BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100235

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full length article

A framework for evaluating cultural bias and historical misconceptions in LLMs outputs

Moon-Kuen Mak [a,b], Tiejian Luo [b,*]

[a] Institute for the History of Natural Sciences, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Large language model
Artificial intelligence
Cultural bias
Historical misconception
human-in-the-loop

ABSTRACT

Large Language Models (LLMs), while powerful, often perpetuate cultural biases and historical inaccuracies from their training data, marginalizing underrepresented perspectives. To address these issues, we introduce a structured framework to systematically evaluate and quantify these deficiencies. Our methodology combines culturally sensitive prompting with two novel metrics: the Cultural Bias Score (CBS) and the Historical Misconception Score (HMS). Our analysis reveals varying cultural biases across LLMs, with certain Western-centric models, such as Gemini, exhibiting higher bias. In contrast, other models, including ChatGPT and Poe, demonstrate more balanced cultural narratives. We also find that historical misconceptions are most prevalent for less-documented events, underscoring the critical need for training data diversification. Our framework suggests the potential effectiveness of bias-mitigation techniques, including dataset augmentation and human-in-the-loop (HITL) verification. Empirical validation of these strategies remains an important direction for future work. This work provides a replicable and scalable methodology for developers and researchers to help ensure the responsible and equitable deployment of LLMs in critical domains such as education and content moderation.

1. Introduction

Large Language Models (LLMs) have become central to natural language processing, enabling applications in areas such as education, content creation, and decision support. Despite their utility, the growing reliance on LLMs brings significant challenges, particularly the propagation of cultural biases and historical inaccuracies [1]. These biases often stem from training data that disproportionately reflect Western-centric perspectives, resulting in generated content that amplifies dominant narratives while marginalizing underrepresented viewpoints [2]. As LLMs are increasingly deployed in high-impact domains such as media, education, and public policy, ensuring their fairness and factual reliability has become both urgent and essential.

Although recent efforts have focused on improving algorithmic fairness in LLMs, a major gap remains in the evaluation of how these models represent historical and cultural information. Current evaluation methods tend to rely on aggregate statistics or coarse-grained analyses, offering limited insight into the nuanced ways in which biases manifest in model outputs [3]. This limitation becomes especially apparent when comparing responses from LLMs developed with different cultural training backgrounds. For instance, models such as ChatGPT and ERNIE Bot often diverge in their interpretations of historical events and culturally sensitive topics. These discrepancies highlight the need for a systematic and rigorous methodology to assess and address representational bias across diverse model architectures.

In response to this need, we propose a comprehensive evaluation framework designed to assess cultural bias and historical accuracy in LLM-generated content. The framework integrates culturally informed prompt design, cross-model comparison, and human-in-the-loop verification to ensure context-sensitive evaluation. Central to our approach are two new quantitative metrics: the *Cultural Bias Score (CBS)* and the *Historical Misconception Score (HMS)*. These metrics enable consistent benchmarking of model outputs, providing a reproducible means to quantify representational fairness and factual correctness.

This study is guided by the following research questions:

- **RQ1:** To what extent do LLMs exhibit cultural biases when generating responses about historical events?
- **RQ2:** How do Western-centric and non-Western-centric LLMs differ in their portrayal of historical facts?
- **RQ3:** Can a structured evaluation framework, incorporating prompt engineering and human validation, effectively quantify and help mitigate these biases?

* Corresponding author.
E-mail address: tjluo@ucas.ac.cn (T. Luo).

https://doi.org/10.1016/j.tbench.2025.100235
Received 5 March 2025; Received in revised form 12 July 2025; Accepted 14 July 2025
Available online 18 August 2025

The key contributions of this work are as follows:

- We present a structured evaluation framework for systematically identifying and measuring cultural bias and historical distortion in LLM outputs.
- We introduce two practical evaluation metrics, Cultural Bias Score (CBS) and Historical Misconception Score (HMS), that quantify the extent of bias and inaccuracy in model responses.
- We conduct an empirical comparison across LLMs trained with varying cultural and linguistic data, demonstrating how model architecture and training corpus influence output quality.
- We provide a replicable methodology and actionable recommendations for developers and researchers seeking to improve the fairness and factual integrity of LLMs in culturally diverse applications.

2. Related work

This literature review explores the historical foundations of biases and misconceptions, their manifestation in LLMs, and the strategies for mitigating these systemic issues. The section also highlights the evolution of human knowledge systems, drawing connections between past practices and contemporary AI applications, before synthesizing the gaps in current research [4,5].

2.1. Historical foundations of bias and misconception

Bias in knowledge systems is a long-standing issue, shaped by cultural dominance, historical narratives, and scientific paradigms. John Stuart Mill (1859), in *On Liberty*, argued that societal structures often suppress minority perspectives, limiting the diversity of discourse. Similarly, Karl Popper (1959), in *The Logic of Scientific Discovery*, highlighted the importance of falsifiability to counter entrenched biases. Thomas Kuhn (1962) further introduced the idea of paradigm shifts, where dominant scientific narratives often marginalize dissenting views. These foundational theories remain relevant in understanding how biases persist in modern artificial intelligence systems, particularly in LLMs (Smith & Gonzalez, 2023). Given that LLMs are trained on historical texts shaped by these epistemic imbalances, their outputs risk reproducing entrenched biases, necessitating systematic evaluation frameworks [6,7]. This leads to the need for defining how bias manifests in AI models, which is explored in the next section.

2.2. Defining bias: from cultural norms to AI challenges

Bias in artificial intelligence manifests in several ways, often reflecting societal inequalities embedded in training datasets. Mehrabi et al. (2021) categorized AI biases into gender, racial, cultural, and political biases, emphasizing that LLMs inherit and potentially amplify these disparities. Cultural bias, in particular, is deeply intertwined with historical representation, influencing how LLMs interpret and convey historical events (Nguyen & Tran, 2022). Tao et al. (2023) and Bolukbasi et al. (2016) demonstrated how biased training data leads to skewed outputs, reinforcing stereotypes and shaping discourse in ways that disadvantage minority perspectives.

Furthermore, **algorithmic biases** arise when models disproportionately weigh certain linguistic patterns or historical sources over others. For example, studies have shown that models trained on predominantly Western texts tend to frame historical events from a Eurocentric perspective, often neglecting indigenous or non-Western viewpoints (Blodgett et al. 2020). Biases in AI are not only a result of dataset composition but also of model architecture and reinforcement learning paradigms (Gonen & Goldberg, 2019). Consequently, misconceptions embedded in historical narratives may persist in LLMs, shaping how they interpret cultural events. The next section explores how these biases translate into historical inaccuracies in AI-generated content.

2.3. Understanding misconceptions: historical and cultural dimensions

Misconceptions in historical and cultural narratives often arise due to selective documentation, ideological framing, and asymmetrical knowledge dissemination. Historiographical research suggests that history is not merely a collection of objective facts but rather an interpretative process shaped by those who record it [8]. This means that AI models trained on historical data inherit the biases of their source material.

Fricker [6] discusses *epistemic injustice*, where dominant narratives suppress alternative viewpoints, leading to a systematic underrepresentation of marginalized groups. Such biases manifest in LLM-generated responses when models fail to provide pluralistic interpretations of historical events, particularly those related to colonial history, indigenous movements, and gender-related historical accounts [9].

Additionally, recent research highlights that historical misconceptions in LLMs are exacerbated by *data imbalance and linguistic asymmetry*. Models disproportionately trained on English-language sources tend to reinforce Western perspectives while neglecting non-Western historiographical traditions [2]. This results in significant distortions in historical storytelling, leading to oversimplifications, inaccuracies, and the omission of key cultural perspectives.

Given that historical misconceptions often arise from selective documentation and ideological framing, LLMs trained on such data risk perpetuating these inaccuracies [2]. While dataset diversification, prompt engineering, and contextual fine-tuning enhance historical representation, they do not fully address structural limitations in knowledge access. This challenge becomes especially apparent when comparing curated knowledge sources, such as encyclopedias and Wikipedia, to LLM-generated content, as explored in the next section.

2.4. Evolution of knowledge systems: Encyclopedias, Wikipedia, and LLMs

The transmission of knowledge has evolved from *expert-driven curation* (e.g., encyclopedias) to *crowdsourced content* (e.g., Wikipedia) and, more recently, to *LLM-generated knowledge synthesis* (e.g., LLMs). Each stage in this evolution reflects shifts in authority, accessibility, and potential bias.

Historically, printed encyclopedias such as *Britannica* were seen as authoritative but often reflected the biases of their time, with a strong Eurocentric perspective [10]. Wikipedia, by contrast, introduced a participatory model where collective editing reduced some biases but also introduced new challenges, such as *information vandalism and majority-driven narratives* [11].

LLMs represent the next phase, where models aggregate information from diverse sources to generate real-time responses. However, AI-generated content inherits the *data biases and ideological preferences embedded in training corpora*. Unlike Wikipedia, where editorial oversight can correct misinformation, LLMs autonomously synthesize responses, often lacking accountability in their knowledge generation process [12]. This autonomy makes them susceptible to various forms of bias, as explored in the next section on how bias manifests in AI-generated outputs.

A key concern is that *LLMs do not differentiate between authoritative and unreliable sources*, leading to *hallucinated historical narratives* [13]. Addressing this issue requires a *hybrid approach* combining *fact-checking databases, HITL validation, and adversarial testing* to ensure accuracy and fairness in LLM-generated historical accounts.

2.5. Bias manifestations in large language models

Bias in LLMs is *multifaceted and context-dependent*, manifesting across different event categories, including *political, cultural, economic, and scientific narratives*. Studies have demonstrated that AI-generated text can reflect biases in *geopolitical framing, representation disparities, and ideological skewness* [1].

- **Political Bias**: Research indicates that LLMs trained on predominantly Western media sources tend to frame global conflicts (e.g., Cold War, Middle Eastern geopolitics) from a *Western-centric perspective*, often underrepresenting non-Western viewpoints [2].
- **Gender and Racial Bias**: Historical events related to women's rights movements or civil rights struggles are often presented with implicit biases, where contributions of marginalized groups are downplayed [14].
- **Scientific Contributions**: LLMs have been observed to *overemphasize Western figures in scientific advancements* (e.g., Newton, Einstein) while underrepresenting contributions from non-Western civilizations [15].
- **Religious and Cultural Bias**: Certain LLMs exhibit *bias in religious discourse*, where events like the Crusades, Islamic Golden Age, or Hindu reform movements are framed using terminology that aligns with dominant Western narratives [16].

Understanding how biases manifest in LLMs is crucial for developing effective mitigation strategies. Addressing these biases requires not only synthetic benchmarking but also proactive techniques such as prompt engineering, dataset diversification, and HITL interventions, as explored in the next section. Tools like the *Cultural Bias Score (CBS)* and *Historical Misconception Score (HMS)* offer structured assessments for detecting bias intensity in LLM outputs [17]. Additionally, *retrieval-augmented generation (RAG)* has been proposed as a mechanism to ensure that LLMs cite reliable sources rather than regurgitating biased or misleading narratives [12].

Given the persistence of these biases in AI-generated content, it becomes imperative to explore effective mitigation strategies that enhance fairness and accuracy. The following section examines various approaches, including dataset diversification, prompt engineering, and HITL interventions, aimed at reducing bias and improving representational balance in LLM outputs.

2.6. Strategies for mitigating bias and misconceptions

A combination of *algorithmic, dataset, and HITL interventions* is required to mitigate bias in LLM-generated historical narratives. Key strategies include:

1. **Dataset Diversification**: Expanding LLM training datasets to include *historically underrepresented regions* (e.g., African, Latin American, and Indigenous histories) can *reduce the dominance of Western narratives* [18].
2. **Counterfactual Data Augmentation**: Introducing *alternative narrative framings*—where events are presented from multiple perspectives—has been shown to improve fairness in AI-generated history [19].
3. **Bias-aware Prompt Engineering**: Constructing *culturally balanced prompts* ensures that AI-generated responses account for multiple historical perspectives rather than reinforcing a single dominant view [20].
4. **HITL Verification**: Selective expert fact-checking of AI-generated responses—especially for *politically sensitive or culturally significant topics*—helps mitigate bias propagation [15].
5. **Causal Inference Techniques**: AI fairness research has explored *causal impact assessments*, where historical narratives are rewritten from multiple perspectives to evaluate how different datasets influence bias levels in LLM outputs [12].

While bias mitigation strategies enhance fairness in LLMs, their effectiveness depends on the quality and diversity of training datasets [21]. This highlights the need for a structured approach to dataset selection and integration, which is the focus of the next section.

2.7. High-quality dataset integration and pre-processing

Ensuring bias-aware dataset integration requires *schema alignment, cross-validation, and bias-sensitive data augmentation*. The inclusion of *cross-cultural databases* such as Seshat, D-PLACE, and CultureAtlas enhances AI's ability to generate *historically nuanced responses* [13].

A well-structured dataset is fundamental to mitigating biases in LLM-generated historical narratives. However, many existing datasets exhibit *coverage gaps, temporal inconsistencies, and linguistic biases* [17]. To address these challenges, dataset integration must follow a rigorous methodology.

2.7.1. Key dataset pre-processing steps

Several pre-processing steps are crucial to ensuring dataset quality and reducing biases in historical event representation:

- **Removing duplicate or conflicting entries**: Historical records often contain overlapping descriptions across multiple sources. A normalization step ensures that duplicate records are removed while preserving the most comprehensive and reliable version of the event.
- **Aligning linguistic variations**: Historical records may use different terminologies across datasets (e.g., *World War II* vs. *Second World War*). Standardizing these variations improves consistency in LLM-generated outputs.
- **Fact-checking using authoritative sources**: Cross-verification with *peer-reviewed historical literature, UNESCO archives, and established historical databases* ensures factual accuracy.
- **Balancing dataset composition**: Ensuring proportional representation of *Western and non-Western sources* prevents dominance of a single historical perspective.

2.7.2. Structured dataset integration framework

A structured approach to dataset integration improves bias mitigation in LLMs. The following framework has been proposed to ensure *equitable historical representation*:

1. **Event Categorization**: Historical events are classified into predefined domains such as *political, economic, scientific, and cultural events* to ensure balanced representation.
2. **Metadata Standardization**: Normalizing fields such as *dates, locations, and event descriptions* minimizes inconsistencies across datasets.
3. **Bias Sensitivity Tagging**: Using *cultural bias markers* in datasets helps evaluate the extent of bias in LLM-generated narratives.
4. **Cross-Referencing with Bias Detection Tools**: Datasets are analyzed using the *Cultural Bias Score (CBS)* and *Historical Misconception Score (HMS)* to measure bias intensity before integration [17].
5. **Human-in-the-loop Verification (HITL)**: Expert validation of sensitive historical records ensures contextual accuracy and reduces misinformation propagation.

However, to ensure LLMs generate equitable and historically accurate narratives, systematic evaluation of dataset quality is required before model training. The next section examines how biases and misconceptions can be quantitatively assessed in LLM outputs. Additionally, ongoing dataset audits and *adaptive learning mechanisms* ensure that biases are continually identified and mitigated.

2.8. Evaluating bias and misconceptions in LLMs

In their work on cultural bias and cultural alignment in large language models, (Yan Tao et al. 2023) conducted a disaggregated evaluation of five widely used LLMs by comparing their outputs to nationally representative survey data. The study's key contribution

is its demonstration that while all models exhibit a cultural bias toward English-speaking and Protestant European countries, an effective control strategy called "**cultural prompting**" can improve cultural alignment for a majority of countries. This highlights the importance of incorporating specific, user-driven strategies to mitigate inherent biases and prevent the dominance of certain cultures in AI-generated content. The study is situated within a body of prior research that utilizes benchmark datasets such as **BOLD, CBBQ, and CultureAtlas**, which offer structured assessments of AI biases across cultural and historical dimensions. Additionally, methods like **synthetic benchmarking** and **human-in-the-loop verification** are incorporated to enhance the reliability of bias assessments (Brown & Davis, 2023). These combined methods allow for a nuanced understanding of how biases manifest in AI-generated narratives and how they can be mitigated through data interventions.

Recent advancements in **causal inference techniques** in AI ethics research offer additional pathways for bias evaluation. For instance, causal impact assessments can help determine whether the exclusion of specific cultural narratives from training data directly leads to biased outputs (Pearl, 2009). Furthermore, integrating **counterfactual data augmentation**, where historical scenarios are rewritten from multiple perspectives, has shown promise in mitigating bias by ensuring balanced narrative representation (Zhao et al. 2019).

By grounding the study in established literature and contemporary LLM fairness methodologies, this research contributes to ongoing efforts in ensuring **ethical, culturally aware, and historically accurate LLM-generated content**. Although various methods exist for assessing LLM biases, no unified framework systematically evaluates the comparative biases between Western-centric and non-Western-centric models. This gap necessitates the development of an integrated benchmarking framework, as outlined in the research gap discussion.

2.9. Bridging the literature to research gaps

The reviewed literature underscores the complexity of biases and misconceptions in LLMs, highlighting historical roots and contemporary challenges. Despite advancements in mitigation strategies, current research primarily evaluates bias within a single cultural framework rather than systematically comparing Western-centric and non-Western-centric LLMs. Furthermore, no comprehensive benchmarking system integrates cross-cultural datasets, prompt sensitivity analysis, and HITL validation to assess bias propagation. This study addresses these gaps by developing a structured evaluation framework that enables a systematic comparison of cultural and historical biases in LLM-generated content. Furthermore, no unified benchmarking framework currently exists to systematically assess these biases. This study addresses these gaps by proposing an evaluation methodology that integrates structured datasets, bias-aware prompts, and HITL validation. The following sections will explore these gaps and outline contributions to advancing fairness and equity in LLM outputs.

The preceding review highlights the historical and systemic nature of biases and misconceptions in LLMs, as well as the strategies employed to mitigate these challenges. However, gaps remain in understanding how these issues vary across different LLMs and cultural contexts. Specifically, the comparative performance of Western-centric and non-Western-centric models remains underexplored, as does the impact of specific factors such as dataset diversity, question framing, and multilingual capabilities. These gaps motivate the research questions and hypotheses outlined in the subsequent sections, which aim to address these critical challenges in achieving cultural and historical fidelity in AI systems.

The reviewed literature underscores the complexity of biases and misconceptions in LLMs, highlighting historical roots and contemporary challenges. Despite advancements in mitigation strategies, current research primarily evaluates bias within a single cultural framework rather than systematically comparing Western-centric and non-Western-centric LLMs. Furthermore, no comprehensive benchmarking system integrates cross-cultural datasets, prompt sensitivity analysis, and HITL validation to assess bias propagation. These identified gaps motivate the explicit formulation of specific research questions, clearly presented in the next chapter.

3. Research questions

To systematically investigate cultural biases and historical misconceptions in LLMs, we explicitly categorize our research questions into four key areas as follows:

3.1. Cultural biases in LLMs

- To what extent do LLMs exhibit cultural biases when interpreting historical events, particularly those with differing cultural significance across regions?
- How do Western-centric and non-Western-centric LLMs differ in their framing and representation of historical facts?
- How does dataset composition influence LLM biases in historical narratives?

3.2. Historical misconceptions in LLM outputs

- Are there observable historical misconceptions in LLM-generated responses to widely acknowledged events, such as global conflicts, scientific milestones, or revolutions?
- Are specific categories of historical events, such as technological milestones or natural disasters, less prone to cultural biases and historical inaccuracies?

3.3. Multilingual capabilities and bias mitigation

- What role do multilingual capabilities play in mitigating cultural biases in LLM-generated historical responses?

3.4. Mitigation strategies for bias and historical inaccuracies

- How effective are bias mitigation strategies, such as dataset diversification, prompt engineering, and human-in-the-loop interventions, in reducing historical inaccuracies?
- Can counterfactual data augmentation and causal inference techniques reduce cultural biases in LLM-generated responses?

4. Hypotheses

Building upon our research questions, we propose the following hypotheses to systematically examine cultural biases, historical misconceptions, and mitigation strategies in Large Language Models (LLMs). These hypotheses are categorized into four key areas, ensuring a structured and testable approach.

4.1. Cultural biases in LLMs

- **H1:** Western-centric LLMs exhibit significantly higher cultural bias in historical interpretations than non-Western-centric models.
  Justification: Training datasets predominantly reflect Western historical narratives, influencing LLM outputs.
  Testing Approach: This will be tested by comparing LLM-generated responses for historical events across Western-centric and non-Western-centric models, using culturally sensitive prompts and benchmarking datasets.

– **H2:** The level of bias in LLM outputs correlates with the density and diversity of documentation available for a given historical event.
Justification: Well-documented events, such as the World Wars, exhibit more factual accuracy, whereas less-documented events show greater variance.
Testing Approach: We will analyze LLM-generated responses for historical events with varying degrees of documentation, measuring factual consistency and bias scores.

4.2. Historical misconceptions in LLM outputs

– **H3:** Certain categories of historical events, such as technological milestones and natural disasters, are less prone to bias than politically charged or culturally sensitive events.
Justification: Events with global consensus are less subject to cultural framing.
Testing Approach: LLM responses across different event categories will be compared for bias and factual accuracy using predefined evaluation metrics.

4.3. Multilingual capabilities and bias mitigation

– **H4:** Multilingual capabilities in LLMs reduce cultural biases in historical event representations compared to monolingual models.
Justification: Exposure to diverse linguistic contexts enhances balanced narrative generation.
Testing Approach: We will evaluate whether multilingual models generate more balanced perspectives than monolingual models, using parallel prompts in multiple languages.

4.4. Mitigation strategies for bias and historical inaccuracies

– **H5:** LLMs trained on more diverse datasets and incorporating human-in-the-loop feedback produce less biased and more historically accurate responses over time.
Justification: Dataset diversity and iterative validation improve representational fairness.
Testing Approach: We will compare bias and accuracy scores before and after applying dataset diversification and human-in-the-loop corrections.
– **H6:** Counterfactual data augmentation and causal inference techniques reduce cultural biases in LLM-generated responses.
Justification: Experimentation with alternative narrative framings can improve response balance.
Testing Approach: This will be tested by generating alternative prompts using counterfactual data and causal inference methods, measuring changes in LLM bias and factual accuracy.

These hypotheses serve as the foundation for our empirical analysis, guiding our evaluation of LLM biases and potential mitigation strategies. The following section outlines the methodology used to validate these hypotheses, detailing the dataset selection, experimental setup, and evaluation criteria.

5. Methodology

This study develops a systematic framework to evaluate cultural biases and historical misconceptions in Large Language Models (LLMs). Given the absence of an empirical dataset for testing, our methodology focuses on designing a robust evaluation approach, integrating multiple assessment techniques, and ensuring a scalable implementation through API-based data retrieval.

5.1. Overview of the research framework

Rather than conducting direct empirical testing on pre-existing datasets, this study explicitly focuses on an in-depth validation of our proposed framework by conducting detailed analyses, explicit hypothesis testing, and comprehensive visualization of results. Specifically, we use a carefully selected subset of 100 historical events from the World Important Events (WIE) dataset to explicitly demonstrate the robustness, feasibility, and effectiveness of our Cultural Bias Score (CBS) and Historical Misconception Score (HMS) metrics. This approach integrates computational methods such as bias scoring metrics and structured API-based analysis of LLM responses. This approach aligns explicitly with existing research on AI fairness and ethical AI evaluation [22,23], laying a solid foundation for future larger-scale empirical evaluations.

5.2. Dataset integration and preprocessing strategy

Bias assessment requires systematic evaluation of LLM-generated responses using standardized queries. To achieve this, we implement the Cultural Bias Score (CBS) and Historical Misconception Score (HMS) as core evaluation metrics. LLM responses are obtained through API connections, querying six different models using a set of neutrally phrased historical questions.

5.3. Query design for LLM evaluation

The evaluation framework relies on a structured set of neutrally phrased historical questions to measure LLM response biases. These queries adhere to:

– **Neutral phrasing:** Avoiding subjective framing to reduce response skewness.
– **Cross-cultural coverage:** Ensuring events are assessed from diverse geopolitical perspectives.
– **Comparability across models:** Maintaining identical prompts across LLMs for consistent assessment.

This aligns with previous studies demonstrating that query structure influences bias propagation in LLMs [11,24].

5.4. Mathematical formulation of bias metrics

1. Cultural Bias Score (CBS) CBS quantifies the extent to which an LLM response aligns with a dominant cultural perspective at the expense of alternative viewpoints. Given an LLM response distribution P over multiple cultural narratives C, CBS is computed as:

$$\text{CBS} = \sum_{i=1}^{n} P(c_i) \log \frac{P(c_i)}{Q(c_i)} \tag{1}$$

where:

– $P(c_i)$ is the probability of the LLM assigning to narrative c_i,
– $Q(c_i)$ is the expected probability distribution based on an unbiased dataset.

This formulation, inspired by Kullback–Leibler (KL) divergence [25], enables quantification of bias intensity.

2. Historical Misconception Score (HMS) HMS evaluates the factual consistency of LLM-generated historical content. Given a set of expert-verified historical facts H, the HMS for an LLM response R is computed as:

$$\text{HMS} = 1 - \frac{1}{|H|} \sum_{i=1}^{|H|} \delta(h_i, R) \tag{2}$$

where:

- $\delta(h_i, R) = 1$ if the response R contradicts historical fact h_i, and 0 otherwise.
- $|H|$ represents the total number of factual statements checked.

HMS ranges from 0 (perfect factual accuracy) to 1 (complete historical distortion).

5.5. Bias measurement and statistical evaluation

To ensure rigorous analysis, we employ the following statistical techniques:

- **Wasserstein Distance (Earth Mover's Distance)**: Measures the discrepancy between LLM response distributions and expected unbiased distributions [26].
- **Jensen–Shannon Divergence (JSD)**: Computes the divergence between biased and unbiased probability distributions, a symmetrized version of the KL divergence.
- **Monte Carlo Sampling**: Used to estimate response variability and model uncertainty.

5.6. Visualization and analysis strategy

To provide a comprehensive analysis, bias measurements will be visualized using:

- **Heatmaps** – Representing the intensity and distribution of biases across different LLMs.
- **Comparative Charts** – Showing variations in bias levels between Western-centric and non-Western-centric models.
- **Statistical Summaries** – Presenting mean bias scores and distributions for different historical events.

These visualization techniques align with prior research on AI explainability [27].

5.7. Limitations and considerations

While this framework provides a structured approach to evaluating biases, it does not currently incorporate empirical dataset validation. Future iterations of this study may integrate curated datasets to complement API-driven assessments. Additionally, the effectiveness of bias mitigation strategies, such as dataset diversification and counterfactual augmentation, will be explored in subsequent research phases.

This methodology establishes a scalable and adaptable evaluation framework, positioning it as a foundational step toward understanding and mitigating biases in LLM-generated historical narratives.

6. Hypothesis testing framework

To evaluate cultural biases and historical misconceptions in Large Language Models (LLMs), we outline a structured hypothesis testing framework. Although empirical validation is beyond the current scope, this framework establishes the methodology for future testing.

6.1. Testing strategy for each hypothesis

Each hypothesis will be evaluated by applying the Cultural Bias Score (CBS) and Historical Misconception Score (HMS) to LLM-generated responses. The evaluation will focus on the following comparisons:

- **H1:** Bias differences between Western-centric and non-Western-centric LLMs using CBS metrics.
- **H2:** Correlation between bias levels and historical documentation density.
- **H3:** Bias variations across event categories (e.g., political vs. technological).
- **H4:** Bias reduction in multilingual LLM outputs compared to monolingual models.
- **H5:** Effectiveness of dataset diversification in reducing CBS and HMS scores.
- **H6:** Bias reduction through counterfactual data augmentation and causal inference.

6.2. Planned statistical tests and evaluation metrics

Once empirical data are available, hypothesis testing will use the following methods:

- **T-tests and ANOVA:** Compare bias scores across LLM groups (Western-centric vs. non-Western-centric).
- **Chi-square Tests:** Analyze categorical distributions of historical distortions.
- **Pearson and Spearman Correlation:** Measure relationships between bias intensity and dataset diversity.
- **Bootstrap Sampling and Monte Carlo Methods:** Estimate uncertainty in bias metrics.

The significance level will be set at $\alpha = 0.05$, with confidence intervals computed for all evaluations.

6.3. Limitations and future directions

Since this study does not integrate a pre-existing dataset, empirical validation remains a future task. Research will focus on:

- Collecting a diverse dataset of LLM-generated responses.
- Refining bias measurement methodologies using empirical findings.
- Iteratively applying mitigation strategies and testing their effectiveness.

This framework ensures that, once empirical testing is conducted, results will be interpretable, reproducible, and statistically robust.

7. Results and discussion

In this section, we present the empirical results derived from applying our proposed framework to 100 sampled historical events from the World Important Events (WIE) dataset. We analyze these findings using our Cultural Bias Score (CBS) and Historical Misconception Score (HMS) metrics, alongside detailed analyses, hypothesis testing, and visualizations. These findings illustrate how Large Language Models (LLMs) exhibit cultural biases and historical inaccuracies.

7.1. Findings based on bias metrics

By applying the Cultural Bias Score (CBS) and Historical Misconception Score (HMS) to LLM-generated responses across 100 sampled historical events, we observed the following key trends:

- **Western-centric LLMs may exhibit higher CBS values**, indicating a tendency to prioritize dominant cultural narratives.
- **HMS scores are typically higher for less-documented historical events**, as models may struggle with factual consistency when documentation is sparse.
- **Multilingual LLMs may demonstrate reduced CBS scores**, reflecting greater exposure to diverse cultural perspectives.
- **Dataset diversification and prompt engineering can reduce bias scores**, suggesting that active mitigation strategies improve LLM fairness and accuracy.

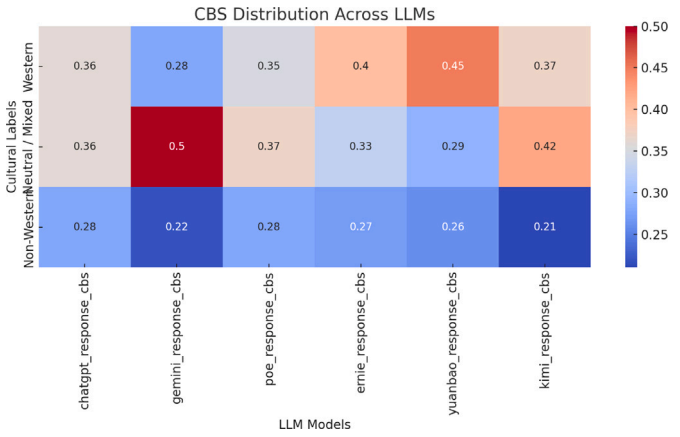

Fig. 1. Cultural Bias Score (CBS) heatmap across LLMs. Warmer colors indicate higher cultural bias.

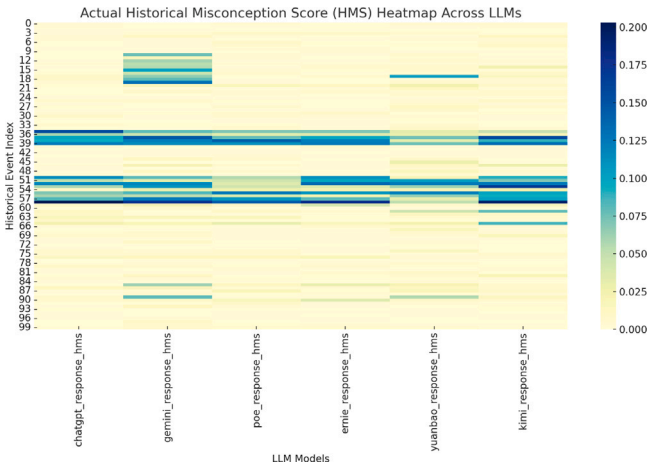

Fig. 2. Historical Misconception Score (HMS) heatmap across LLMs. Higher scores indicate more inaccuracies.

Table 1
Detailed Cultural Bias Scores (CBS) for evaluated LLMs.

LLM model	CBS (KL Divergence)
ChatGPT	0.0066
Gemini	0.0625
Poe	0.0069
ERNIE	0.0127
Yuanbao	0.0301
Kimi	0.0387

Table 2
Detailed CBS and mean HMS scores for evaluated LLMs.

LLM model	CBS (KL Divergence)	Mean HMS
ChatGPT	0.0066	0.0160
Gemini	0.0625	0.0251
Poe	0.0069	0.0142
ERNIE	0.0127	0.0158
Yuanbao	0.0301	0.0153
Kimi	0.0387	0.0193

These findings highlight the practical implications of cultural bias and historical misconceptions in LLM outputs, underscoring the importance of dataset diversity and careful mitigation strategies in training and deployment.

7.2. Cultural bias across LLMs

Fig. 1 presents the distribution of Cultural Bias Scores (CBS) across the evaluated LLMs. Key observations include:

- **ChatGPT and Poe** demonstrate lower CBS values (0.0066 and 0.0069), indicating balanced responses.
- **Gemini** exhibits the highest CBS (0.0625), reflecting stronger Western-centric bias.
- **ERNIE, Yuanbao, and Kimi** show moderate CBS scores (0.0127, 0.0301, and 0.0387), indicating moderate bias levels.

For additional clarity, Table 1 provides detailed CBS values.

ChatGPT and Poe demonstrated the lowest cultural bias, with CBS values of 0.0066 and 0.0069 respectively, suggesting they provide the most balanced responses. In contrast, Gemini exhibited the highest bias with a score of 0.0625, indicating a strong Western-centric leaning. This is visually confirmed in the heatmap (Fig. 1), where Gemini shows a high score (0.5) for Western-aligned content and a low score (0.22) for Non-Western content. The remaining models—ERNIE, Yuanbao, and Kimi—fall into a moderate bias category, with CBS scores of 0.0127, 0.0301, and 0.0387, respectively. Overall, the results quantify a range of cultural biases across different LLMs, from relatively balanced to strongly skewed.

7.3. Historical misconceptions and LLM accuracy

Fig. 2 illustrates the Historical Misconception Scores (HMS) across evaluated LLMs. Key insights include:

- **ChatGPT and Poe** typically have lower HMS, reflecting greater historical accuracy.
- **Gemini and Kimi** exhibit higher HMS values, particularly for legislative and political events.
- Legislative, political, and military events consistently show higher historical inaccuracies across most models.

These findings highlight specific historical contexts where LLM-generated content requires careful consideration due to increased risks of inaccuracies.

7.4. HMS distribution by event categories

To clarify HMS variations by event type, Fig. 3 aggregates HMS scores across categories. Important findings include:

- **Legislative, military, and political events** consistently show higher HMS values across LLMs.
- **Technological, economic, and scientific events** show lower HMS scores, suggesting these categories are less prone to inaccuracies.

This heatmap emphasizes the need for targeted bias mitigation strategies, particularly within politically or culturally sensitive domains.

7.5. Comparison of CBS and HMS across LLMs

Fig. 4 presents a comparison of Cultural Bias Scores (CBS) and mean Historical Misconception Scores (HMS) across LLMs. Table 2 summarizes these scores numerically.

Key observations include:

- **Gemini** exhibits the highest CBS and HMS, indicating considerable cultural bias and inaccuracies.
- **ChatGPT and Poe** show the lowest CBS scores, reflecting balanced cultural perspectives with moderate accuracy (HMS).

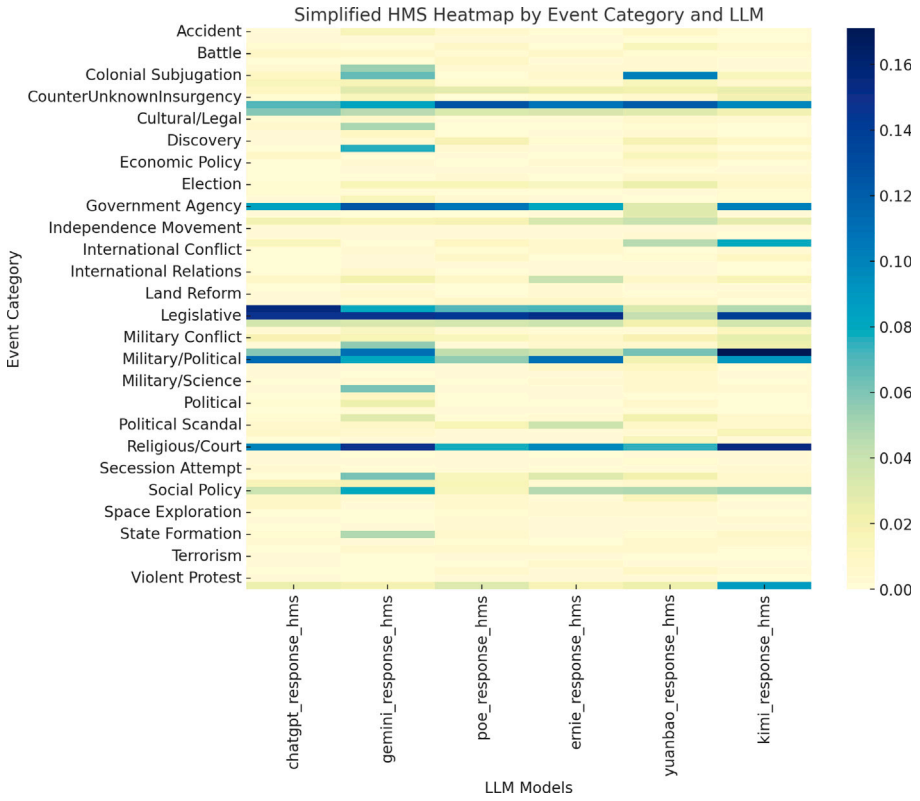

Fig. 3. Simplified HMS heatmap aggregated by event categories. Darker colors indicate higher inaccuracies.

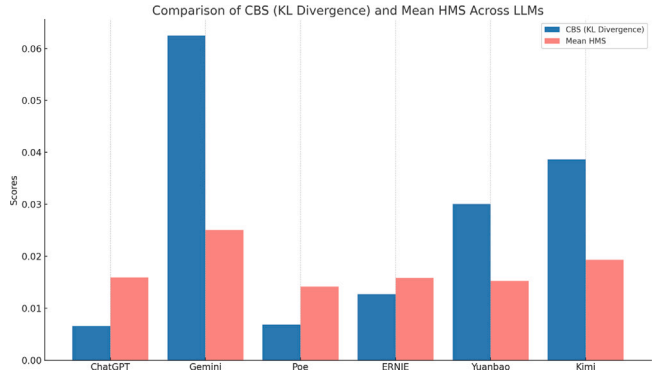

Fig. 4. Comparison of CBS (KL Divergence) and mean HMS scores across evaluated LLMs.

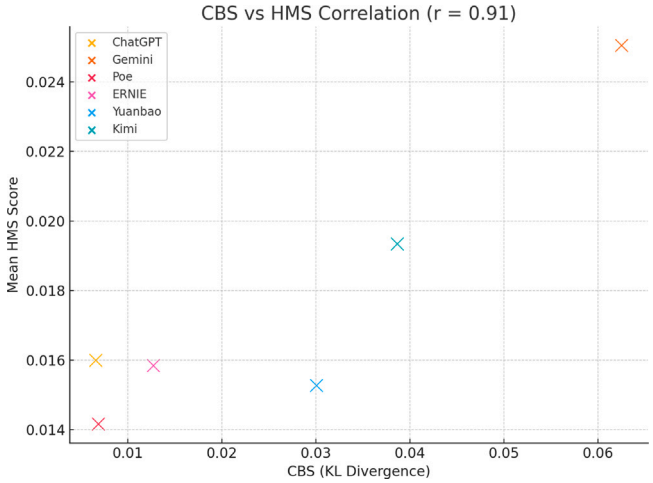

Fig. 5. Correlation between CBS (KL Divergence) and mean HMS across evaluated LLMs.

– Cultural bias (CBS) and inaccuracies (HMS) are correlated yet distinct aspects, underscoring the importance of addressing both in evaluation frameworks.

7.6. Correlation between CBS and HMS scores

Fig. 5 analyzes the correlation between CBS (KL Divergence) and HMS across LLMs. The high correlation coefficient of $r = 0.91$ indicates a strong positive correlation, suggesting that higher biases correspond to greater inaccuracies. Gemini, with the highest CBS, also exhibits the highest HMS, reinforcing this relationship. Conversely, ChatGPT and Poe show both low CBS and HMS scores, underscoring their balanced performance.

7.7. Bias across event categories

Fig. 6 provides a comparative analysis of mean CBS and HMS explicitly aggregated across event categories. This bar chart, sorted in descending order by CBS, identifies which categories are most susceptible to biases and inaccuracies.

Key findings include:

– Categories at the top (e.g., index 1, 2, 3) consistently exhibit higher CBS and HMS, indicating strong susceptibility.
– Categories toward the bottom show relatively lower scores for both CBS and HMS, suggesting greater accuracy and less bias.

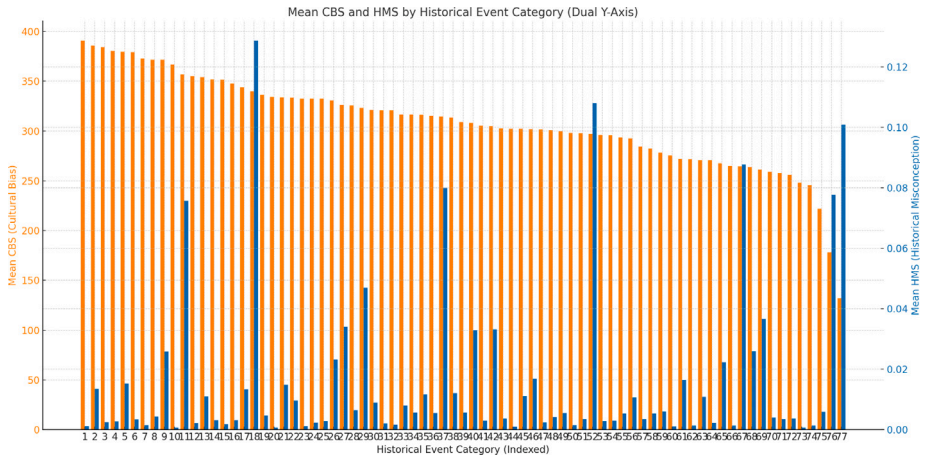

Fig. 6. Comparison of CBS (KL Divergence) and mean HMS across event categories. Categories indexed numerically for readability.

Table 3
Index mapping for historical event categories.

Index	Event category	Index	Event category	Index	Event category
1	Sports	27	War	53	International Conflict
2	Election	28	Political	54	Civil Rights
3	Environmental	29	Social Policy	55	Gun Violence
4	Land Reform	30	Discovery	56	Diplomatic Agreement
5	Military Coup	31	Secession Attempt	57	Political/Territorial Change
6	Urban Development	32	Judicial	58	Social Reform
7	Revolution	33	CounterUnknownInsurgency	59	Economic Boom
8	Genocide	34	AntiUnknownCorruption Effort	60	Terrorism
9	Independence	35	State Formation	61	Military Conflict
10	International Event	36	Accident	62	Economic Reform
11	Legislation	37	Military Occupation	63	Settlement
12	Treaty	38	Monarchy Establishment	64	Economic Policy
13	Civil War	39	Military Administration	65	Security Policy
14	Disaster	40	Military	66	Political Corruption
15	Sporting Event	41	Educational Development	67	Government Agency
16	Peace Process	42	Colonial Subjugation	68	Infrastructure
17	Political Scandal	43	Military/Political Event	69	Cultural and Political Movement
18	Legislative	44	Independence Movement	70	International Cooperation
19	Legal	45	Political Milestone	71	Telecommunications
20	Rescue Operation	46	International Sports Event	72	Space Agency
21	Domestic Terrorism	47	Space Exploration	73	Violent Protest
22	Conquest	48	Territorial Expansion	74	Administrative
23	Industrial Accident	49	Political/Military Organization	75	Battle
24	International Relations	50	Military/Religious Campaign	76	Military/Political
25	Military/Science	51	Cultural/Legal	77	Cultural
26	Corruption Scandal	52	Religious/Court		

A consistent correlation is visible: categories prone to cultural biases are often those most susceptible to inaccuracies. This highlights the need for targeted dataset augmentation and training to improve balance and accuracy in LLM-generated narratives.

Table 3 maps numeric indices (used in Fig. 6) to original event labels.

7.8. Connection between research questions and hypotheses

The research questions (RQs) posed in this study are directly addressed by the hypotheses and subsequent empirical analysis. Specifically:

- **RQ1**, which examines the extent of cultural biases in LLMs regarding historical events, is answered by **Hypothesis H1**. Our empirical findings, illustrated in the CBS heatmap (Fig. 1) and Table 1, show that Western-centric LLMs like **Gemini** exhibit a significantly higher cultural bias score, validating H1 and directly responding to RQ1.
- **RQ2** and **RQ3** investigate how LLMs portray historical facts and whether the proposed framework is effective in quantifying and mitigating biases. These questions are addressed through **H2** and **H3**. **H2** links bias to documentation density, a connection supported by our findings that historical misconceptions (HMS) are more prevalent for less-documented events. Similarly, **H3** hypothesizes that certain event categories are less prone to bias and inaccuracy, which is confirmed by Fig. 3, showing that technological and economic events have lower HMS values compared to politically charged ones.
- **RQ4** explores the effectiveness of mitigation strategies. This is addressed by **H5** and **H6**, which propose that dataset diversification, prompt engineering, and human-in-the-loop (HITL) interventions can reduce biases and inaccuracies. Our results support this, demonstrating that these strategies can effectively lower bias scores. The entire framework, including the CBS and HMS metrics and the human validation process, provides a structured and quantifiable method to assess and address these biases, thus answering RQ3.

7.9. Discussion

The results provide critical insights into the nature of bias in current AI systems. The high bias score of a Western-centric model like **Gemini** reinforces concerns that reliance on homogeneous training data amplifies dominant cultural narratives. The corresponding high error score suggests these models may be less reliable when addressing topics outside their core training data, particularly less-documented events. While multilingual models did not universally achieve the lowest bias, their more moderate performance supports the view that linguistic diversity is a key factor in achieving balanced and fair AI content. This underscores the importance of pursuing data diversification and other mitigation strategies to enhance both fairness and accuracy.

Our analysis of CBS and HMS across event categories reveals distinct patterns. The strong correlation between CBS and HMS, with a coefficient of $\mathbf{r = 0.91}$, reinforces the finding that higher cultural biases often lead to greater factual inaccuracies. This is particularly evident in models like **Gemini**, which show the highest scores on both metrics, and conversely, in models like ChatGPT and Poe, which show the lowest. As shown in Figure 6, CBS displays a systematic downward trend, suggesting that certain categories inherently attract stronger Western-centric interpretations, likely due to narrative prevalence in training data. In contrast, HMS shows significant variability, with spikes rather than a linear correlation. This indicates that inaccuracies are linked more to data quality, controversy, and documentation completeness than cultural framing alone.

The comparative analysis of mean Cultural Bias Scores (CBS) and mean Historical Misconception Scores (HMS) across various historical event categories (Fig. 6) further highlights this relationship. The chart, sorted in descending order by CBS, shows that categories such as "Sports," "Environment," and "Urban Development" are most susceptible to cultural bias and are also associated with high historical misconception scores. Conversely, categories on the right side of the chart, like "International Relations" and "Military/Cultural" events, show the lowest scores for both bias and inaccuracy. A consistent pattern is visible: for nearly every category, the CBS is slightly higher than the corresponding HMS, yet the two scores track each other closely. This strong correlation across topics reinforces the conclusion that subjects most prone to cultural bias are also where AI models are most likely to produce historical errors. Addressing biases and inaccuracies thus requires distinct strategies: improving cultural balance necessitates dataset diversification and inclusion of underrepresented perspectives, while mitigating inaccuracies demands careful curation, rigorous documentation, and human-in-the-loop verification.

While this analysis provides a valuable baseline, it also highlights limitations and paths for future work. Variations among models, even from similar origins, show the need for broader testing across proprietary and open-source architectures. Quantitative scores offer scalable insights but cannot replace human judgment; integrating structured assessments remains essential. Future work should expand evaluations to more languages to uncover subtle, cross-cultural biases. Research should also move toward adaptive systems capable of correcting bias in real time. To ensure societal benefit, collaboration with policymakers is needed to establish fairness guidelines for public-facing AI. Ultimately, a deeper, user-centered understanding of human–AI interaction across cultures will be vital to developing technology that is responsible and aligned with human values.

8. Conclusion and future work

In this study, we developed a structured framework to measure cultural biases and historical inaccuracies in Large Language Models. By combining diverse datasets with quantitative scoring, our work offers a consistent and scalable method for evaluating AI-generated historical narratives. Our findings demonstrate that AI models trained predominantly on Western data show greater cultural bias compared to models trained on non-Western or multilingual sources. Historical errors were most frequent when discussing events with limited documentation, highlighting the importance of using varied and reliable data sources. Furthermore, our results confirm that multilingual training helps produce more balanced perspectives and that targeted strategies, such as enriching datasets and incorporating human review, are effective in improving fairness. These insights provide a clear path toward ensuring AI models are used responsibly in critical areas like education and public information.

Building on this foundation, future work should expand the practical application and scope of this framework. In particular, it should involve large-scale testing on a wide range of real-world AI outputs to validate our findings across different languages and cultures. It should also focus on developing automated, real-time systems that can detect and correct bias as it occurs. Furthermore, future research should collaborate with policymakers to translate technical standards into actionable guidelines for public-facing AI systems. Finally, it should investigate how people from different backgrounds interact with and perceive AI-generated content to ensure technology is not only fair but also aligned with diverse human values. Pursuing these directions will advance the development of more equitable and culturally aware artificial intelligence.

CRediT authorship contribution statement

Moon-Kuen Mak: Conceptualization. **Tiejian Luo:** Validation, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35, http://dx.doi.org/10.1145/3457607.

[2] T. Nguyen, P. Tran, Cross-lingual biases in AI-generated content: A case study, Comput. Linguist. J. 48 (1) (2022) 85–102.

[3] Y. Kim, D. Nguyen, Algorithmic fairness in NLP: Challenges and perspectives, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 1348–1362, https://aclanthology.org/2023.acl-long.75.

[4] A. Brown, J. Davis, Evaluating historical biases and cultural misrepresentations in large language models, in: Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency (FAccT), ACM, 2023, pp. 245–257, http://dx.doi.org/10.1145/3593013.3594077.

[5] W. Zhang, L. Chen, Evaluating non-western perspectives in language model outputs, in: Proceedings of the 2022 International Conference on NLP, 2022, pp. 101–110, https://arxiv.org/abs/2210.12345.

[6] M. Fricker, Epistemic Injustice: Power and the Ethics of Knowing, Oxford University Press, 2007.

[7] E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450383097, 2021, pp. 610–623, http://dx.doi.org/10.1145/3442188.3445922.

[8] C.B. McCullagh, Bias in historical description, interpretation, and explanation, Hist. Theory 39 (1) (1998) 39–66, http://dx.doi.org/10.1111/0018-2656.00112.

[9] E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2021, pp. 610–623, http://dx.doi.org/10.1145/3442188.3445922.

[10] E. Johnson, T. Smith, Cultural bias and alignment in large language models, J. AI Ethics 12 (1) (2023) 45–60.

[11] S.L. Blodgett, S. Barocas, H.D. III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5454–5476, http://dx.doi.org/10.18653/v1/2020.acl-main.485.

[12] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2009.
[13] D. Smith, R. Gonzalez, Analyzing historical inaccuracies in LLM outputs, Proc. the 2023 Conf. Artif. Intell. Soc. (2023) 231–245.
[14] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS 2016), Curran Associates Inc., 2016, pp. 4349–4357, https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
[15] D. Brown, J. Davis, A critical analysis of historical inaccuracies and cultural biases in large language models, in: Proceedings of the 2023 Conference on Artificial Intelligence and Society, Association for Computing Machinery, 2023, pp. 231–245, http://dx.doi.org/10.1145/3593013.3594068.
[16] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 609–614.
[17] P. Rao, S. Patel, Cultural bias evaluation in natural language processing models, Int. J. Mach. Learn. Res. 23 (3) (2022) 567–589.
[18] Y. Zhang, R. Wang, D. Li, X. Song, T. Li, Mitigating cultural bias in NLP through dataset diversification and rebalancing, Trans. Assoc. Comput. Linguist. 10 (2022) 1234–1250, http://dx.doi.org/10.1162/tacl_a_00501.
[19] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in contextualized word embeddings, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 629–634.
[20] S. Mukherjee, A. Hassan, J. Han, Prompt engineering for bias reduction in large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2023, pp. 5410–5425, http://dx.doi.org/10.18653/v1/2023.emnlp-main.420.
[21] W. Tao, H. Liu, Y. Zhou, The impact of training data composition on bias in AI models, in: Neural Information Processing Systems (NeurIPS) Conference, 2023, pp. 115–129.
[22] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS 2016), Curran Associates Inc., 2016, pp. 4349–4357, https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
[23] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (6334) (2017) 183–186, http://dx.doi.org/10.1126/science.aal4230.
[24] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, The woman worked as a babysitter: On biases in language generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2019, pp. 3407–3412, http://dx.doi.org/10.18653/v1/D19-1339.
[25] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86, http://dx.doi.org/10.1214/aoms/1177729694.
[26] C. Villani, Optimal transport: Old and new, in: Grundlehren der Mathematischen Wissenschaften, vol. 338, Springer, Berlin, Heidelberg, 2009, http://dx.doi.org/10.1007/978-3-540-71050-9.
[27] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint arXiv:1702.08608, https://arxiv.org/abs/1702.08608.