AccessEval: Benchmarking Disability Bias in Large Language Models

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

AccessEval: Benchmarking Disability Bias in Large Language Models

- **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

ORACLE