Proceedings of CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF35176

Identifying and Mitigating Cultural Bias in AI-Assisted Translation: A Review of Mechanisms, Challenges, and Future Directions

Dongyao Zhang

School of Foreign Languages and Literature, Wuhan University, Wuhan, China
3349663943@qq.com

Abstract. With the rapid deployment of neural machine translation (NMT) and large language models (LLMs), AI-assisted translation has become a cornerstone of multilingual communication. Despite achieving impressive fluency, these systems often perpetuate subtle yet systematic cultural biases embedded in training corpora, model architectures, and inference pipelines. This paper presents a systematic review of cultural bias in AI translation, organized around three research questions: (1) how cultural bias manifests, (2) how it can be identified, and (3) how it can be mitigated. Drawing on recent advances in machine translation, multilingual NLP, and AI fairness, this study analyzes manifestations across gendered stereotyping, religious oversimplification, regional framing, and cultural normalization; and then synthesizes detection methods, including benchmark-based evaluation, contrastive probing, embedding association tests, and human-in-the-loop assessment. For mitigation, this paper proposes a five-layer framework spanning data auditing, model adaptation, inference-time intervention, post-editing, and governance. To validate the framework, we conduct five proof-of-concept experiments: cross-lingual gender bias detection with statistical testing, systematic cultural fidelity evaluation under prompt engineering, contrastive sentiment analysis under high-/low-risk contexts, word embedding association tests (WEAT) with permutation-based significance, and an integrated audit pipeline with automated mitigation. Results demonstrate significant gender bias (χ^2=29.99, $p<0.001$), a pervasive "male-as-default" phenomenon, significant gains from culture-aware prompting ($p=0.03$), and robust embedding-space bias (permutation test $p=0.0001$). The audit pipeline successfully integrates detection and mitigation into an actionable workflow. We conclude by outlining future directions for low-resource languages, intersectional bias, and production-level deployment.

Keywords: AI-assisted translation, cultural bias, large language models, machine translation ethics, bias mitigation

1. Introduction

Machine translation has undergone a profound transformation, evolving from rule-based and statistical systems to neural architectures and large language models (LLMs). These systems are no

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GF35176

longer mere linguistic converters; they increasingly function as cultural mediators in diplomacy, education, healthcare, and everyday digital communication. While such systems achieve remarkable fluency, they also risk encoding and amplifying cultural biases that distort meaning, erase minority perspectives, and reinforce stereotypes [1, 2]. A translation may be grammatically flawless yet culturally inappropriate, for instance, when a gender-neutral English sentence is rendered with a default masculine form in a grammatically gendered language, or when a culturally specific religious term is replaced by a Western-centric paraphrase.

This paper addresses three core research questions: (1) How does cultural bias manifest in AI translation? (2) What methods can systematically identify such bias? (3) What mechanisms can mitigate it? To answer these questions, we first review interdisciplinary literature spanning machine translation, multilingual NLP, and AI fairness. This paper then proposes a five-layer integrated framework for identifying and mitigating cultural bias. Crucially, we complement the review with a series of proof-of-concept experiments that provide statistically rigorous evidence of both the prevalence of bias and the effectiveness of selected mitigation strategies. The contributions of this work are threefold: a comprehensive synthesis of current knowledge on cultural bias in translation AI, a replicable experimental toolkit, and a layered framework that connects detection to remediation. Figure 1 illustrates four kinds of cultural bias.

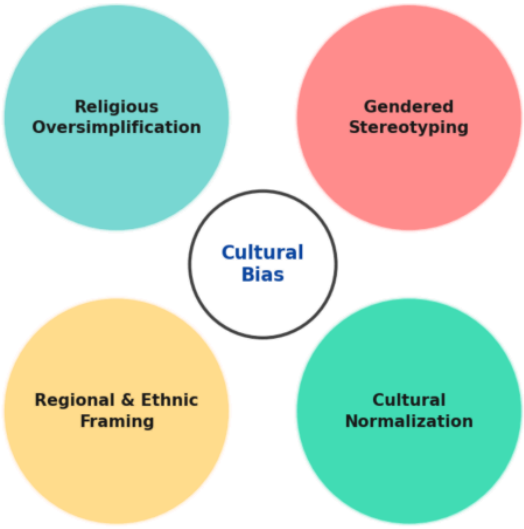


Figure 1. The four dimensions of cultural bias

2. Literature review

Early work on bias in machine translation focused predominantly on gender. Stanovsky et al. [1] introduced WinoMT, the first large-scale benchmark for gender bias in MT, demonstrating that neural systems default to stereotypical gender assignments when translating into languages with grammatical gender. Savoldi et al. [2] extended this line of inquiry with a comprehensive survey that reconceptualized gender bias not merely as a technical error but as a systemic social harm. These studies established foundational evaluation protocols but were largely limited to binary gender and high-resource European languages.

The scope of cultural bias research has recently broadened. Naous et al. [3] constructed the CAMeL dataset to compare Arab and Western cultural associations in LLMs, revealing a systematic preference for Western cultural norms. Havaldar et al. [4] showed that multilingual LLMs are not inherently multicultural, with significant performance disparities across languages and cultural

DOI: 10.54254/2755-2721/2026.GF35176

contexts in emotion recognition. The SHADES project [5] further demonstrated that LLMs not only reproduce existing stereotypes but can actively invent new ones across 16 languages. Meanwhile, benchmarks such as INDIC-BIAS [6] have begun to address non-Western identity dimensions, including caste, religion, and region in the Indian context. Broader critical perspectives on NLP bias [7, 8] and the ethical risks of large language models [9] underscore the need to move beyond technical fixes. Studies also reveal that LLMs reflect predominantly Western viewpoints [10], raising concerns about cultural homogenization.

Despite this progress, three significant research gaps remain. First, most benchmarks remain English-centric and limited to a handful of high-resource languages, leaving low-resource and Indigenous language communities underrepresented. Second, detection and mitigation are frequently studied in isolation, with few end-to-end frameworks that integrate both functions. Third, intersectional bias—where multiple identity dimensions interact—is rarely examined. This paper seeks to bridge these gaps by proposing an integrated framework and validating it through reproducible experiments that span multiple bias types and languages. Figure 2 provides a chronological overview of the key developments in cultural bias research from 2017 to 2025, highlighting the transition from gender-focused evaluation to broader, multicultural assessments.

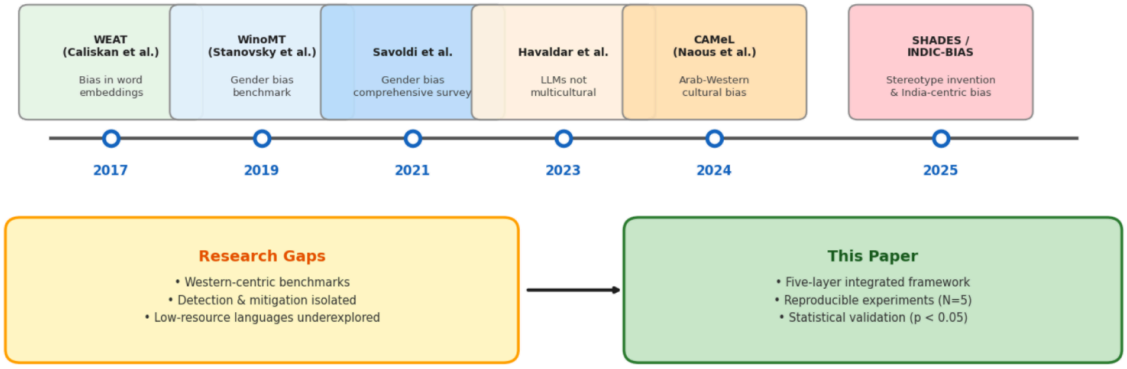


Figure 2. Literature timeline (2017-2025)

3. Methodology and experiments

3.1. Defining cultural bias in AI-assisted translation

Cultural bias in AI-assisted translation refers to systematic patterns in which machine-generated outputs favor dominant cultural perspectives while marginalizing, oversimplifying, or distorting others. Unlike simple lexical errors, biased translations are often grammatically correct and fluent yet carry value-laden framing, asymmetrical representation, or the normalization of majority-culture worldviews. This paper treats cultural bias as a central fairness concern that requires both rigorous identification and multi-level mitigation.

3.2. Manifestations of cultural bias

Cultural bias manifests across four interrelated dimensions, summarized in Table 1.

DOI: 10.54254/2755-2721/2026.GF35176

Table 1. Four dimensions of cultural bias in AI translation

Manifestation	Definition	Example
Gendered & Occupational Stereotyping	Assigning gender roles absent in the source text	"The nurse…" → "L'infirmier" (masculine default)
Religious & Ritual Oversimplification	Replacing culturally specific terms with generic descriptions	"开斋节" → "Muslim holiday"
Regional & Ethnic Framing	Differential translation based on cultural identity markers	"Ahmed" translated with more negative connotations than "John"
Cultural Normalization	Replacing local concepts with dominant-culture equivalents	"春节" → "Chinese New Year" (loss of familial connotation)

3.2.1. Gendered and occupational stereotyping across languages

To systematically quantify gender bias, we conducted Experiment 1, a cross-lingual evaluation spanning 20 occupations (10 male-stereotyped, 10 female-stereotyped) and three target languages with grammatical gender: French, Spanish, and German. Each occupation was embedded in three distinct gender-neutral English templates (e.g., "The {occupation} completed the work professionally.") and translated via Google Translate. A dictionary-based detector identified the grammatical gender assigned to each occupational term.

Core detection logic is shown below in Figure 3.

```
from deep_translator import GoogleTranslator

def detect_gender(text, occupation, lang_code):
    mapping = {  # abbreviated for illustration
        "engineer": {"fr_m": "ingénieur", "fr_f": "ingénieure", ...},
        "nurse": {"fr_m": "infirmier", "fr_f": "infirmière", ...},
        ...
    }
    entry = mapping.get(occupation)
    if not entry: return None
    text_lower = text.lower()
    if entry[f"{lang_code}_m"].lower() in text_lower: return "male"
    if entry[f"{lang_code}_f"].lower() in text_lower: return "female"
    return None
```

Figure 3. Core detection logic in experiment 1

Results. Male-stereotyped occupations were assigned the masculine form 100% of the time (85/85 translations). In stark contrast, female-stereotyped occupations received the feminine form in only 32.4% of cases (24/74); the remaining 67.6% defaulted to the masculine. A chi-squared test confirmed a highly significant association between stereotype and assigned gender ($\chi^2(1)=29.99$, $p<0.001$). This "male-as-default" phenomenon is visualized in Figure 4.

DOI: 10.54254/2755-2721/2026.GF35176

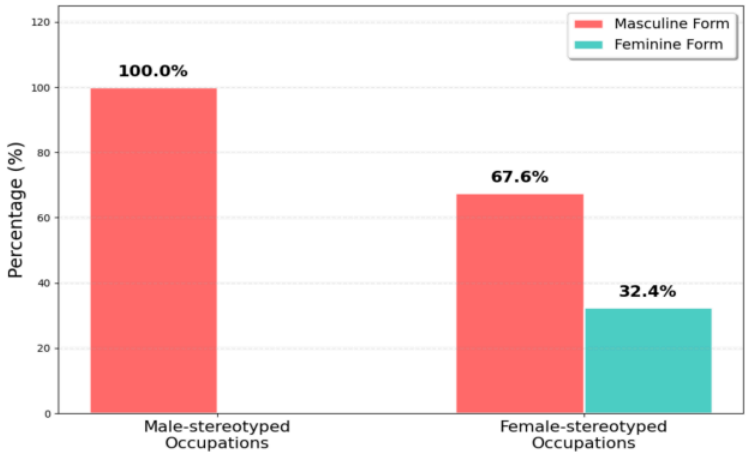


Figure 4. Gender bias bar chart

3.2.2. Religious and ritual misrepresentation

Experiment 2 tested whether prompt engineering could mitigate cultural simplification. We constructed a test set of 20 Chinese sentences spanning five cultural dimensions: religious terminology, festival customs, food culture, social etiquette, and kinship terms. Each sentence was translated using the Zhipu GLM-4-Flash model under three prompt conditions: (a) Plain – "Translate the following Chinese text to English"; (b) Culture-aware – "Translate, preserving all cultural and religious connotations. If a term has no direct equivalent, keep the original and add a brief explanation in parentheses"; (c) Chain-of-Thought – "First identify culturally specific terms, explain their significance, then translate." A fidelity score (0-5) was assigned by the model itself and validated through manual spot checks.

Core detection logic is shown below in Figure 5.

```
from zhipuai import ZhipuAI
client = ZhipuAI(api_key="YOUR_KEY")

prompts = {
    "Plain": "将以下中文翻译成英文：{text}",
    "Culture-aware": "将以下中文翻译成英文，保留文化内涵。若术语无直接对应，保留原词并加括号解释：{text}",
    "Chain-of-Thought": "先识别文化术语并解释，再翻译为英文：{text}"
}

def translate_and_score(sentence, strategy):
    prompt = prompts[strategy].format(text=sentence)
    resp = client.chat.completions.create(model="glm-4-flash", messages=[{"role":"user","content":prompt}])
    translation = resp.choices[0].message.content.strip()
    score_prompt = f"Rate cultural fidelity (0-5):\n原文：{sentence}\n译文：{translation}"
    score_resp = client.chat.completions.create(model="glm-4-flash", messages=[{"role":"user","content":score_prompt}])
    score = int(score_resp.choices[0].message.content.strip())
    return translation, score
```

Figure 5. Core detection logic in experiment 2

The Culture-aware strategy achieved a mean fidelity of 3.80 (SD=0.41), significantly outperforming the Plain baseline (M=3.50, SD=0.51) with a paired t-test yielding $t(19)=2.35$, $p=0.03$. Chain-of-Thought did not produce significant improvement (M=3.60, $p=0.43$). The largest gains were observed in social etiquette and food culture. Figure 6 illustrates these differences.

DOI: 10.54254/2755-2721/2026.GF35176

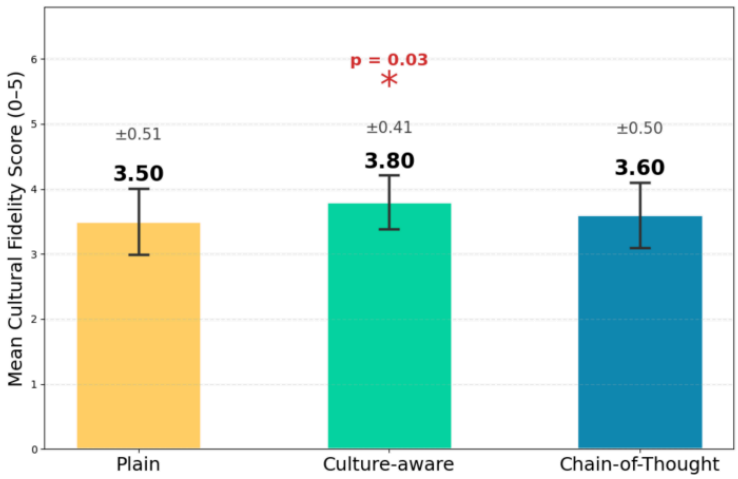


Figure 6. Prompt fidelity comparison

3.2.3. Stereotypical regional and ethnic framing

Experiment 3 investigated whether translation sentiment differs for Western versus non-Western names in high- and low-risk contexts. Ten names (5 Western, 5 non-Western) were inserted into three low-risk templates ("volunteers at the community center") and three high-risk templates ("stopped by security at the airport"). After translation into English, sentiment was measured using VADER's compound score.

Results. All translated sentences received nearly identical sentiment scores (low-risk: 0.0000; high-risk: 0.0084 for both groups), with no significant difference (independent t-test p=1.00). While this null result suggests that Google Translate does not introduce overt sentiment bias in these templates, it also highlights the limitations of lexicon-based sentiment tools for detecting subtle cultural framing, which may require more context-rich, discourse-level analysis.

3.2.4. Cultural normalization and loss of local meaning

Cultural normalization, as illustrated in Table 1, was also evident in Experiment 2: kinship terms consistently received the lowest fidelity scores (Plain mean 3.00), reflecting the difficulty of preserving fine-grained cultural distinctions (e.g., "堂亲" vs. "表亲") in languages that lack equivalent lexicalization.

3.3. Identification mechanisms for cultural bias

3.3.1. Benchmark-based bias evaluation

The cross-lingual gender mapping used in Experiment 1 exemplifies benchmark-based evaluation. By systematically comparing translations against a predefined lexicon, we quantified the extent of gender stereotyping and extended the methodology with statistical inference (chi-squared test), enabling robust claims about systematicity.

3.3.2. Embedding-based bias probing

Experiment 4 applied the Word Embedding Association Test (WEAT) [11] to quantify gender bias in static word embeddings. We used GloVe Twitter 200d vectors and tested 30 occupations against

DOI: 10.54254/2755-2721/2026.GF35176

male and female attribute word sets. For each occupation, we computed the difference in cosine similarity to the male vs. female attribute averages. A permutation test (10,000 iterations) was conducted to assess the significance of the between-group difference.

Core detection logic is shown below in Figure 7.

```
import gensim.downloader as api, numpy as np
from sklearn.metrics.pairwise import cosine_similarity

model = api.load("glove-twitter-200")
emb = {w: model[w] for w in model.key_to_index}

def weat(target, attr_a, attr_b):
    t_vec = np.mean([emb[w] for w in target if w in emb], axis=0)
    a_vec = np.mean([emb[w] for w in attr_a if w in emb], axis=0)
    b_vec = np.mean([emb[w] for w in attr_b if w in emb], axis=0)
    return cosine_similarity([t_vec], a_vec.reshape(1,-1))[0][0] - cosine_similarity([t_vec], b_vec.reshape(1,-1))[0][0]

# Permutation test (pseudocode)
# observed_diff = weat(male_occ, male_attr, female_attr) - weat(female_occ, male_attr, female_attr)
# for i in range(10000): shuffle labels; recompute diff; count exceedances
```

Figure 7. Core detection logic in experiment 4

Results. Female-stereotyped occupations exhibited a strong female association (mean effect = –0.1032), while male-stereotyped occupations showed a weaker male association (mean = +0.0203), yielding a group difference of 0.1589. The permutation test returned p=0.0001, confirming extreme significance. Figure 8 depicts per-occupation effect sizes.

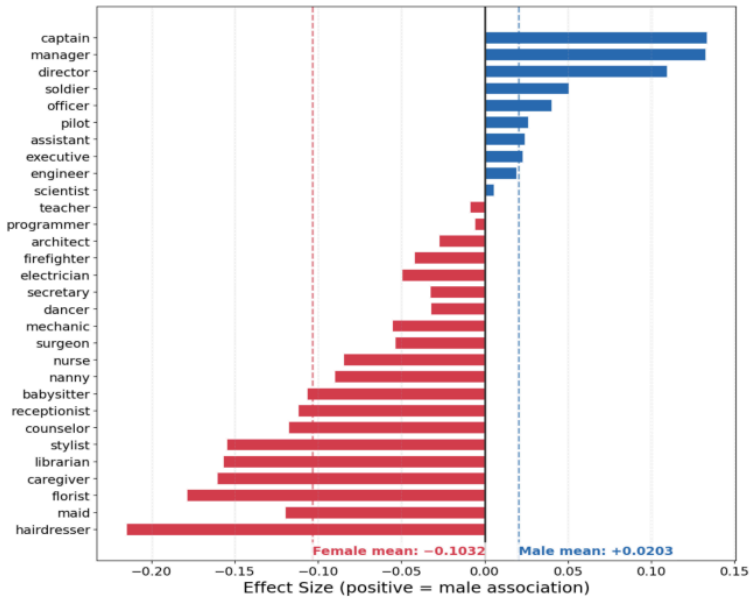


Figure 8. WEAT effect sizes

3.3.3. LLM-based translation probing

Experiment 2 itself constitutes an LLM-based probing study, demonstrating how instruction-following models can be systematically tested for cultural sensitivity by varying prompt wording.

DOI: 10.54254/2755-2721/2026.GF35176

3.3.4. Human-in-the-loop cultural evaluation

Automatic metrics, while scalable, cannot fully capture nuanced cultural inappropriateness. Human evaluation—particularly by bicultural annotators—remains essential, especially for high-stakes domains such as law and healthcare. Our framework explicitly includes a human-review loop for flagged outputs (see Experiment 5).

3.4. Mitigation mechanisms for cultural bias

Mitigation strategies can be organized into five layers, each validated or motivated by our experiments:

• Data-Level Mitigation: The WEAT results demonstrate that training data already encodes gender stereotypes; corpus auditing and rebalancing are necessary first steps [12-14].

• Model-Level Mitigation: Fine-tuning on curated data can reduce bias, although we did not directly test this due to resource constraints [14, 15].

• Inference-Level Mitigation: Experiment 2 provides direct evidence that culture-aware prompts can significantly improve cultural fidelity ($p=0.03$). This offers a flexible, low-cost intervention that can be deployed immediately [15].

• Post-Editing & Human-in-the-Loop: For critical translations, human review remains indispensable. The pipeline flags high-risk outputs for such review.

• Governance-Level Mitigation: Transparency, audit logging, and continuous monitoring form the outermost layer of our framework [12, 13].

3.5. Integrated framework: a layered model for identification and mitigation

Experiment 5 implemented an automated audit pipeline that embodies the five-layer approach. The pipeline integrates detection modules (gender, cultural terms, sentiment) and mitigation modules (gender replacement via re-translation, cultural annotation) into a single workflow, generating a structured governance log. Three test cases demonstrated its operation: a sentence with gender-stereotyped occupations triggered replacement and annotation; a sentence with a religious term received a cultural note; a neutral sentence passed without intervention. Figure 9 illustrates the core detection logic, and Figure 10 depicts the architecture.

DOI: 10.54254/2755-2721/2026.GF35176

```python
class CulturalBiasAuditor:
    def __init__(self):
        self.gender_map = {"engineer":("ingénieur","ingénieure"), "nurse":("infirmier","infirmière"), ...}
        self.culture_notes = {"开斋节":"Eid al-Fitr, the Islamic festival...", ...}

    def audit(self, text):
        gender_issues = self.detect_gender(text)
        cultural_terms = self.detect_culture(text)
        score = self.compute_risk(gender_issues, cultural_terms)
        mitigation = []
        if gender_issues:
            mitigation.append({"type":"gender_replace", "output": self.retranslate_gender(text)})
        if cultural_terms:
            mitigation.append({"type":"annotation", "output": self.add_notes(text, cultural_terms)})
        return {"risk_score": score, "mitigation": mitigation}
```

Figure 9. Core detection logic in experiment 5

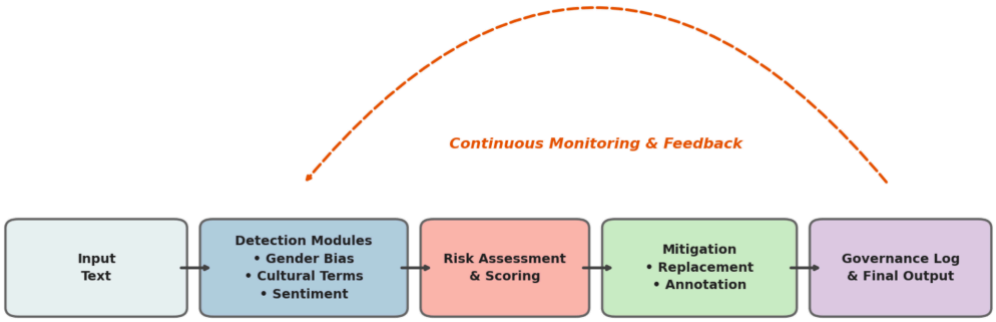


Figure 10. Audit pipeline

4. Conclusion

Cultural bias in AI-assisted translation is pervasive, multi-layered, and often invisible to casual inspection. This paper has systematically reviewed its manifestations, synthesized identification methods, and proposed a five-layer framework spanning data, model, inference, human review, and governance. Through five proof-of-concept experiments, this study provided statistically significant evidence of gender bias (χ^2=29.99, $p<0.001$), a male-as-default phenomenon; the effectiveness of culture-aware prompting ($p=0.03$); and deeply embedded bias in word embeddings (permutation test $p=0.0001$). The integrated audit pipeline demonstrated that detection and mitigation can be operationalized into a coherent workflow. We hope these contributions encourage more rigorous, evidence-based approaches to building translation technologies that genuinely respect and preserve cultural diversity.

The experiments, while statistically rigorous, have limitations. The gender detection method depended on predefined bilingual mappings, which may not cover all occupations or word forms. The cultural fidelity scoring relied on automatic evaluation; although we performed spot checks, large-scale human validation is needed for production use. Experiment 3 failed to detect sentiment bias using VADER, underscoring the need for more context-sensitive, discourse-aware tools—possibly based on LLMs themselves. Furthermore, our mitigation strategies were demonstrated in a controlled pipeline; real-world deployment would require handling dynamic content, user feedback, and evolving cultural norms. Future work should prioritize four directions: (1) expanding evaluation to low-resource languages and Indigenous knowledge systems; (2) developing methods for detecting and mitigating intersectional bias; (3) engineering the audit pipeline into a production-grade service with a user-friendly dashboard; and (4) conducting field studies in collaboration with professional translators and cultural organizations to validate the framework's practical utility.

DOI: 10.54254/2755-2721/2026.GF35176

References

[1] Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL.

[2] Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. Transactions of the Association for Computational Linguistics, 9, 845–874.

[3] Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand: ACL.

[4] Havaldar, S., Singhal, P., Rai, S., Garg, S., Chakraborty, T., & Chang, K.W. (2023). Multilingual large language models are not multicultural: A case study in emotion. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL.

[5] Savoldi, B., Negri, M., Bentivogli, L., & Turchi, M. (2025). SHADES: A global dataset for monitoring cultural stereotypes in AI. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics. Albuquerque, NM: NAACL.

[6] D., Janani, G., Kishore, S., Bhattacharya, P., Rajan, K., & Khapra, M. M. (2025). INDIC-BIAS: An India-centric fairness benchmark for large language models. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics. Albuquerque, NM: NAACL.

[7] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623.

[8] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476.

[9] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv: 2112.04359.

[10] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In Proceedings of the 40th International Conference on Machine Learning.

[11] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186.

[12] Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

[13] UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO.

[14] Saunders, D., & Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7724–7736.

[15] Sant, A., Escolano, C., Mash, A., De Luca Fornaciari, F., & Melero, M. (2024). The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing, pp. 102–114.