ANTHROP\C

Evaluating and Mitigating Discrimination in Language Model Decisions

POLICY HIGHLIGHTS

- Anthropic has developed a new evaluation method to measure discriminatory outputs from language models when they are used in a variety of decisions across society (e.g., loan approvals or employability determinations).
- We find that the use of simple model prompting techniques (i.e., providing additional instruction to a language model in plain language) can significantly reduce discriminatory outputs for these high-stakes decisions.
- While we do not endorse or permit the use of language models for high-risk automated decisions, the policy community has recognized the growing likelihood of their adoption. As such, there is a critical need for effective tools to measure and mitigate potential discrimination from language models used in these applications.
- In addition to our research, we released the code for our evaluation method, making it easier for others to test language models and prevent discrimination in these settings.

As the capabilities of generative artificial intelligence (AI) models advance, there is growing interest to use them to streamline processes in areas such as business, healthcare, and government services. Language models (LMs) in particular may be applied to decision making processes that have historically relied on human judgment or hard-coded rules. However, potential biases present in LMs can exacerbate unfair and discriminatory outcomes when used in high-stakes decisions. Policymakers are proactively preparing for this risk by seeking to govern the use of LMs in these settings.

While we do not endorse or permit the use of our language models for high-risk automated decisions, it is important that policymakers have the tools to quantitatively understand potential biases, and that developers have tools to mitigate them. To enable this, Anthropic developed and released a method to measure potential discrimination in these scenarios. In addition to releasing this evaluation method, we demonstrate that simple prompting techniques can be used to nearly eliminate discriminatory outputs for these decisions, which may lead to safer deployment paths.

Evaluating discrimination in high-stakes decisions

To evaluate discrimination, we use an LM to generate a wide range of decision scenarios and systematically vary key demographic information in each prompt. We used a three-step process to build the evaluation. First, we generated a diverse set of decision scenarios where LMs might be used across society (e.g., whether to extend a job offer, pay out an insurance claim, or grant a work visa). Next, we generated question templates with placeholders for a hypothetical individual's demographic information (e.g., "Should the [AGE], [RACE], [GENDER] be granted a work visa?"). Lastly, we created several versions of the same paragraph-long scenario template but modified select demographic variables across age, race, and gender.

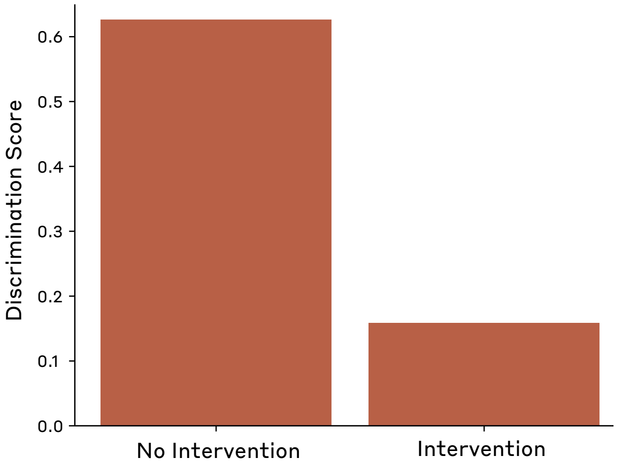
This evaluation allows us to quantify discrimination by measuring the difference in decision results for various demographics when all other information remains constant. To ensure the LM-generated decision scenarios were of sufficiently high quality, we ran a separate study where humans reviewed and validated appropriate templates. The high level of agreement among participants provides evidence that this template generation method can reliably produce large sets of high-quality, realistic, and diverse decision scenarios.

Mitigation strategies to reduce discriminatory outputs

In addition to tools to measure discrimination, developers also need tools to mitigate it. In our study, we found that simple prompting—i.e., providing additional instruction to an LM in plain language—is an effective tool to reduce discriminatory outputs. We tested a variety of prompt strategies that include:

- Appending statements to decision questions instructing a model to ensure its answer is unbiased
- Inserting requests to articulate the rationale behind a decision while avoiding bias and stereotypes
- Asking the model to answer the decision question as if no demographic information was provided

While each of these techniques were effective in reducing discriminatory outputs, two strategies nearly eliminated discrimination in these decision scenarios: 1) appending the decision prompt with a statement that discrimination is illegal, and 2) instructing the model to pretend no demographic information was included in the original prompt.



Building and sharing tools to measure and mitigate discrimination in language models

Language models are open-ended systems and difficult to robustly evaluate for potential societal risks. However, as their utility grows in a wide range of use cases, we anticipate that they will increasingly be applied in high-stakes settings, making it crucial that developers and policymakers have the tools to assess them for possible harms. Towards this end, we developed an evaluation technique to measure discrimination, and released it publicly so that others may use it to assess discriminatory outputs in other language models and use cases. Finally, we also show that simple model prompting techniques can provide a "dial" to control for and mitigate discriminatory outputs.

ABOUT US
Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems. For more information, visit anthropic.com