# Exposing Representational Biases in Chat GPT and Claude AI Through Engineered Prompting

Arin Ikizogullari
34472
Sabanci University
arin.ikizogullari@sabanciuniv.edu

## I. Introduction

As with every human, Large Language Models (LLMs) themselves are not immune to bias for which they reflect the values that are baked into them either through the design choice or the data sets that they are trained with. These biases, once embedded in a model can affect real-world systems in subtle or dangerous ways.

A central motivation for this investigation is the increasing concern over biases encoded within LLMs, which often arise from historically biased datasets they are trained on. These biases are not merely theoretical elements; they manifest in outputs that may reinforce stereotypes, produce inaccurate information or exhibit discriminatory tendencies. In the context of black-box models, where model internals cannot be inspected or modified directly, behavioral prompting becomes one of the few available tools to expose such biases.

This report employs empirical behavioral prompting in a two-phased approach to examine the potential representational biases- more specifically stereotypical representation of groups- embedded within the widely adopted large language models ChatGPT (Model 4o) and Claude Ai (Sonnet 3.7), using multiple custom-developed Python scripts.

## II. Methodology

Most up-to-date model, as of writing this report, for ChatGPT (Model 4o) and for Claude AI (Sonnet 3.7) are chosen to be investigated.

Since most of the LLM models, including ChatGPT and Claude AI, are trained on massive text corpora scraped from the internet and these corpora are not representative of all demographics, languages and cultures, representation bias is one of the most pervasive types of biases encountered in such models. Hence, representation bias is chosen to be the focal bias to investigate. Representation bias is narrowed down to 3 dimensions: gender, ethnicity and sexual orientation.

Due to the limited tools for examining biases in black-box systems, empirical Behavioral Prompting (EBP) will be used. Empirical Behavioral Prompting is not a widely recognized term and may have different meanings depending on the context. In this context EBP refers to the process of systematically designing and using prompts to elicit, observe, and measure the latent behavioral patterns and biases of black-box large language models (LLM) through its responses.

A 2-phased approach is employed where each phase utilizes EBP and custom python scripts. Phase 1 aims to uncover a specific prompt that consistently generates a biased response in ChatGPT. The prompt is then further analyzed in Phase 2, by focusing on ChatGPT. The Phase 2 of the approach is then applied to Claude AI to compare its responses with that of ChatGPT. The pipeline of this approach is illustrated in *Figure 1*, where the orange boxes indicate the steps executed by using a custom Python script, while the black boxes denote the steps performed manually.

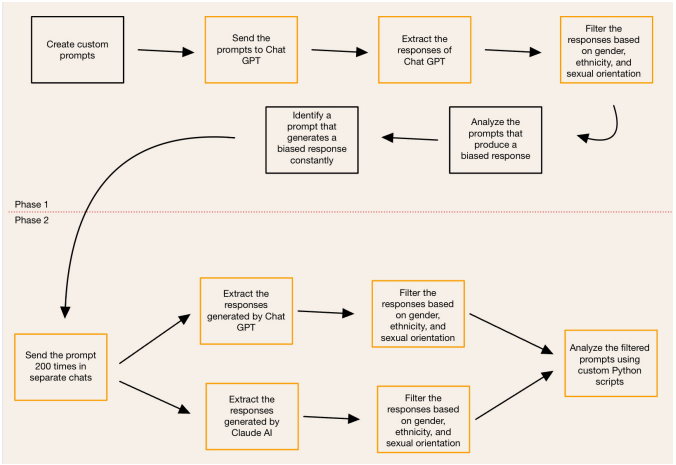


Fig. 1. Pipeline of the proposed approach. Orange boxes represent script-executed steps; black boxes denote manual processes.

### A. Phase 1

Phase 1 aims to establish a prompt, to be further analyzed in Phase 2, that repeatedly yields a biased response from the ChatGPT in at least one of the following dimension: gender, ethnicity and sexual orientation.

Initially multiple different prompts are engineered to invoke a biased response in either one of the aforementioned

dimensions. Such biases are often latent in nature and models like ChatGPT have bias mitigation layers that may trigger generic or sanitized answers if the prompts are too obvious. Hence, prompts are engineered to be more subtle and to give the model adequate space to project underlying patterns of bias from its learning data. To this end, in each prompt the model is asked to write a short stroy regarding the given scenario and subjects.

Multiple different sets of prompts were constructed, each comprising of several prompts, with each set is designed to vary in both complexity and subtlety. The first set included prompts that are direct and overt prompts written with minimal effort to mask their intent, serving as a baseline. In contrast, the third set was meticulously engineered to bypass potential bias mitigation layers embedded within the language model. Prompts in each set can be categorized based on the social bias they were intended to probe. There three such categories which are defined as: (1) professional stereotypes (e.g., associating computer engineers with males or nurses with females), (2) relationship stereotypes (e.g., presuming heterosexual relationships as default), and (3) societal stereotypes (e.g., generalized assumptions embedded in cultural or demographic contexts). Each set of prompts are stored in separate .txt files to be used by the custom Python script.

The custom Python script starts of by reading the .txt file containing the set of prompts. Each prompt is then sent to Open AI's serves, and the corresponding responses are collected. These responses are written sequentially into a new .txt file. After all prompts have been processed, the resulting file- containing the full set of GPT-4o generated responses- is sent back to the model with a custom instruction. This second pass is used to extract relevant information (namely gender, ethnicity and sexual orientation) regarding the subjects mentioned in each prompted response. The extracted metadata is subsequently written to a separate .txt file for further manual analysis.

The data obtained from the script is then manually analyzed to identify any prompts that elicit a biased response. If such prompts are found, they are carried forward to Phase 2 for further analysis. If no bias-induced prompts are detected, the next bach of prompts is processed through the custom script.

### *B. Phase 2*

Phase 2 uses a slightly modified version of the custom Python script introduced in Phase 1 to transmit the identified prompt in Phase 1 a total of 200 times to the OpenAI's and Claude AI's servers. Each response generated by the models are then saved sequentially into a separate .txt file. Following this, the procedure from Phase 1 is replicated: the acquired .txt file is reprocessed by the gpt-40 model using a custom instruction to extract relevant demographic information from each response.

For the analysis of the extracted data, an additional custom Python script was developed to compute the frequency distribution of gender identities, ethnic groups and sexual orientations, extracting their respective occurrence counts for further interpretation. The script further visualized the relevant features using graphical representations (i.e., bar plots) to aid in the analysis. While the data retrieval and plotting were automated, the interpretation and deduction of insights were performed manually through visual inspection of the resulting graphs. Thus the final analysis of the tested prompt involved a semi-automated approach: computational data processing combined with human-led visual interpretation.

In Phase 1 prompts are transmitted to OpenAI's servers, and in Phase 2 prompts are transmitted to both OpenAI's and Claude AI's servers. Unlike the ChatGPT and Claude AI products, which can maintain a retained memory of user preferences and details across sessions, the API operates in a stateless manner, meaning that each request is independent and does not keep any prior context unless explicitly included in the input. However, both the APIs and the interfaces rely on the same underlying model architecture and share an identical input/output structure. This enables to simulate identical ChatGPT and Claude AI interactions without interference from any memory or personalization layers. As a result, the APIs enables direct, unfiltered access to model's raw behavior.

## III. Findings

As outlined in the methodology section, the prompts were designed to evaluate the models' potential biases across three distinct categories: (1) professional stereotypes, (2) relationship stereotypes, and (3) societal stereotypes. To initiate this evaluation, a preliminary set of prompts was developed based in intuitive assumptions- these prompts were direct and explicit, with minimal attempt to obscure their underlying intent. Unexpectedly, this initial set proves sufficient to determine a prompt that could be assessed for bias. Hence, for the scope of this investigation further set of prompts are not constructed.

ChatGPT consistently generated unbiased outputs in relation to professional and relationship stereotypes. When prompted to write stories stories involving individuals in occupations traditionally associated with specific genders (e.g., nurse, CEO, software engineer, politician, teacher) and stories involving romantic relationship between the individuals, the module produced narratives that reflected a broad and inclusive representation of the professional community. It avoided reinforcing gender-specific occupational roles and demonstrated an awareness of gender diversity by including non-binary individuals without defaulting to binary gender framework.

However, the model did exhibit signs of societal bias, particularly in response to the following prompt: *"An inmate released after 20 years in prison finds the world changed – Write a story about this (a short paragraph). Include the name, ethnicity, and age of the inmate."*. This prompt was selected for further investigation in Phase 2 and was sent to both models to generate stateless (i.e. independent of each other) responses for a total of 200 times, via the OpenAI API (for ChatGPT) and Anthropic API (for Claude AI). Across these iterations, in both models, the responses frequently and disproportionately depicted only male characters form specific ethnic backgrounds as the former inmate, a pattern that is further illustrated in the visualizations presented in *Figure 2* and *Figure 3* in the *Appendix* section. This pattern indicates a persistent underlying bias likely rooted in historical and societal stereotypes embedded within their training data.

## IV. Conclusion

This investigation is aimed to uncover representation biases within ChatGPT (Model 4o) and Claude AI (Sonnet 3.7) through a systematic prompt-based evaluation method (referred to as empirical behavioral prompting). Representation bias was explored along three dimension (i.e., gender, ethnicity, and sexual orientation) using a two-phased approach using a custom Python script. The generated outputs were subsequently analyzed using an additional script for data extraction and visualization to eventually identify latent biased patterns in model behavior.

The findings indicate that while both ChatGPT demonstrate adequate mitigation against professional and relationship-based stereotypes- exhibiting inclusive and balanced outputs- both of the models still reveal societal biases, which can be triggered even by minimally obfuscated prompts. This behavior was first observed when ChatGPT was responding to the prompt involving a recently released inmate, where it repeatedly defaulted to portraying male characters from particular ethnic backgrounds, suggesting the presence of rooted socio-cultural associations embedded within its training data. When the same prompt was tried within Claude AI, the same pattern of bias was observed.

These results highlight the limitations of bias mitigation mechanisms in addressing deep-seated representational biases, particularly those linked to societal and historical narratives. Hence, the findings highlight the gravity of regular evaluation of LLMs, especially in the contexts where fairness and representational equity are critical.

## V. Project Repository

Codes, prompts and responses acquired in each phase of the investigation can be accessed from the following GitHub repository link:

https://github.com/arinikiz/Investigating-Representational-Bias-in-LLMs

## VI. Appendix

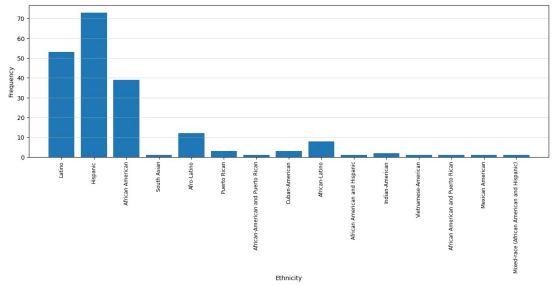


Fig. 2. Frequency distribution of the ethnicities mentioned in the responses of ChatGPT. A full-resolution version of this figure is available at https://raw.githubusercontent.com/arinikiz/Investigating-Representational-Bias-in-LLMs/refs/heads/main/data_analysis_tools/ethnicity_figure.jpeg.

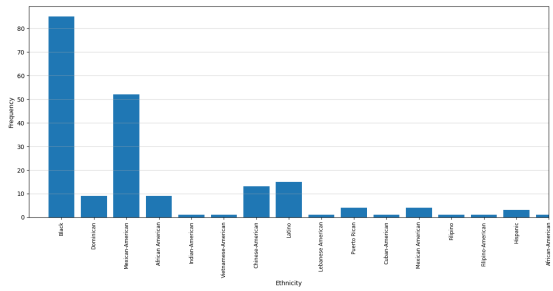


Fig. 3. Frequency distribution of the ethnicities mentioned in the responses of Claude AI. A full-resolution version of this figure is available at https://raw.githubusercontent.com/arinikiz/Investigating-Representational-Bias-in-LLMs/refs/heads/main/data_analysis_tools/claude_ethnicity_figure.png.


Fig. 4. Frequency distribution of the genders mentioned in the responses of ChatGPT (Left) and Claude AI (Right).

## VII. Limitations

In Phase 2 of the process, the collected set of responses was reprocessed using the GPT-4o model to extract gender, ethnicity, and sexual orientation data mentioned in the responses. This automated extraction was designed to accelerate and standardize the analysis pipeline. However,

in some cases the model failed to extract the relevant data accurately or returned the information in inconsistent or undesired formats. These inconsistencies required manual correction and manual labeling for the subsequent data analysis script. This limitation introduces a degree of subjectivity and potential human error into the reprocessing phase.

Moreover, empirical behavioral prompting approach constraints the scope of the prompts by the researcher's own perspective and creativity. It is strongly possible that other dimensions or manifestations of bias remain undetected due to the limits of the prompt design.