When "Government-Approved" LLMs Silence Citizen Voices: Evidence of Systematic Demographic Bias in Federal Comment Analysis

Abstract

Motivation: Federal agencies are rapidly deploying Large Language Models to process millions of public comments for efficiency gains. At least 10 agencies—including the Department of the Interior, USDA, CDC, Centers for Medicare & Medicaid Services, and Department of Justice—now use FedRAMP-authorized LLMs to analyze citizen input on critical policies from climate change to healthcare. FedRAMP (Federal Risk and Authorization Management Program) certifies these systems for security compliance, but crucially does not evaluate algorithmic fairness. This creates a dangerous gap where "secure and compliant" authorization masks potential discrimination in democratic participation.

Objective: This is the first systematic evaluation of FedRAMP-authorized LLMs for fairness in democratic participation. Federal agencies deploy these "secure" systems (Amazon Bedrock, Anthropic Claude, Meta Llama, OpenAI Azure), assuming security compliance ensures fair representation. We reveal that FedRAMP authorization masks systematic discrimination in how citizen voices are weighted and represented to policymakers.

Method: We evaluate FedRAMP-authorized LLMs using controlled permutation testing on 10,000 federal public comments from active dockets: identical comments are summarized under (1) baseline conditions without demographics, then (2) with systematically varied demographic profiles (age, gender, education, political affiliation, location). By holding content constant while only changing attributed demographics, we isolate the pure effect of demographic information on summarization. We measure bias through semantic drift (Wasserstein distance), SHAP-based salience scoring, and framing analysis.

Results: Initial testing reveals systematic demographic bias in FedRAMP-authorized LLMs. Identical comments receive different treatment: gendered language appears only for women ("She urges"), terminology weakens ("strategic flaw" vs "incoherence"), and framing shifts by political affiliation. Our Democratic Representation Index (DRI) quantifies this bias on a 0–1 scale where 1.0 represents demographic neutrality. FedRAMP models average DRI scores below 0.5, meaning citizen voices are represented at less than half of parity. The mechanism is linguistic profiling: models infer education levels from writing style and systematically privilege professional discourse, filtering community voices out of policy relevance. This algorithmic filtering potentially violates the Administrative Procedure Act's mandate for equal consideration of all public input.

Impact: By demonstrating that FedRAMP security compliance does not ensure fairness, we propose extending federal procurement standards to include mandatory DRI thresholds alongside existing security requirements. Without this expansion, agencies unknowingly deploy systems that silence the very communities most affected by their policies—transforming a legal obligation for equal consideration into algorithmic discrimination. The implications extend beyond individual agencies: as federal LLM adoption accelerates, security-only compliance threatens to institutionalize bias at the core of participatory governance. Our work provides both the evidence and the metrics needed for immediate policy intervention.