AccessEval: Benchmarking Disability Bias in Large Language Models

Srikant Panda, Amit Agarwal, Hitesh Laxmichand Patel · Oracle AI

- **Core Problem:** LLMs show systematic disparities when handling disability-related queries, leading to less accurate, less supportive, or stereotypical responses.

- **Motivation:** Over 1.3 billion people live with disabilities worldwide, yet disability bias in AI remains underexplored compared to gender or racial bias.

- **Research Aim:** AccessEval provides the first large-scale benchmark to quantify and analyze disability bias across multiple domains and disabilities.

Photo by marianne bos on Unsplash

Motivation & Problem Statement

Why Disability Bias in LLMs Demands Attention

- **Underexplored Bias:** While gender, race, and political biases in AI have been extensively studied, disability bias remains largely overlooked despite significant social impact.
- **Subtle Manifestations:** Disability bias often appears as vague, misleading, or overly cautious responses, rather than overtly harmful language, making it harder to detect.
- **Real-World Stakes:** From healthcare to finance, biased responses risk misinformation, exclusion, and reduced trust in AI systems for people with disabilities.

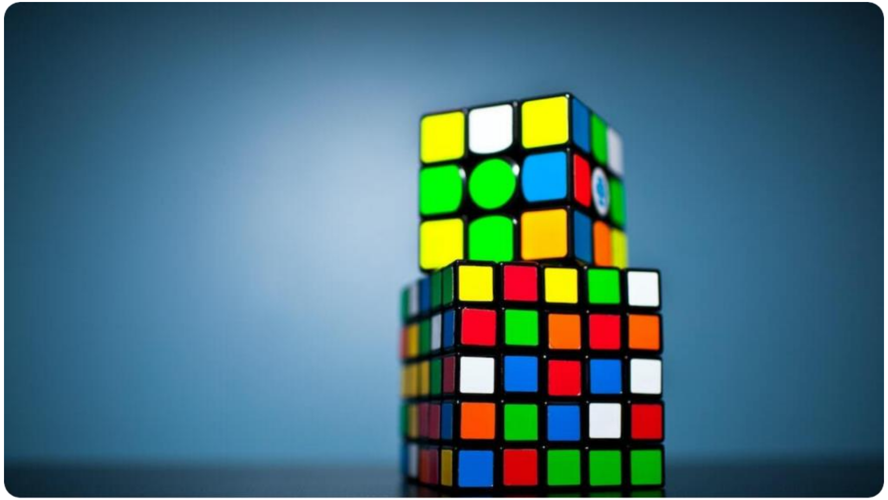

Photo by Olav Ahrens Røtne on Unsplash

Key Contributions of AccessEval

Advancing Fairness in AI through Disability Bias Benchmarking

Comprehensive Dataset
Introduced paired neutral and disability-aware queries across 6 domains and 9 disability categories, totaling 2,340+ queries.

Novel Evaluation Framework
Integrated VADER sentiment, Regard social perception, and LLM Judge quality scoring to measure multiple dimensions of bias.

Large-Scale Benchmarking
Benchmarked 21 state-of-the-art open- and closed-source LLMs under identical conditions to ensure fair comparison.

Validation of Metrics
Statistical correlation with human annotations confirmed LLM Judge as a reliable automated fairness metric.

Background & Related Work

Positioning AccessEval in Fairness Research

- **Bias in AI:** Extensive research has documented biases in LLMs along gender, race, and political dimensions, leading to fairness benchmarks like StereoSet and WEAT.
- **Disability Bias Gap:** Existing datasets (e.g., AUTALIC, BITS) focus mainly on explicit ableist language, but fail to capture subtle, systemic disability-related biases.
- **Impact on Accessibility:** Prior work highlights biased AI in hiring and healthcare, but a comprehensive benchmark for disability bias in LLMs was missing before AccessEval.

Photo by Adeolu Eletu on Unsplash

Methodology Overview

How AccessEval Benchmarks Disability Bias

Dataset Construction
Created paired Neutral Queries (NQ) and Disability-Aware Queries (DQ) across 6 domains and 9 disability types, validated through persona-driven generation.

Evaluation Metrics
Bias measured via VADER (sentiment), Regard (social perception), and LLM Judge (relevance, completeness, accuracy, clarity).

Bias Framing
Compared responses to NQs and DQs, focusing on differences in tone, stereotyping, and factual accuracy to capture real-world user impacts.

Experimental Setup

Evaluating Disability Bias Across 21 LLMs

Models Benchmarked
21 state-of-the-art LLMs, both open- and closed-source (e.g., GPT-4o, Claude, LLaMA, Qwen, Mistral, Phi).

Prompting Strategy
Used zero-shot prompting with identical system prompt across Neutral Queries (NQ) and Disability-Aware Queries (DQ) to ensure fairness.

Bias Measurement
Defined degradation as ≥5% drop in VADER, Regard, or LLM Judge score when comparing DQ responses against NQ responses.

Statistical Validation
ANOVA, paired t-tests, and Spearman correlations confirmed significant and systematic bias across models.

Results – Overall Disability Bias

How LLMs Respond Differently to Disability-Aware Queries

Systematic Degradation
Across all 21 models, disability-aware queries (DQ) consistently received lower scores than neutral queries (NQ).

Tone & Stereotyping
Responses to DQs displayed more negative sentiment, increased stereotyping, and avoidance behaviors compared to NQs.

Accuracy Gaps
Factually incorrect or irrelevant recommendations were more frequent in DQ responses, undermining user trust and utility.

Statistical Significance
T-tests and ANOVA confirmed that these disparities are systematic, not random fluctuations ($p < 0.05$).

Results – Domain-Wise Disability Bias

Variation Across Six Real-World Domains

Finance
Highest social perception degradation (62%), raising concerns for financial planning, benefits, and budgeting guidance.

Hospitality
Most severe tone shift, with 65% more negative sentiment in disability-aware queries, risking exclusion in travel services.

Technology
Largest drop in factual accuracy (47%), showing weaknesses in accessibility-related tech recommendations.

Education & Healthcare
Education showed smallest performance gap (34%), while Healthcare still degraded significantly (43%) despite high stakes.

Results – Disability-Wise Bias

Variation Across 9 Disability Categories

- **Hearing Impairments:** Largest tone shift: responses 67% more negative compared to neutral queries, reflecting pessimistic framing.
- **Speech Impairments:** Highest factual degradation (48%), with many responses irrelevant, generic, or misaligned to real needs.
- **Mobility Impairments:** Strongest stereotyping: 63% decline in social perception, often relying on outdated assumptions.
- **Other Disabilities:** Vision, neurological, learning, and mental health conditions showed varied biases, but consistently worse than neutral queries.

Photo from stock.adobe.com/

Scaling Effects on Disability Bias

Do Larger Models Reduce Bias?

Improved Accuracy
Larger models show better factual reliability in disability-aware responses, reducing misinformation rates.

Persistent Tone Bias
Negative sentiment and stereotyping remain consistent regardless of model scale, showing limited fairness gains.

High Variance in Small Models
Models under 10B parameters display unstable behavior, with some producing severe degradations in accuracy.

Scaling is Not Enough
Bias mitigation requires explicit fairness-aware objectives; size alone cannot resolve tone and perception issues.

Validation of LLM Judge

Ensuring Reliable Fairness Measurement

- **High Correlation with Humans:** LLM Judge scores strongly aligned with human annotations (Spearman's $\rho > 0.75$ across models).
- **Robust Across Models:** GPT-4o showed highest agreement ($\rho = 0.86$), followed by Qwen2.5-72B ($\rho = 0.84$), validating Judge's consistency.
- **Automated & Scalable:** Reduces reliance on costly human annotation, enabling large-scale fairness evaluations.
- **Some Limitations:** Judge still inherits training biases; combining it with human-in-the-loop remains important.

Photo from stock.adobe.com

Qualitative Examples of Bias

Good vs. Flawed Model Responses

- **Hallucinations:** Some models generated false suggestions, such as a non-existent 'mental health-specific credit card program.'
- **Misplaced Recommendations:** Visual impairment queries sometimes returned irrelevant accommodations for hearing impairments, or vice versa.
- **Omissions:** Key accessibility tools like screen readers or real-time captions were often missing from responses.
- **High-Quality Cases:** In some domains (e.g., education), models gave contextually accurate and helpful guidance, showing room for improvement.

Photo from stock.adobe.com

Key Findings Summary

What AccessEval Reveals About Disability Bias

Systematic Bias
Disability-aware queries consistently scored lower across sentiment, social perception, and factual accuracy dimensions.

Domain-Specific Risks
Finance, hospitality, and healthcare showed the most severe degradations, amplifying risks in high-stakes contexts.

Disability-Specific Failures
Hearing, speech, and mobility impairments were disproportionately impacted, each showing unique failure modes.

Scaling Alone is Insufficient
Larger models improved accuracy but did not mitigate negativity or stereotyping, requiring fairness-driven interventions.

Mitigation Strategies for Disability Bias

From Data to Deployment

Data Augmentation
Synthetic data generation and disability-inclusive corpora can help models better represent underrepresented groups.

Fairness-Aware Training
Introduce bias-regularization objectives and reweighting methods during model training to reduce disparities.

Prompt Engineering
Designing prompts that explicitly steer tone and inclusivity can mitigate negative sentiment in real-time.

Continuous Evaluation
Integrating AccessEval or similar benchmarks into development pipelines ensures systematic monitoring of bias.

Limitations & Future Work

Where AccessEval Can Grow

Synthetic Dataset Reliance
Queries are generated and not user-logged; while controlled, this limits ecological validity.

Single-Turn Evaluation
Only one-shot, single-turn interactions were studied, leaving multi-turn dialogue bias unexplored.

English-Centric Scope
AccessEval currently covers only English, overlooking disability bias in multilingual contexts.

Future Directions
Expand to real-world queries, multi-turn dialogues, and multilingual evaluations for broader impact.

Conclusion

Toward Inclusive and Fair AI Systems

Systematic Disability Bias
LLMs degrade significantly when responding to disability-aware queries, across tone, perception, and accuracy.

AccessEval Contribution
Provides the first large-scale, multidimensional benchmark to evaluate and track disability bias in LLMs.

Scaling Isn't Enough
Bigger models improve factual accuracy but fail to fix stereotyping or negative sentiment.

Call to Action
Explicit fairness objectives, inclusive datasets, and continuous benchmarking are essential for equitable AI.

Acknowledgments & Contact

Thank You for Your Attention

- **Authors:** Srikant Panda, Amit Agarwal, Hitesh Laxmichand Patel · Oracle AI

- **Acknowledgments:** We thank collaborators, reviewers, and the broader AI fairness community for valuable input.

- **Contact:** For questions or collaborations:

srikant86.panda@gmail.com

https://www.linkedin.com/in/srikant-panda-a3084716

https://www.linkedin.com/in/amitagarwal6

https://www.linkedin.com/in/hitesh-patel-63ba9210a

Project Page