The Genealogy of Social Bias in LLMs: A Survey

Anonymous ACL submission

Abstract

Large language models are increasingly evaluated for social bias, but most evaluations treat bias as an output-level disparity rather than asking how it enters, changes form, and becomes consequential across the LLM lifecycle. This survey addresses that gap by asking how socially and historically situated hierarchies become textual traces, how LLM pipelines transform those traces, and how models reproduce, amplify, suppress, transform, or newly produce bias. We synthesize 257 papers published between 2016 and 2026 and introduce a genealogical lifecycle framework that traces bias across data, preprocessing, training, alignment, retrieval, and downstream deployment. Using this framework, we organize prior work on bias detection, explanation, mitigation, multilingual and cultural evaluation, alignment and refusal behavior, retrieval, and high-stakes deployment. The key insight of our survey is that social bias in LLMs is not merely a benchmark score, but a lifecycle diagnosis problem: progress requires evaluations that connect observed disparities to source context, transformation patterns, pipeline stages, and interventions that actually change the implicated system component.

1 Introduction

Large language models (LLMs) are increasingly evaluated for social bias, but much of this work still begins from an output-level question: does model behavior differ systematically across socially meaningful groups? Prior work has shown that such disparities appear in embedding associations, sentence-completion benchmarks, underspecified question answering, open-ended generation, multilingual evaluation, refusal behavior, retrieval, and downstream decision-making (Caliskan et al., 2017; Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022; Dhamala et al., 2021; Smith et al., 2022; Hofmann et al., 2024; Zhang et al., 2025b). These findings establish that social bias in LLMs is systematic rather than anecdotal. However, detecting a disparity is not the same as understanding how it came to exist, how it changed form across the model pipeline, or what kind of intervention would address it. This distinction matters because the same surface behavior can have different genealogies. A stereotyped association may be inherited from training data, amplified during optimization, suppressed by alignment, transformed through proxy cues such as names or dialects, or introduced by retrieval and deployment settings (Tal et al., 2022; Tang et al., 2023; Hofmann et al., 2024; Khorramrouz and Levy, 2025; Zhang et al., 2025b). These are not merely different explanations for the same benchmark score. They imply different evaluation designs, different kinds of evidence, and different mitigation targets. A corpus imbalance, a tokenization artifact, an alignment-induced refusal pattern, a retrieval-ranking effect, and a downstream hiring disparity require different forms of diagnosis.

This survey develops a *genealogical understanding* of social bias in LLMs. We define *social bias* as historically and socially situated patterns of unequal treatment, representation, exclusion, or hierarchy across socially meaningful groups. (Hida et al., 2025) We use *model bias* to refer to systematic model behavior that reproduces, amplifies, suppresses, transforms, legitimizes, or newly produces such unequal treatment or representation. A genealogical account asks how social bias becomes available in text, how LLM pipelines transform it, and when it becomes consequential in use. This framing is especially important because textual data are not neutral records of society: they reflect the institutions, genres, categories, omissions, and power relations through which they were produced (Blodgett et al., 2020; Banaji et al., 2021; Winling and Michney, 2021; Wetts and Willer, 2019; Borenstein et al., 2023). Existing surveys provide the foundation for this shift in framing. Blodgett et al.

(2020) critique vague uses of "bias" in NLP; Sun et al. (2019) review gender-bias mitigation; Gallegos et al. (2023) consolidate LLM fairness metrics, datasets, and mitigation methods; Guo et al. (2024) organize LLM bias by origins, evaluation, and mitigation; Lin and Li (2025) focus on implicit bias; Cignarella et al. (2025) survey stereotype detection; and Gamboa et al. (2025) survey multilingual bias. These surveys establish the foundations for studying social bias in NLP and LLMs. However, they leave open a question that becomes central in downstream settings: *what evidence do we have about where bias enters the LLM pipeline, how it changes form, and which points of intervention are appropriate?*

Our paper makes four contributions:

- We introduce *lifecycle provenance reporting*, a five-part framework for evaluating social bias in LLMs. For any observed disparity, the framework asks researchers to report: observed disparity, plausible source evidence, lifecycle stage, transformation pattern, and intervention implication.
- We synthesize 257 papers on social bias in NLP and LLMs and analyze how existing work provides different forms of lifecycle evidence, including behavioral, source-level, representation-level, alignment and retrieval evidence.
- We distinguish five recurring transformation patterns linking social bias to model bias: inherited, amplified, suppressed, transformed, and emergent model-specific bias.
- We show that existing bias evaluations often provide strong local evidence but incomplete lifecycle diagnosis, and we propose reporting practices for connecting measured disparities to source context, pipeline stage, and appropriate intervention.

2 Background

What Counts as Social Bias? The term *bias* is used in many ways in NLP. It can refer to statistical skew in a dataset, harmful stereotypes, unequal error rates, group-dependent refusal behavior, disparate downstream performance, a mismatch between model behavior and a normative goal, or the reproduction of social hierarchies (Lauscher et al., 2020; Blodgett et al., 2020; Schwartz et al., 2022; Gallegos et al., 2023; Guo et al., 2024; Lin and Li, 2025). In this survey, we distinguish between *social bias* and *model bias*. ***Social bias*** refers to historically and socially situated patterns of unequal treatment, representation, exclusion, or hierarchy across socially meaningful groups. ***Model bias*** refers to systematic model behavior that reproduces, amplifies, obscures, legitimizes, transforms, or newly produces such unequal treatment or representation. This distinction has three implications. First, model bias is **systematic**: the question is not whether one harmful output exists, but whether model behavior changes in patterned ways across comparable social cues (Czarnowska et al., 2021; Parrish et al., 2022; Marchiori Manerba et al., 2024). Second, model bias is **socially embedded**: a disparity matters when it relates to social bias, including histories of inequality, exclusion, and unequal power. Third, model bias is **not only inherited**: LLMs can generate biased artifacts shaped by technical design, prompt format, alignment policy, retrieval, and deployment architecture rather than artifacts directly copied from society.

Search and Selection Protocol. We conducted a structured literature search for work on social bias in NLP and LLMs published between 2016 and 2026 following the **PRISMA-ScR** Screening Flow (Page et al., 2021). We searched arXiv, DBLP, Semantic Scholar, and the ACL Anthology using queries grouped into seven themes: bias detection, interpretability and explanation, benchmarks, mitigation, safety and alignment, multilingual and cultural bias, and prior surveys. The search initially retrieved 6,645 candidate papers. We then applied automated relevance filtering and cross-source deduplication, which removed 5,520 records and left 1,125 unique records for eligibility assessment. We then used an LLM-assisted (Gemini-3.1-Pro (Gemini Team, 2025)) screening step to score each record against four review pillars: bias detection, explanation, mitigation, and resources or benchmarks. The LLM was used only for relevance triage, not for final inclusion decisions; the scoring prompt is provided in Appendix A.7. This step excluded 686 papers that were outside the central scope of the survey, leaving 439 high-relevance papers for manual review. We then conducted manual filtering and full-text review, excluding a further 182 papers that were redundant, only tangentially related, or insufficiently focused on social bias in NLP and LLMs. The final survey corpus consists

Survey / framework	Defs.	Data	Mitig.	Sources	Diachronic	Deploy.	Main organizing principle
Blodgett et al. (2020)	✓	●	●	●	✗	●	Normative clarity and critique of vague "bias" usage.
Sun et al. (2019)	●	●	✓	●	✗	✗	Gender bias recognition and mitigation in NLP.
Gallegos et al. (2023)	✓	✓	✓	●	✗	●	LLM bias datasets, metrics, and mitigation.
Guo et al. (2024)	✓	✓	✓	✓	✗	●	Origins, evaluation, mitigation, and implications.
Lin and Li (2025)	✓	●	●	●	✗	●	Implicit bias concepts and measurements.
Gamboa et al. (2025)	●	✓	●	●	✗	●	Multilingual bias evaluation and mitigation.
OUR SURVEY	✓	✓	✓	✓	✓	✓	Genealogy and lifecycle provenance: social structure → textual trace → model transformation → intervention.

Table 1: Comparison with prior surveys. ✓ indicates central coverage, ● indicates partial coverage, and ✗ indicates that the issue is not a main organizing focus.

of 257 papers. Figure 1 summarizes the yearly growth of this final corpus across lifecycle stages. The distribution shows a sharp increase after 2023, with the largest concentration of papers in 2025, reflecting the rapid expansion of LLM-focused bias research. It also shows that algorithmic processing and alignment-related work account for much of this growth, while retrieval and deployment studies remain smaller but increasingly visible.

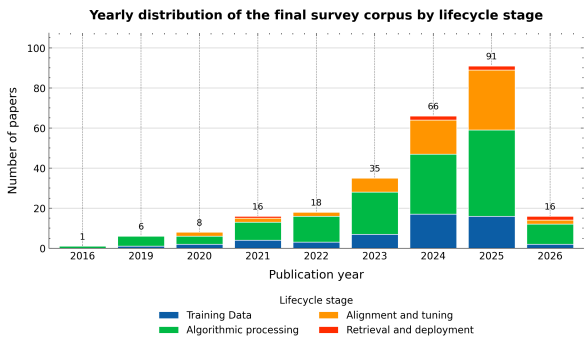

Figure 1: Yearly distribution of the final survey corpus ($n = 257$) by lifecycle stage.

How This Survey Differs from Existing Surveys. Table 1 positions our survey relative to prior work. Existing surveys define bias, organize harms, consolidate datasets and metrics, or identify broad sources of bias in NLP systems (Blodgett et al., 2020; Sun et al., 2019; Gallegos et al., 2023; Guo et al., 2024; Lin and Li, 2025; Cignarella et al., 2025; Gamboa et al., 2025). Our focus is as follows: we ask how social bias becomes visible in textual evidence, how LLM pipelines transform that evidence into model bias, and how lifecycle provenance can guide evaluation and intervention.

3 The Genealogy of Social Bias in LLMs

Genealogy and lifecycle provenance. The literature offers many ways to measure social bias in LLMs, but rarely explains its origins, how it evolves across the model lifecycle, or where best to intervene. We use ***genealogy*** to name this source-tracing problem: connecting observed model behavior to the underlying social, textual, and technical conditions. Genealogy acts as an organizing perspective for studying how social hierarchies become textual traces, and how LLM pipelines transform those traces into downstream consequences. To structure this analysis, we use ***lifecycle provenance*** to refer to the necessary evidentiary trail, specifically, where a bias enters, how it mutates, and which system component an intervention would actually affect. A genealogical evaluation therefore goes beyond asking whether a model is biased; it asks where the bias originates, what evidence supports that diagnosis, and how to effectively address it.

This perspective bridges two bodies of work. First, critical NLP research argues that datasets and benchmarks often treat "bias" too abstractly, failing to specify the relevant harm or normative basis (Blodgett et al., 2020; Weidinger et al., 2021; Schwartz et al., 2022). Second, historical and sociological work shows why this matters for language data: historical records and modern digital platforms rarely preserve all groups equally or describe them from the same standpoint (Banaji et al., 2021; Wetts and Willer, 2019; Winling and

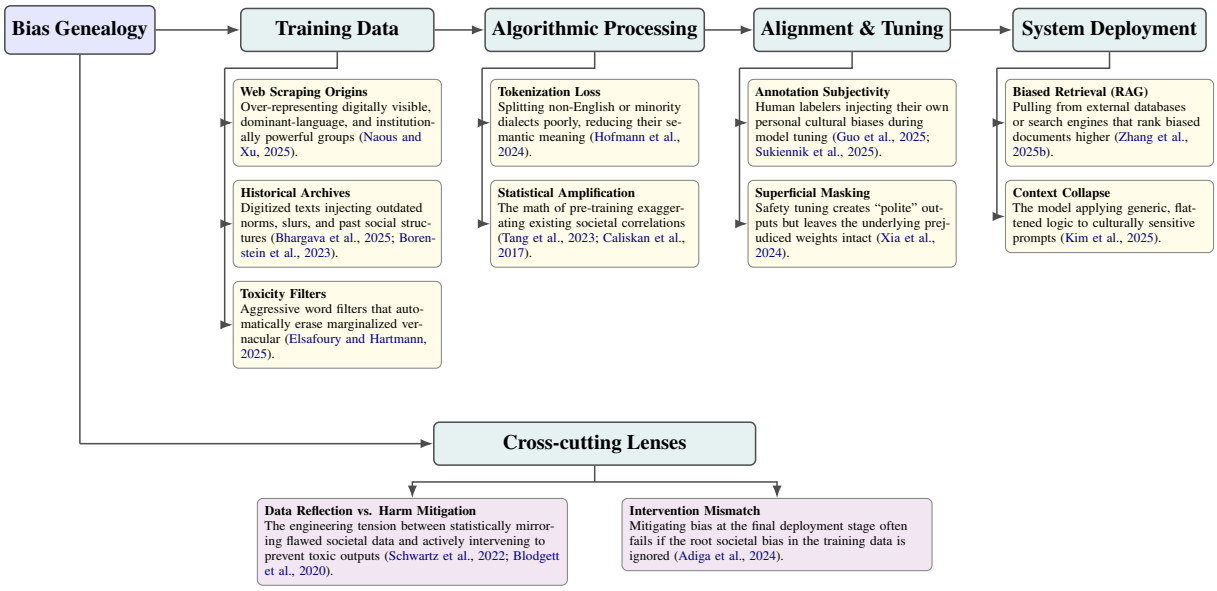

Figure 2: This framework shows how social bias can be introduced, amplified, suppressed, or transformed across the LLM lifecycle. It traces the origins of prejudice from unequal data ingestion, through algorithmic amplification (like tokenization and pre-training), to the flaws in human alignment and deployment systems. The lenses highlight core engineering tensions, such as the struggle between reflecting reality and mitigating harm.

Michney, 2021; Borenstein et al., 2023). These sources often encode institutional viewpoints, outdated classifications, and broader epistemic injustices (Fricker, 2007). Consequently, LLM training and retrieval corpora contain not just explicit stereotypes, but also systemic absences and historically skewed labels. LLM development then transforms these socially embedded[1] textual traces through technical choices made across the model lifecycle.

The LLM lifecycle and bias. Lifecycle provenance is necessary because social bias is not confined to training data; textual traces are amplified, suppressed, or fundamentally transformed at multiple stages of an LLM system. For example, web corpora and historical archives overrepresent digitally visible, dominant communities (Tao et al., 2024; Naous and Xu, 2025; Gamboa et al., 2025; Penedo et al., 2024, 2025; Borenstein et al., 2023). During preprocessing, filtering can mistakenly remove minority dialects by flagging them as toxic (Mozafari et al., 2020; Halevy et al., 2021; Elsafoury and Hartmann, 2025), while tokenization often fragments names and low-resource languages into rare sequences that degrade performance (Hofmann et al., 2024; Bhatia et al., 2025a; Zahraei and Emami, 2025). In pretraining and alignment, models may strengthen harmful associations to improve prediction (Tang et al., 2023; Caliskan et al., 2017), or mask explicit stereotypes while leaving implicit biases and selective refusals intact (Bai et al., 2024; Zhao et al., 2025; Lin and Li, 2025; Khorramrouz and Levy, 2025; Xia et al., 2024). Finally, downstream retrieval and deployment can reintroduce bias when models are integrated into sensitive workflows like hiring or healthcare (Zhang et al., 2025b; Salinas et al., 2023; Glazko et al., 2024; Rao et al., 2025; Mousavian et al., 2025; Suenghataiphorn et al., 2025). Figure 2 maps these lifecycle entry points. Rather than acting as a universal causal diagram, it highlights how different methods, ranging from corpus analysis to downstream audits, provide different pieces of genealogical evidence. Because biases cannot always be fully traced (especially in closed systems), the strength of an evaluation depends on connecting these methodological pieces across the pipeline (see Table 8 in Appendix A.5).

Five bias transformation patterns. This lifecycle view clarifies the relation between social bias and model bias. We identify five recurring transformation patterns that describe how socially embedded textual traces appear in model behavior. **Inherited bias** occurs when models reproduce socio-historical associations present in data, such as gendered occupation stereotypes (Boluk-

[1]Embeddedness in the context of this paper refers to Mark Granovetter's (Granovetter, 1985) use of the term. The concept acknowledges that systems cannot be detached from the social contexts under which they operate and emerge.

basi et al., 2016; Caliskan et al., 2017; Kurita et al., 2019; Nangia et al., 2020; Nadeem et al., 2021). In lifecycle terms, the strongest evidence points to source composition or training-data distribution. **Amplified bias** occurs when optimization or generation strengthens an association to improve prediction or retrieval salience (Tal et al., 2022; Tang et al., 2023; Salinas et al., 2023). Here, the relevant evidence concerns transformation during optimization, ranking, or decoding. **Suppressed bias** occurs when alignment or safety tuning reduces explicit stereotypes while implicit associations, unequal helpfulness, or selective refusals remain (Bai et al., 2024; Zhao et al., 2025; Lin and Li, 2025; Khorramrouz and Levy, 2025). This pattern points to post-training as a stage where bias may become less visible without being removed. **Transformed bias** occurs when one social cue becomes another kind of unequal treatment, as when dialect is treated as a proxy for competence or names are treated as proxies for nationality, religion, ethnicity, or class (Hofmann et al., 2024; Mozafari et al., 2020; Halevy et al., 2021; Wagh and Srivastava, 2025). The relevant evidence concerns proxy formation across data, representation, and task framing. **Emergent model-specific bias** occurs when unequal behavior arises from technical design choices, ranging from tokenization methods to the final deployment architecture (Hofmann et al., 2024; Hida et al., 2025; Kohankhaki et al., 2024; Zhang et al., 2025b). Because identical surface biases can stem from entirely different lifecycle stages, ranging from historical corpus imbalances to model alignment policies, *genealogy frames bias evaluation as evidence matching*. Rather than treating detection, explanation, and mitigation as isolated categories, this approach integrates them to trace a bias's specific origin, its transformation pattern, and the most effective point for intervention (see Table 10 in Appendix A.8).

Lifecycle provenance reporting. We operationalize genealogical bias analysis through *lifecycle provenance reporting*, summarized in Table 2. For each observed bias, an evaluation should report five components: the *observed disparity*, or the measurable difference in model behavior across social cues or groups; the *source evidence*, or the available evidence about where the disparity may originate; the *lifecycle stage* where the disparity becomes visible; the *transformation pattern* linking social bias to model behavior; and the *intervention implication*, or the system component that would need to change to address the diagnosed source rather than only the surface output. This framework does not require full causal reconstruction in every study, especially for closed models where lifecycle evidence is often incomplete. Instead, it asks researchers to make the evidentiary scope explicit: what is directly supported, what is only partially supported, and what remains uncertain. Appendix A.6 applies this checklist to representative benchmark studies in Table 9, while Appendix A.8 illustrates it through dialect bias, gender–occupation bias, and historical immigration bias in Table 10. These examples show that bias categories are not interchangeable labels: a single aggregate score may reveal that a disparity exists while obscuring its social meaning, technical entry point, and appropriate intervention.

Component	Diagnostic question	Example	Reporting (%)
Observed disparity	What unequal behavior is measured?	Translation defaults "doctor" to male and "nurse" to female.	63
Source evidence	What suggests where bias came from?	Skewed demographic counts or word embedding distances.	28
Lifecycle stage	Where does the disparity surface?	Web scraping, RLHF human feedback, or downstream fine-tuning.	70
Transformation pattern	How does bias become behavior?	Marginalized identities frequently co-occurring with toxic contexts.	24
Intervention implication	What should be changed?	Counterfactual data augmentation or debiasing embeddings.	47

Table 2: Lifecycle provenance reporting framework and survey corpus coding ($N = 257$). Percentages reflect the proportion of papers providing evidence for each component.

4 Operationalizing Genealogical Bias Understanding

Bias in NLP is a sociotechnical problem shaped by language, institutions, data, and deployment contexts (Blodgett et al., 2020; Schwartz et al., 2022; Weidinger et al., 2021). Building on our genealogical framework, this section reviews the resources needed to make social bias visible and

actionable across high-stakes domains like hiring and healthcare. This section categorizes the papers included in our survey corpus according to the lifecycle framework introduced in § 3. Figure 2 maps where bias enters or changes form across the LLM lifecycle, while Table 3 shows how different methods act as lenses on those stages. Rather than treating detection, interpretability, and mitigation as isolated tasks, we treat them as interconnected evidence. The critical question is not just whether a method detects bias, but which lifecycle stage it illuminates. Our corpus coding supports this lifecycle imbalance. As seen in Figure 1 of the 257 papers reviewed, 136 papers, or 52.9%, primarily address algorithmic processing, while 62 papers, or 24.1%, focus on alignment and tuning. By contrast, only 52 papers, or 20.2%, are centered on training data, and just 7 papers, or 2.7%, directly address retrieval and deployment. This distribution suggests that the literature is richest where bias is measured or modified inside model behavior, but thinnest where bias becomes system-level, source-dependent, or consequential in use.

Training data and textual traces. Bias may enter through diverse sources, ranging from historical archives and translated corpora to modern web scraping. Detection benchmarks and corpus audits provide an initial lens on this stage by showing whether model behavior differs systematically across groups, prompts, identities, languages, or contexts. Standard resources such as CrowS-Pairs, StereoSet, BBQ, BOLD, and related probing frameworks are valuable because they make such disparities measurable (Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022; Dhamala et al., 2021; Li et al., 2020; Marchiori Manerba et al., 2024). However, in genealogical terms, they usually provide behavioral evidence rather than source evidence: they show that a disparity exists, but not whether it originates from historical data, model filtering, or the deployment context. Appendix A.1 lists the main datasets and benchmarks used for this kind of behavioral evidence in Table 4. Benchmarks often encode particular cultural assumptions, protected categories, and stereotype inventories. Recent work across diverse multilingual and non-Western settings provides evidence that relevant social categories and harms vary across languages, histories, and communities (Jin et al., 2024; Huang and Xiong, 2023; Sahoo et al., 2024; Zulaika and Saralegi, 2025; Kamruzzaman et al., 2025; Satheesh et al., 2025; Farsi et al., 2025; Gamboa et al., 2025). As a lens on the training-data stage, these resources help prevent the mistaken assumption that bias categories transfer unchanged across sociolinguistic contexts.

Algorithmic processing. Bias is not simply copied from data into outputs. It can be fragmented by tokenization, strengthened by pretraining objectives, encoded in hidden representations, or converted into proxy behavior through task framing and decoding. Detection methods, including embedding association tests, counterfactual perturbations, and downstream evaluations, provide behavioral evidence that bias is present after algorithmic processing, but they usually do not by themselves identify which processing step produced it (Caliskan et al., 2017; May et al., 2019; Nangia et al., 2020; Nadeem et al., 2021; Li et al., 2020; Zhao et al., 2018; Hida et al., 2025). Appendix A.2 compares these detection methods in Table 5. Explanation and interpretability methods provide a second lens on this same stage. Mechanistic interpretability and probing methods (e.g., token attribution, hidden-state analysis, and activation patching) show where bias-related information appears inside the model (Attanasio et al., 2023; Gamboa and Lee, 2024; Koksal et al., 2023; Prakash and Lee, 2023; Yang et al., 2025; Liu et al., 2024; Chandna et al., 2025). These methods are especially useful for distinguishing inherited bias from model-specific bias. For example, unequal treatment of dialects or culturally specific expressions may reflect social stereotypes in the data, but it could equally stem from technical factors like tokenization or representational geometry (Hofmann et al., 2024; Bhatia et al., 2025a; Zahraei and Emami, 2025; Naous and Xu, 2025). Appendix A.3 summarizes explanation methods in Table 6. However, localization is not genealogy. Identifying an influential word or demographic feature in a hidden state does not explain its origin. Processing evidence must therefore be combined with corpus evidence and behavioral evaluation before making claims about origin or transformation.

Alignment and tuning. Here, the relevant question is not only whether bias is reduced, but whether it is removed, hidden, reframed, or shifted into another behavior. Alignment studies show that models can reduce explicitly biased outputs while retaining implicit biases, selective refusals, or unequal helpfulness (Bai et al., 2024; Zhao et al.,

Evidence lens	Training data and textual traces	Algorithmic processing	Alignment and tuning	Retrieval and deployment
Genealogical question	Which histories, sources, corpus choices, or absences make the trace available?	How does processing preserve, amplify, fragment, suppress, or transform bias?	Does tuning remove bias, suppress bias, or shift bias into refusals, helpfulness, or proxies?	When does bias become consequential, and which system component must change?
Detection and benchmarks	Corpus and prompt tests reveal disparities, but rarely identify where they enter the lifecycle (Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022; Dhamala et al., 2021).	Association tests, probability scoring, counterfactuals, and prompt variation expose bias after representation learning or decoding (Caliskan et al., 2017; May et al., 2019; Li et al., 2020; Hida et al., 2025).	Explicit/implicit bias tests, refusal audits, and helpfulness comparisons show whether post-training reshapes visible behavior (Bai et al., 2024; Zhao et al., 2025; Lin and Li, 2025; Pan et al., 2025).	Generation, RAG, ranking, hiring, clinical, and moderation audits show whether bias becomes consequential in use (Jin et al., 2025; Lan et al., 2025; Zhang et al., 2025b; Rao et al., 2025).
Explanation and interpretability	Documentation, metadata, annotation analysis, and culturally grounded benchmarks clarify what categories and traces are operationalized (Blodgett et al., 2020; Borenstein et al., 2023; Bhatt et al., 2022).	Attribution, probing, logit-lens, attention, neuron, and mechanistic methods localize bias-related computation (Yang et al., 2025; Liu et al., 2024; Chandna et al., 2025).	Post-tuning representation checks reveal whether alignment suppresses visible bias while retaining internal associations (Xia et al., 2024; Adiga et al., 2024).	Retrieved-document attribution, source ranking, interface analysis, and workflow audits separate model effects from system effects (Zhang et al., 2025b; Mousavian et al., 2025; Suenghataiphorn et al., 2025).
Mitigation and intervention	Corpus balancing, filtering audits, counterfactual augmentation, and culturally grounded annotation intervene on available text (Zmigrod et al., 2019; Han et al., 2024; Elsafoury and Hartmann, 2025).	Representation debiasing, adversarial training, steering, and editing intervene on model-internal transformations (Li et al., 2025; Xu et al., 2025; Yu and Ananiadou, 2025).	Instruction tuning, preference optimization, safety alignment, and refusal calibration reshape post-training behavior (Schick et al., 2021; Xia et al., 2024; Zhang et al., 2025a; Gallegos et al., 2025).	Retrieval-aware debiasing, source re-ranking, deployment constraints, and human review intervene on system evidence and workflows (Rehman and Muzammil, 2025; Zhang et al., 2025b).

Table 3: Evidence lenses across the LLM lifecycle. Figure 2 identifies where social bias can enter or change form; this table shows how evidence types inspect those stages.

2025; Lin and Li, 2025; Khorramrouz and Levy, 2025; Xia et al., 2024). In the mapping between Figure 2 and Table 3, alignment is where mitigation and explanation evidence must be read together. Mitigation methods may show that an intervention changes visible outputs, while explanation or implicit-bias evaluations can reveal whether the underlying association remains. This distinction is central to the suppressed-bias pattern introduced above: an aligned model may appear less biased at the surface while still preserving unequal internal representations, selective refusals, or proxy discrimination.

Retrieval and deployment. The final lifecycle stage concerns system deployment, including retrieval-augmented generation (RAG) and high-stakes decision-making workflows. At this stage, bias may be reintroduced by external sources, ranking functions, or downstream decision rules. Retrieval-grounded generation can correct unsupported stereotypes when evidence is reliable and well contextualized, but it can also reintroduce bias when the retrieval corpus overrepresents historically skewed, institutionally dominant, outdated, or culturally narrow sources (Zhang et al., 2025b). Deployment audits therefore provide a system-level lens on the lifecycle. They show when a benchmark disparity or representation-level association becomes consequential in practice. For instance, analyzing institutional records with LLMs can distort findings by over-detecting toxicity in minority dialects or reproducing dominant historical narratives (Mozafari et al., 2020; Halevy et al., 2021; Borenstein et al., 2023; Horych et al., 2024). If LLMs are embedded in recommendation, hiring, or public-sector systems, hidden biases can affect which candidates, services, or viewpoints are prioritized (Salinas et al., 2023; Glazko et al., 2024;

Rao et al., 2025; Mousavian et al., 2025; Suenghataiphorn et al., 2025). In genealogical terms, deployment evidence extends the analysis beyond model behavior to the sociotechnical setting in which that behavior acts.

Intervention evidence across the lifecycle. The final operational question is what an intervention actually changes. Mitigation methods span the lifecycle, from data augmentation and representation debiasing to inference-time steering and model editing (Zmigrod et al., 2019; Ravfogel et al., 2020; Dayanik and Pado, 2021; Schick et al., 2021; Xia et al., 2024; Zhang et al., 2025a; Li et al., 2025; Xu et al., 2025). Because each method targets a specific lifecycle stage, Table 7 in Appendix A.4 maps these interventions accordingly. A genealogical account cautions that mitigation can hide, shift, or transform bias rather than remove it. Prior work shows that debiasing may fail to transfer across tasks, reduce overall utility, or fix one bias dimension while leaving others intact (Meade et al., 2021; Kaneko et al., 2024; Hida et al., 2025; Chand et al., 2025; Yamamoto et al., 2025).

Taken together, Figure 2 and Table 3 define the operational structure of genealogical bias evaluation.

5 Discussion: From Bias Scores to Lifecycle Diagnosis

Existing work shows that LLMs reproduce, amplify, suppress, transform, and newly produce social bias, but evidence is often reported in isolation. A genealogical approach treats benchmark scores as starting points for lifecycle diagnosis: identifying the implicated stage and the intervention it requires.

Alignment and tuning as suppression rather than removal. Alignment and safety tuning make this issue especially visible. Safety tuning can suppress explicit stereotypes while preserving implicit biases, selective refusals, or unequal helpfulness (Bai et al., 2024; Zhao et al., 2025; Lin and Li, 2025; Khorramrouz and Levy, 2025; Xia et al., 2024). Because apparent surface-level improvements may merely reflect behavioral masking, post-alignment evaluations must compare explicit outputs against implicit associations and downstream task performance. This makes it possible to distinguish removal from suppression and to identify whether the relevant intervention belongs at the level of data, representation, alignment policy, retrieval, or deployment design.

Open system-deployment considerations for evaluation. At the retrieval and deployment stage, many socially situated harms are difficult to observe in short, template-based prompts. Longer-context, open-ended, retrieval-augmented, and task-specific evaluations are needed when bias depends on framing, source selection, institutional context, or historically specific categories (Dhamala et al., 2021; Smith et al., 2022; Pan et al., 2025; Zhang et al., 2025b; Parihar and Cheng, 2026). For example, a model may not produce an explicit stereotype in a sentence-completion benchmark but may still frame a group negatively across a generated summary, or select retrieved evidence that over-represents institutional viewpoints. These cases are not merely longer versions of existing benchmarks; they require evaluation protocols that record sources, context, retrieval behavior, and downstream use.

Interdisciplinary interpretation. A genealogical account also clarifies why bias evaluation cannot be fully specified by NLP metrics alone. Finally, metrics identify disparities, but interdisciplinary interpretation determines their meaning. Deciding whether an intervention removes harm or inappropriately erases meaningful group-specific context requires moving beyond standard benchmarks and integrating historical, legal, and community expertise (Jin et al., 2024; Huang and Xiong, 2023; Sahoo et al., 2024; Tomar et al., 2025; Kamruzzaman et al., 2025; Zulaika and Saralegi, 2025; Satheesh et al., 2025; Farsi et al., 2025; Gamboa et al., 2025). Future work should therefore move beyond translating English benchmarks and combine NLP methods with community knowledge, historical context, legal expertise, and domain-specific evaluation.

6 Conclusion

This survey reframes social bias in LLMs as a genealogical lifecycle problem. Across benchmarks, interpretability, mitigation, alignment, retrieval, deployment, and multilingual evaluation, we show that bias can enter and change form at different pipeline stages. Progress therefore requires provenance-aware evaluation that connects observed behavior to source evidence, transformation patterns, and lifecycle-matched interventions.

Limitations

This survey is limited by the scope and unevenness of the existing literature. Although we followed a structured search and screening protocol, relevant work may be missing from adjacent fields, non-English venues, policy reports, practitioner audits, or unpublished community-led evaluations. The final corpus is also unevenly distributed across lifecycle stages: algorithmic processing and alignment are better represented than retrieval, deployment, and historical analysis, so our conclusions are stronger for some stages than others. Finally, lifecycle provenance is an evidentiary framework, not a guarantee of full causal reconstruction; for closed models especially, training data, filtering rules, alignment data, retrieval indices, and deployment logs are often unavailable, so researchers must report what is directly supported, partially supported, and uncertain.

Ethical Statement

This paper examines social bias in LLMs as a sociotechnical problem involving representation, exclusion, unequal treatment, and institutional harm. We rely only on previously published work and do not collect new human-subject data, but the reviewed literature includes sensitive social categories, stereotypes, and discriminatory language. We therefore avoid reproducing unnecessary harmful examples and emphasize that social categories are historically and culturally situated rather than universal or fixed. We also note that our framework should not be used to legitimize superficial audits: reporting a lifecycle stage or transformation pattern does not establish that a system is fair or safe, and mitigation may suppress, shift, or obscure bias rather than remove it.

Broader Impact

This survey aims to improve accountability in LLM bias evaluation by shifting attention from isolated benchmark scores to lifecycle diagnosis. The proposed framework can help researchers, auditors, and practitioners connect observed disparities to source evidence, transformation patterns, pipeline stages, and lifecycle-matched interventions, especially in multilingual, retrieval-augmented, and high-stakes deployment settings. Its broader benefit is to make visible how social hierarchies become textual traces and how technical systems transform those traces into consequential behavior. However, provenance-aware evaluation is not a substitute for participatory review, legal accountability, domain expertise, or careful governance when LLMs are used in sensitive settings such as hiring, healthcare, education, moderation, or public-sector decision-making.

References

Rishabh Adiga, Besmira Nushi, and Varun Chandrasekaran. 2024. Attention speaks volumes: Localizing and mitigating bias in language models. *Preprint*, arXiv:2410.22517.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *2402.04105v2*.

Mahzarin R. Banaji, Susan T. Fiske, and Douglas S. Massey. 2021. Systemic racism: individuals and interactions, institutions and society. *Cognitive Research: Principles and Implications*, 6(1).

Rahul Bhargava, Malene Hornstrup Jespersen, Emily Boardman Ndulue, and Vivica Dsouza. 2025. Impacts of racial bias in historical training data for news ai. *arXiv preprint*.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025a. Swan and arabicmteb: Dialect-aware, arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. *Preprint*, arXiv:2411.01192.

Mehar Bhatia, Shravan Nayak, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, Vered Shwartz, and Siva Reddy. 2025b. Value drifts: Tracing value alignment during llm post-training. *Preprint*, arXiv:2510.26707.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in nlp: The case of india. *2209.12226v5*.

Guanqun Bi, Yuqiang Xie, Lei Shen, and Yanan Cao. 2025. A group fairness lens for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8117–8139, Suzhou, China. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *1607.06520v1*.

Nadav Borenstein, Karolina Stanczak, Thea Rolskov, Natacha Klein Kaefer, Natalia da Silva Perez, and Isabelle Augenstein. 2023. Measuring intersectional biases in historical documents. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Trenton Bricken and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Preprint*, arXiv:2309.08600.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Shireen Chand, Faith Baca, and Emilio Ferrara. 2025. No free lunch in language model bias mitigation? targeted bias reduction can exacerbate unmitigated llm biases. *2511.18635v1*.

Bhavik Chandna, Zubair Bashir, and Procheta Sen. 2025. Dissecting bias in llms: A mechanistic interpretability perspective. *Preprint*, arXiv:2506.05166.

Alessandra Teresa Cignarella, Anastasia Giachanou, and Els Lefever. 2025. A survey on stereotype detection in natural language processing. *2505.17642v2*.

Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *2106.14574v1*.

Erenay Dayanik and Sebastian Pado. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872. Association for Computing Machinery.

Fatma Elsafoury and David Hartmann. 2025. Out of sight out of mind, out of sight out of mind: Measuring bias in language models against overlooked marginalized groups in regional contexts. *Preprint*, arXiv:2504.12767.

Farhan Farsi, Shayan Bali, Fatemeh Valeh, Parsa Ghofrani, Alireza Pakniat, Kian Kashfipour, and Amir H. Payberah. 2025. Pbbq: A persian bias benchmark dataset curated with human-ai collaboration for large language models. *2510.19616v1*.

Eve Fleisig and Christiane Fellbaum. 2022. Mitigating gender bias in machine translation through adversarial learning. *Preprint*, arXiv:2203.10675.

Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Isabel O. Gallegos, Ryan Aponte, Ryan A. Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, Nedim Lipka, Deonna Owens, and Jiuxiang Gu. 2025. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *2309.00770v3*.

Lance Calvin Lim Gamboa, Yue Feng, and Mark Lee. 2025. Social bias in multilingual language models: A survey. *2508.20201v2*.

Lance Calvin Lim Gamboa and Mark Lee. 2024. A novel interpretability metric for explaining bias in language models: Applications on multilingual models from southeast asia. *Preprint*, arXiv:2410.15464.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gemini Team. 2025. A new era of intelligence with Gemini 3.

Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and improving disability bias in gpt-based resume screening. *2402.01732v2*.

Mark Granovetter. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, 91(3):481–510.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. Care: Multilingual human preference learning for cultural awareness. *Preprint*, arXiv:2504.05154.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *2411.10915v1*.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. *2109.13137v1*.

Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. ChatGPT based data augmentation for improved parameter-efficient debiasing of LLMs. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity, Inclusion, and Accessibility*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2025. Social bias evaluation for large language models requires prompt variations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14507–14530, Suzhou, China. Association for Computational Linguistics.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Tomas Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2024. The promises and pitfalls of llm annotations in dataset labeling: a case study on media bias detection. *2411.11081v2*.

Yufei Huang and Deyi Xiong. 2023. CBBQ: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *Preprint*, arXiv:2306.16244.

Jiho Jin, Woosung Kang, Junho Myung, and Alice Oh. 2025. Social bias benchmark for generation: A comparison of generation and QA-based evaluations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11215–11228, Vienna, Austria. Association for Computational Linguistics.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Mahammed Kamruzzaman, Abdullah Al Monsur, Shrabon Kumar Das, Enamul Hassan, and Gene Louis Kim. 2025. BanStereoSet: A dataset to measure stereotypical social biases in LLMs for bangla. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3450–3460, Vienna, Austria. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2024. The gaps between pre-train and downstream settings in bias evaluation and debiasing. *2401.08511v1*.

Adel Khorramrouz and Sharon Levy. 2025. Characterizing selective refusal bias in large language models. *2510.27087v1*.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When tom eats kimchi: Evaluating cultural bias of multimodal large language models in cultural mixture contexts. *Preprint*, arXiv:2503.16826.

Farnaz Kohankhaki, D. B. Emerson, Jacob-Junqi Tian, Laleh Seyyed-Kalantari, and Faiza Khan Khattak. 2024. Template-based probes are imperfect lenses for counterfactual bias evaluation in llms. *2404.03471v5*.

Abdullatif Koksal, Omer Yalcin, Ahmet Akbiyik, Merve Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *1906.07337v1*.

Tian Lan, Jiang Li, Yemin Wang, Xu Liu, Xiangdong Su, and Guanglai Gao. 2025. F^2Bench: An open-ended fairness evaluation benchmark for LLMs with factuality considerations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2031–2046, Suzhou, China. Association for Computational Linguistics.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8131–8138. AAAI Press.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. FairSteer: Inference time debiasing for LLMs with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Xinru Lin and Luyang Li. 2025. Implicit bias in llms: A survey. *2503.02776v1*.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in pre-trained language models. *Preprint*, arXiv:2406.10130.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628.

Holly J. McCammon, Sandra C. Arch, and Erin M. Bergner. 2014. A radical demand effect: Early us feminists and the married women's property acts. *Social Science History*, 38(1-2):221–250.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *2110.08527v3*.

Maryam Mousavian, Zahra Abbasiantaeb, Mohammad Aliannejadi, and Fabio Crestani. 2025. Towards fair rankings: Leveraging llms for gender bias detection and measurement. *2506.22372v1*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *2008.06460v2*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pretraining data to linguistic phenomena. *Preprint*, arXiv:2501.04662.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. *CoRR*, abs/2406.07243.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Jinhao Pan, Chahat Raj, Ziyu Yao, and Ziwei Zhu. 2025. What's not said still hurts: A description-based evaluation framework for measuring social bias in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1438–1459, Suzhou, China. Association for Computational Linguistics.

Shweta Parihar and Lu Cheng. 2026. Evaluating social bias in rag systems: When external context helps and reasoning hurts. *Preprint*, arXiv:2602.09442.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Nirmalendu Prakash and Roy Ka-Wei Lee. 2023. Layered bias: Interpreting bias in pretrained large language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 284–295, Singapore. Association for Computational Linguistics.

Pooja S. B. Rao, Laxminarayen Nagarajan Venkatesan, Mauro Cherubini, and Dinesh Babu Jayagopi. 2025. Invisible filters: Cultural bias in hiring evaluations using large language models. *Preprint*, arXiv:2508.16673.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Mughees Ur Rehman and Saleha Muzammil. 2025. Towards fairer ai: Multi-agent debiasing of llms with online evidence retrieval.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Abel Salinas, Parth Vipul Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Revealing demographic bias through job recommendations. *2308.02053v2*.

Shalaka Satheesh, Katrin Klug, Katharina Beckh, Hector Allende-Cid, Sebastian Houben, and Teena Hassan. 2025. GG-BBQ: German gender bias benchmark for question answering. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing 2025*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. *Towards a standard for identifying and managing bias in artificial intelligence*.

Enze Shi, Lei Ding, Linglong Kong, and Bei Jiang. 2024. Debiasing with sufficient projection: A general theoretical framework for vector representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thanathip Suenghataiphorn, Narisara Tribuddharat, Pojsakorn Danpanichkul, and Narathorn Kulthamrongsri. 2025. Bias in large language models across clinical applications: A systematic review. *2504.02917v1*.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *Preprint*, arXiv:2504.08863.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *1906.08976v1*.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. *2206.09860v1*.

Raphael Tang, Xinyu Zhang, Jimmy Lin, and Ferhan Ture. 2023. What do llamas really think? revealing preference biases in language model representations. *Preprint*, arXiv:2311.18812.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. 2025. BharatBBQ: A multilingual bias benchmark for question answering in the Indian context. *Transactions of the Association for Computational Linguistics*, 13:1672–1692.

Aarushi Wagh and Saniya Srivastava. 2025. ""the dentist is an involved parent, the bartender is not"": Revealing implicit biases in qa with implicit bbq. *2512.06732v1*.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš, Anne Lauscher, and Simone Paolo Ponzetto. 2024. Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries*, JCDL '21, page 51–60. IEEE Press.

Kevin Wang and 1 others. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *Preprint*, arXiv:2211.00593.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *2112.04359v1*.

Rachel Wetts and Robb Willer. 2019. Who is called by the dog whistle? experimental evidence that racial resentment and political ideology condition responses to racially encoded messages. *Socius: Sociological Research for a Dynamic World*, 5.

LaDale C Winling and Todd M Michney. 2021. The roots of redlining: Academic, governmental, and professional networks in the making of the new deal lending regime. *Journal of American History*, 108(1):42–69.

Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. 2024. Aligning as debiasing: Causality-aware alignment via reinforcement learning with interventional feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. 2025. BiasEdit: Debiasing stereotyped language models via model editing. In *Proceedings of the Workshop on Trustworthy Natural Language Processing 2025*.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2025. Bias mitigation or cultural commonsense? evaluating llms with a japanese dataset. *2509.24468v1*.

Yi Yang, Hanyu Duan, Ahmed Abbasi, John P. Lalor, and Kar Yan Tam. 2025. Bias a-head? analyzing bias in transformer-based language model attention heads. In *Proceedings of the Workshop on Trustworthy Natural Language Processing 2025*.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in LLMs via interpretable neuron editing. *Preprint*, arXiv:2501.14457.

Pardis Sadat Zahraei and Ali Emami. 2025. Translate with care: Addressing gender bias, neutrality, and reasoning in large language model translations. *2506.00748v1*.

Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James R. Foulds, and Shimei Pan. 2025a. GenderAlign: An alignment dataset for mitigating gender bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Tianhui Zhang, Yi Zhou, and Danushka Bollegala. 2025b. Evaluating the effect of retrieval augmentation on social biases.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 15–20.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *2501.02295v4*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *Preprint*, arXiv:1906.04571.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA benchmark for assessing social biases in LLMs for basque, a low-resource language. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics.

A Appendix

A.1 Datasets and Benchmark Resources

Table 4 summarizes datasets and benchmark resources used to evaluate social bias in language models and LLMs. We include both established English-language benchmarks and more recent culturally grounded, multilingual, and open-ended resources. This table supports the main text's argument that benchmarks provide important behavioral evidence, but that their interpretation depends on the social categories, languages, and contexts they operationalize.

A.2 Detection Methods

Table 5 compares major families of bias-detection methods. These methods are useful for showing that a disparity exists, for example through embedding associations, probability scoring, prompt variation, or task-specific audits. In the lifecycle-provenance framework, detection methods provide the first layer of evidence, but they usually need to be combined with source, explanation, or deployment evidence to diagnose where the bias enters or changes form.

A.3 Explanation and Interpretability Methods

Table 6 summarizes methods that localize or explain bias-related behavior inside models. These include attribution, probing, attention analysis, neuron-level methods, and mechanistic interpretability. In our framework, explanation methods act as evidence lenses on the algorithmic-processing stage: they can show where bias becomes visible in representations or computations, but they do not by themselves identify the social or textual source of that bias.

A.4 Mitigation Methods

Table 7 compares mitigation strategies across the LLM lifecycle. These methods intervene at different points: data augmentation changes the training or tuning distribution, representation debiasing acts on model-internal encodings, alignment methods reshape post-training behavior, and inference-time or retrieval-aware methods alter outputs or evidence access. This table supports the paper's central claim that mitigation should be matched to the diagnosed lifecycle stage rather than applied only to the final output.

Dataset	Description	Type	Size	License	Lang	Ref
CrowS-Pairs	Paired stereotype vs. anti-stereotype sentences for likelihood comparison.	Paired sentences	1,508 pairs	CC-BY-SA 4.0	EN	(Nangia et al., 2020)
StereoSet	Contexts with stereotype, anti-stereotype, and unrelated continuations.	Sentence completion	4,316 ctx / ∼17k cand.	CC-BY-SA 4.0	EN	(Nadeem et al., 2021)
UnQover	Underspecified question templates for revealing stereotyped role attributions.	QA / perturbation benchmark	30k+ inst.	n/r	EN	(Li et al., 2020)
WinoBias	Coreference minimal pairs built around occupational gender stereotypes.	Downstream task	3,168 sent.	MIT	EN	(Zhao et al., 2018)
BBQ	Ambiguous and disambiguated QA contexts for measuring stereotype reliance.	QA benchmark	58,492 ex.	n/r	EN	(Parrish et al., 2022)
BOLD	Open-ended prompts for measuring bias in language generation.	Generative benchmark	23,679 prompts	CC-BY-SA 4.0	EN	(Dhamala et al., 2021)
HolisticBias	Descriptor–template prompt bank spanning many demographic groups and harm types.	Prompt benchmark	491,373 prompts	CC-BY-SA 4.0	EN	(Smith et al., 2022)
Multilingual HolisticBias	HolisticBias-style descriptors and templates extended across many languages.	Multilingual prompt benchmark	20,459 sent.	n/r	50 langs	(Costa-jussà et al., 2023)
CBBQ	Chinese culturally adapted BBQ-style benchmark curated with human–AI collaboration.	QA benchmark	100k+ q.	CC-BY-SA 4.0	ZH	(Huang and Xiong, 2023)
KoBBQ	Korean culturally adapted BBQ benchmark with local stereotype categories.	QA benchmark	76,048 samples	CC-BY 4.0	KO	(Jin et al., 2024)
IndiBias	English/Hindi benchmark for Indian-context and intersectional social biases.	Paired + intersectional benchmark	800 pairs + 300 tuples	n/r	EN/HI	(Sahoo et al., 2024)
BharatBBQ	Multilingual Indian-context BBQ-style benchmark covering culturally grounded and intersectional social biases.	Multilingual QA benchmark	49,108 ex. (392,864 expanded)	n/r	8 langs	(Tomar et al., 2025)
MBBQ	Cross-lingual BBQ-style benchmark with shared stereotypes and control items.	Multilingual QA benchmark	20,176 ex.	CC-BY 4.0	4 langs	(Neplenbroek et al., 2024)
BasqBBQ	Basque low-resource BBQ-style benchmark across eight social-bias domains.	QA benchmark	n/r	n/r	EU	(Zulaika and Saralegi, 2025)
GG-BBQ	German gender-bias QA benchmark with group-term and proper-name subsets.	QA benchmark	n/r	CC-BY 4.0	DE	(Satheesh et al., 2025)
BanStereoSet	Bangla stereotype benchmark localized from English-centric resources.	Sentence benchmark	1,194 sent.	n/r	BN	(Kamruzzaman et al., 2025)
DBB	Description-based benchmark targeting subtle, semantic, and contextually hidden bias.	Semantic QA benchmark	∼103,649 pairs	n/r	EN	(Pan et al., 2025)
BBG	Long-form story-generation benchmark adapted from BBQ for open-ended bias evaluation.	Generative benchmark	n/r	n/r	EN/KO	(Jin et al., 2025)
F^2Bench	Open-ended fairness benchmark with factuality-aware generation and reasoning tasks.	Generative benchmark	2,568 inst.	n/r	EN	(Lan et al., 2025)
GFair	Hierarchical target–attribute resource with open-ended statement-organization evaluation.	Open-ended fairness benchmark	20,000 eval. pts.[a]	n/r	EN	(Bi et al., 2025)

Table 4: Datasets and benchmark resources for evaluating social bias in language models and LLMs.

Note: **Size** and **License** are reported from the public paper, repository, or release consulted; *n/r* = not clearly reported. We expanded the table to include recent culturally adapted, multilingual, and open-ended generative resources, since these are central to current LLM-bias evaluation.

[a] The full GFair resource contains 575,134 target–attribute combinations; the paper reports sampling 20,000 evaluation points for experiments.

A.5 Lifecycle Evidence Types

Table 8 synthesizes the kinds of evidence provided by different areas of the surveyed literature. Detection work shows that disparities exist; explanation work localizes where bias-related information appears; mitigation work tests interventions; multilingual and cultural work shows that categories and harms vary across communities; and retrieval or deployment studies show when bias becomes consequential in systems. The table makes explicit that the gap is not a lack of evidence, but that these forms of evidence are rarely connected into a lifecycle diagnosis.

A.6 Checklist Examples

Table 9 applies the proposed reporting checklist to representative benchmark studies. The goal is not to rank these studies, but to show how the checklist can be used in practice. The examples illustrate which checklist items are directly reported, such as observed disparity, lifecycle stage, or transformation pattern, and which items future evaluations may need to make more explicit, such as source evidence or intervention implication.

A.7 LLM-Based Relevance Filtering Prompt

For each candidate paper, we provided the LLM with the paper title, abstract, venue or source when available, and publication year. The model was

instructed to assess whether the paper was centrally relevant to a survey on social bias in NLP and large language models. The exact prompt used for relevance filtering was as follows:

> You are assisting with a systematic scoping review on social bias in NLP and large language models.
>
> Given the title, abstract, venue/source, and publication year of a candidate paper, assess whether the paper is relevant to the review.
>
> A paper should be marked as relevant if it substantially addresses one or more of the following pillars:
>
> 1. **Bias detection**: methods, analyses, or empirical studies that identify, measure, evaluate, or characterize social bias, stereotyping, fairness harms, representational harms, allocational harms, toxicity, hate speech, or discriminatory behavior in NLP systems or language models.
> 2. **Explanation and interpretability**: work that explains, attributes, diagnoses, visualizes, or interprets biased behavior in NLP models, language models, embeddings, datasets, prompts, representations, or generated outputs.
> 3. **Bias mitigation**: methods or interventions that reduce, debias, control, align, detoxify, filter, fine-tune, prompt, evaluate, or otherwise mitigate social bias or fairness harms in NLP systems or language models.
> 4. **Resources and benchmarks**: datasets, benchmarks, evaluation protocols, taxonomies, metrics, challenge sets, multilingual resources, cultural resources, or survey papers related to social bias, fairness, stereotypes, safety, or responsible evaluation in NLP and language models.
>
> Consider the paper relevant only if social bias, fairness, discrimination, stereotyping, toxicity, hate speech, demographic harms, cultural bias, multilingual bias, or related responsible-AI concerns are central to the paper rather than mentioned only in passing.
>
> Exclude papers that are primarily about unrelated NLP topics, general model efficiency, generic interpretability without a bias or fairness component, general safety without social bias relevance, general machine learning fairness outside NLP, or applications where bias is not a central analytical focus.
>
> Return your judgment in the following format:
>
> Decision: Relevant / Not Relevant / Unclear
>
> Pillars: - Bias detection: Yes/No - Explanation: Yes/No - Bias mitigation: Yes/No - Resources or benchmarks: Yes/No
>
> Confidence: High / Medium / Low

A.8 Operationalizing Temporal Provenance

The genealogical view can be operationalized by asking how a particular bias moves across the historical and technical chain. This perspective furthermore points at the contingent nature of the outputs generated by LLM. A bias does not enter an LLM as a single object. It may begin as a social hierarchy, become visible through textual traces, be selected or filtered during corpus construction, be transformed by tokenization and training, and then appear in model outputs, refusals, rankings, or downstream decisions. Table 10 gives representative examples of this chain. The examples are not intended as an exhaustive taxonomy of social bias. Rather, they show how different forms of bias require different genealogical questions, and why the same surface-level disparity may require different forms of evaluation and intervention.

Table 10 illustrates two consequences of a provenance-centered account. First, bias categories are not interchangeable labels. Gender, dialect, migration status, caste, race, religion, and language are shaped by different histories of classification, documentation, exclusion, standardization, and institutional authority. Various forms of exclusion and hierarchization are furthermore deeply interconnected exacerbating the issue further. A single benchmark score can therefore show that a disparity exists while obscuring why that disparity is socially

meaningful and where it entered the system. Second, mitigation depends on the point of transformation. A bias produced by retrieval ranking requires a different intervention from a bias produced by tokenizer fragmentation, historical corpus imbalance, annotation subjectivity, or alignment behavior.

This operational view links the historical account in this section to the pipeline analysis in the next section. If bias is understood as a chain of transformations, then final outputs are only one audit target. Training corpora, filtering rules, tokenizers, annotation guidelines, alignment data, retrieval indices, ranking functions, prompts, interfaces, and downstream decision procedures can each preserve, suppress, amplify, or transform historically produced textual traces. A provenance-aware account of LLM bias therefore asks not only whether a model is biased, but where the bias becomes visible, how it has been transformed, and what an intervention actually changes.

Method	Core idea	Lang	Time	Ctx	Evid.	Strengths	Limitations	Refs
Embedding association tests	Measure association strength between target and attribute terms in word, sentence, or hidden-state spaces.	Med	Med	Low	Rep	Cheap, model-agnostic, and easy to compare across models.	Weak contextual grounding; highly sensitive to lexicon design and prompt wording.	(Caliskan et al., 2017; May et al., 2019)
Paired continuation scoring	Compare model preference for stereotype vs. anti-stereotype continuations under matched prompts.	Med	Low	Med	BB	Simple, widely used, and directly tied to next-token behavior.	Template-sensitive; can conflate bias with fluency or plausibility.	(Nangia et al., 2020; Nadeem et al., 2021)
Counterfactual perturbation tests	Swap names, pronouns, or social-group markers while holding the rest of the context fixed.	High	Med	High	BB	Strong prompt-level control; reveals asymmetric treatment under matched contexts.	Hard to make natural in morphologically rich languages or culture-specific settings.	(Li et al., 2020; Zhao et al., 2018)
QA benchmarks with ambiguity control	Evaluate whether models rely on stereotypes when context is ambiguous, then compare against disambiguated variants.	Med	Low	High	BB	Separates bias from ordinary task success better than raw accuracy alone.	Annotation-heavy; often localized to one culture or language.	(Parrish et al., 2022; Neplenbroek et al., 2024)
Open-ended generation audits	Use identity-conditioned prompts and score outputs for toxicity, sentiment, associations, or refusal patterns.	Med	Med	High	BB	Captures richer harms and interactional behavior than multiple-choice tests.	High variance across prompts, decoding choices, and scoring metrics.	(Dhamala et al., 2021; Smith et al., 2022)
Downstream task audits	Measure disparate errors or stereotyped outputs in tasks such as coreference, MT, retrieval, or ranking.	Med	Low	High	BB	Connects measured bias to concrete task consequences.	Hard to disentangle model bias from task design and training data artifacts.	(Zhao et al., 2018; Attanasio et al., 2023)
Prompt robustness audits	Repeat evaluations across paraphrases, roles, system prompts, and instruction formats.	Med	Med	High	BB	Reveals instability in measured bias and improves ranking reliability.	Expands evaluation cost and requires careful template balancing.	(Hida et al., 2025; Marchiori Manerba et al., 2024)
Multilingual transfer benchmarking	Compare matched prompts or adapted benchmarks across languages and sociocultural settings.	High	Med	High	BB	Exposes transfer distortions and language-specific stereotype patterns.	Translation equivalence is difficult; social categories do not map one-to-one.	(Costa-jussà et al., 2023; Koksal et al., 2023)
Temporal anchoring (emerging)	Re-run prompts with explicit historical anchors such as era labels or present-day vs. historical framing.	Med	High	Med	BB	Best fit for studying temporally evolving stereotypes and social vocabularies.	Few standardized benchmarks; easy to confound with general world-knowledge gaps.	(Borenstein et al., 2023; Hida et al., 2025)

Table 5: Bias-detection methods, evaluated across the survey's cross-cutting lenses.

Legend: **Lang** = support for multilingual or cross-lingual analysis; **Time** = support for temporally anchored or diachronic analysis; **Ctx** = sensitivity to discourse, cultural, or interactional context; **Evid.** = typical evidence level (**Data** = data intervention, **BB** = black-box behavioral evidence, **Rep** = representation-level analysis, **Mech** = mechanistic or causal intervention). High / Med / Low indicate how naturally the method family supports each cross-cutting lens.

Method	Core idea	Lang	Time	Ctx	Evid.	Strengths	Limitations	Refs
Token attribution / saliency	Attribute biased outputs to influential input tokens or prompt fragments using gradients or related scores.	Med	Low	Med	Rep	Easy to attach to black-box evaluations and useful for prompt analysis.	Usually correlational; unstable across attribution choices and reparameterizations.	(Attanasio et al., 2023; Gamboa and Lee, 2024)
Logit lens / layerwise decoding	Decode intermediate residual states to see where stereotyped preferences emerge across layers.	Med	Med	Med	Rep	Good first-pass localization of when bias signals sharpen during computation.	Still correlational and sensitive to the decoding probe used at each layer.	(Prakash and Lee, 2023)
Hidden-state probing	Train probes to recover protected attributes or stereotype signals from internal representations.	High	Med	Med	Rep	Scales well across models, layers, and languages.	Probe success does not by itself establish causal use by the model.	(Koksal et al., 2023; Marchiori Manerba et al., 2024)
Attention-head localization	Identify heads whose attention patterns systematically track biased associations or demographic cues.	Med	Med	Med	Rep/Mech	Useful bridge between representation analysis and intervention.	Attention is not a complete explanation and may miss distributed effects.	(Yang et al., 2025; Adiga et al., 2024)
Neuron-level analysis	Localize neurons associated with biased features, stereotypes, or gendered predictions.	Med	Med	Med	Rep/Mech	Supports targeted ablation or editing experiments.	Localization can be model-specific and effects are often distributed rather than sparse.	(Liu et al., 2024; Yu and Ananiadou, 2025)
Causal interventions / patching	Intervene on activations, heads, or neurons to test whether localized components causally mediate biased outputs.	Med	Med	Med	Mech	Stronger causal evidence than passive observation alone.	Experiment design is labor-intensive and usually limited to open-weight models.	(Adiga et al., 2024; Xu et al., 2025)
Subspace analysis	Characterize protected-attribute directions, stereotype subspaces, or separability in hidden representations.	Med	Low	Low	Rep	Provides compact geometric summaries of internal bias structure.	Often oversimplifies nonlinear and context-dependent bias effects.	(Ravfogel et al., 2020; Shi et al., 2024)
Sparse autoencoders	Decompose activations into sparse, more human-interpretable features that may isolate bias-related concepts.	Med	Med	Med	Mech	Promising route to feature-level interpretation beyond neurons or heads.	Application to social bias is still emerging and feature semantics remain imperfect.	(Bricken et al., 2023; Chandna et al., 2025)
Circuit discovery	Trace multi-component pathways that transform prompts into biased intermediate predictions and outputs.	Low	Med	Med	Mech	Most direct route to explaining where and how bias is computed inside a model.	Technically demanding and still rare in applied social-bias work.	(Wang et al., 2023; Chandna et al., 2025)

Table 6: Bias-explanation methods, evaluated across the survey's cross-cutting lenses.

Legend: **Lang** = support for multilingual or cross-lingual analysis; **Time** = support for temporally anchored or diachronic analysis; **Ctx** = sensitivity to discourse, cultural, or interactional context; **Evid.** = typical evidence level (**Data** = data intervention, **BB** = black-box behavioral evidence, **Rep** = representation-level analysis, **Mech** = mechanistic or causal intervention). High / Med / Low indicate how naturally the method family supports each cross-cutting lens.

Method	Core idea	Lang	Time	Ctx	Evid.	Strengths	Limitations	Refs
Counterfactual data augmentation	Augment training data with identity-swapped or norm-balanced examples.	Med	Med	High	Data	Easy to integrate into fine-tuning pipelines; often improves robustness to lexical swaps.	Can create unnatural examples and may miss culture-specific or historical asymmetries.	(Zmigrod et al., 2019; Han et al., 2024)
Projection / nullspace removal	Estimate protected-attribute directions in representation space and remove or suppress them.	Low	Low	Low	Rep	Interpretable and lightweight once attribute directions are identified.	May remove useful signal; struggles with nonlinear, contextual, or multilingual bias.	(Ravfogel et al., 2020; Shi et al., 2024)
Adversarial / invariance training	Train models so protected attributes are harder to recover from representations or outputs.	Med	Low	Med	Rep	Targets latent encodings directly and can generalize beyond lexical templates.	Optimization is unstable and fairness gains may not transfer across tasks or languages.	(Dayanik and Pado, 2021; Fleisig and Fellbaum, 2022)
Alignment / debiasing tuning	Use instruction tuning, RL, or preference optimization to steer models toward less biased behavior.	Med	Med	Med	BB	Compatible with modern chat models and behavioral objectives.	Can become over-refusal or hide rather than remove underlying associations.	(Xia et al., 2024; Zhang et al., 2025a)
Self-debiasing decoding	Use self-diagnosis, control prompts, or decoding constraints to reduce biased generations at inference time.	High	Med	High	BB	No retraining required and easy to deploy across proprietary models.	Often brittle; gains depend heavily on prompt format and decoding settings.	(Schick et al., 2021; Gallegos et al., 2025)
Retrieval-grounded generation	Condition outputs on external evidence or retrieved documents to counter unsupported stereotypes.	High	High	High	BB	Naturally supports temporal updates and provenance-aware responses.	Can inherit bias from retrieved corpora and depends on retrieval quality.	(Rehman and Muzammil, 2025; Zhang et al., 2025b)
Activation steering	Steer activations or attention patterns at inference time using learned bias directions or controller vectors.	Med	Med	Med	Mech	Fine-grained control with minimal additional training.	Transfer across models and languages is uncertain; side effects can be hard to predict.	(Li et al., 2025; Adiga et al., 2024)
Model / neuron editing	Edit localized parameters, neurons, or facts associated with stereotyped behavior.	Med	Low	Low	Mech	Potentially targeted and interpretable when localization is reliable.	Edits may be brittle, non-local, or erase valid culturally grounded distinctions.	(Xu et al., 2025; Yu and Ananiadou, 2025)

Table 7: Bias-mitigation methods, evaluated across the survey's cross-cutting lenses.

Legend: **Lang** = support for multilingual or cross-lingual analysis; **Time** = support for temporally anchored or diachronic analysis; **Ctx** = sensitivity to discourse, cultural, or interactional context; **Evid.** = typical evidence level (**Data** = data intervention, **BB** = black-box behavioral evidence, **Rep** = representation-level analysis, **Mech** = mechanistic or causal intervention). High / Med / Low indicate how naturally the method family supports each cross-cutting lens.

Literature area	Typical evidence	What it shows	Provenance limitation	Representative refs.
Detection and benchmarks	Output probabilities, QA accuracy, stereotype preferences, toxicity, sentiment, refusal, or generation differences across groups	Whether model behavior differs systematically across social cues or groups	Usually identifies disparity, but not whether the source is data, tokenization, alignment, retrieval, or deployment	Caliskan et al. (2017); May et al. (2019); Nangia et al. (2020); Nadeem et al. (2021); Parrish et al. (2022); Dhamala et al. (2021)
Explanation and interpretability	Attribution, probing, logit-lens analysis, attention-head localization, neuron analysis, activation patching, or circuits	Where bias-related information appears or becomes active inside the model	Often localizes bias without linking it back to corpus history or source documents	Attanasio et al. (2023); Koksal et al. (2023); Prakash and Lee (2023); Yang et al. (2025); Liu et al. (2024); Chandna et al. (2025)
Mitigation and debiasing	Counterfactual data augmentation, adversarial training, projection, alignment tuning, steering, self-debiasing, editing, or retrieval-aware mitigation	Whether an intervention changes biased behavior at a specific stage	May suppress visible bias while leaving latent associations, proxy discrimination, or deployment harms intact	Zmigrod et al. (2019); Ravfogel et al. (2020); Schick et al. (2021); Xia et al. (2024); Li et al. (2025); Xu et al. (2025)
Multilingual and cultural evaluation	Localized benchmarks, multilingual stereotype resources, translated or culturally adapted QA tasks	How stereotypes, protected categories, and harms vary across languages and communities	Often captures cultural variation without tracing the historical processes that produced the textual evidence	Huang and Xiong (2023); Jin et al. (2024); Sahoo et al. (2024); Tomar et al. (2025); Neplenbroek et al. (2024); Gamboa et al. (2025)
Alignment and refusal	Explicit vs. implicit bias, refusal rates, safety behavior, helpfulness, and group-dependent moderation	How safety tuning changes visible bias, refusal, and helpfulness	Can reveal suppression or over-refusal without identifying whether the source is data, annotation, policy, or prompt design	Bai et al. (2024); Zhao et al. (2025); Lin and Li (2025); Khorramrouz and Levy (2025); Pan et al. (2025)
Retrieval and deployment	RAG audits, ranking audits, resume screening, clinical use, recommendation, moderation, or public-sector applications	How bias can be reintroduced or amplified by external sources and workflows	Difficult to separate model behavior from retrieval quality, interface design, institutional context, and task framing	Zhang et al. (2025b); Salinas et al. (2023); Glazko et al. (2024); Rao et al. (2025); Mousavian et al. (2025); Suenghataiphorn et al. (2025)
Historical and diachronic work	Historical corpora, temporal anchors, archival text analysis, value drift, changing social categories, and diachronic stereotype measurement	Why source date, genre, and historical context matter for interpreting model behavior	Still underdeveloped in LLM evaluation; evidence is often case-specific rather than standardized	Borenstein et al. (2023); Winling and Michney (2021); McCammon et al. (2014); Garg et al. (2018); Walter et al. (2024); Bhatia et al. (2025b)

Table 8: Types of lifecycle-provenance evidence in the surveyed literature. Existing work provides partial evidence for different parts of the bias genealogy. Detection work shows that disparities exist; explanation work localizes where bias-related information appears; mitigation work tests interventions; multilingual and cultural work shows that bias categories vary across communities; alignment and refusal studies show how safety tuning can reshape visible behavior; retrieval and deployment studies show how bias becomes consequential in systems; and historically situated work shows why source context matters. The gap is that these forms of evidence are rarely connected.

Study	What it shows	Disparity	Source	Stage	Pattern	Intervention
CrowS-Pairs / StereoSet (Nangia et al., 2020; Nadeem et al., 2021)	Paired sentence and continuation benchmarks reveal stereotyped preferences in pretrained language models, but primarily as behavioral evidence.	Yes	No	Yes	Yes	No
BBQ (Parrish et al., 2022)	QA models rely on stereotypes in underinformative contexts and can perform differently when correct answers align or conflict with social stereotypes.	Yes	No	Yes	Yes	No
BOLD / HolisticBias (Dhamala et al., 2021; Smith et al., 2022)	Open-ended generation benchmarks expose toxicity, sentiment, regard, and descriptor-conditioned harms beyond multiple-choice settings.	Yes	No	Yes	Yes	Partial
BharatBBQ / KoBBQ (Tomar et al., 2025; Jin et al., 2024)	Culturally grounded benchmarks show that bias categories, stereotypes, and harms vary across languages and sociocultural settings.	Yes	Partial	Yes	Yes	No
DBB / BBG (Pan et al., 2025; Jin et al., 2025)	Description-based and long-form generation evaluations show that bias can persist when explicit identity terms are absent or when evaluation moves beyond QA.	Yes	No	Partial	Yes	No
Historical bias (Borenstein et al., 2023)	Historical-document analysis links intersectional bias to archival sources and shows why source date, genre, and documentation context matter.	Yes	Yes	Partial	Yes	Partial
Dialect bias (Hofmann et al., 2024)	Dialect-marked language can trigger covertly discriminatory model behavior, suggesting that social bias may be transformed through linguistic proxies.	Yes	Partial	Yes	Yes	Partial
RAG bias (Zhang et al., 2025b)	Retrieval augmentation can reduce or reintroduce social bias depending on retrieved evidence, ranking, and source composition.	Yes	Partial	Yes	Yes	Yes
Hiring deployment (Rao et al., 2025)	Task-level audits show that cultural bias can become consequential in hiring evaluations, where model outputs affect downstream decisions.	Yes	Partial	Yes	Yes	Yes
Alignment/refusal (Xia et al., 2024)	Alignment can reduce visible biased outputs, but evaluation must test whether it removes bias or shifts it into refusal, helpfulness, or latent associations.	Yes	No	Yes	Yes	Yes

Table 9: Representative studies viewed through the proposed lifecycle reporting checklist. The columns correspond to five reporting components: observed disparity, plausible source evidence, lifecycle stage, transformation pattern, and intervention implication. "Yes" indicates that the study substantially reports the item; "Partial" indicates that the item is addressed indirectly or for a limited setting; "No" indicates that the item is not a main reporting focus. The table is not intended to rank studies, but to show that existing work often provides strong local evidence while leaving gaps in full lifecycle diagnosis.

Bias case	Historical or social source	Textual trace	Model transformation	Evaluation limit	Lifecycle-matched response
Dialect bias	Racialized language ideologies and standard-language hierarchies	Dialect-marked text may be treated as informal, toxic, noisy, or low quality	Filtering, tokenization, and alignment may penalize dialectal forms or treat them as less professional	Output tests may reveal negative behavior without identifying whether the source is filtering, tokenization, annotation, or safety tuning	Audit filtering decisions, evaluate dialect robustness, and use dialect-aware annotation and community-informed evaluation
Gender–occupation bias	Gendered labor-market segregation and historical occupational roles	Gendered terms may co-occur with occupations in books, web text, resumes, job ads, and news	Pretraining may strengthen predictive associations between gender and occupation	Paired examples can reveal stereotyped preferences but do not show whether the relevant intervention is data, representation, decoding, or task design	Use corpus audits, counterfactual evaluation, representation analysis, and task-specific mitigation in hiring or recommendation settings
Historical immigration bias	Immigration regimes, nationalist discourse, and racialized policy debates	Older news, legal, and policy texts may describe migrants through threat, burden, illegality, or assimilation frames	Pretraining or retrieval-augmented generation may reproduce historically situated language as neutral present-day description	A generated answer may appear factual while importing outdated, source-specific, or politically situated framing	Attach temporal and source metadata, distinguish historical evidence from present-day claims, and audit retrieval sources

Table 10: Operationalizing lifecycle provenance for representative social-bias cases in LLMs. The table shows how historical and social hierarchies become textual traces, how LLM pipelines transform those traces, and why evaluation and mitigation should be matched to the relevant lifecycle stage rather than only to the final output.