A Review of the Evolution of Bias Detection Pathways and Fairness Evaluation Dilemmas in Large Language Models: From Surface-Level Auditing to Mechanism-Driven Contextualized Evaluation

Bowen Deng
Dongguan University of Technology, Dongguan, Guangdong, China

Abstract: As large language models (LLMs) become deeply integrated into various sectors of society, their potential algorithmic bias and unfairness have emerged as core risks constraining safe deployment. From the distinctive perspective of the "evolution of detection pathways," this review systematically examines the transition of LLM bias evaluation from early surface-level output auditing to mechanism-driven contextualized evaluation. First, it develops a comprehensive taxonomy covering endogenous and exogenous bias, as well as representational and allocational harms, and compares the fundamental differences between LLMs and traditional NLP systems in how bias is manifested. Second, it analyzes five core detection pathways-surface-level output auditing, counterfactual consistency testing, internal mechanism analysis, attitude probing, and adversarial elicitation-in terms of their measurement dimensions, explanatory power, and limitations. Through analysis of CBBQ, TWBias, and cross-cultural benchmarks, this review identifies three major gaps in current evaluation infrastructure: cultural coverage, dynamic interaction, and harm linkage. The findings show that a significant "harm gap" exists between detection scores and real-world social harm, and that there is a deep circularity challenge between detection methods and definitions of bias. Finally, this review proposes future research directions, including the construction of a layered bias ontology, a tiered fairness evaluation system, and full-lifecycle monitoring, thereby providing theoretical support for building a more inclusive and culturally responsive AI governance framework.

Keywords: Large Language Models; Bias Detection; Fairness Evaluation; Cross-Cultural Context; Harm Gap; Mechanism Analysis; Contextualized Evaluation

1. Introduction

1.1 Current State and Governance Challenges in Research on Bias in Large Language Models

With the global-scale deployment of large language models (LLMs), these models have demonstrated strong generative capabilities while also showing the risk of perpetuating and even amplifying harmful biases in human society [1]. Unlike the static statistical associations observed in the traditional natural language processing (NLP) stage, bias in LLMs emerges as a systemic risk that dynamically arises across multiple stages, including pretraining, fine-tuning, alignment, and user interaction [2]. This shift from "static representation" to "dynamic behavior" has turned bias into a critical governance issue affecting digital fairness, social justice, and system safety [1]. However, as the object of detection has moved from simple semantic associations to complex dynamic interactions, the research community has encountered a serious "evaluation dilemma": fairness conclusions for the same model often fail to converge, and may even contradict one another, across different benchmarks [3]. The root of this dilemma lies in the lack of a unified definition of bias and the logical disconnect between existing detection pathways and frameworks of social harm [2]. Accordingly, this review aims to systematically analyze the evolutionary logic of bias detection pathways. By unpacking the validity boundaries and limitations of different probing methods, it extracts an evaluation logic that accounts for both cultural differences and dynamic interaction contexts, thereby providing

theoretical support for the construction of a scientific fairness evaluation system for LLMs.

1.2 Complexity of LLM Bias Detection Compared with Traditional NLP

The rise of large language models (LLMs) has fundamentally changed the research paradigm of bias studies. The progressive logic underlying the complexity of bias detection can be summarized as follows: first, the object of detection has shifted from the discrete and static semantic space of traditional NLP to the continuous, probabilistic, and dynamic generative behavior of LLMs [1]; further, surface-level alignment mechanisms such as SFT and RLHF construct a "compliance shell" through instruction tuning, suppressing explicit harmful outputs while inducing a "masking effect" through which bias remains latently retained [4]; ultimately, this transformation creates an observability crisis at the evaluation level, causing models to exhibit greater defensiveness and randomness during auditing and leaving traditional single-metric auditing approaches with serious validity gaps [3].

Table 1 summarizes a systematic comparison between the traditional NLP and LLM paradigms across key dimensions, including detection objects, influencing mechanisms, and evaluation consequences.

Table 1. Systematic Comparison of Bias Detection Dimensions between Traditional NLP and LLMs

Evolutionary stage	Traditional NLP paradigm (Static/Encoder)	LLM paradigm (Generative/Aligned)	Core complexity shift
Stage 1: Detection object	Static word/sentence embeddings, focusing on geometric associations in space	Dynamic generated text streams, focusing on shifts in probability distributions	Dimensional expansion from the "representation layer" to the "behavior layer"
Stage 2: Influencing mechanism	Direct statistical projection of pretrained weights	Masking effects induced by safety alignment (SFT/RLHF)	Shift from "explicit release" to "instructional masking"
Stage 3: Evaluation consequence	Deterministic metrics, such as WEAT, with high observability	Evaluation shortcuts, defensive refusals, and reduced sensitivity	Validity gap between "direct auditing" and "adversarial evaluation"

1.3 Insufficient Explanatory Power of Existing Reviews Regarding Evaluation Divergence

Existing reviews have systematically summarized bias taxonomies, metrics, and datasets [1-2], but several limitations remain:

- **From "static inventories" to "conflict explanation":** Existing reviews rarely explain why the same model may yield sharply contradictory fairness conclusions under different detection pathways, such as direct questioning versus indirect elicitation, and they lack an analysis of the underlying logic behind evaluation divergence.
- **Insufficient attention to the "masking effect" induced by alignment mechanisms:** Alignment is often treated as a means of debiasing, while the possibility that safety training, such as RLHF and SFT, may function as a "masking mechanism" is overlooked. As a result, models may exhibit "pseudo-fairness" in benchmark tests while retaining bias under specific distributions [3].
- **Insufficient justification for the necessity of detection-pathway evolution:** Existing reviews often present detection methods in parallel, but fail to systematically justify the causal evolutionary logic from "shifts in bias forms" to "migration of detection objects" and then to "expansion of detection pathways" [4].

1.4 Research Scope and Conceptual Boundaries

To ensure focus and scientific rigor, this review concentrates on the evolutionary logic of social-group bias detection in text-based general-purpose large language models. It does not cover multimodal inputs such as images or speech, general toxicity detection, or the governance of factual hallucinations. The evidence base mainly covers publicly available academic literature and benchmarks from 2013 to 2025. By precisely defining the "research scope," this review aims to examine the underlying tensions between alignment mechanisms and fairness performance.

1.5 Core Propositions and Analytical Logic of This Review

Based on the deconstruction of the complexity of bias in large language models, this review proposes three core propositions as the logical

framework and argumentative thread of the paper.

Proposition 1: A paradigm shift in the ontology of bias from "static priors" to "dynamic emergence"

In the context of LLMs, bias has shifted from the quantifiable geometric displacement of word vectors in traditional NLP to behavioral emergence that is highly dependent on prompts and context. With the introduction of alignment mechanisms, such as RLHF and SFT, bias exhibits a clear form of "encryption," leading traditional static probing paradigms to face structural failures of observability when applied to deeply aligned models. Bias is no longer merely the statistical sediment of pretraining data; rather, it is a dynamic probability drift produced by the coupling of endogenous priors and exogenous interaction conditions.

Proposition 2: The heterogeneity of detection pathways is the epistemological root of conflicting evaluation conclusions

The five mainstream detection pathways-surface-level auditing, consistency testing, mechanism analysis, attitude probing, and adversarial elicitation-essentially reflect mutually non-overlapping operational definitions and measurement standards of fairness. Systematic divergence in evaluation conclusions is not a matter of random experimental error, but results from imbalanced validity weights across different pathways when capturing models' "explicit responses" and "latent associations." The lack of complementarity among pathways and the one-sided characterization of measurement dimensions constitute the technical root of the current "validity trap" in fairness evaluation.

Proposition 3: The evaluation dilemma is structurally constrained by "alignment masking" and multi-objective games

The validity crisis of evaluation conclusions is deeply constrained by the underlying value conflicts and zero-sum trade-offs among "safety, usefulness, and fairness" in LLMs. While alignment training suppresses explicit harmful outputs, it also induces secondary forms of bias, such as "masked fairness" and "differential refusals." This "alignment masking effect," driven by model compliance, means that existing benchmarks often capture only a model's ability to follow safety-oriented instructions rather than its actual social fairness performance.

Following the analytical logic of "transformation in the form of the object-evolution of detection pathways-divergence in evaluation conclusions," this review seeks to construct a contextualized evaluation framework capable of penetrating models' "alignment masking."

2. Foundations of Bias Research: Concepts, Types, and Harm Frameworks

2.1 Conceptual Distinctions among Bias, Fairness, and Harm

In the evaluation of large language models, it is necessary to clarify the hierarchical relationship among bias, fairness, and harm, so as to avoid ambiguity in research objectives [1].

This review defines bias as a statistical deviation in model outputs relative to a reference distribution or protected attributes. In essence, it reflects the model's systematic representation of probability distributions in the training data [5]. Fairness, by contrast, is defined as a normative evaluation criterion, involving judgments about whether model deviations conform to social ethics and value orientations [5]. Harm refers to the negative social consequences produced by bias and unfairness in the real world, covering multiple dimensions such as representational harm, allocational harm, and loss of service quality [1,39]. As an underlying technical deviation, bias may, under the scrutiny of fairness norms and through specific mechanisms of influence, ultimately be transformed into substantive social harm [6,39].

Together, these three concepts constitute a layered mapping system from "behavioral statistics" to "value evaluation" and then to "real-world consequences." Bias, as an underlying statistical deviation, is examined through fairness norms and may ultimately be translated into social harm that substantively affects specific groups [39]. This hierarchical logic provides theoretical support for understanding the emergence of risks in LLMs across different interaction contexts.

2.2 Sources of Bias: Endogenous Bias and Exogenous Bias

To accurately determine whether bias is a "background tendency" embedded in the model's parameter space or a "contextual manifestation" arising during interaction, this review categorizes bias into endogenous and exogenous forms.

2.2.1 Endogenous bias: statistical imprints from

pretraining

Endogenous bias originates from the model's statistical inheritance and feature internalization of large-scale pretraining corpora. It manifests as default reasoning pathways rooted in parameter weights, which may lead to value-laden tendencies and imbalanced expressions [5]. It corresponds to data bias and algorithmic bias in classical taxonomies, and constitutes the model's underlying behavioral baseline when it is not guided by specific interaction instructions [39].

2.2.2 Exogenous bias: interaction-based emergence during deployment

Exogenous bias arises during the application stage and is reflected in the model's high sensitivity to prompt framing, the position of system instructions, or contextual examples. This type of bias corresponds to interaction bias and prompt-induced bias, revealing the dynamic fluctuations in model generation caused by perturbations in the external environment [7]. It forms the conceptual basis of prompt bias and interaction bias.

2.2.3 Theoretical significance of the distinction

The core distinction between the two lies in the following: endogenous bias is a stable "background tendency" derived from training data, whereas exogenous bias is a form of "dynamic emergence" triggered by prompt framing or interaction contexts. This distinction helps penetrate the masking effect introduced by alignment mechanisms and enables a logical transition from static background measurement to dynamic contextualized evaluation.

2.3 From Model Deviation to Social Harm: Transformation Logic and Core Variables

To better understand the risk transmission mechanism of bias in LLMs, this section aims to clarify the evolutionary process through which underlying technical deviations migrate toward macro-level social harm.

2.3.1 A three-step logical chain from bias to harm

The social impact of bias follows a three-step transformation chain of "bias → scenario exposure → social consequences." First, bias, as a statistical deviation in model outputs, constitutes a potential source of risk. Further, when such deviations are activated in specific application scenarios, such as recruitment screening, medical advice, or judicial assessment, and exposed to users, they complete the transition from "parameter space" to "social space" [8]. Finally, this exposure is transformed into substantive allocational or representational harm by influencing resource distribution or reinforcing social bias [8], [9].

2.3.2 Core variables in the transformation process

Along this transformation pathway, the severity of harm is controlled by the sensitivity of the interaction context and the weight of the decision-making stage. Bias is not equivalent to immediate harm; its ultimate impact depends on the depth of model-output involvement in downstream tasks. By identifying this dynamic transformation logic, detection paradigms can shift from single-output auditing toward more context-aware harm evaluation, thereby more precisely locating the mechanisms through which risks emerge across different levels of exposure [6,39].

2.4 Prompt Bias and Interaction Bias

The introduction of prompt engineering and multi-turn dialogue mechanisms has shifted the manifestation of bias in LLMs from "static parameter imprints" to "dynamic behavioral evolution." Prompt bias can be understood as behavioral variation triggered by input framing. Influenced by system instructions, example distributions, and input order, it exhibits significant positional sensitivity [7]. As dialogue progresses, interaction bias may cause response tendencies to change with the accumulation of context, thereby inducing bias drift effects [7,9].

In cross-cultural contexts, this form of bias displays more complex characteristics: cultural context profoundly influences which outputs are perceived as biased or harmful [5,8]. Therefore, effective bias governance must move beyond a single filtering mechanism and shift toward sensitivity probing across the entire interaction process.

3. Evolution of Bias Measurement Objects in LLMs

3.1 Representation-Level Bias Measurement: Semantic-Space Geometry and Static Association

In early research, bias was primarily regarded as a form of static representation or probabilistic deviation. Studies at this stage defined bias as geometric displacement in the embedding space, namely statistical association bias at the level of word or sentence representations [10-11].

3.1.1 WEAT: a standardized mathematical paradigm for bias measurement

The Word Embedding Association Test (WEAT) quantifies social bias by calculating differences in cosine similarity between two sets of target words and two sets of attribute words [11]. Its core test statistic and normalized effect size are defined as follows:

$$d=\frac{\text{mean}_{x\in X}s(x,A,B)-\text{mean}_{y\in Y}s(y,A,B)}{\text{std\_dev}_{w\in X\cup Y}s(w,A,B)} \quad (1)$$

This metric reflects the degree of systematic deviation between different target-word sets in terms of attribute association, thereby laying the paradigmatic foundation for evaluating bias based on distances in vector space.

3.1.2 From words to sentences: the extension of seat and template dependence

The Sentence Encoder Association Test (SEAT) evaluates bias by embedding target words into "semantically bleached" sentence templates, such as "This is [target]," and computing the similarity of contextual vectors. Although SEAT attempts to capture contextual representations, it still follows the logic that "bias equals vector distance," and its evaluation validity is easily affected by model anisotropy and sensitivity to template selection [10].

3.1.3 "Lipstick on a pig": a fundamental critique of representation-level debiasing

Gonen and Goldberg argued that traditional geometric debiasing methods are merely "Lipstick on a Pig": even when the explicit gender direction is mathematically removed, implicit bias clustering relationships still remain in the word-vector space [17]. This finding sharply reveals the "masking effect": debiasing at the static representation level often only smooths out observable statistical indicators, while failing to address the deeper social-bias imprints embedded in the model.

3.2 Generation-Level Bias Measurement: Probability Distribution Shifts and Architectural Differentiation

As large language models have shifted from discriminative learning to generative pretraining, the object of bias detection has evolved from "positions in vector space" to "conditional probability distributions over token-sequence generation" [12-13]. Because different architectures differ fundamentally in probabilistic modeling, detection paradigms must be differentiated according to model families.

3.2.1 Masked Language Models (MLMs): conditional probability estimation based on Pseudo-Log-Likelihood (PLL)

For bidirectional encoder models such as BERT and RoBERTa, researchers have introduced pseudo-log-likelihood (PLL) metrics to evaluate bias because these models can observe both left and right contexts simultaneously [13]. This method measures the model's probabilistic preference for specific social associations in "fill-in-the-blank" tasks by accumulating the log probability of each token under a given context and comparing the probability differences between stereotypical and non-stereotypical descriptions.

3.2.2 Autoregressive models (CLMs) and Seq2Seq models: shifts in continuation probability

For GPT-series autoregressive models and sequence-to-sequence models such as T5 and BART, bias detection has evolved into measuring "shifts in continuation probability after a given prompt" [12-13]. Autoregressive models compute the product of sequence probabilities based on the chain rule, whereas Seq2Seq models measure the log-likelihood distribution of generating specific biased outputs within an encoder–decoder mapping [12].

3.2.3 Beyond probability: semantic metrics in open-ended generation contexts

In addition, researchers have introduced more macro-level measurement metrics, such as Regard, which aims to capture tendencies in social evaluation; differences in output distributions across groups, such as KL divergence [13]; and multidimensional evaluations of toxicity and affective polarization, such as the BOLD benchmark [12]. These metrics are used to more comprehensively capture bias characteristics in open-ended generation tasks [12-13].

By moving from low-level token probabilities to higher-level narrative tendencies, generation-level measurement can more realistically reflect the discrimination risks of LLMs in complex interactive tasks.

3.3 How Alignment Changes Bias Visibility: Surface-Level Suppression and Latent Retention

As supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) have become standard components of large language models, the object of bias

detection has undergone a profound structural transformation [14-15]. This process does not simply eliminate bias; rather, by constructing a set of "defensive strategies" at the upper layers of the model, it transforms originally observable explicit bias into more challenging latent bias [15-16].

3.3.1 Explicit suppression: alignment as a semantic "filter"

Through SFT and RLHF, models show a substantial reduction in explicit inappropriate outputs in standard tests [14-15]. In essence, this process adds a "semantic filter" at the probabilistic output layer, selectively lowering the generation probability of high-risk tokens and enabling the model to remain highly compliant under explicit auditing [16]. However, this explicit suppression is more accurately understood as "distributional pruning" targeted at specific evaluation metrics. The compliance displayed by the model may stem from its "evaluation awareness" of test templates rather than the genuine elimination of the roots of bias [15-16].

3.3.2 The "masking effect" and the latent retention of social bias

Although alignment effectively reduces explicit harmful outputs, social bias often remains present in the underlying representations [15,17]. This is consistent with the earlier critique of "lipstick on a pig": even when a model appears fair in explicit tests, it may still expose persistent systematic associations when computing the conditional probabilities of attribute words or responding to adversarial elicitation [15,17]. This "masking effect" means that, under a suppressed output space, social bias can still be latently retained through narrative tendencies and interaction logic.

3.3.3 Secondary unfairness introduced by differential refusals

Alignment also introduces new secondary fairness risks, namely "differential refusals" across different groups or scenarios [14,18]. To avoid risk, models often show a higher probability of refusal when facing certain identity backgrounds, such as Muslim names or specific minority groups, or when handling sensitive topics [14,15]. This uneven distribution of refusals caused by "over-defensiveness" essentially deprives certain groups of the right to access high-quality information services, thereby creating a new form of service-related unfairness.

3.4 Summary of the Evolutionary Logic: Paradigm Shift from Static Space to Interactive Behavior

The review of bias detection pathways in large language models reveals an evolutionary pattern from static representations to dynamic interactive behaviors [13,15,18].

As shown in Table 2, model evaluation has undergone four key paradigmatic transitions.

Table 2. Evolution Matrix of Bias Detection in Large Language Models

Evolutionary stage	Core object of examination	Representative measurement methods	Essential form of bias
Representation-level stage	Word/sentence embedding vectors	WEAT / SEAT / geometric displacement [10-11]	Statistical distance in semantic space [11]
Generation-level stage	Token-sequence generation probability	PLL / log-likelihood / perplexity [12-13]	Probability drift in token distributions [13]
Instruction-alignment stage	Task following and compliance	Explicit output auditing / alignment testing [14-16]	Masked latent associations [15,17]
Contextual-interaction stage	Multi-turn dialogue and dynamic behavior	Stress testing / interaction drift detection [14,18]	Emergent behavior in complex contexts [18]

Therefore, bias detection methods have shifted from static vector-space measurement to dynamic behavioral evaluation. The main trends include:

1) from geometric measurement to probabilistic measurement;

2) from single-turn question answering to multidimensional interactive tasks;

3) from single bias metrics to multi-objective integrated balancing.

This evolutionary pathway indicates that bias governance must move beyond simple static feature erasure and toward in-depth supervision of models' emergent behaviors in complex social contexts.

4. Expansion of Bias Detection Pathways: What Existing Methods Actually Measure

4.1 Surface-Level Output Auditing Pathway

Surface-level output auditing evaluates bias by measuring the distribution of a model's explicit

outputs under specific protocols, such as answer choices, polarity shifts, and stereotypical associations. Its core logic is to operationalize bias as countable differences in textual behavior, rather than directly revealing the internal parameter-level sources of bias [1]. Representative tools include BBQ, which evaluates selection bias in question-answering scenarios [19], and BOLD, which measures bias in open-ended generation [12]. Owing to its intuitiveness and scalability, this pathway has expanded from English-language settings to multilingual and multitask Chinese benchmarks such as CBBQ [20], TWBias [21], and McBE [22], enabling horizontal comparisons across models.

From the historical lineage of evaluation paradigms, bias measurement has migrated from early representation-level metrics, such as WEAT, and probability-level metrics, such as StereoSet based on vocabulary-level probability shifts, toward comprehensive generation-level benchmarks [1]. Although current auditing methods have moved beyond simple word counting and have broadly expanded into a multidimensional system covering sentiment polarity, toxicity scores [12], distributional difference metrics, and even LLM-as-a-Judge-based automatic evaluation, their essence remains descriptive auditing of the compliance of models' explicit behavior [1].

This pathway faces significant threats to validity. The first is prompt sensitivity: minor template changes, such as variations in prompt format or order, may cause large fluctuations in a model's bias score on the same dimension [7,23], and may even change the ranking of models, thereby undermining the robustness of evaluation conclusions [24]. The second is the trap of **post-alignment "surface neutralization"**: supervised fine-tuning and reinforcement learning from human feedback (RLHF) often merely add a "semantic filter" at the probabilistic output layer, suppressing explicit discriminatory language [25]. As a result, models may appear neutral in descriptive tests, yet still exhibit significant "inconsistency between words and actions" in simulated decision-making tasks involving resource allocation, due to inherent bias imprints in their underlying parameters [7]. This masking effect indicates that surface-level auditing has difficulty accurately capturing latent background risks inside the model.

4.2 Counterfactual and Consistency Testing Pathway

Counterfactual and consistency testing measures the sensitivity and stability of model outputs by applying a controlled-variable approach. Under the premise that the semantic background remains largely unchanged, this pathway perturbs sensitive attribute labels, such as gender, ethnicity, and names [1-2]. Typical methods include minimal-pair tests in gender coreference resolution and question-answering formats, in which sensitive attribute terms are replaced while the contextual semantics remain unchanged, such as replacing "he" with "she" or changing ethnically marked names, and the degree of output reversal in classification, selection, or generation probability is then observed [19,26].

The core advantage of this testing paradigm lies in its capacity for causal isolation. By constructing strictly aligned contrastive samples, it can effectively remove interference caused by superficial variations in input text and directly reveal the conditional dependence of model decisions on specific demographic attributes [1]. Compared with surface-level auditing, consistency testing can better reflect the latent weight of stereotypes embedded in complex logical reasoning [1].

However, the limitations of this pathway are mainly reflected in coverage and validity. Because the tests rely heavily on predefined sentence templates or lexicons, their evaluation conclusions are often limited to a restricted set of stereotype dimensions and cannot exhaustively cover the intersecting and hidden forms of social bias in the real world [1]. In addition, this pathway is also affected by prompt effects: even minimal-pair samples may fluctuate sharply with positional effects in guiding instructions [7] or changes in prompt format [23], resulting in a lack of robustness in evaluation conclusions [24]. At the same time, because some benchmarks lack clear definitions of power asymmetries, they are often questioned for capturing only statistical correlations rather than genuine bias rooted in deeper social structures.

4.3 Internal Mechanism Analysis Pathway

Internal mechanism analysis aims to explain the formation mechanism of bias at its source by analyzing model weights, activations, or neural layers. Typical methods include using probing

techniques to identify the geometric structure of semantic space [1], locating key bias-related components through causal intervention analysis [26], and providing explanatory clues for subsequent bias governance [1]. Research on these techniques reveals a central warning: internal interpretability does not equate to external expressiveness. This means that even if a model can be internally interpreted as containing a salient bias signal, it does not necessarily express discriminatory tendencies in actual outputs. Similarly, surface-level neutralized outputs may conceal bias imprints that remain active in the underlying parameters [1-2]. Because internal probing metrics often lack strong correlations with models' external performance in downstream tasks, researchers emphasize that internal probing results alone should not be used to determine the social-bias risk of a model; rather, they should be regarded as an explanatory supplement to external behavioral auditing [1].

Core boundary: Internal interpretability ≠ external expressiveness

It must be made clear that "detectable" does not mean "expressed under normal deployment conditions" [1]. Mechanistic studies show that bias-related information may remain dormant in the later layers of a model or in specific neurons [26], but whether it is ultimately transformed into explicit social harm depends on specific prompt triggers and decoding strategies [7,23]. Therefore, the contribution of internal mechanism analysis lies in improving the explanatory power and intervention precision of bias governance, rather than independently replacing fairness evaluation at the deployment level.

4.4 Value and Cultural Alignment Probing

This pathway focuses on the consistency of models' responses to values and cultural knowledge. By simulating diverse value-oriented prompts and examining the coverage of cultural knowledge, it evaluates whether models possess cross-cultural adaptability [27-28]. In brief, by comparing models' response preferences across different cultural contexts, this pathway probes whether models tend to favor the value system of particular groups [27]. However, the evaluation validity of this pathway faces serious criticism regarding stability. Röttger et al. found that when models simulate different personas or handle culturally sensitive topics, their value expressions are extremely context-sensitive and highly stochastic [24]. This means that even when explicit harmful content is suppressed through alignment, the consistency of models across different cultural expressions remains difficult to guarantee. The apparent "alignment" they display is often a product of random fluctuation rather than a stable internal representation [24,27].

4.5 Adversarial Elicitation and Stress Testing Pathway

This pathway aims to probe the robustness of models at the boundaries of alignment defenses and to uncover potential latent bias risks by constructing extreme or adversarial inputs.

4.5.1 Jailbreak Attacks and Latent Harmful Generation

Jailbreak attacks refer to the deliberate construction of specific input prompts to induce safety-aligned models to generate harmful information that would otherwise be prohibited [29]. The core of this evaluation method lies in identifying particular prompting strategies or textual formats, such as role-playing or logical induction, that can bypass the model's built-in compliance constraints [29]. Although various automated test sets have been developed, the key to evaluation lies in understanding how attack methods bypass instruction-following mechanisms and thereby reveal the underlying social biases and safety risks that remain in models under extreme interaction contexts [1,29].

4.5.2 Over-Refusal and Service Unfairness

Research shows that, in order to avoid potential bias risks, aligned models often exhibit "over-refusal" when responding to high-risk or semantically ambiguous inputs, and this refusal behavior shows significant systematic differences across groups or task types [2]. For example, tests using XSTest and OR-Bench have found that models generally show a higher probability of refusing to answer questions involving specific socially sensitive topics or protected groups, such as race and religion [25,30]. Such differential refusal not only reflects the excessive sensitivity of model safety strategies, but may also lead to de facto unfair treatment in access to information services for specific groups [1,30].

4.6 Task-Format Effects and Integration of Detection Pathways

Research on task-format effects has found that the degree of bias exhibited by models varies

significantly across different testing formats [1-2]. For example, the bias tendencies observed in multiple-choice formats, such as BBQ [19], may not be consistent with those observed in open-ended question answering or free-form text generation tasks, such as BOLD [1,12]. This phenomenon indicates that the manifestation of model bias is dynamically reshaped by task requirements, prompt instructions, and interaction formats [7,23], and that evaluation results based on a single pathway are often partial [2,24].

To systematically integrate the detection pathways discussed above, we propose the five-dimensional comparative analytical framework shown in Table 3, which summarizes the roles and complementary relationships of each pathway in bias governance.

Table 3. Comparative Analytical Framework of Five Bias Detection Pathways

Detection pathway	Core objective	Observation level	Core advantage	Main limitation
Surface-level output auditing	Quantify the distribution of explicit bias	Generated-output level	Intuitive and suitable for large-scale automated testing [1-19]	Easily affected by prompts [7]; difficult to detect latent bias [1]
Counterfactual consistency testing	Evaluate sensitivity to group labels	Input–output comparison	Enables causal isolation and removes semantic interference [1]	Strong template dependence and limited coverage [1]
Internal mechanism analysis	Probe the bias baseline in parameter space	Model weights / activation layers	Penetrates alignment masking and reveals formation mechanisms [1,26]	Internal interpretability ≠ external expressiveness [1]
Value and cultural alignment probing	Test consistency in value-oriented responses	Contextual-preference level	Evaluates globalization and cross-cultural adaptability [27-28]	Strongly affected by context, with relatively poor stability [24]
Adversarial elicitation and stress testing	Probe the boundaries of safety defenses	Adversarial-interaction level	Reveals risks in extreme or latent scenarios [29]	Complex to construct and difficult to exhaust all attack vectors [1-29]

The manifestation of bias in LLMs results from the joint effects of the base model's capabilities, alignment defenses, and task contexts. The core finding is that bias evaluation should not rely on a single metric; instead, it should establish a dynamic monitoring system that integrates multiple pathways in order to address threats to evaluation validity introduced by task-format effects.

5. Root Causes of the Evaluation Dilemma: Why Bias Evaluation Conclusions Often Diverge

The evaluation dilemma essentially stems from the misalignment between constructs and measurements [31-32], rather than from purely technical bottlenecks. Its core tensions lie in the following aspects:

- **Confusion between constructs and operationalization:** Researchers often fail to distinguish theoretical concepts, namely what bias actually is, from their specific measurement indicators. As a result, phenomena with different ontological foundations are compressed into the same evaluation system, obscuring the heterogeneity of evaluation conclusions [31,33].
- **Lack of normative reasoning:** Most studies do not clearly define the nature of harm or the groups affected by it, and they lack a logical explanation of why a given bias phenomenon is "harmful" [5].
- **Heterogeneity of measurement objects:** Although these studies are all labeled as "bias evaluation," they in fact measure distinct and non-equivalent dimensions, including stereotypical associations, generative tendencies [12], refusal behavior [25], and cultural mismatch [20,34].

5.1 Inconsistencies between Definitions of Bias and Harm Frameworks

This section adopts a three-level hierarchical framework of "construct level (the concept of bias)-operational level (measurement indicators)-harm level (actual impact)" to analyze the evaluation dilemma [5,32,39]. Within this framework, statistical deviation, social bias, and cultural difference are not parallel evaluation objects, but belong to different logical levels:

1) Statistical deviation mainly occurs at the operational level and is manifested as statistical skewness in measurement data;

2) Social bias corresponds to the construct level

and reflects differences in underlying values and beliefs;

3) Cultural difference is more deeply embedded in the harm level and broader context, involving normative judgments about substantive impacts on specific groups.

Because these three types of bias are fundamentally disconnected in terms of ontology and measurement logic, and because they lack a unified consistency-based explanation, researchers find it difficult to achieve a one-size-fits-all fairness evaluation through a single metric [6], [35].

5.2 Tension between Benchmark Testing and Ecological Validity

There is a significant tension between benchmark testing and real-world application scenarios. Existing bias evaluations mostly rely on simplified closed-loop benchmarks [3], [31], which often attempt to decontextualize complex forms of social bias. However, studies have shown that such benchmarks are limited in ecological validity: evaluation frameworks and harm definitions often show a strong "U.S.-centric" tendency, lacking a global perspective and cross-cultural sensitivity [19] [20], [36]. Because the manifestation of bias in model outputs is highly related to the cultural values in which the model is embedded, mainstream models often exhibit clear cultural bias or value misalignment when handling non-Western contexts [20], [36].

Scholars have pointed out that bias evaluation must consider adaptation to different cultural backgrounds; otherwise, the "fairness" conclusions derived from benchmarks are difficult to generalize to real-world multicultural scenarios [34], [36]. Cultural differences are reflected in multiple deep dimensions, including linguistic habits, social norms, and moral customs, which makes decontextualized general-purpose evaluation metrics prone to failure [20], [34]. Therefore, bias evaluation is undergoing a transition from single-metric assessment toward culturally adaptive pathways. In summary, all analyses emphasize that cultural adaptation is essential: the measurement and definition of bias must account for the cultural context of the target groups in order to ensure the validity of evaluation results in real-world ecological settings [20], [34], [36].

5.3 Objective Conflicts among Safety, Usefulness, and Fairness

There are inherent conflicts among safety, namely avoiding harm, usefulness, namely effectiveness, and fairness. Studies have pointed out that these three objectives often cannot be maximized simultaneously during optimization [35], [37]. This tension is particularly evident in the HHH framework, which emphasizes helpfulness, honesty, and harmlessness [37]: strengthening safety, such as adopting a highly cautious stance toward potentially sensitive or controversial content, often triggers "over-refusal" behavior in models, thereby directly impairing their usefulness and service effectiveness [25], [30], [37].

In addition, tightening safety strategies may cause biased responses to topics related to specific groups, unintentionally exacerbating unfairness while improving safety defenses [15], [35]. Overall, under different alignment objectives, optimizing a single-dimensional metric often comes at the cost of other dimensions. This persistent tension among multiple objectives makes it difficult for bias evaluation to be unified through a single optimal solution.

5.4 Side Effects of Alignment: From Content Neutrality to Service Accessibility

Safety alignment strategies suppress harmful outputs, but they also introduce significant side effects [25], [37]. Through benchmarks such as XSTest [25] and OR-Bench [30], researchers have found that alignment is shifting fairness issues from "content neutrality" toward "service accessibility." These side effects are mainly reflected in three typical forms of refusal:

- **Over-refusal:** As revealed by XSTest, models show a generally higher refusal rate for benign but superficially sensitive questions, such as literary discussions involving violent terms, displaying a pattern of "over-defensiveness" [16], [25].
- **Differential refusal:** Tests such as OR-Bench show that models exhibit uneven distributions of refusal across different topics or domains, causing some legitimate needs to be systematically blocked [30], [38].
- **Identity-triggered refusal:** When prompts mention specific identity markers, such as religious or ethnic terms, the probability of triggering a refusal mechanism increases significantly, resulting in de facto identity discrimination [14], [15].

Overall, this indicates that alignment makes fairness no longer only a matter of whether output content is neutral, but more centrally a matter of whether model services are accessible. When the cost of service refusal is unevenly distributed across different user groups, this "secondary unfairness" constitutes substantive harm [14], [38].

5.5 Layered Explanatory Framework

The layered explanatory framework aims to systematically reveal the root causes of inconsistent bias evaluation results in large language models. This framework decomposes the evaluation dilemma into four interrelated but logically distinct dimensions:

1) **Conceptual level:** This level concerns the normative definition of bias and fairness, namely clarifying the underlying values of "what constitutes harm" and "who is harmed" [5]. Because different studies differ ontologically in their bias constructs, the lack of unified normative reasoning leads to a fundamental divergence in evaluation objectives [32].

2) **Operational level:** This level focuses on how abstract concepts are transformed into concrete measurement indicators and benchmark datasets. Inconsistencies in evaluation often arise from measurement bias in the process of operationalization, causing statistical association bias to become disconnected from actual social bias [6], [39].

3) **System level:** This level focuses on the model's internal alignment mechanisms and their multi-objective trade-offs, especially the inherent tension among safety, usefulness, and fairness. As revealed by XSTest [25] and OR-Bench [30], system-level safety alignment strategies often introduce side effects such as over-refusal, shifting fairness from a content-related attribute to an issue of service accessibility.

4) **Contextual level:** This level emphasizes that bias evaluation must be embedded in specific cultural, linguistic, and social backgrounds. Because the definition of bias is strongly culture-dependent, decontextualized general-purpose benchmarks are difficult to use for capturing substantive injustice in multicultural settings. Cultural adaptation has therefore become a core variable in evaluation validity [20], [34].

6. Future Research Directions and Conclusion

6.1 Future Research Directions

1) Promote a methodological transition in evaluation from "score-oriented" approaches to "construct-oriented" approaches;

2) Build context-sensitive evaluation systems that combine static closed benchmarks with dynamic multi-turn interactions;

3) Realize a paradigm shift from general-purpose bias measurement toward cultural localization and adaptation to social contexts;

4) Expand the scope of evaluation from content-neutrality auditing alone to full-scope service-accessibility auditing;

5) Explore a shift from phenomenon identification toward closed-loop governance and actionable intervention pathways.

6.2 Conclusion

This review systematically traces the evolution of bias detection in LLMs, showing that the object of detection has expanded from static representational associations to generative behavior, prompt-based interaction, and post-alignment performance. The core contribution of this review is not simply to list bias detection methods, but to reorganize existing research according to the logic of "evolution of detection objects-expansion of detection pathways-explanation of evaluation dilemmas." It further clarifies what different detection pathways actually measure and why the same model may yield inconsistent conclusions under different evaluation frameworks. This review argues that such divergence mainly stems from misalignments among definitions of bias, harm frameworks, task protocols, cultural contexts, and alignment objectives. Overall, future bias evaluation for LLMs should move beyond single-score comparison and toward a more context-sensitive integrated framework that can unify theoretical constructs, system behavior, and actual social harm.

References

[1] I. O. Gallegos et al., "Bias and Fairness in Large Language Models: A Survey," Computational Linguistics, vol. 50, no. 3, pp. 1097–1179, Jan. 2024.

[2] Z. Chu, Z. Wang, and W. Zhang, "Fairness in Large Language Models: A Taxonomic Survey," ACM SIGKDD Explorations Newsletter, vol. 26, no. 1, pp. 34–48, July 2024.

[3] R. Hida, M. Kaneko, and N. Okazaki,

"Social Bias Evaluation for Large Language Models Requires Prompt Variations," arXiv, July 2024.

[4] Y. Zhao et al., "Mind vs. Mouth: On Measuring Re-judge Inconsistency of Social Bias in Large Language Models," arXiv, Aug. 2023.

[5] S. L. Blodgett, S. Barocas, H. Daumé, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," pp. 5454–5476, Jan. 2020.

[6] H. Cyberey, Y. Ji, and D. Evans, "Do Prevalent Bias Metrics Capture Allocational Harms from LLMs?," pp. 34–45, Jan. 2025.

[7] A. Neumann, E. Kirsten, M. B. Zafar, and J. Singh, "Position is Power: System Prompts as a Mechanism of Bias in Large Language Models (LLMs)," pp. 573–598, June 2025,.

[8] R. Shelby et al., "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction," pp. 723–741, Aug. 2023.

[9] K. Chehbouni et al., "From Representational Harms to Quality-of-Service Harms: A Case Study on Llama 2 Safety Safeguards," pp. 15694–15710, Jan. 2024, doi: 10.18653/v1/2024.findings-acl.927.

[10] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On Measuring Social Biases in Sentence Encoders," pp. 622–628, Jan. 2019.

[11] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, no. 6334, pp. 183–186, Apr. 2017.

[12] J. Dhamala et al., "BOLD," pp. 862–872, Mar. 2021.

[13] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," pp. 5356–5371, Jan. 2021.

[14] N. Sivakumar et al., "Bias after Prompting: Persistent Discrimination in Large Language Models," arXiv, Sept. 2025.

[15] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths, "Measuring Implicit Bias in Explicitly Unbiased Large Language Models," arXiv, Feb. 2024.

[16] X. Yang, R. Zhan, D. F. Wong, S. A. Yang, J. Wu, and L. S. Chao, "Rethinking Prompt-based Debiasing in Large Language Models," arXiv, Mar. 2025.

[17] H. Gonen and Y. Goldberg, "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them," arXiv, Mar. 2019.

[18] X. Lin, L. Li, and X. Liu, "Inconsistency Between Words and Actions: Implicit Bias in Decision-making with Large Language Models." Aug. 11, 2025.

[19] A. Parrish et al., "BBQ: A hand-built bias benchmark for question answering," in Findings of the Association for Computational Linguistics: ACL 2022 , Jan. 2022, pp. 2086–2105.

[20] Y. Huang and D. Xiong, "CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models," arXiv, June 2023.

[21] H.-Y. Hsieh, S.-C. Huang, and R. T. Tsai, "TWBias: A Benchmark for Assessing Social Bias in Traditional Chinese Large Language Models through a Taiwan Cultural Lens," pp. 8688–8704, Jan. 2024.

[22] T. Lan et al., "McBE: A Multi-task Chinese Bias Evaluation Benchmark for Large Language Models," pp. 6033–6056, Jan. 2025.

[23] Z. Xu, K. Peng, L. Ding, D. Tao, and X. Lu, "Take Care of Your Prompt Bias! Investigating and Mitigating Prompt Bias in Factual Knowledge Extraction," arXiv, Mar. 2024.

[24] P. Röttger et al., "Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models," pp. 15295–15311, Jan. 2024.

[25] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models," pp. 5377–5400, Jan. 2024.

[26] J. Vig et al., "Investigating Gender Bias in Language Models Using Causal Mediation Analysis," in Neural Information Processing Systems , Jan. 2020, pp. 12388–12401. Accessed: Oct. 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf

[27] E. Karinshak et al., "LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output," arXiv, Nov. 2024.

[28] Y. Wang et al., "CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models," arXiv, Nov. 2023.

[29] E. Perez et al., "Red Teaming Language Models with Language Models," pp. 3419–3448, Jan. 2022.

[30] J. Cui, W.-L. Chiang, I. Stoica, and C. Hsieh, "OR-Bench: An Over-Refusal Benchmark for Large Language Models," arXiv, May 2024.

[31] O. van der Wal, D. Bachmann, A. Leidinger, L. van Maanen, W. Zuidema, and K. Schulz, "Undesirable Biases in NLP: Addressing Challenges of Measurement," Journal of Artificial Intelligence Research, vol. 79, pp. 1–40, Jan. 2024.

[32] A. Z. Jacobs, S. L. Blodgett, S. Barocas, H. Daumé, and H. Wallach, "The meaning and measurement of bias," pp. 706–706, Jan. 2020.

[33] P. Delobelle, G. Attanasio, D. Nozza, S. L. Blodgett, and Z. Talat, "Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP," pp. 21669–21691, Jan. 2024.

[34] A. Khan, S. Casper, and D. Hadfield-Menell, "Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs," pp. 2151–2165, June 2025.

[35] J. R. Anthis, K. Lum, M. Ekstrand, A. Feller, A. D'Amour, and C. Tan, "The Impossibility of Fair LLMs," arXiv, May 2024.

[36] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," PNAS Nexus, vol. 3, no. 9, Sept. 2024.

[37] L. Ouyang et al., "Training language models to follow instructions with human feedback," arXiv, Mar. 2022.

[38] M. Dabas, S. Chen, C. A. Fleming, M. Jin, and R. Jia, "Just Enough Shifts: Mitigating Over-Refusal in Aligned Language Models with Targeted Representation Fine-Tuning," arXiv, July 2025, Accessed: Oct. 2025. [Online]. Available: https://arxiv.org/abs/2507.04250

[39] S. Dev et al., "On Measures of Biases and Harms in NLP," pp. 246–267, Jan. 2022.

