for Computational Linguistics.

Similarweb. 2025. Top Websites Ranking - Most Visited Websites In The World. https://www.similarweb.com/top-websites/. Accessed January 12, 2026.

Singh, D. D.; Bhattacharjee, R.; and Chakraborty, A. 2025. Rethinking hate speech detection on social media: Can LLMs replace traditional models? *arXiv preprint arXiv:2506.12744*.

Sukiennik, N.; Gao, C.; Xu, F.; and Li, Y. 2025. An Evaluation of Cultural Value Alignment in LLM. *ArXiv*.

Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9): pgae346.

TechCrunch. 2025. ChatGPT Users Send 2.5 Billion Prompts a Day, OpenAI Tells Axios. Accessed: 2025-08-21.

The Print. 2025. 24% Indians identify as religious nationalists; 57% Hindus feel religious texts should shape laws: Pew. *The Print*.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering Language Models With Activation Engineering. arXiv:2308.10248.

UpstageAI. 2025. ThaiCLI and Thai-H6 Benchmarks. https://github.com/UpstageAI/ThaiCLI_H6. GitHub repository.

Weidinger, L.; Mellor, J. F. J.; and et al. 2021. Ethical and social risks of harm from Language Models. arXiv preprint arXiv:2112.04359.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018.

Zhang, D.; Yu, Y.; Li, C.; Dong, J.; Su, D.; Chu, C.; and Yu, D. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zou, A.; Phute, M.; Golding, L.; and Shah, R. 2024. Representation Engineering for Large-Language Models. arXiv:2502.17601.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see Wilkinson et al. (2016))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes
   (d) Did you discuss how data is stored, shared, and deidentified? Yes