March 31, 2026
National Institute of Standards and Technology (NIST)
100 Bureau Drive
Gaithersburg, MD 20899

Re: Comment Letter on NIST Draft Guidance on Automated Benchmark Evaluations of Language Models

The undersigned civil rights advocacy and civil society organizations appreciate the opportunity to submit comments in response to the National Institute of Standards and Technology (NIST) draft guidance on Practices for Automated Benchmark Evaluations of Language Models. We commend NIST and the Center for AI Safety and Innovation (CAISI) for developing a framework to improve the validity and reproducibility of AI evaluations, and we hope our comments will help strengthen and expand this important work.

General Feedback

We appreciate NIST's commitment to building a rigorous, reproducible framework for evaluating AI models. Evaluation standards are foundational to responsible AI development and deployment, they define what gets measured, how it gets measured, and ultimately what AI systems get built and deployed. We strongly support the effort to bring consistency to this process. Particularly at a time of rapid AI development, NIST has the opportunity to significantly enhance AI system evaluation. Additionally, in the current AI Governance environment, where no comprehensive federal AI framework exists, NIST is uniquely positioned, by virtue of its technical expertise and institutional credibility, to bring coherence to a fragmented landscape.

NFHA evaluated the draft guidance in the context of a housing recommendation use case that considered the propensities of multiple Large Language Models to create or amplify racial bias and steering when tasked with recommending houses to potential homebuyers. We discovered that the highest-stakes decisions lie in Section 1 of the guidance: defining what to measure and why. When that foundational design work lacks explicit civil rights guidance, even a well-executed evaluation can miss the harms that matter most. Given the relevance of NFHA's findings, the recommendations are largely focused on Section 1 of the draft guidance, defining evaluation objectives, and selecting benchmarks that meet the evaluation objective.

Our recommendations center on the following key priorities:

1. Incorporate civil rights principles, including antidiscrimination measures such as disparate treatment and disparate impact testing as explicit requirements for AI evaluation frameworks.

2. Encourage and incentivize the development of domain-specific benchmarks, with particular attention to housing, lending, criminal justice, child welfare, education, and employment as priority domains for benchmark development.

I. Incorporate Civil Rights Principles into Evaluation Objective Design

We are concerned that the current draft guidance does not sufficiently provide guidance on how the proposed framework can be used to test the civil rights impacts of AI systems, in particular discrimination in treatment, discrimination in outcome (disparate impact), and broader harms to protected classes. AI systems are increasingly deployed in domains where flawed or incomplete evaluation can directly harm communities.[1][2] Government agencies are deploying AI models in housing authorities, healthcare systems, benefits agencies, courts, and social service systems right now, often without the evaluation of infrastructure that this guidance seeks to build.[3] Furthermore, evaluation frameworks are influential for downstream AI development and often become targets for model optimization;[4] evaluation frameworks that fail to incorporate civil rights principles risk enabling discriminatory outcomes at scale. NIST has an opportunity to ensure that evaluation frameworks developed under its guidance do not enable discriminatory systems and instead promote reliable performance across communities.

We urge NIST to:

- Amend Practice 1.1 to recommend that evaluation objectives for AI systems deployed in high-stakes domains explicitly include assessment of disparate treatment (differential outputs for identical inputs across protected classes under civil rights laws) and disparate impact across protected classes centered on fairness evaluation techniques[5] applicable to generative AI systems.
- Require that benchmark selection under Practice 1.2 includes documentation of whether selected benchmarks are capable of detecting and remediating sector-specific output disparity across protected class variables.

---

[1] *The New York Times*, "When DOGE Unleashed ChatGPT on the Humanities," March 7, 2026, https://www.nytimes.com/2026/03/07/arts/humanities-endowment-doge-trump.html.

[2] *The Guardian*, "Experts Find Flaws in Hundreds of Tests That Check AI Safety and Effectiveness," November 4, 2025, https://www.theguardian.com/technology/2025/nov/04/experts-find-flaws-hundreds-tests-check-ai-safety-effectiveness.

[3] Devin Windelspecht, "How ProPublica Exposed a Flawed AI Tool 'Munching' Hundreds of Veteran Contracts," *Global Investigative Journalism Network*, September 9, 2025, https://gijn.org/stories/how-the-did-it-propublica-ai-tool-cut-veterans-affairs-contracts/

[4] Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gómez, and David Fernandez-Llorca, "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation," arXiv preprint arXiv:2502.06559 (May 25, 2025), https://arxiv.org/html/2502.06559v2.

[5] Emily Black et al, "Towards Effective Discrimination Testing for Generative AI," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2025), 1028–1047, https://doi.org/10.1145/3715275.3732067.

- Emphasize the cross-functional nature of evaluation design. An effective evaluation framework is not only a technical product and so must involve the perspectives of multiple teams beyond just the technical staff. The guidance is intended for technical staff, but should emphasize that collaboration with other non-technical teams such as legal and compliance teams is necessary to create robust evaluation objectives.
- Establish a tiered evaluation framework that recommends enhanced standards for AI systems deployed in high-stakes domains, including but not limited to housing, lending, criminal justice, employment, healthcare, child welfare, and education. The framework should consider including: civil rights benchmarks, disaggregated performance reporting by protected class, documented human oversight protocols, and mandatory post-deployment monitoring.
- Explicitly define discriminatory outputs as a category of worst-case behavior requiring mandatory assessment under Practice 1.1 for any AI system deployed in a domain governed by federal or state civil rights law.
- Require post-deployment monitoring as a standard evaluation lifecycle component for AI systems used in high-stakes domains. NFHA's housing evaluation identified endpoint volatility – the risk that the judge model is deprecated, breaking the evaluation pipeline – as a significant operational challenge. Post-deployment monitoring must be designed with this fragility in mind, including documented contingency procedures for judge model replacement.
- Recommend that AI system evaluations in high stakes domains incorporate multiple methods, including human review, red-teaming, and field testing, not only automated benchmarks. Our housing evaluation used automated scoring, but manual QA audits of raw data were essential for identifying hallucination and prompt template failures that automated scoring missed.
- Provide recommended courses of action in the event that the evaluation discovers harms from the system, or that a system is unfit for its purpose.

II. Invest in Benchmark Development for High-Stakes Domains

The AI field has made significant progress measuring model capability. However, the field needs an equally rigorous infrastructure for measuring whether models navigate domain complexity responsibly, engage in discrimination, and respond to the lived realities of the communities they affect. Nowhere is this gap more consequential than in the civil-rights-sensitive domains of housing, lending, criminal justice, employment, healthcare, child welfare, and education..

Developing benchmarks in these domains is particularly vital, because their effects reverberate across sectors. Housing is a powerful example of this. A housing-specific AI benchmark could address inequities across multiple sectors and catalyze the adoption of evaluation frameworks that place human well-being at their center. As the National Fair Housing Alliance's *Where You*

Live Matters framework makes clear, housing determines our access to fresh air, clean water, well-resourced schools, healthcare facilities, reliable transportation, good jobs, quality internet service, and much more. It impacts one's credit score, one's chance of attending college, health and educational outcomes, one's ability to buy a home, even lifespan. In the U.S., many people live in communities that are poorly resourced: places that may impede rather than promote their ability to thrive. AI benchmarks designed with fair housing principles at their core would therefore have spillover benefits: improving the equity of tools used not only in housing, but in the health, education, and finance sectors that are inextricably linked to where someone lives.

Our evaluation of LLMs in housing reflects a structural limitation in housing-related AI benchmarks. Current benchmarks were not designed to detect the kinds of discrimination AI systems can produce in regulated domains, not because any particular benchmark is poorly constructed, but because the field has not yet translated the full range of human rights concerns, cultural context, and community impact into evaluation criteria.[6] Discriminatory steering, geographic bias, and proxy discrimination rooted in socioeconomic conditions have gone largely unmeasured as a consequence. To build AI systems worthy of public trust in high-stakes domains, benchmarks must capture domain-specific concerns.

We urge NIST to:

- Identify housing, lending, criminal justice, child welfare, education, healthcare, and employment as priority domains for benchmark development and allocate NIST and CAISI resources to support the creation of publicly available, domain-specific evaluation frameworks.
- Address the benchmark contamination risk (Practice 1.2.4c) specifically in civil-rights-sensitive contexts. AI systems may perform well on existing benchmarks not because they are fair, but because benchmarks do not test the full range of discriminatory patterns the systems may produce.
- Encourage that high-stakes domain benchmarks test AI systems for patterns consistent with historical redlining, geographic steering, or other forms of location-based discrimination.
- Require that benchmark development for high-stakes domains include disaggregated performance assessment across protected classes including but not limited to race, color, national origin, religion, sex, familial status, and disability.

A framework that incorporates our recommendations will set expectations for the field, for developers, deployers, and institutions that use AI to serve the public. We urge NIST to update

---

[6] Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gómez, and David Fernandez-Llorca, "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation," arXiv preprint arXiv:2502.06559 (May 25, 2025), https://arxiv.org/html/2502.06559v2.

their evaluation guidance to lead AI development and deployment in a direction centered on civil rights principles.

Thank you for considering our recommendations.

Sincerely,

1. National Fair Housing Alliance (NFHA)
2. The Feminist Majority Foundation
3. Electronic Privacy Information Center (EPIC)
4. National Action Network (NAN)
5. Welcoming America
6. United Church of Christ Media Justice Ministry
7. National Consumer Law Center, on behalf of its low-income clients
8. Common Cause
9. Hispanic Federation
10. CDD - Center for Digital Democracy
11. Center for AI and Digital Policy (CAIDP)
12. Center for Democracy and Technology
13. Bazelon Center for Mental Health Law
14. UnidosUS
15. Asian Americans Advancing Justice | AAJC
16. Open MIC (Open Media and Information Companies Initiative)
17. HTTP - Hispanic Tech & Telecommunications Partnerships
18. Investor Alliance for Human Rights