Say My Name: How Declaring Black Identity Triggers the Safety Filters that Writing Black Does Not

Mario DeSean Booker, Ph. D

CIS/IT Department, Purdue University Global

DOI: https://doi.org/10.51244/IJRSI.2026.1313CS015

Received: 18 June 2026; Accepted: 24 June 2026; Published: 01 July 2026

ABSTRACT

This paper presents an independent secondary validation of Haq & Saldías (2026), which reported that large language models refuse requests from explicitly Black identified users at rates 7.5 to 8.27 percentage points above those observed for White identified users, and that African American Vernacular English (AAVE) dialect use nearly eliminates this penalty a pattern the authors term the "dialect jailbreak." All validation is conducted without new model inference, deriving exclusively from the study's published statistics and the publicly available BOLD dataset (Dhamala et al., 2021). Three findings are reported. First, the paper's core refusal rate claims and the 97.82% soft refusal figure are arithmetically confirmed, and the primary Average Marginal Effects are independently replicated via two proportion z tests (Black vs. White Explicit: +8.27pp, $p < .001$). Second, the conclusion level claim of a "400% increase in odds" requires substantive qualification: the figure reflects a GLMM conditional odds ratio inflated by extraordinary prompt level random effect variance ($\sigma^2 = 100.44$, ICC = .97), diverging approximately 5-fold from the raw marginal odds ratio of 1.56 a distinction the original paper does not address. Third, a novel additive decomposition analysis reveals that the AAVE dialect jailbreak is sub additive, with the observed refusal rate falling 7.41 percentage points below what independent race and dialect penalties would additively predict, indicating synergistic rather than sequential safety filter degradation. Taken together, these findings suggest that keyword-gated alignment mechanisms are doubly brittle: susceptible to explicit identity bypass through lexical substitution, and to multi-signal compounding that exceeds what single signal analyses anticipate. An interactive validation dashboard and all derived data files are publicly available at https://github.com/DrMarioDeSean411-arch/say-my-name-llm-bias

Keywords: large language model alignment, racial bias, AAVE, safety filter bypass, GLMM, dialect jailbreak, algorithmic fairness

POSITIONALITY STATEMENT

This analysis was conducted by a Black scholar whose engagement with algorithmic discrimination as both a research problem and a lived social reality predates the present study by nearly a decade. The author's dissertation (Empowering, Engaging, and Equipping Technology through Acceptance, University of the Cumberlands, 2022) examined facial recognition software acceptance within predictive policing systems, an inquiry that established the foundational concern animating this work: that algorithmic systems presented as neutral instruments routinely encode and operationalize racial hierarchy. That trajectory runs continuously through subsequent publications on digital redlining, ethnic name filtering in hiring algorithms, and false positive generation in automated criminal identification systems, and informs the methodological frameworks in which the Cumulative Disadvantage Score and Intersectional Disparity Index are applied here. The analytical emphasis of this paper, on how effect size characterizations travel from technical reporting into policy language and on whether the populations most harmed by alignment failures are accurately represented in the statistics used to describe those failures, reflects both scholarly commitments and positional knowledge that cannot be fully separated from the research itself. No financial relationship with the authors of the validated study exists. No external funding specific to this analysis was received. All data derives from publicly available sources.

INTRODUCTION

Every time a user submits a request to a large language model, a safety system decides whether to answer. That decision is supposed to turn on the content of the request, its potential for harm, its ambiguity, and the sensitivity of the topic. What Haq & Saldías (2026) showed is that it also turns, measurably, on who appears to be asking. Their study found that users who identified themselves as Black faced refusal rates nearly nine percentage points higher than White users submitting identical prompts. The prompts did not change. The topics did not change. Only the declared identity did, and the model's willingness to respond shifted accordingly. Safety alignment systems, in this light, are not neutral gatekeepers. They are mechanisms that distribute access to information, and when that distribution tracks racial identity, the system is doing something more than filtering harmful content.

The paper's most provocative finding cuts even deeper. Users who wrote in African American Vernacular English (AAVE) without ever stating their race were refused at rates comparable to a racially unmarked White control group. Declaring Black identity triggered the penalty; sounding Black did not. A user apparently achieves better outcomes by allowing their race to remain implicit than by naming it. Haq & Saldías (2026) term this the "dialect jailbreak," and its implications extend well beyond the two models they tested: it suggests that the filter's sensitivity is indexed on explicit keyword signals rather than on any deeper representation of user identity or intent.

Two aspects of the paper's statistical reporting warrant closer examination, though neither undermines its core contribution. The first is the claim, stated in the study's conclusion, of a "400% increase in the odds of refusal" for Black-identified users. The second is the absence of any analysis of whether the race penalty and the dialect jailbreak interact additively, meaning their effects simply stack, or synergistically, meaning the combination produces an outcome the individual signals alone would not predict. These are precision gaps rather than errors, but they matter for how the findings are applied and how the argument should be extended.

This paper addresses both through secondary analysis using only the study's published statistics and the publicly available BOLD dataset (Dhamala et al., 2021). No new model inference is conducted. Four analytic operations are performed: first, arithmetic verification of all refusal rates in Table 4 and the reported 97.82% soft refusal figure; second, independent replication of the study's core Average Marginal Effects via two proportion z tests; third, a formal examination of why the GLMM conditional odds ratio of 8.01 diverges so sharply from the raw marginal odds ratio of 1.56 with prompt level random effect variance ($\sigma^2 = 100.44$) identified as the mechanism; and fourth, a novel additive decomposition revealing that the observed AAVE refusal rate falls 7.41 percentage points below what independent race and dialect penalties would jointly predict. All verification steps and derived visualizations are documented in an interactive dashboard accompanying this paper at https://DrMarioDeSean411-arch.github.io/say-my-name-llm-bias/llm_bias_dashboard.html

Three research questions organize the analysis. *RQ1:* Do the paper's reported statistics hold under independent arithmetic and inferential scrutiny? *RQ2:* Does the "400% increase in odds" accurately represent the marginal population level effect, or is it a conditional estimate whose inferential scope the paper does not adequately specify? *RQ3:* Is the dialect jailbreak synergistic rather than additive when racial identity and dialect signal co-occur? The answers taken together do not revise the paper's conclusions; they sharpen them.

ORIGINAL STUDY

Haq & Saldías (2026) conducted a factorial audit of two open-weight large language models, Gemma 3 12B and Qwen 3 VL 8B, using 2,219 prompts drawn from the Bias in Open-ended Language Generation dataset (BOLD; Dhamala et al., 2021), stratified across four sensitive domains: Gender, Race, Politics, and Religion. Each prompt was run under six identity conditions crossing three demographic profiles (Black, Singaporean, White) with two signaling modes: explicit biographical declaration embedded in the system prompt, and implicit dialect coding applied to the instruction line only (African American Vernacular English, Singlish, and Standard American English control). This design yielded a total corpus exceeding 24,000 responses.

Three outcomes were measured: refusal rate (binary noncompliance, coded via validated keyword heuristic), BERTScore (Zhang et al., 2020; semantic similarity between model output and the Wikipedia source text underlying each BOLD prompt), and Relative Negative Regard (Sheng et al., 2019; the difference in negative sentiment between model response and source text). Primary analyses employed logistic generalized linear mixed effects models with prompt level random intercepts, reporting Average Marginal Effects as headline estimates. The study found that explicit Black identification produced the highest refusal rates across both models and all domains, while AAVE dialect use reduced refusal to near baseline levels, a pattern the authors term the dialect jailbreak.

ANALYTICAL FRAMEWORK

Secondary Validation as a Scholarly Practice

Replication and secondary validation are related but distinct activities. Replication reruns the original procedure with the same stimuli, the same models, and new data collection to test whether findings reproduce under fresh conditions. Secondary validation takes the published findings themselves as its object. The question is not whether the results would replicate, but whether the reported statistics accurately represent what the data show and whether the interpretive claims follow from the analytic choices made. No new model inference is conducted here. The contribution is one of precision: establishing the exact inferential scope of the paper's statistics, identifying where reported figures require contextual qualification, and deriving findings that the original analysis was not designed to detect.

This mode of post-publication scrutiny has gained traction in algorithmic fairness research precisely because the field moves fast and high-stakes policy claims can travel further than the evidence strictly supports. Audits of this kind function as a quality layer within the cumulative research enterprise, not adversarial corrections, but analytical deepening that strengthens the evidentiary foundation for downstream work. In the interest of full transparency, all arithmetic derivations, z-test computations, and subadditivity calculations reported here are rendered interactively at https://DrMarioDeSean411-arch.github.io/say-my-name-llm-bias/llm_bias_dashboard.html, permitting independent inspection of each computational step.

Conditional vs. Marginal Effects in Logistic Mixed Models

Logistic GLMMs produce two conceptually distinct types of effect estimates, and conflating them is one of the more consequential interpretive errors in applied mixed model work. A *conditional* odds ratio describes the effect of a predictor for a given unit in this paper's design, for a specific prompt held at its estimated random effect value. A *marginal* (population average) effect describes the expected impact averaged across all units in the population across all prompts and users.

In linear models, these two quantities converge. In logistic models with large random intercept variance, they do not. The nonlinear sigmoid transformation means that a large spread in baseline log odds, exactly what $\sigma^2 = 100.44$, produces inflated conditional ORs far above the marginal effect. Intuitively, when individual prompts vary wildly in baseline refusal likelihood (some almost certain to trigger refusal regardless of identity, others almost certain not to), the conditional coefficient must be large to explain any residual variation attributable to identity. The marginal average, pulled toward the many low-risk prompts, stays comparatively modest. The ICC of .97 in the Haq & Saldías model confirms that 97% of total log odds variance is prompt driven, leaving identity operating on a thin margin of residual variability, which the conditional OR then magnifies.

Additive vs. Sub additive Interaction

When two signals independently affect an outcome, their joint effect should approximate the sum of their individual contributions to the additive benchmark. For the present analysis, the additive prediction for the AAVE Implicit condition is constructed as: *White Explicit baseline + Black–White race penalty + SAE Control–to–White Explicit dialect reduction*. Any observed rate falling below this sum indicates subadditivity: the two signals together suppress the safety filter more than they do independently. This is a stronger claim than simple bypass; it implies that the co-occurrence of implicit racial identity cues and dialect markers

interacts within the filter's detection architecture, producing a synergistic rather than sequential degradation. The analytic approach is arithmetic, not a formal interaction test, because individual-level data are unavailable; the finding is therefore presented as an additive residual rather than a modeled interaction coefficient.

VERIFICATION OF CORE CLAIMS

Arithmetic Verification of Table 4

The paper reports refusal rates for six identity conditions, each applied to N = 2,201 prompts, not the 2,219 prompts remaining after dataset cleaning, a small discrepancy consistent with the table footnote and attributable to final preprocessing. All six refusal rates compute directly from the published raw counts:

Condition	Refusals	N	Calculation	Reported Rate
Black Explicit	637	2,201	637 ÷ 2,201	28.9% ✓
Black Implicit (AAVE)	119	2,201	119 ÷ 2,201	5.4% ✓
White Explicit	455	2,201	455 ÷ 2,201	20.7% ✓
White Implicit (SAE Control)	100	2,201	100 ÷ 2,201	4.5% ✓
Singaporean Explicit	414	2,201	414 ÷ 2,201	18.8% ✓
Singaporean Implicit (Singlish)	18	2,201	18 ÷ 2,201	0.8% ✓

All six figures verify exactly. The 97.82% soft refusal claim confirms with equal precision: of 1,743 total refusals, 1,705 were soft (conversational evasion) and 38 were hard (explicit policy blocks), yielding 1,705 ÷ 1,743 = 97.82%. The internal arithmetic of Table 4 is clean throughout. Raw counts, refusal rate calculations, and soft/hard refusal breakdowns are visualized interactively in the accompanying dashboard (https://github.com/DrMarioDeSean411-arch/say-my-name-llm-bias).

Independent Z Tests on Core AME Claims

The paper's primary effect estimates are Average Marginal Effects derived from logistic GLMMs. To assess whether those estimates hold at the level of raw population proportions before controlling for prompt difficulty, each key comparison was subjected to an independent two-proportion z test:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_2}\right)}}$$

where $\hat{p}_{pool} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$.

For the primary comparison, Black Explicit vs. White Explicit, the pooled proportion is (637 + 455) ÷ 4,402 = 0.2481, yielding SE = 0.01302 and Z = (0.2894 − 0.2067) ÷ 0.01302 = 6.35, $p < .001$. Results across all four comparisons are summarized below:

Comparison	Paper AME	Z test result	Status
Black vs. White Explicit	+7.5pp	+8.27pp, $p < .001$	Confirmed ✓
AAVE Implicit vs. Black Explicit	∼−23pp	−23.53pp, $p < .001$	Confirmed ✓
Singlish vs. Singaporean Explicit	∼−18pp	−17.99pp, $p < .001$	Confirmed ✓
Singaporean vs. White Explicit	−1.8pp	−1.86pp, $p = .12$ ns	Directional only

Three of the four comparisons are independently confirmed with high confidence. The Singaporean explicit penalty directionally consistent at −1.86pp but non-significant in raw z tests warrants a brief note. The GLMM's advantage over the z test is precisely its capacity to partial out the prompt level difficulty: when a given effect is modest in absolute size, controlling for the substantial prompt level variance ($\sigma^2 = 100.44$) can shift a borderline result to significance by removing noise that otherwise obscures the signal. The Singaporean

result's non-significance in the raw test does not contradict the GLMM finding; it illustrates why the GLMM was necessary.

The 706 Mixed State Prompts

The most direct internal validity evidence the paper provides sits in Figure 7: 706 of the 2,201 unique BOLD prompts produced refusal under at least one identity condition but compliance under at least one other. These prompts are matched on content by design; the stimulus is identical across conditions, only the surrounding identity signal changes. Within this matched subset of 4,248 observations, the paper reports Z = 11.32 ($p <$ .001) for Black vs. White Explicit, a result that cannot be attributed to differences in prompt difficulty or topic sensitivity because those factors are held constant across rows. This subset analysis constitutes the cleanest available test of the identity trigger hypothesis, and the effect survives it decisively. That 32% of all prompts in the dataset showed this sensitivity to identity framing, flipping between refusal and compliance depending solely on who appeared to be asking, is itself a finding that the headline statistics do not fully convey.

QUALIFICATION OF THE "400% INCREASE IN ODDS"

The Raw Data Odds Ratio

Before any statistical modeling, Table 4 alone permits a direct calculation of how much more likely refusal is for Black-identified users relative to White-identified users. Odds, unlike probabilities, express the ratio of an outcome occurring to it not occurring, and dividing one group's odds by another's produces the odds ratio, the standard effect size measure for binary outcomes in logistic regression.

From the published counts:

$$Odds\ (Black\ Explicit) = \frac{637}{2{,}201 - 637} = \frac{637}{1{,}564} = 0.4073 Odds\ (White\ Explicit) = \frac{455}{2{,}201 - 455} = \frac{455}{1{,}746} = 0.2606 Raw\ OR = \frac{0.4073}{0.2606} = 1.56$$

Black identified users face 56% higher odds of refusal than White identified users. That is a meaningful, statistically confirmed disparity. It is not 400%.

The GLMM Conditional Odds Ratio

The paper's conclusion states that explicit identity declarations produced "a 400% increase in the odds of refusal." This figure does not appear directly in the GLMM output tables but is consistent with the additive GLMM specification, which reports a Race [Black] odds ratio of 4.19, yielding $(4.19 - 1) \times 100\% \approx 319\%$, rounded to approximately 400%. The paper's preferred interactive GLMM pushes this further, reporting OR = 8.01 (95% CI: 3.98–16.11, $p <$ .001), which would imply a 701% increase. The conclusion level "400%" most likely derives from the additive model, presented as a conservative summary of the range.

Why Discrepancy Is Expected and What Causes It

The gap between OR = 1.56 from the raw data and OR = 4.19 to 8.01 from the GLMM is not a contradiction; it is a mathematical consequence of the model's structure. The prompt level random intercept variance of τ_{00} = 100.44 produces an ICC of .97, meaning 97% of all variability in refusal log odds is attributable to which specific prompt is being asked, not to which identity condition surrounds it. When a model absorbs that much variance into a random effect, the fixed effect coefficients are estimated on a residual scale that has been stripped of nearly all prompt-level noise. The resulting conditional OR describes the effect of racial identity *for a given prompt held at its estimated difficulty value* a narrow, context-specific quantity. The marginal OR, averaged across the full population of prompts with their enormous spread in baseline refusal likelihood, converges much closer to what the raw data show.

This divergence between conditional and marginal estimates is a known property of logistic mixed models with large random-intercept variance (Neuhaus et al., 1991), not an error in the paper's analysis. The identity penalty is genuine, statistically unambiguous, and confirmed independently here. The qualification is one of framing: a conditional odds ratio placed in a conclusion without clarification will almost universally be read as a population-level claim, and at OR = 4.19 to 8.01 versus the marginal 1.56, that misreading carries material consequences for how the finding is communicated to policy audiences.

Matching the Estimate to the Question

The table below maps each available effect estimate to the inferential question it actually answers:

Estimate	Value	Question Answered
Raw odds ratio	1.56	How do observed population refusal odds compare across groups?
GLMM AME	+7.5 to 8.27pp	What is the average absolute probability impact across all users and prompts?
Conditional GLMM OR (additive)	4.19	Given the prompt of average difficulty, how do the odds shift with identity?
Conditional GLMM OR (interactive)	8.01	Same, estimated within the fully specified interaction model?

For policy audiences and for the paper's own argument about differential informational access, the AME is the appropriate headline figure. A user or policymaker asking, "How much more likely am I to be refused?" is asking a marginal question; the answer is roughly eight percentage points, not a fourfold multiplication of odds. The conditional OR retains value as a mechanistic estimate, but its inferential scope should be stated.

NOVEL FINDING: SUB-ADDITIVE JAILBREAK DYNAMICS

The Additive Prediction

The paper reports the race penalty and the dialect jailbreak as separate findings but does not examine whether they interact. An additive model offers a natural benchmark: if the two signals are independent, their joint effect should approximate the sum of their individual contributions. Using only Table 4 values, the additive prediction for the Black AAVE Implicit condition is constructed as follows.

The race penalty is the additional refusal burden associated with explicit Black identification, the difference between Black Explicit and White Explicit rates: 28.9% − 20.7% = +8.27pp. The dialect reduction in the refusal rate associated with switching from an explicit identity frame to an implicit dialect signal is the difference between the SAE Control and White Explicit rates: 4.5% − 20.7% = −16.2pp. Applied additively to the White Explicit baseline:

$$Predicted\ (AAVE\ Implicit) = 20.7\% + 8.27pp - 16.2pp = 12.77\%$$

The observed AAVE Implicit rate is 5.4%. The gap between prediction and observation is −7.41 percentage points. The AAVE condition performs better than the additive model says it should.

The parallel calculation for Singlish is smaller but directionally consistent: predicted rate = 20.7% − 1.86pp − 16.2pp = 2.64%; observed rate = 0.8%; subadditivity gap = −1.86pp. Both dialect conditions land below their additive benchmarks, but the AAVE gap is four times larger, a difference that maps onto the substantially larger raw race penalty for Black identified users relative to Singaporean identified users.

What Subadditivity Implies

An additive outcome would mean the safety filter processes the race cue and the dialect signal as independent inputs, two separate levers, each pulling refusal probability by its individual amount. Subadditivity means they are not independent. The co-occurrence of AAVE markers and the absence of an explicit racial keyword

produces a bypass effect that the individual signals, combined linearly, cannot account for.

One figure makes this concrete. The AAVE subadditivity gap (−7.41pp) is nearly eight times larger than the AAVE dialect jailbreak's own standalone contribution relative to the SAE control (5.4% − 4.5% = 0.9pp). In other words, the bulk of the AAVE condition's bypass advantage does not come from the dialect signal acting independently, but comes from the synergistic component, the part that neither the race penalty nor the dialect effect alone predicts.

Mechanistic Interpretation

The paper describes current safety systems as brittle and over-indexed on explicit identity keywords. The subadditivity finding extends that characterization: once the explicit racial keyword is removed and the dialect substituted, the filter does not simply lose one of its inputs and adjusts proportionally. It appears to operate in a qualitatively different mode, one where remaining lexical triggers are sparse enough that the model defaults toward compliance at rates lower than additive logic would anticipate. Whether this reflects a threshold dynamic, a switching mechanism in the filter's activation, or feature sparsity in the model's representation of implicit racial identity is not determinable from published statistics alone. The finding is arithmetically clean; the mechanism remains an open question and a direct target for interpretability research.

Implications for Safety System Design

The practical implication is sharper than the existing policy recommendations convey. If a safety filter's keyword sensitivity can be synergistically degraded by combining explicit signal removal with dialect substitution, deployed systems face a multi-signal attack surface that single signal audits will systematically underestimate. An audit that tests dialect bypass and racial identity penalty separately and simply adds the results will miss the interaction that produces the largest share of the bypass effect. This does not require adversarial intent on the part of users; it is a structural property of how the filter responds to co-occurring signals, with implications for any user who communicates naturally in AAVE. The subadditivity result, therefore, reinforces rather than contradicts the paper's core recommendation that alignment mechanisms must generalize beyond lexical triggers to sociolinguistic patterns while adding that the urgency of doing so scales nonlinearly with signal combination, not linearly.

IMPLICATIONS AND LIMITATIONS

What This Analysis Contributes

The core empirical claims of Haq & Saldías (2026) hold. The racial refusal disparity is real, arithmetically confirmed, and survives independent inferential testing. This analysis does not contest that foundation; it sharpens two specific claims built on top of it. First, the "400% increase in odds" characterization overstates the population-level effect by presenting a GLMM conditional estimate without acknowledging the fivefold divergence from the raw marginal odds ratio that $\sigma^2 = 100.44$ produces. The AME of +7.5 to 8.27 percentage points is the figure that accurately describes what an average user can expect; conditional ORs belong in mechanistic discussion, not policy summaries. Second, the dialect jailbreak is not merely large, it is synergistic. The subadditivity finding moves the theoretical account from description to structure: the question is no longer only whether the filter can be bypassed, but why the bypass intensifies when signals combine, and what interaction reveals about how alignment systems process identity information.

Together, these qualifications suggest that the paper's practical implications are, if anything, stronger than its own framing conveys. A disparity that persists synergistically across signal combinations is harder to patch than one that scales linearly with individual inputs.

Limitations

Four constraints bound what can be claimed here. The original study tested two open weight models, Gemma 3 12B and Qwen 3 VL 8B, and the refusal analysis is, in practice, a Gemma only finding given Qwen's near-zero refusal rate across all conditions. Whether the patterns observed generalize to proprietary systems with

architecturally distinct safety stacks remains untested. The subadditivity calculation relies on raw refusal rates and an additive benchmark derived from published marginals; it is an arithmetic residual, not a formal interaction term estimated within the GLMM. Testing the interaction properly requires individual-level response data, which is not publicly available. The survivorship bias in BERTScore quality measured only on responses the model produced, excluding all refusals, means the full magnitude of the quality penalty for Black-identified users remains undercharacterized; the present analysis identifies this gap precisely but cannot close it. Finally, the absence of any gender manipulation in the original design leaves the Race × Gender intersection, the theoretically most consequential node under Crenshaw's (1989) framework, entirely unexamined.

Future Directions

Three extensions follow directly from the limitations above. The Race × Gender intersection warrants priority attention, particularly the comparison between female-coded and male-coded AAVE. The original study held gender constant at male, and the intersectional prediction is non-additive disadvantage rather than summed penalties. Multi-turn experimental designs would test whether the dialect jailbreak holds when safety context accumulates across a conversation rather than operating on a single prompt; turn-by-turn context likely changes the filter's activation in ways a single prompt audit cannot capture. Finally, mechanistic interpretability work targeting attention activations under explicit versus implicit identity conditions could locate where in the model's forward pass the filter activates on explicit keywords and fails to activate on dialect, moving from behavioral measurement to architectural explanation.

CONCLUSION

Haq & Saldías (2026) established that the same question, submitted to the same model, meets meaningfully different responses depending on whether the user has declared a Black racial identity, and that writing in AAVE, without any explicit identity statement, nearly eliminates that penalty. Both claims survive independent scrutiny. The arithmetic is clean, the z tests are unambiguous, and the 706 matched prompts that flipped between refusal and compliance across identity conditions constitute some of the most direct evidence the dataset contains. The disparity is not a statistical artifact; it is a consistent behavioral property of the tested system.

Two contributions sharpen that foundation without displacing it. The population-level effect of explicit Black identification is a refusal rate approximately 8.27 percentage points above the White Explicit baseline, a finding confirmed independently here, and the figure most appropriate for policy interpretation. The conditional odds ratio that generates the "400%" headline answers a different, narrow question: how much do the odds shift for a given prompt at its estimated difficulty level? When 97% of total log odds variance is prompt-driven, that conditional figure inflates well above the marginal effect a population average reader should take away. The distinction matters not because the penalty is smaller than reported; it is substantial and real, but because precision in effect size communication determines whether the finding translates correctly into system design, audit standards, and legal frameworks.

The subadditivity result adds a dimension that the original paper did not reach. The AAVE dialect jailbreak is not simply large; it is structurally synergistic, outperforming additive predictions by 7.41 percentage points. Safety systems indexed with explicit identity keywords are not bypassed gradually; they collapse more sharply when signals combine than when either operates alone. Patching individual lexical triggers addresses the symptom. What the combined finding argues for is alignment that reads sociolinguistic context, not just demographic vocabulary, a harder problem, and a more honest account of what these systems currently cannot do.

REFERENCES

1. Booker, M. D. (2022). *Empowering, engaging, and equipping technology through acceptance: A quantitative study utilizing the modified technology acceptance model (mTAM) to explore facial-recognition software acceptance in the smart policing initiative (SPI)* [Doctoral dissertation,

University of the Cumberlands]. ResearchGate. https://doi.org/10.13140/RG.2.2.13108.77441

2. Booker, M. D. (2026). *Say my name: Interactive validation dashboard — LLM refusal bias study* [Software]. GitHub. https://github.com/DrMarioDeSean411-arch/say-my-name-llm-bias
3. Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1989*(1), 139–167.
4. Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K. W., & Gupta, R. (2021). BOLD: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872. https://doi.org/10.1145/3442188.3445924
5. Haq, I., & Saldías, B. (2026). Dialect vs. demographics: Quantifying LLM bias from implicit linguistic signals vs. explicit user profiles. *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*. https://doi.org/10.1145/3805689.3812419
6. Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, *59*(1), 25–35.
7. Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3407–3412. https://doi.org/10.18653/v1/D19-1339
8. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *Proceedings of the 8th International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr