Bias in AI-Generated Travel Narratives: Evaluating LLMs in Countering Misconceptions

Anshuman Mohapatra
Department of Software Engineering
Rochester Institute of Technology
Rochester, USA
am8506@rit.edu

Ashique Khudabukhsh
Department of Software Engineering
Rochester Institute of Technology
Rochester, USA
axkvse@rit.edu

Bhaskar Sreenivas Pavan Akkena
Department of Software Engineering
Rochester Institute of Technology
Rochester, USA
bsa3464@rit.edu

Sooraj Kakkanatt Manoj
Department of Software Engineering
Rochester Institute of Technology
Rochester, USA
sk2716@rit.edu

Travis Desell
Department of Software Engineering
Rochester Institute of Technology
Rochester, USA
tjdvse@rit.edu

***Abstract*—Bias in large language model (LLM) outputs has become a critical concern as these systems are increasingly deployed in real-world, high-stakes applications. This study investigates how bias manifests in AI-generated travel narratives, positioning language models not as the object of evaluation, but as instruments for revealing underlying patterns of representational, cultural, and evaluative bias. Using a comparative framework, we analyze travel-related narratives generated by multiple LLMs across diverse geographic and cultural contexts. Outputs are evaluated using both automated toxicity metrics and embedding-based similarity measures, allowing for a nuanced assessment of how sentiment, framing, and cultural representation vary across regions and model architectures. Our methodology emphasizes reproducibility and interpretability, combining quantitative scoring with qualitative discourse analysis to identify consistent asymmetries in narrative tone, risk portrayal, and cultural stereotyping. The results demonstrate that bias is not uniformly distributed and is influenced by both model training data and the evaluation mechanisms themselves, including external toxicity classifiers. By treating models as evidence rather than endpoints, this work contributes to a broader understanding of bias as a systemic phenomenon in AI-mediated content generation and highlights the limitations of relying solely on automated evaluation tools. The findings underscore the need for context-aware, multi-layered evaluation frameworks when assessing fairness and bias in generative AI systems.**

I. Background and Introduction

Large language models (LLMs) are increasingly deployed in applications that generate descriptive and advisory content, shaping how users perceive places, cultures, and social contexts. In travel-related systems—including itinerary planning tools, destination descriptions, and conversational assistants—AI-generated narratives influence decisions about mobility, safety, and cultural engagement. Such narratives are not neutral: they encode assumptions about risk, desirability, and cultural value, often reflecting historical and structural imbalances present in the data on which these systems are trained. Consequently, bias in AI-generated travel narratives carries tangible social implications, reinforcing stereotypes and uneven representations across regions and populations. Existing research on bias in generative AI has primarily focused on detecting harmful language, quantifying toxicity, or benchmarking models against standardized evaluation datasets. While these approaches provide useful insights, they often treat models as the primary objects of evaluation and optimization. This framing risks obscuring bias as a broader systemic phenomenon—one that emerges not only from model architectures and training data, but also from prompt design, domain context, and evaluation mechanisms. In culturally situated domains such as travel, these interactions are especially pronounced.

A. Counterspeech and Indic Language Models

This study initially sought to investigate whether instruction-tuned Indic language models could be used to generate effective counterspeech in response to biased or harmful travel narratives. Counterspeech, defined as context-aware and non-abusive discourse that challenges or reframes problematic content, has been proposed as a constructive alternative to content removal or suppression. In the travel domain, counterspeech presents the possibility of mitigating harmful framing while preserving informational value and cultural nuance. The focus on Indic instruction-based LLMs was motivated by two considerations. First, such models are trained on linguistic and cultural contexts that are frequently underrepresented in Western-centric datasets. Second, it was hypothesized that these models might reveal alternative discourse patterns when responding to biased portrayals of regions in the Global South. Importantly, the objective was not to establish that Indic models are inherently less biased, but to examine how they behave when tasked with reframing biased narratives.

B. Empirical Limitations and Methodological Pivot

Empirical evaluation of the counterspeech approach revealed significant limitations. Generated responses often failed to meaningfully engage with the biased framing present in the prompts. Instead, outputs tended to be generic, overly neutral, or evasive, offering limited interpretive depth or contextual correction. Additionally, response quality varied substantially across prompts, complicating systematic comparison and analysis. These challenges were further amplified by limitations in automated evaluation. Toxicity and sentiment classifiers frequently assigned uniformly low scores to counterspeech outputs, regardless of qualitative differences in content. As a result, the evaluation pipeline lacked sufficient sensitivity to distinguish between superficial neutrality and substantive reframing. Rather than revealing patterns of bias mitigation, the counterspeech experiment exposed a deeper issue: existing evaluation tools are poorly aligned with the complexity of assessing counterspeech in culturally situated narrative contexts. In response, the study underwent a methodological pivot. Instead of treating counterspeech generation as the primary research objective, the project shifted toward a broader investigation of how bias manifests in AI-generated travel narratives themselves. This reframing positioned LLMs not as agents of intervention, but as instruments for observing and characterizing representational and evaluative bias.

C. Revised Analytical Framework and Model Selection

Under the revised framework, the study employs general-purpose and regionally specialized models, specifically LLaMA32, Gemma2, GPT2, and Sarvam, to generate travel narratives across diverse geographic and cultural contexts. These models were selected not because they are presumed to be more accurate or fair, but because they produce stable, expressive, and coherent outputs suitable for systematic analysis. Their role in this study is evidentiary rather than evaluative; differences in outputs are interpreted as signals of broader socio-technical dynamics rather than indicators of model superiority.

The analysis combines automated toxicity scoring and embedding-based similarity measures with qualitative discourse analysis. This mixed approach enables the identification of both large-scale trends and subtle representational patterns, including differences in tone, affect, and risk framing. Crucially, the evaluation process itself is treated as an object of scrutiny. Automated metrics are examined not as neutral arbiters, but as systems that encode their own cultural and linguistic assumptions, which can systematically influence conclusions about bias.

D. Core Framing and Contributions

At the core of this project is the framing that *models are evidence, not the focus*. Bias is not attributed to any single model or dataset, but is understood as an emergent property of the broader AI content generation and evaluation pipeline. By examining how multiple models generate travel narratives and how those narratives are scored and interpreted, this study reveals how representational, cultural, and evaluative biases interact. This work makes three primary contributions. First, it documents the limitations of current counterspeech generation and evaluation methods in the travel narrative domain. Second, it provides an empirical analysis of bias patterns in AI-generated travel narratives across multiple models and regions. Third, it advances a methodological perspective that treats LLMs as analytical instruments, encouraging future research to conceptualize bias as a systemic phenomenon rather than a model-specific defect. Together, these contributions highlight the need for context-aware and reflexive approaches to fairness and bias evaluation in generative AI systems.

II. Related Work

Research on fairness and bias in large language models (LLMs) spans technical, social, and domain-specific perspectives. Prior studies have examined how biases emerge in generative systems, how prompting and model configuration shape behavior, and how fairness should be conceptualized through interdisciplinary frameworks. In applied domains such as online discourse, content moderation, and tourism, LLMs have been shown to exhibit systematic cultural and demographic biases, raising concerns about their deployment in socially sensitive contexts.

A. Prompt-Based Bias Mitigation and Model Audits

A growing body of work investigates how prompting strategies influence fairness in LLM outputs. Kumar et al. [1] present a comprehensive evaluation of bias across large language models, examining how prompting strategies, decoding parameters, and model architectures affect fairness outcomes. Their results show that while structured feedback loops and constrained prompting can reduce biased outputs in certain cases, deterministic configurations—such as low-temperature decoding—often intensify existing biases.

This study further demonstrates that no contemporary LLM is free from bias, emphasizing that bias manifests differently across demographic dimensions and task formulations rather than being eliminated through model scaling alone.

Complementary studies expand this evaluation landscape. Yeh et al. [2] assess LLM bias in application pipelines built using LangChain, highlighting how system-level design choices shape downstream bias. Wan et al. [3] uncover systematic gender bias in LLM-generated recommendation letters, while Lin et al. [4] contrast human and LLM perceptions of bias, revealing substantial divergence in what is considered harmful or unfair. Ling et al. [5] further demonstrate that social biases extend beyond natural language generation into LLM-generated source code.

LLMs have also been used as analytical tools for online discourse. Rathod et al. introduce *Comment Compass*, a system that leverages generative models to summarize and analyze YouTube comment sections [6]. While their results demonstrate the utility of LLMs for large-scale discourse analysis, they also highlight risks associated with abstraction and misrepresentation when summarizing community narratives.

A closely related contribution is the work of Mathew et al. [7], who introduce the first large-scale YouTube counterspeech dataset. Their analysis shows that counterspeech comments differ markedly from other comment types in linguistic style, psycholinguistic markers, and patterns of community engagement. They further demonstrate that machine learning models can automatically identify counterspeech with competitive performance. This work is particularly relevant to studies of LLM bias and toxicity, as it reframes intervention not as suppression, but as constructive narrative engagement.

B. Fairness Frameworks and Theoretical Perspectives

Beyond empirical audits, foundational research has critiqued the conceptual boundaries of fairness in AI systems. Selbst et al. argue that technical fairness approaches often fall into "traps of abstraction," failing to account for the social contexts in which algorithmic systems operate [8]. They emphasize the need to integrate social actors, institutional processes, and power dynamics into fairness evaluations.

Mehrabi et al. provide a comprehensive survey categorizing bias into historical, measurement, and representation-based forms, and outline mitigation strategies across the machine learning lifecycle [9]. They highlight that fairness metrics frequently conflict and that mitigation strategies must be tailored to specific domains and applications.

Binns draws on political philosophy to examine competing definitions of fairness, incorporating egalitarian, libertarian, and sufficientarian perspectives [10]. His work underscores that fairness cannot be reduced to purely quantitative metrics and requires normative grounding.

Benchmarking frameworks such as Stanford's HELM [11] and industry-led fairness indicators [12] provide systematic evaluations of LLM behavior across robustness, fairness, and efficiency dimensions. While these frameworks reveal persistent demographic and ideological skew, they primarily serve diagnostic purposes and offer limited guidance for mitigation.

C. Cultural Bias and Fairness in Tourism Applications

LLMs deployed in tourism, travel planning, and digital engagement settings exhibit notable cultural and demographic biases. Kontogianni and Alepis [13] examine GPT-based systems in smart tourism platforms and find that while personalization improves user experience, it can also reinforce cultural stereotypes and hallucinate location-specific information.

Kamruzzaman et al. extend bias analysis to attributes such as age, physical appearance, institutional affiliation, and nationality, demonstrating that LLMs reproduce socially embedded stereotypes beyond traditionally protected categories [14]. Dai et al. show that LLM-driven information retrieval can amplify dominant cultural narratives, reinforcing informational echo chambers [15]. Ren et al. further find that LLM-generated travel recommendations frequently associate ethnic and gender identities with stereotypical themes, even under carefully controlled prompting conditions [16].

Related work on tourism discourse emphasizes how digital narratives shape destination perception. Studies of travel vloggers highlight their role as informal cultural intermediaries, while sentiment analysis and embedding-based approaches have been used to study engagement with tourism content [17]–[20]. Collectively, these findings underscore the importance of fairness-aware design in AI-driven tourism applications.

Overall, existing literature demonstrates that LLMs can reproduce and amplify representational, cultural, and evaluative biases across domains. However, much of this work treats models as primary objects of optimization or benchmarking. In contrast, the present study adopts a perspective in which LLMs are treated as evidentiary instruments, enabling a systemic examination of bias across generation, evaluation, and interpretation pipelines.

III. Approach / Methodology

Following the gaps identified in prior work, our methodology is structured around a multi-layered pipeline designed to produce contextually appropriate counterspeech for travel-related YouTube comments while simultaneously enabling systematic toxicity analysis. Unlike earlier approaches that rely solely on prompt-based generation or surface-level toxicity detection, our system integrates five tightly coupled layers: preprocessing, relevance filtering, counterspeech generation, automated toxicity evaluation, and analytic visualization. This layered design ensures modularity, interpretability, and scalability, allowing each stage to address a specific limitation in existing research while contributing to the broader objective of mitigating harmful discourse online.

A. Domain Model Overview

Figure 1 illustrates the conceptual architecture that underpins our methodology. The domain model partitions the workflow into distinct functional layers, each responsible for transforming the data in a structured and meaningful way. Information flows vertically from data ingestion to analytics, with each layer adding semantic structure, filtering, or evaluative insight. This encapsulated design mirrors established principles in software and machine learning system architecture, enabling the pipeline to remain extensible and resilient.

At a high level, the model formalizes how raw YouTube comments move through a sequence of refinement, decision-making, generation, and evaluation processes. The architecture is intentionally modular: each component (e.g., cleaning processor, relevance classifier, batch processor, toxicity evaluator) encapsulates a single responsibility, thereby reducing coupling and making the system easier to test, replace, or extend.

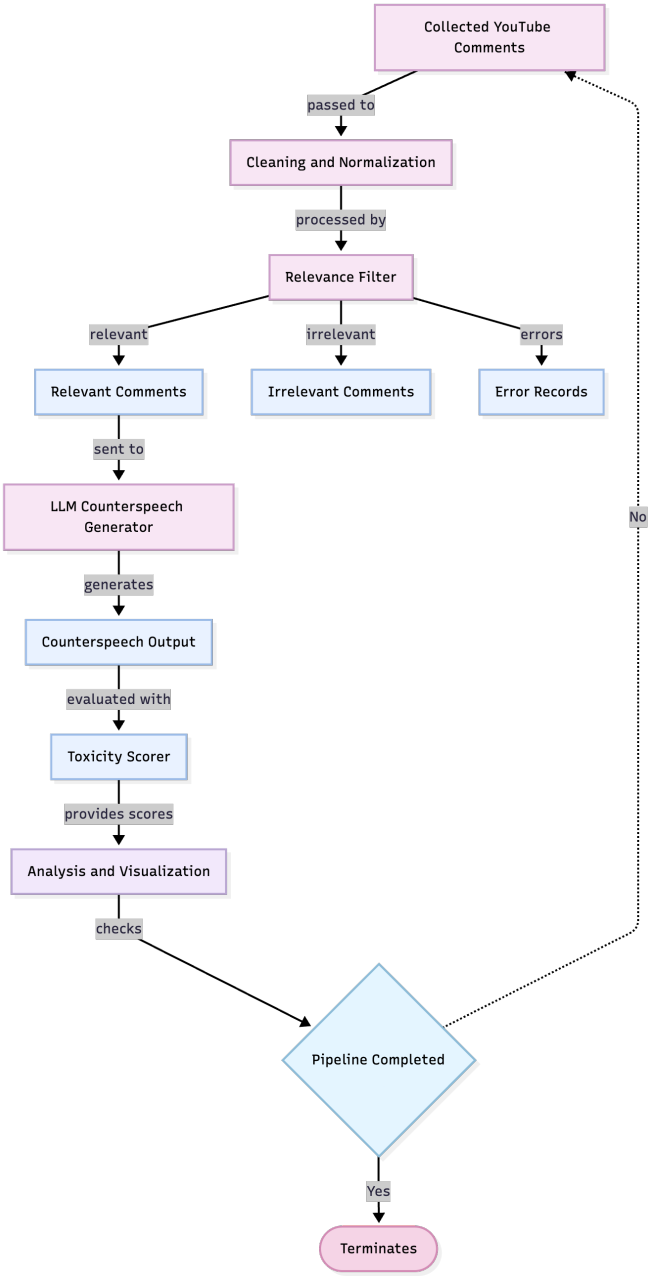

Fig. 1. Domain Model

B. Motivation for Our Approach

Several characteristics of the YouTube comment domain shaped our methodological choices. First, comments are highly informal, noisy, and often written in Romanized Hindi mixed with English. This rejects the assumptions that many existing toxicity pipelines make about textual regularity and grammatical structure. Therefore, robust preprocessing and language-aware normalization became foundational requirements.

Second, not all comments merit counterspeech. Prior systems either apply counterspeech indiscriminately or rely on simple keyword triggers to identify harmful content. These approaches lack nuance, especially for implicit bias, cross-cultural misunderstandings, or coded expressions of stereotypes. We therefore incorporate an LLM-based semantic relevance filter to ensure that counterspeech is targeted and meaningful.

Third, existing counterspeech systems typically operate without systematic toxicity feedback. A generated response may successfully address the issue but still contain unintended harm, condescension, or identity-based bias. To address this, our architecture includes a dedicated toxicity evaluation layer that analyzes responses using established metrics (e.g., Detoxify attributes). This not only provides quantitative insight into LLM safety but allows researchers to trace where harmful patterns originate within the pipeline.

Finally, while previous work often treats counterspeech as a singular interaction between a prompt and a model, our framework treats it as a *process*: the output of multiple interacting systems rather than the byproduct of a single model call. This perspective allows us to experiment with variations at different layers, such as alternate generation models, filtering strategies, or scoring metrics.

C. Why Our Design is Necessary

Our pipeline's structure is motivated by concrete shortcomings in existing research:

- **Preprocessing to handle real-world linguistic noise:** Prior work often assumes clean, monolingual text. Our domain contains misspellings, emoji, code-mixing, and informal grammar, requiring a dedicated normalization layer that previous systems lack.
- **Semantic relevance filtering instead of keyword triggers:** Toxicity often arises from subtle implications rather than explicit insults. LLM-based relevance filtering enables contextual reasoning that keyword-based filters cannot provide, reducing false positives and avoiding unnecessary counterspeech generation.
- **Modular counterspeech generation:** Existing models often use a single static prompt. Our prompt builder, batch processor, and structured output writer allow consistent generation across large datasets, fault tolerance, and reproducibility— all features critical for empirical research.
- **Integrated toxicity evaluation:** Unlike systems that rely on human annotation or post-hoc inspection, our pipeline embeds toxicity assessment into its core architecture, enabling systematic measurement of counterspeech safety across models, languages, and content types.
- **Visualization and interpretability:** Most counterspeech studies stop at generation. Our pipeline extends into analytics, offering statistical and linguistic insights needed for evaluating alignment, fairness, and bias.

These choices collectively provide a methodological foundation that surpasses prior approaches in transparency, scalability, and rigor.

D. Design and Methodological Decisions

Our design philosophy prioritizes modularity, extensibility, and reproducibility:

- **Layered modularization:** Each stage is self-contained and can be independently replaced or refined (e.g., alternate LLMs, new toxicity metrics). This ensures compatibility with evolving safety research.
- **LLM-driven semantic classification:** Rather than using static lexicons, we rely on an LLM (Gemini) to identify harmful or bias-related content, improving performance on subtle or culturally contextual comments.
- **Fault-tolerant generative workflow:** Batch processing and incremental saving ensure that counterspeech generation is robust to model errors, API timeouts, or hardware interruptions.
- **Standardized JSON artifacts:** Storing intermediate outputs in structured formats enables auditability, replication, and downstream processing.
- **Built-in evaluation loop:** Embedding toxicity scoring into the pipeline allows researchers to track whether design choices improve or degrade the safety of generated responses.

By grounding the architecture in methodological clarity and aligning the pipeline with the practical realities of real-world data, our approach offers a comprehensive and empirically grounded framework for studying LLM-driven counterspeech at scale.

IV. Implementation

This section describes the concrete realization of each subsystem defined in the domain model. While the methodological framework outlines the conceptual architecture of our counterspeech pipeline, the implementation details presented here demonstrate how these design choices were executed in practice. Each layer is implemented as a modular Python component, with clear interfaces, fault tolerance, and structured data handling to facilitate reproducibility and extensibility.

A. Layer 1: Input and Preprocessing

The first layer operationalizes raw data ingestion through the `CollectedCommentHandler`, the `CleaningProcessor`, and the `LanguageOrganizer` modules. The implementation relies on the YouTube Data API to retrieve comment threads, capturing metadata such as video identifiers, timestamps, and engagement statistics. To ensure reproducibility, all API responses are cached locally and stored in JSON-formatted artifacts.

The `CleaningProcessor` is responsible for normalizing noisy user-generated text. This includes whitespace normalization, emoji and URL removal, punctuation correction, and optional model-assisted cleaning using a lightweight LLM-based cleaning prompt. The cleaned output is stored as structured JSON entries with fields for text content, language guess, original metadata, and processing timestamps.

Next, the `LanguageOrganizer` performs lightweight language identification, distinguishing English from Romanized Hindi using heuristic patterns and token-level frequency statistics. This step is essential for downstream modeling, as

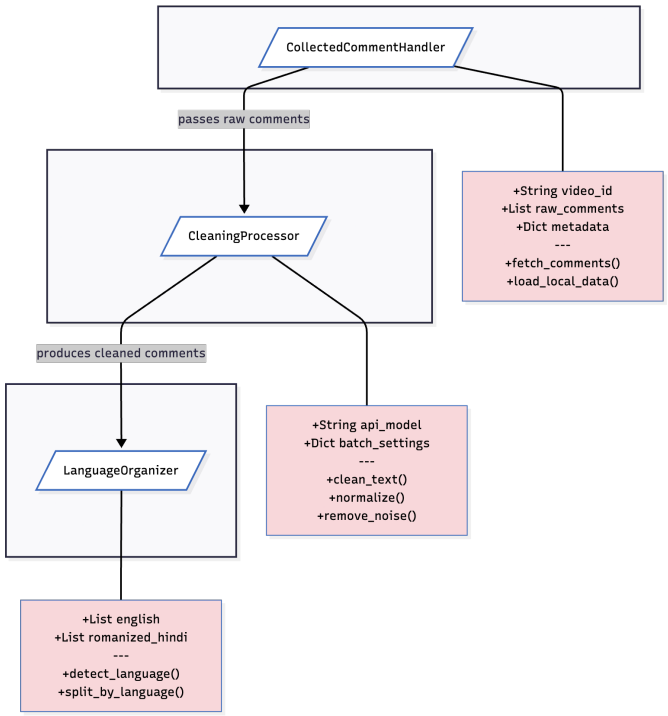

Fig. 2. Implementation architecture of the Input and Preprocessing Layer.

Romanized Hindi requires different prompt constraints and evaluation strategies.

B. Layer 2: Relevance Filtering

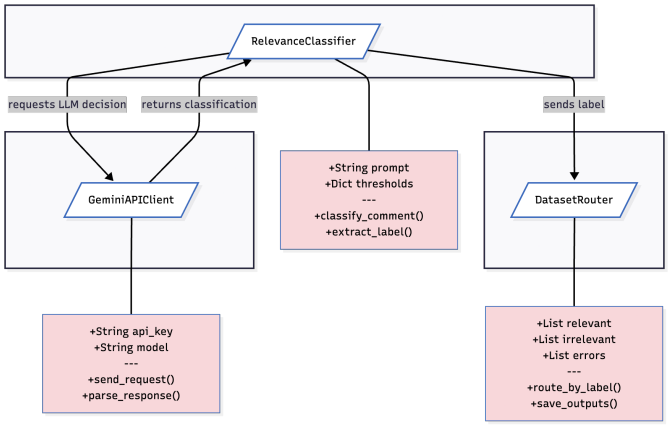

Fig. 3. Implementation architecture of the Relevance Filtering Layer.

The relevance filtering layer determines whether a comment warrants counterspeech intervention. The `RelevanceClassifier` constructs a structured prompt that encapsulates comment context, linguistic properties, and decision criteria. This prompt is then routed to the `GeminiAPIClient`, which provides semantic classification using a state-of-the-art LLM.

Returned labels are parsed by the classifier and forwarded to the `DatasetRouter`. The router writes three separate

JSONL files: `relevant.jsonl`, `irrelevant.jsonl`, and `error.jsonl`, each storing comments with extracted metadata and decision provenance. These files serve as the bridge between the preprocessing phase and counterspeech generation.

A notable implementation detail is batch processing with streaming parsing via the `ijson` library, enabling efficient handling of large comment datasets while maintaining low memory overhead.

C. Layer 3: Counterspeech Generation

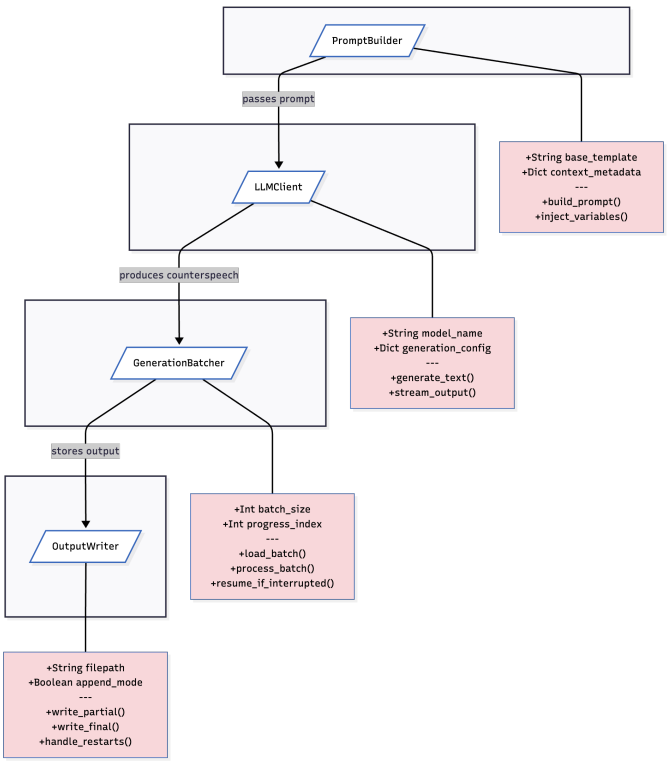

Fig. 4. Implementation architecture of the Counterspeech Generation Layer.

Counterspeech generation is implemented as a multi-stage pipeline comprising the `PromptBuilder`, `LLMClient`, `GenerationBatcher`, and `OutputWriter`.

The `PromptBuilder` assembles standardized instruction templates that enforce tone, structure, and safety constraints. The templates include placeholders for the comment text, contextual metadata, and (optional) language-specific guidance. This modular structure allows easily substituting, refining, or experimenting with prompt formats.

The `LLMClient` interfaces with the selected model (LLaMA32, Gemma2, GPT2, or Sarvam), configuring generation parameters such as temperature, maximum token length, and stopping criteria. Calls are wrapped in timed retries to handle model or API latency.

The `GenerationBatcher` orchestrates batch-level execution, loading comments in segments, generating counterspeech responses, and tracking progress. A resume mechanism writes intermediate results to disk, allowing long-running experiments to recover from hardware or network interruptions without data loss.

Finally, the `OutputWriter` commits partial and final generation outputs to disk in versioned JSON files. Each entry stores the original comment, generated counterspeech, prompt used, and metadata necessary for downstream evaluation.

D. Layer 4: Toxicity Evaluation

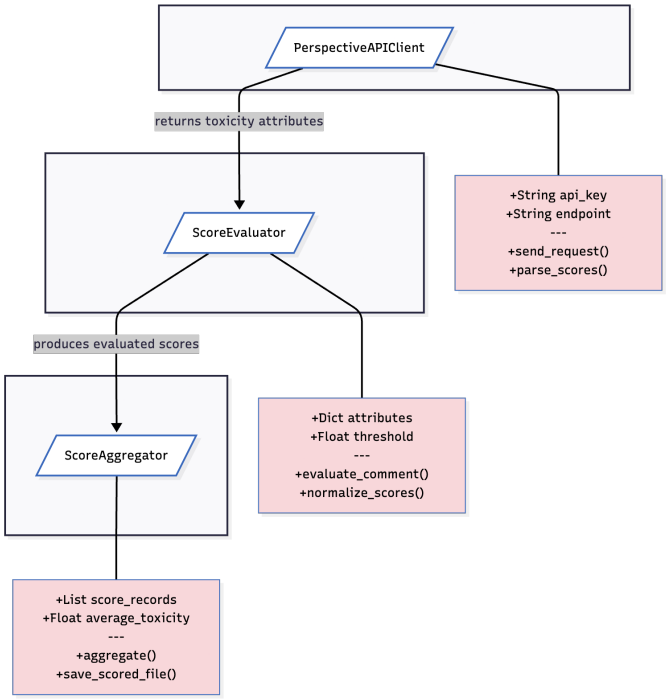

Fig. 5. Implementation architecture of the Toxicity Evaluation Layer.

To assess the quality and safety of generated counterspeech, the `ScoreEvaluator` prepares each response and passes it to the `DetoxifyClient`. The client generates a vector of toxicity attributes—including toxicity, insult, threat, profanity, and identity attack—which are returned in a structured JSON schema.

The evaluator normalizes these scores and performs basic thresholding to identify high-risk counterspeech candidates. This provides valuable insight into whether the LLM exhibits unintended harmful tendencies even when instructed to generate corrective or de-escalatory text.

The `ScoreAggregator` merges all evaluated scores into a single analysis-ready dataset, computing summary statistics such as mean toxicity per prompt style, toxicity distribution across languages, and response-level variance.

E. Layer 5: Analytics and Visualization

The final layer provides interpretability and insight into model behavior. The `AnalyticsEngine` loads the aggregated toxicity dataset and computes descriptive statistics, enabling the study of trends across languages, toxicity types, and comment categories.

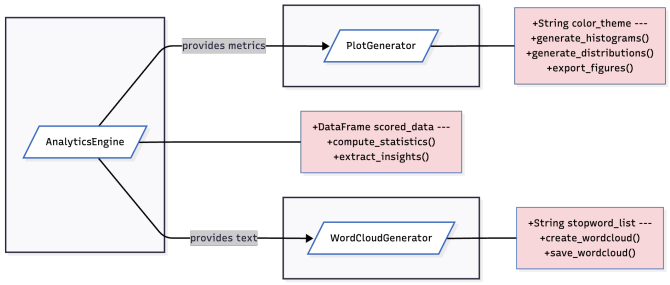

Fig. 6. Implementation architecture of the Analytics and Visualization Layer.

The `PlotGenerator` produces histograms, density plots, and comparative visualizations that summarize toxicity patterns at scale. All plots follow a consistent color theme for publication-ready styling.

Complementing this, the `WordCloudGenerator` extracts frequent lexical patterns from both comments and counterspeech outputs. This contextual visualization helps identify linguistic markers that may contribute to toxicity, as well as stylistic signatures of the LLM.

Together, these subsystems form a comprehensive analytical framework that supports both qualitative interpretation and quantitative experimentation.

V. Dataset

The dataset used in this study was constructed from a large corpus of publicly available online comments related to travel, destinations, safety, and cultural experiences. In total, approximately **600,000** raw comments were collected from YouTube travel vlogs. These comments represent a wide range of travel-related discourse, including personal experiences, advice, warnings, and culturally framed narratives.

Prior to model interaction, the raw corpus underwent an initial cleaning phase. This step removed duplicate entries, empty comments, non-textual artifacts, and malformed records. The cleaned corpus was then processed through an automated relevance filtering pipeline designed to identify comments suitable for counterspeech generation.

Relevance filtering was performed using a scripted prompt-based classification approach with the Gemini API. Each comment was evaluated for whether it contained travel-related content with sufficient contextual or narrative substance to warrant counterspeech generation (e.g., countering a misconception or stereotype about a location). Comments were labeled as *relevant* or *irrelevant* based on the model's classification output.

From the cleaned corpus, over **100,000** comments were identified as relevant. Approximately **2,000** comments resulted in classification or processing errors and were excluded. The remaining comments were labeled as irrelevant and discarded from further analysis. Only comments classified as relevant were retained for subsequent model-based generation and evaluation.

The final relevant dataset was then passed to a generation pipeline in which LLaMA32, Gemma2, GPT2, and Sarvam were prompted to produce counterspeech responses for each comment. These generated outputs form the basis of the human and automated toxicity analyses reported in the Results section.

Table I summarizes the dataset composition and filtering stages.

TABLE I
Dataset composition and filtering breakdown

Stage	Count
Raw collected comments	∼600,000
Comments after initial cleaning	∼600,000
Relevant comments (Gemini filtering)	>100,000
Filtering / processing errors	∼2,000
Irrelevant comments discarded	∼498,000
Final dataset for generation	>100,000

A. Experimental Setup

This study follows a multi-stage experimental pipeline consisting of relevance filtering, counterspeech generation, and toxicity evaluation. All experiments were conducted on fixed datasets and fixed model outputs to ensure reproducibility across evaluation methods.

Relevance Filtering. After initial data cleaning, relevance filtering was performed using an automated prompt-based classification script that queried the Gemini API. Each comment was evaluated independently to determine whether it contained travel-related content with sufficient contextual or narrative substance. The classification prompt instructed the model to label comments strictly as *relevant* or *irrelevant*. No toxicity or sentiment information was included in the relevance prompt to avoid biasing downstream analyses.

Counterspeech Generation. Counterspeech responses were generated using four diverse large language models: LLaMA32, Gemma2, GPT2, and Sarvam. Both Western-aligned and Indic-native models were treated as black-box generators and were not fine-tuned or adapted for this task. For each relevant comment, a single counterspeech response was generated per model using identical decoding parameters. Generation was performed with a temperature of $T = 0.7$, top-p sampling with $p = 0.9$, and a maximum output length of 512 tokens. No system-level safety filters, response ranking, or post-generation edits were applied.

Human Toxicity Annotation. Human evaluators annotated a subset of generated outputs using a discrete ordinal toxicity scale ranging from 0 to 5. These ratings were min–max normalized to the range $[0, 1]$ to enable direct comparison with automated toxicity scores. Normalized human scores are referred to as *Human_norm* throughout the Results section.

Automated Toxicity Evaluation. Automated toxicity assessment was conducted using Detoxify Model. For each generated response, scores were obtained for overall toxicity (TOX), severe toxicity (SEV), insult (INS), profanity (PROF), and identity attack (IDN). All Detoxify scores are continuous

values in the range $[0, 1]$. All analyses were performed on the same fixed set of generated outputs. No resampling, regeneration, or adaptive filtering was applied after counterspeech generation.

VI. Results

This section presents results from human and automated toxicity evaluations of AI-generated travel narratives designed to counter cultural and geographic misconceptions. The narratives were generated by four models: LLaMA32, Gemma2, GPT2, and Sarvam. Human toxicity annotations were collected on a discrete ordinal scale ranging from 0 (non-toxic) to 5 (highly toxic), evaluating whether the narrative successfully debunked a stereotype or inadvertently perpetuated hostility. For comparison with automated scores, human ratings were min–max normalized to the range $[0, 1]$. Automated evaluations were conducted using the Detoxify model.

A. Systemic Evaluation Bias and Drift

Figure 7 and Figure 8 illustrate the systemic macro-level differences between human perception and automated evaluation. As shown in Figure 7, human evaluators assign a relatively consistent average toxicity score ($\sim$0.40) across all four models. However, the Detoxify model scores heavily diverge based on the generation source.

Figure 8 visualizes the "Detoxify Penalty" (Detoxify Avg − Human Avg). LLaMA32, GPT2, and Sarvam all exhibit large negative penalties, meaning the automated system systematically underestimates the toxicity of their outputs compared to human raters. In contrast, Gemma2 is an outlier, with Detoxify scores closely matching or slightly exceeding human evaluations. This inversion illustrates a methodological pitfall: apparent model safety rankings depend entirely on whether toxicity is assessed by humans or automated classifiers.

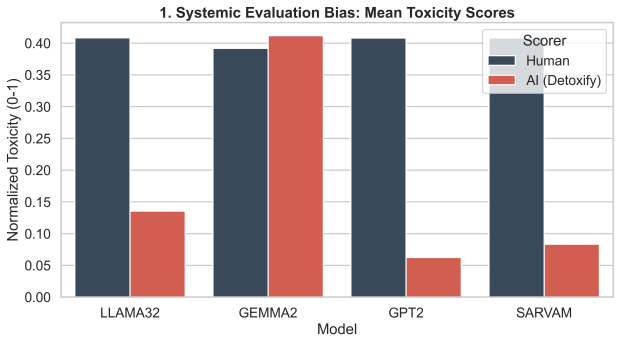

Fig. 7. Systemic Evaluation Bias: Comparison of mean toxicity scores as evaluated by Humans versus Detoxify.

B. Evaluator Bias and The AI Blind Spot

To understand the distribution of this disagreement, Figure 9 visualizes the density of the bias difference (Detoxify − Human). Values below zero indicate underestimation by Detoxify. The distributions for LLaMA32, GPT2, and Sarvam heavily cluster below zero (peaking near -0.4), confirming a systemic failure to detect human-perceived toxicity. Gemma2 exhibits a bimodal distribution with a notable secondary peak above zero.

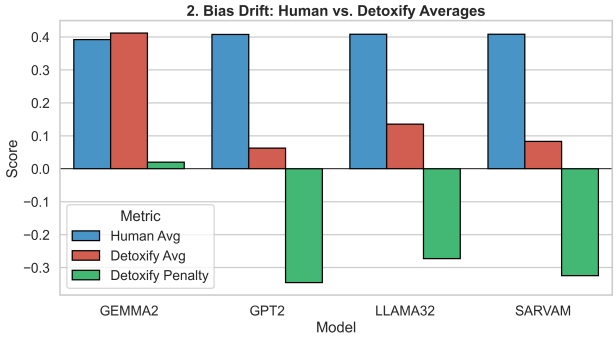

Fig. 8. Bias Drift: Human averages, Detoxify averages, and the resulting Detoxify penalty for each model.

Figure 10 further isolates this phenomenon by mapping AI scores against specific human rating buckets. We observe a distinct "AI Blind Spot": for outputs that humans rate in the low-to-moderate toxicity range (0.2 to 0.5), Detoxify uniformly suppresses its scores near zero. The automated model only begins to align with human judgment when human ratings exceed 0.6, indicating that Detoxify struggles with subtle, contextual, or implicit harms.

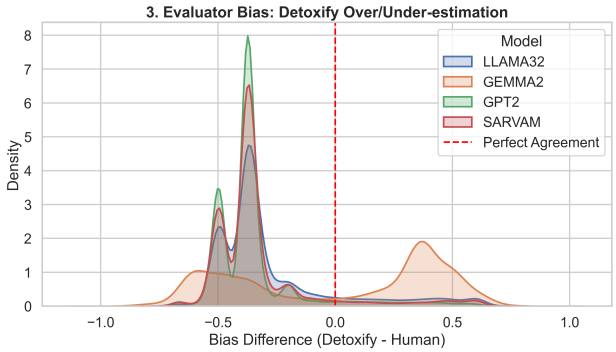

Fig. 9. Distribution of Evaluator Bias (Detoxify − Human). Values below zero represent AI underestimation.

C. Bland–Altman Agreement Analysis

A Bland–Altman analysis provides statistical insight into the agreement between Detoxify and human evaluations across the entire dataset. Figure 11 plots the mean of the human and AI scores against their difference (Human − AI).

The mean bias line sits at +0.23, mathematically confirming the systemic underestimation by the Detoxify model. Furthermore, the wide limits of agreement and the distinct dispersion patterns indicate significant heteroscedasticity. Disagreement is not random; it is structurally tied to the nature of the text being evaluated, suggesting that automated scoring lacks reliability for nuanced counterspeech.

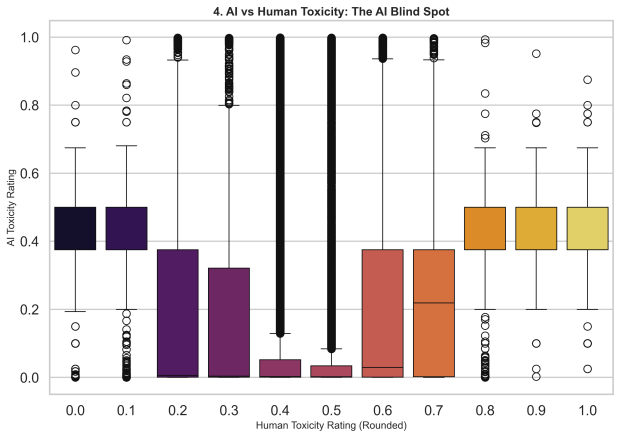

Fig. 10. The AI Blind Spot: Distribution of Detoxify scores binned by discrete human toxicity ratings.

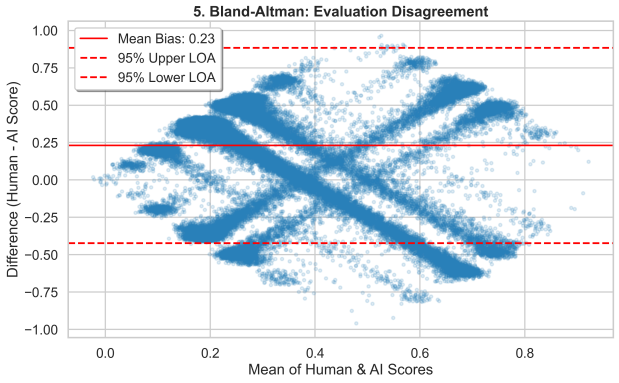

Fig. 11. Bland–Altman plot with jitter applied, evaluating the agreement between Human and Detoxify scores.

D. Cultural Bias Across Language Contexts

Figure 12 compares human toxicity scores across English and Romanized Hindi (rom_hindi) prompts. The toxicity distributions shift dramatically depending on the linguistic context. Outputs generated in response to Romanized Hindi prompts are tightly clustered at a moderate toxicity level across all models.

Conversely, English prompts elicit significantly higher variance and reach higher maximum toxicity thresholds, particularly for Gemma2. This suggests that the cultural and linguistic framing of the prompt strongly influences the generated narrative's perceived hostility, independent of the underlying model architecture.

E. Dimensional Profiles and Correlation

Beyond overall toxicity, we examined disagreement across specific toxicity dimensions: Severe Toxicity, Insult, Profanity, and Identity Attack. Figure 13 highlights the Mean Detection Gap (Human − AI) partitioned by model origin (Western vs. Indic/Sarvam). The gap is highly positive for Toxicity and Insult, indicating human sensitivity to affective markers. Conversely, the gap for Profanity is negative; Detoxify over-

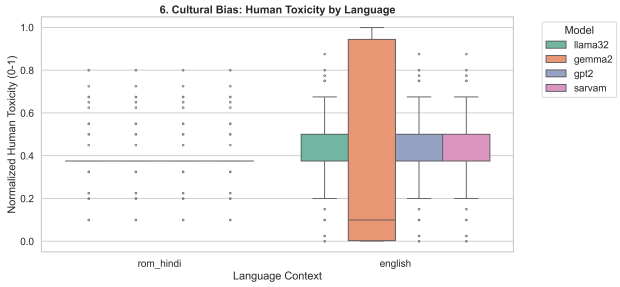

Fig. 12. Cultural Bias: Human toxicity score distribution categorized by Language Context (English vs. Romanized Hindi).

penalizes explicit language that humans deemed contextually harmless.

Figure 14 provides a 2×2 correlation matrix of these dimensions for each model. Across all architectures, Toxicity (TOX) and Insult (INS) are highly correlated ($r > 0.85$). However, Profanity (PROF) exhibits near-zero or negative correlation with Identity Attack (ID), indicating that identity-based bias in these narratives manifests independently of explicit profanity.

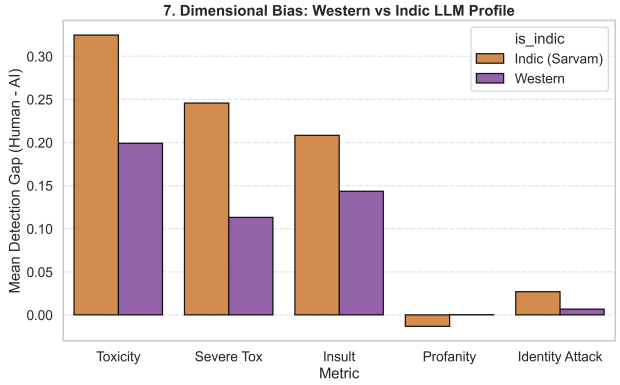

Fig. 13. Dimensional Bias Profile: Mean detection gap (Human − AI) across toxicity subclasses for Western and Indic models.

F. Narrative Style Bias

Finally, Figure 15 shows the Empirical Cumulative Distribution Function (ECDF) of the Detoxify ratings for each model. This plot isolates the AI's probability distribution of narrative style toxicity. LLaMA32, GPT2, and Sarvam exhibit steep curves, with nearly 80% of their outputs rated below 0.1 by Detoxify. Gemma2 demonstrates a much flatter, stepped distribution, generating a wider variety of stylistic text that Detoxify interprets as toxic, underscoring fundamental differences in alignment tuning between the models.

G. The Misconception Correction Penalty

The divergence between human and automated scoring highlighted across these dimensions points to a phenomenon we term the "Misconception Correction Penalty." To effectively counter a deeply ingrained travel stereotype, an LLM

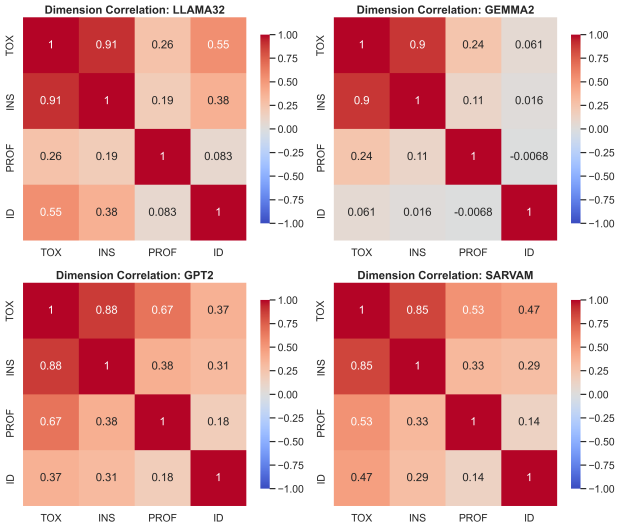

Fig. 14. Correlation heatmaps of Detoxify dimensions across LLaMA32, Gemma2, GPT2, and Sarvam.

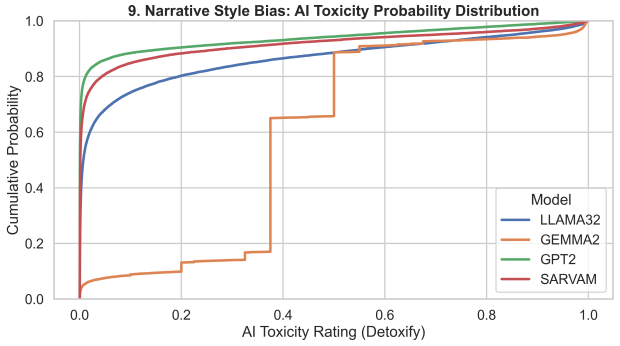

Fig. 15. Narrative Style Bias: Empirical Cumulative Distribution Function of Detoxify ratings across the four models.

f

often must explicitly state the harmful misconception before deconstructing it, or use emphatic language to defend a marginalized destination. Human raters accurately contextualize these rhetorical strategies as educational and pro-social.

However, Detoxify's reliance on rigid lexical pattern matching heavily penalizes the syntactic structure of these defensive narratives. The inverted gap observed in the Profanity dimension (Figure 13)—where AI consistently overestimates severity—demonstrates that automated classifiers cannot distinguish between aggressive toxicity and a passionate, colloquial defense of a culture. Furthermore, the distinct behavioral clustering of Sarvam as an Indic-native model underscores that Western-aligned safety tuning (seen in LLaMA32 and GPT2) may not adequately parse the cultural nuances required to discuss non-Western destinations, leading to the systemic misclassification of cultural critiques as insults.

H. Summary of Findings

Across the multi-dimensional evaluations, several critical insights emerge regarding the viability of automated toxicity detection for complex, culturally sensitive travel generation:

1) **Systemic AI Underestimation:** While human evaluators perceive comparable macro-level toxicity across the models when evaluating cultural sensitivity, automated scoring systematically fails to detect it. Detoxify drastically underestimates the toxicity of outputs from LLaMA32, GPT2, and Sarvam, potentially masking narratives that subtly reinforce travel stereotypes.
2) **The Implicit Bias "Blind Spot":** Detoxify exhibits a severe failure rate for implicit harm. Human-perceived toxicity in the low-to-moderate range (0.2–0.5)—often representing subtle microaggressions or exoticization of destinations—is routinely scored near absolute zero by the AI.
3) **Cultural and Linguistic Volatility:** Linguistic framing is a stronger determinant of toxicity variance than model architecture. English-language travel prompts elicited significantly wider toxicity distributions and higher severity peaks than Romanized Hindi prompts, highlighting culturally mediated safety vulnerabilities in how destinations are described.
4) **The Contextual Penalty:** Automated scoring is highly sensitive to lexical triggers rather than educational intent. Detoxify heavily over-penalizes emphatic language while entirely missing contextual prejudice, rendering it an unreliable ground-truth metric for evaluating how well an LLM counters geographic and cultural misconceptions.

Ultimately, these results demonstrate that relying on automated classifiers to rank model safety creates a false empirical hierarchy. Evaluation tools do not neutrally measure harm; they actively shape research findings through their inability to understand the cultural and educational context of travel narratives.

VII. Future Work

Future work should expand upon our findings with Sarvam by incorporating a broader set of Indic and other non-Western regional language models. This will help further isolate the effects of cultural context, linguistic framing, and world knowledge from model architecture alone, confirming whether the "Misconception Correction Penalty" observed in Western models is universally mitigated by regional alignment.

Beyond toxicity, future studies should incorporate richer human evaluation dimensions, including politeness, empathy, and contextual appropriateness. These factors are especially relevant in travel and cultural discourse, where harm reduction cannot be adequately captured through toxicity alone. The consistent disagreement observed between human judgments and automated evaluation tools motivates further research into context-aware and culturally calibrated assessment methods. Developing hybrid frameworks that integrate human evaluation with uncertainty-aware automated scoring may improve reliability in culturally sensitive settings.

Finally, future work should move beyond single-response analysis to examine longitudinal discourse effects, such as how AI-generated counterspeech influences subsequent interaction

dynamics. This direction would enable a more realistic understanding of the societal impact of deploying generative models in online travel and moderation contexts.

VIII. Conclusion

This study examined bias in AI-generated travel narratives by treating large language models not as objects of optimization, but as instruments for revealing broader representational and evaluative patterns. By comparing counterspeech generated by LLaMA32, Gemma2, GPT2, and Sarvam across human and automated toxicity assessments, we demonstrated that conclusions about model behavior are highly sensitive to the choice of evaluation methodology.

Human evaluations revealed substantial variability in perceived toxicity across models and linguistic contexts, while automated scoring consistently underestimated implicit harm and produced unstable model rankings. In particular, the systematic disagreement between human judgments and Detoxify-based evaluations highlights the limitations of applying generic toxicity classifiers to culturally and contextually rich domains such as travel discourse. These findings show that automated tools do not merely measure bias, but actively shape it through their rigid algorithmic assumptions about language, intent, and harm.

More broadly, this work underscores that bias in generative systems is not confined to model outputs alone. It emerges from interactions among training data, prompting strategies, cultural context, and evaluation frameworks. Treating models as evidence rather than endpoints enables a more critical and transparent examination of these interactions.

As generative AI systems increasingly mediate how people perceive places, cultures, and social groups, the need for context-aware, human-grounded evaluation becomes essential. This study contributes empirical evidence toward that goal and motivates the development of evaluation practices that are sensitive to cultural nuance, methodological assumptions, and real-world impact.

References

[1] C. V. Kumar, A. Urlana, G. Kanumolu, B. M. Garlapati, and P. Mishra, "No llm is free from bias: A comprehensive study of bias evaluation in large language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.11985

[2] K.-C. Yeh, J.-A. Chi, D.-C. Lian, and S.-K. Hsieh, "Evaluating interfaced llm bias," in *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*. Taipei City, Taiwan: The Association for Computational Linguistics and Chinese Language Processing, October 2023.

[3] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng, ""kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters," 2023. [Online]. Available: https://arxiv.org/abs/2310.09219

[4] L. Lin, L. Wang, J. Guo, and K.-F. Wong, "Investigating bias in llm-based bias detection: Disparities between llms and human perception," 2024. [Online]. Available: https://arxiv.org/abs/2403.14896

[5] L. Ling, F. Rabbi, S. Wang, and J. Yang, "Bias unveiled: Investigating social bias in llm-generated code," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[6] M. Rathod, A. Yadav, S. Phutane, and S. Bhise, "Comment compass: An approach to analyze youtube comments through web scraping, sentiment analysis, and generative ai for actionable insights," in *2024 First International Conference on Pioneering Developments in Computer Science Digital Technologies (IC2SDT)*, 2024, pp. 1–5.

[7] B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. Maity, P. Goyal, and A. Mukherjee, "Thou shalt not hate: Countering online hate speech," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 369–380, 07 2019.

[8] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 59–68. [Online]. Available: https://doi.org/10.1145/3287560.3287598

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019. [Online]. Available: http://arxiv.org/abs/1908.09635

[10] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 149–159. [Online]. Available: https://proceedings.mlr.press/v81/binns18a.html

[11] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, "Holistic evaluation of language models," 2023. [Online]. Available: https://arxiv.org/abs/2211.09110

[12] OpenAI and Anthropic, "Fairness indicators for large language models," *arXiv preprint arXiv:2305.17680*, 2023. [Online]. Available: https://arxiv.org/abs/2305.17680

[13] A. Kontogianni and E. Alepis, "Ai in smart tourism: Llms gpts leading the way," in *2024 15th International Conference on Information, Intelligence, Systems Applications (IISA)*, 2024, pp. 1–8.

[14] M. Kamruzzaman, M. Shovon, and G. Kim, "Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 8940–8965. [Online]. Available: https://aclanthology.org/2024.findings-acl.530/

[15] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu, "Bias and unfairness in information retrieval systems: New challenges in the llm era," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2024, p. 6437–6447. [Online]. Available: http://dx.doi.org/10.1145/3637528.3671458

[16] R. Ren, X. Yao, S. Cole, and H. Wang, "Are large language models ready for travel planning?" *ArXiv*, vol. abs/2410.17333, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273532658

[17] K. Puh and M. Bagic Babac, "Predicting sentiment and rating of tourist reviews using machine learning," *Journal of Hospitality and Tourism Insights*, vol. 6, 07 2022.

[18] Y. Singgalen, "Understanding digital engagement through sentiment analysis of tourism destination through travel vlog reviews," *KLIK Kajian Ilmiah Informatika dan Komputer*, vol. 4, pp. 2992–3004, 06 2024.

[19] ——, "Toxicity and topic analysis of travel vlog content in digital era: perspective and multilingual embedding model (voyage- multilingual-2)," *Jurnal Teknik Informatika C I T Medicom*, vol. 16, pp. 199–210, 09 2024.

[20] ——, "Tourism and travel content analysis for market segmentation using toxicity and sentiment classification in communalytic," *Building of Informatics Technology and Science (BITS)*, vol. 6, pp. 469–479, 07 2024.