Available Online: 13 January 2026 DOI: 10.54254/2753-7080/2026.31302

Linguistic and dialectal bias in large language models: a review of evidence, mechanisms, and implications (2018–2025)

Eric Zhao

Gilman School, Baltimore, USA

zhaoeric2009@gmail.com

Abstract. Large Language Models (LLMs) increasingly mediate communication, evaluation, and decision-making across social and institutional domains. While prior work has documented various forms of bias in language technologies, linguistic variation has received comparatively less integrated treatment despite its deep sociolinguistic significance. This review synthesizes recent research on how LLMs reflect and reinforce linguistic bias toward nonstandard dialects, with a focus on African American English and other marginalized varieties. Drawing on studies spanning natural language understanding, reasoning, speech recognition, and conversational systems, we show that LLMs exhibit consistent performance disparities when processing dialectal inputs, even when semantic content is held constant. Beyond accuracy degradation, emerging evidence demonstrates that dialect functions as a social signal that triggers differential judgment, stereotyping, and covertly racist decision-making in high-stakes contexts. We organize the literature around four themes: dialect-based performance gaps, dialect-triggered social evaluation, measurement frameworks for linguistic bias, and proposed mitigation strategies. While recent benchmarks and adaptation methods improve robustness, we argue that many approaches treat dialect bias as a technical deficiency rather than a sociotechnical problem rooted in language ideology and power. By synthesizing research across technical, social, and evaluative dimensions, this review identifies methodological gaps and proposes directions for developing linguistically and culturally robust LLMs.

Keywords: linguistic bias, dialectal variation, large language models, raciolinguistic bias, AI fairness

1. Introduction

Large Language Models (LLMs) have rapidly become foundational components of modern sociotechnical systems, shaping how language is processed, evaluated, and acted upon in various domains, including education, employment, healthcare, and criminal justice. These systems increasingly mediate access to resources and opportunities, often under the assumption that automated language processing is neutral or objective. However, a growing body of research demonstrates that LLMs systematically encode and reproduce social biases present in human language data, raising concerns about fairness, equity, and representational harm [1]. Because language itself is a site of social meaning and power, biases in language technologies have

uniquely far-reaching consequences. A critical yet underexamined dimension of AI bias is linguistic variation, particularly dialectal variation. Sociolinguistic research has long established that nonstandard dialects such as African American English (AAE) and other regional or ethnic varieties are linguistically systematic and rule-governed, yet socially stigmatized. When language technologies privilege standardized varieties as the implicit norm, they risk marginalizing speakers whose language practices differ from institutional standards. Recent work shows that LLMs perform significantly worse on nonstandard dialect inputs across a wide range of tasks, including reasoning, sentiment analysis, natural language inference, and speech recognition [2-5]. Importantly, these disparities persist even when semantic content is held constant, suggesting that dialect alone can trigger degraded system behavior. Beyond performance gaps, emerging evidence indicates that LLMs also engage in differential social evaluation based on dialect. Hofmann et al. demonstrate that LLMs generate covertly racist decisions in contexts such as hiring and sentencing when personas are described using AAE rather than Standard American English (SAE), even without explicit racial markers [6]. Similarly, Fleisig et al. show that conversational agents respond more condescendingly, stereotypically, and dismissively to nonstandard dialect prompts [7]. These findings indicate that LLMs not only reflect linguistic bias but also actively reinforce existing social hierarchies through automated judgment and interaction. Given the rapid deployment of LLMs in high-stakes contexts, a comprehensive review of linguistic and dialect bias is both timely and necessary. While prior surveys have addressed fairness and bias in LLMs broadly [1], there remains a need for a focused synthesis that foregrounds linguistic theory and sociolinguistic insight. This review addresses the following research questions: 1. How do LLMs encode and exhibit bias toward nonstandard dialects? 2. What methodological approaches have been used to measure dialect-based disparities in language technologies? 3. To what extent do existing mitigation strategies address the sociolinguistic roots of linguistic bias? The remainder of this paper is structured as follows. Section 2 outlines the theoretical background, introducing key sociolinguistic and computational frameworks for understanding linguistic bias. Subsequent sections review empirical evidence of dialect bias in LLMs, examine evaluation and mitigation approaches, and discuss open challenges and directions for future research.

2. Theoretical background

Understanding linguistic bias in Large Language Models (LLMs) requires grounding in both sociolinguistic theory and computational perspectives on bias and fairness. Central to this discussion is the concept of linguistic variation: the idea that language naturally varies across social groups, regions, and contexts. Sociolinguistics has long rejected the notion that so-called standard language varieties are linguistically superior. Foundational work demonstrates that dialects such as African American English (AAE) are systematic, rule-governed linguistic varieties with coherent grammatical, phonological, and pragmatic structures [8, 9]. Despite this, institutional norms continue to privilege standardized varieties of English, positioning nonstandard dialects as deficient or incorrect. This hierarchy is sustained through standard language ideology, which Lippi-Green defines as the belief that one homogeneous, standardized language variety is inherently superior and should function as the norm in public life [10]. As Lippi-Green argues, this ideology is not merely descriptive but disciplinary: it enables institutions such as schools, courts, and workplaces to subordinate speakers of nonstandard dialects while framing such subordination as neutral or merit-based. These ideologies produce both material and symbolic inequalities, shaping how speakers are evaluated and treated across social domains. Within this context, the concept of raciolinguistic bias is particularly relevant. Flores and Rosa argue that linguistic features are often racialized, such that speakers are perceived and evaluated through racialized listening practices regardless of their actual linguistic competence

or identity [11]. From this perspective, stigma attaches not to language structure itself, but to socially constructed associations between language and race. Hofmann et al. empirically demonstrate the operation of this theory in LLMs, showing that dialect functions as a proxy for race in automated decision-making systems [6]. Their findings align with sociolinguistic research emphasizing that linguistic prejudice frequently operates covertly, making it difficult to detect through surface-level or explicitly racialized analysis.

From a computational standpoint, bias in NLP systems is typically conceptualized as systematic performance disparity or representational harm. Gallegos identifies training data imbalances, modeling choices, and evaluation practices as primary sources of bias in LLMs [1]. Because these models are trained on large-scale corpora dominated by standardized varieties of English, dialectal language is often underrepresented or implicitly treated as noise. This results in what Kantharuban et al. term the dialect gap: persistent underperformance for nonstandard varieties across tasks, languages, and modalities [5]. Evaluation frameworks such as VALUE and Multi-VALUE operationalize this gap by applying controlled dialect transformations to benchmark datasets, enabling researchers to isolate the effects of linguistic variation on model performance [2, 12]. While these frameworks have been instrumental in quantifying dialect-based disparities, they also reflect an ongoing debate within the field: whether dialect bias should be treated primarily as a robustness issue or as a manifestation of deeper sociotechnical inequities. Mitigation approaches such as Task-Agnostic Dialect Adapters (TADA) demonstrate that architectural interventions can reduce performance gaps, but they do not fully address the social meanings and ideological hierarchies attached to dialect use [13]. A final theoretical distinction concerns reflection versus reinforcement of bias. LLMs reflect bias insofar as they inherit skewed representations from human-generated data shaped by standard language ideology and raciolinguistic hierarchies. They reinforce bias when their outputs influence real-world decisions or user behavior, as documented in hiring, sentencing, and conversational interaction contexts [6, 7]. This feedback loop positions LLMs not merely as passive mirrors of linguistic inequality, but as active participants in the reproduction of linguistic hierarchies. Taken together, these frameworks demonstrate that linguistic bias in AI is not solely a technical flaw, but a sociotechnical problem rooted in language ideology, racialization, and institutional power. Addressing it requires integrating sociolinguistic and raciolinguistic theory directly into model evaluation, design, and deployment.

3. Literature review

Research on linguistic bias in language technologies has expanded rapidly in recent years, reflecting broader concerns about fairness, representation, and social harm in artificial intelligence. This literature review synthesizes prior work on how Large Language Models (LLMs) and related language technologies encode, exhibit, and potentially mitigate bias toward nonstandard dialects. The reviewed studies can be grouped into four interrelated themes: (1) dialect-based performance disparities, (2) dialect as a trigger for social judgment and discrimination, (3) measurement frameworks for linguistic bias, and (4) proposed mitigation and adaptation strategies.

3.1. Dialect-based performance disparities in language technologies

A substantial body of work documents systematic performance gaps between standard and nonstandard dialects across NLP tasks. Early large-scale evidence is provided by Ziems et al., who introduce the VALUE benchmark to measure dialect disparity in Natural Language Understanding (NLU) systems [13]. By applying controlled African American English (AAE) transformations to benchmark datasets across sentiment analysis, natural language inference, and paraphrase detection, they demonstrate consistent accuracy drops for AAE

variants across BERT- and T5-style models. These results establish that dialect features alone—independent of semantic content—can cause model failure. Subsequent studies extend these findings across domains, tasks, and languages. Lin et al. examine dialect fairness in reasoning tasks, including mathematics, logic, and algorithmic reasoning, using parallel query pairs rewritten into AAVE by native speakers [3]. Across multiple LLM families (GPT-4, Claude, LLaMA, Mistral, Phi), the authors find significant performance degradation when models process AAVE inputs compared to Standard English equivalents. Importantly, this work demonstrates that dialect bias is not limited to surface-level tasks such as sentiment analysis but affects higher-order reasoning capabilities, challenging assumptions that reasoning performance is dialect-invariant. Beyond text-based tasks, Markl shows that linguistic bias extends to speech technologies [4]. Analyzing commercial Automatic Speech Recognition (ASR) systems trained on British English, Markl finds that speakers of stigmatized or regional dialects experience substantially higher word error rates. These errors correlate with accent features rather than audio quality, reinforcing existing inequalities in speech-mediated systems. Similarly, Kantharuban et al. broaden the scope cross-linguistically, documenting a persistent dialect gap across multiple languages and technologies, including machine translation and ASR [5]. While economic and data-resource disparities explain part of the gap, the authors emphasize that linguistic structure and sociolinguistic marginalization also play a critical role. Together, these studies establish dialect bias as a systemic and cross-modal phenomenon, rather than an artifact of specific datasets or architectures. However, much of this literature focuses on accuracy-based metrics, leaving open questions about how performance disparities translate into social harm.

3.2. Dialect as a trigger for social judgment and discrimination

More recent research moves beyond performance metrics to examine how LLMs make socially consequential judgments based on dialect. Hofmann et al. represent a landmark contribution in this area [6]. Using matched-guise experiments in which prompts differ only by dialectal features, the authors show that LLMs assign lower-prestige jobs, higher criminal risk, and harsher sentences to personas using AAE compared to those using standard varieties. Crucially, race is never explicitly mentioned, demonstrating that dialect alone functions as a proxy for racialized social categories. The authors characterize this as covert raciolinguistic bias, revealing how deeply embedded linguistic prejudice is within LLM decision-making. Fleisig et al. complement these findings by analyzing conversational interactions with ChatGPT across ten English dialects [7]. Their results indicate that responses to nonstandard dialects exhibit significantly higher levels of stereotyping, demeaning content, and condescension, alongside reduced comprehension. Unlike decision-making tasks, this work highlights interactional harm, showing how everyday user experiences with LLMs can reinforce linguistic stigma. Notably, the model often corrects or standardizes dialectal input, implicitly positioning standard language as normative and nonstandard varieties as deficient. These biases are not merely technical shortcomings—they have tangible, real-world consequences. By systematically disadvantaging speakers of nonstandard dialects, LLMs can restrict access to educational resources, influence hiring and promotion decisions, and shape legal outcomes, thereby reproducing and reinforcing existing social hierarchies. These studies mark an important shift in the literature: dialect bias is no longer framed solely as a robustness or engineering problem, but as a mechanism through which LLMs reproduce social hierarchies. However, this line of research remains relatively small, and further work is needed to examine long-term effects on user behavior, such as self-censorship or linguistic accommodation.

3.3. Frameworks for measuring linguistic bias

To systematically study dialect bias, researchers have developed specialized benchmarks and evaluation frameworks. The VALUE benchmark [2] provides one of the earliest structured approaches, enabling controlled comparisons between standard and dialectal variants across multiple tasks. Building on this, Ziems et al. introduce Multi-VALUE, which expands coverage to 50 English dialects using a rule-based transformation system encompassing 189 linguistic features [13]. By stress-testing models on synthetically generated dialectal data, Multi-VALUE reveals substantial performance degradation for wide nonstandard varieties. A key contribution of Multi-VALUE is its validation against native-speaker gold data, demonstrating that synthetic transformations can capture linguistically meaningful variation. This counters the criticism that synthetic dialect data lacks authenticity. Nevertheless, the reliance on rule-based transformations raises ongoing debates about whether such methods adequately represent the pragmatic, stylistic, and contextual dimensions of real-world dialect use. At a broader level, Gallegos situates dialect bias within a comprehensive taxonomy of bias in LLMs, distinguishing between representational harm, allocational harm, and performance disparity [1]. This survey highlights how dialect bias intersects with other forms of social bias and emphasizes that evaluation practices often normalize standard language, rendering dialectal variation invisible. Similarly, Hovy and Yeh et al. propose a roadmap for culturally aware and adapted NLP, arguing that dialect should be understood as one dimension of culture rather than a peripheral linguistic feature [14]. These frameworks have significantly advanced methodological rigor, yet they also reveal a tension in the literature: whether linguistic bias should be operationalized primarily through technical benchmarks or through socially grounded evaluation criteria.

3.4. Mitigation and adaptation strategies

In response to documented disparities, researchers have proposed various mitigation strategies aimed at improving dialect robustness. Ziems et al. introduce Task-Agnostic Dialect Adapters (TADA), a lightweight architectural approach that separates dialectal adaptation from task-specific learning [12]. By composing dialect adapters with task adapters, TADA improves performance on non-SAE variants across multiple GLUE tasks without requiring retraining for each dialect-task combination [12]. This scalability makes TADA an attractive technical solution. Similarly, Multi-VALUE-based augmentation shows that training or fine-tuning models on dialect-augmented data can reduce performance gaps. These approaches suggest that some dialect bias can be mitigated through representation alignment and data diversification. However, Kantharuban et al. caution that increasing data alone does not fully resolve dialect gaps, particularly when sociolinguistic marginalization shapes language use in ways that models struggle to capture [5]. Moreover, mitigation-focused work often treats dialect bias as a purely technical deficiency, overlooking interactional and decision-based harms documented by Hofmann et al. and Fleisig et al. [6, 7]. While adapters may improve accuracy, they do not necessarily prevent stereotyping, condescension, or discriminatory judgment. This limitation has concrete sociotechnical implications. Even models that perform well on benchmarks may continue to disadvantage speakers of nonstandard dialects in everyday interactions, reinforcing barriers to education, employment, and legal equity. The persistence of these harms underscores that technical fixes alone are insufficient: effective mitigation requires approaches that recognize the social context in which language is produced and evaluated, as well as the real-world consequences for marginalized communities.

3.5. Synthesis and research gaps

Across the reviewed literature, several themes emerge. First, dialect bias is pervasive across models, tasks, and languages, indicating that it is structurally embedded in contemporary language technologies. Second, dialect

functions not only as a linguistic variable but also as a social signal that triggers evaluative and discriminatory behavior. Third, while measurement frameworks have become increasingly sophisticated, they remain largely performance-centric. Significant gaps remain. Few studies integrate sociolinguistic theory directly into model design or evaluation. Interactional harms and downstream behavioral effects are underexplored, and mitigation strategies often fail to address the social meanings attached to dialect use. These gaps motivate the need for interdisciplinary approaches that treat linguistic variation as both a technical and social phenomenon. This review positions linguistic bias in LLMs as a sociotechnical problem requiring collaboration between computational linguistics, sociolinguistics, and AI ethics, setting the stage for future research directions.

4. Discussion

Across the reviewed literature, a striking consistency emerges: Large Language Models systematically disadvantage nonstandard dialects, particularly African American English, across tasks, modalities, and application domains. Studies focused on performance metrics such as VALUE, Lin et al., and Kantharuban et al. consistently demonstrate accuracy degradation when models process dialectal inputs that are semantically equivalent to standard-language prompts [2, 3, 5]. This pattern holds for both surface-level tasks and higher-order reasoning, indicating that dialect bias is not merely a preprocessing issue but reflects deeper representational limitations within models. At the same time, research examining social evaluation and interaction reveals that linguistic bias extends beyond performance disparities. Hofmann et al. and Fleisig et al. show that dialect triggers differential judgments, condescension, and stereotyping, even in the absence of explicit racial or social markers [6, 7]. Notably, Hofmann et al. find that LLMs exhibit stronger bias in covert dialect-based evaluations than in overtly racialized scenarios, suggesting that linguistic prejudice may be more deeply embedded and harder to detect. Together, these studies converge on the conclusion that dialect functions as a powerful social signal within LLMs, shaping both outputs and decisions. Despite this convergence, important differences in methodological focus lead to partial blind spots. Performance-oriented studies often frame dialect bias as a robustness or generalization problem, implicitly treating standard language as the evaluation baseline. In contrast, sociotechnically oriented studies foreground harm, dignity, and discriminatory impact, but typically examine fewer tasks or models. This methodological divergence creates a gap between technical mitigation work and analyses of real-world consequences. For instance, while Task-Agnostic Dialect Adapters and dialect-augmented training demonstrate measurable performance improvements, it remains unclear whether such approaches reduce stereotyping or biased judgment in interactional contexts [13].

Another point of tension concerns the use of synthetic dialect transformations. Frameworks such as VALUE and Multi-VALUE rely on controlled rule-based mappings to isolate dialect effects, offering strong internal validity and reproducibility [2, 13]. However, critics note that synthetic data may fail to capture pragmatic nuance, code-switching, and stylistic variability present in natural dialect use. While Multi-VALUE partially addresses this concern by validating transformations against native-speaker data, the broader question of ecological validity remains unresolved. This tension reflects a broader trade-off in NLP research between experimental control and sociolinguistic realism. Cross-linguistic work further complicates the picture. Kantharuban et al. show that dialect gaps persist across languages and technologies, but also identify correlations with economic and resource inequalities [5]. This suggests that dialect bias arises from an interaction between linguistic structure, social marginalization, and data availability. However, resource disparities alone cannot fully explain observed patterns, as performance gaps persist even in relatively well-resourced languages. This finding challenges simplistic data-centric explanations and underscores the need for

theory-driven approaches. Collectively, these studies reveal a troubling trend: as LLMs become more deeply integrated into institutional decision-making, linguistic bias risks being scaled and automated. While recent mitigation strategies offer promising technical tools, the literature suggests that without sociolinguistic grounding, such solutions may address symptoms rather than underlying causes. The field thus faces a critical choice between treating dialect bias as a narrow engineering problem or confronting it as a sociotechnical issue tied to language ideology and power. Importantly, the implications extend beyond academic analysis. Unchecked dialect bias poses institutional risks, potentially affecting admissions, hiring, legal adjudication, and other high-stakes domains. Policymakers and organizations deploying LLMs must therefore consider not only technical performance but also the broader equity consequences of these systems. Regulatory guidance, auditing practices, and inclusive evaluation standards are necessary to ensure that AI deployment does not reinforce existing social hierarchies. Recognizing and addressing dialect bias is therefore not just about improving model accuracy but is a core issue of institutional accountability and social justice.

5. Conclusion

This review demonstrates that Large Language Models systematically encode and amplify linguistic bias against nonstandard dialects, producing both technical performance disparities and socially consequential harms. Synthesizing research from computational linguistics, sociolinguistics, and AI fairness, the paper demonstrates that dialect bias is a pervasive and structural feature of contemporary language technologies. Empirical studies consistently show that LLMs underperform on dialectal inputs, engage in differential social judgment, and reproduce stigmatizing language ideologies that privilege standardized varieties. The reviewed literature contributes several key insights. First, dialect bias is not limited to surface-level tasks or isolated systems; it affects reasoning, interaction, speech recognition, and high-stakes decision-making. Second, dialect operates as a social signal within LLMs, often serving as a proxy for racialized or marginalized identities. Third, while recent evaluation frameworks and adaptation techniques have improved the measurement and mitigation of performance disparities, they frequently fail to address interactional harm and discriminatory judgment. These findings point to several directions for future research. Methodologically, there is a need to integrate sociolinguistic theory more directly into model evaluation and design, moving beyond accuracy-based metrics to consider dignity, respect, and user experience. Empirically, longitudinal and user-centered studies are needed to understand how linguistic bias affects behavior, trust, and access to resources. Technically, mitigation strategies should be evaluated not only for robustness but also for their impact on social outcomes. Ultimately, addressing linguistic bias in LLMs requires rethinking how language variation is represented, valued, and operationalized in AI systems. Treating dialect as legitimate linguistic structure rather than deviation from a norm is not only a linguistic imperative but a prerequisite for equitable and responsible language technology.

References

[1] Gallegos, I. O. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics, 50*(3), 1097–1153.https: //doi.org/10.1162/tacl_a_00760

[2] Ziems, C., Chen, J., Harris, C., Anderson, J., & Yang, D. (2022). VALUE: Understanding dialect disparity in natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*(pp. 3701–3720).https: //doi.org/10.18653/v1/2022.acl-long.258

[3] Lin, F., et al. (2024). *Assessing dialect fairness and robustness of large language models in reasoning tasks*. arXiv.https: //arxiv.org/abs/2410.11005

[4] Markl, N. (2022). Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*.

[5] Kantharuban, A., Vulić, I., & Korhonen, A. (2023). Quantifying the dialect gap and its correlates across language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

[6] Hofmann, V., et al. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*.https: //doi.org/10.1038/s41586-024-07856-5

[7] Fleisig, D., et al. (2024). *Linguistic bias in ChatGPT*. arXiv.https: //arxiv.org/abs/2406.08818

[8] Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press.

[9] Rickford, J. R. (1999). *African American Vernacular English: Features, evolution, educational implications*. Blackwell.

[10] Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United States* (2nd ed.). Routledge.

[11] Flores, N., & Rosa, J. (2015). Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review, 85*(2), 149–171.https: //doi.org/10.17763/0017-8055.85.2.149

[12] Ziems, C., et al. (2023). Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*(pp. 741–758).https: //doi.org/10.18653/v1/2023.acl-long.44

[13] Ziems, C., Held, W., & Yang, D. (2023). *TADA: Task-agnostic dialect adapters for English*. arXiv.https: //arxiv.org/abs/2305.16651

[14] Hovy, D., Yeh, Y.-T., et al. (2025). Culturally aware and adapted NLP: A taxonomy and a roadmap. *Transactions of the Association for Computational Linguistics*.