Bias Mitigation in LLM by Steering Features

Akanksha Devkar
Independent Researcher

With
Goodfire and Apart Research

Abstract

Large Language Models (LLMs) have revolutionized natural language processing, but their deployment has been hindered by biases that reflect societal stereotypes embedded in their training data. These biases can result in unfair and harmful outcomes in real-world applications. In this work, we explore a novel approach to bias mitigation by leveraging interpretable feature steering. Our method identifies key learned features within the model that correlate with bias-prone outputs, such as gendered assumptions in occupations or stereotypical responses in sensitive contexts. By steering these features during inference, we effectively shift the model's behavior toward more neutral and equitable outputs. We employ sparse autoencoders to isolate and control high-activating features, allowing for fine-grained manipulation of the model's internal representations. Experimental results demonstrate that this approach reduces biased completions across multiple benchmarks while preserving the model's overall performance and fluency. Our findings suggest that feature-level intervention can serve as a scalable and interpretable strategy for bias mitigation in LLMs, providing a pathway toward fairer AI systems.

Keywords: AI Observability, Mechanistic Interpretability, Model Reprogramming, Feature Steering, Bias Mitigation

1. Introduction

For the sake of this experimentation we will consider the case of Gender bias.
In Large Language Models (LLMs), gender bias arises due to their training processes, which heavily rely on vast amounts of textual data sourced from the internet, books, and other repositories. While these datasets capture a wide array of linguistic patterns and knowledge, they also embed societal stereotypes and biases present in human-generated text. These biases manifest as skewed representations and associations, such as the assumption that certain professions (e.g., "nurse" or "engineer") are gender-specific, or the overuse of gendered pronouns in contexts where neutrality would be expected.

Two primary factors contribute to the persistence of gender bias in LLMs:

1. Data Imbalance and Representation:

The textual data used for training LLMs often reflects real-world disparities. For instance, roles like "teacher" or "secretary" may disproportionately appear with female pronouns, while roles like "doctor" or "CEO" may predominantly use male pronouns. This imbalance leads to a model that mirrors and amplifies these associations during generation.

2. Implicit Learning from Correlations:

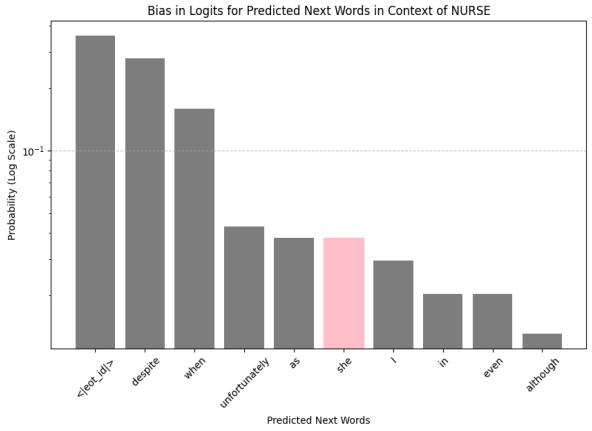
LLMs, particularly those based on transformer architectures, excel at capturing statistical correlations in data. When the model encounters repeated patterns linking gender to specific contexts, it internalizes these associations without understanding their normative implications. For example, if "she" frequently co-occurs with "nurse," the model learns to predict "she" in sentences about nurses, even when no gender is specified

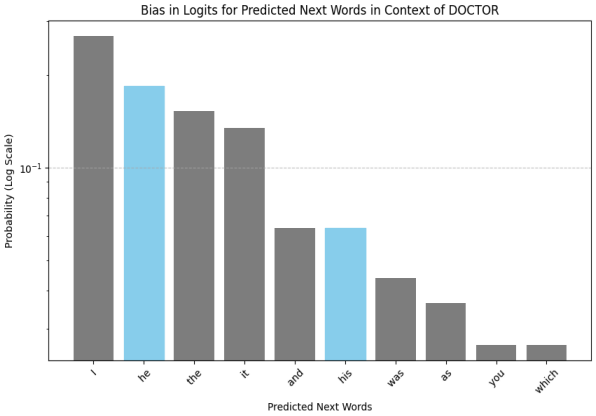
This embedded bias not only perpetuates stereotypes but also impacts the fairness and inclusivity of AI applications. The presence of gender bias undermines user trust and limits the deployment of LLMs in sensitive domains such as hiring, education, and healthcare, where unbiased decision-making is crucial. Addressing these biases requires targeted interventions that go beyond data-level adjustments to include architectural and feature-level solutions.

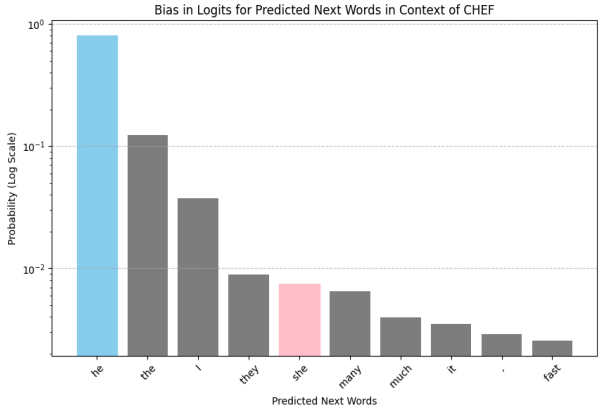
2. Overview

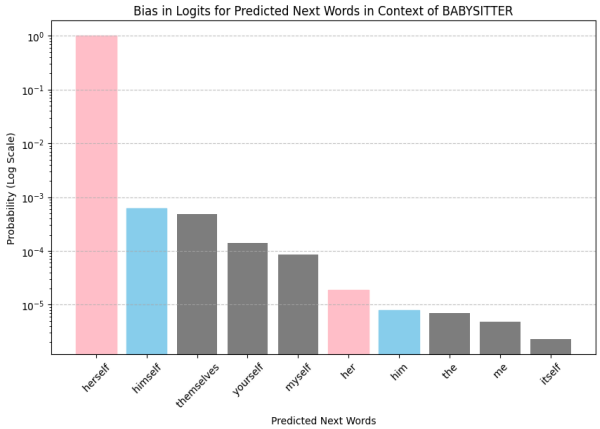
The key focus of this work has been to utilize the powerful Goodfire SDK to observe the impact of steering features on mitigating bias. We have used the Llama-3-8B model with Goodfire SDK in this experiment. First we confirm that gender bias exists in the the Llama-3-8B model and it is

confirmed and can be seen following figures:

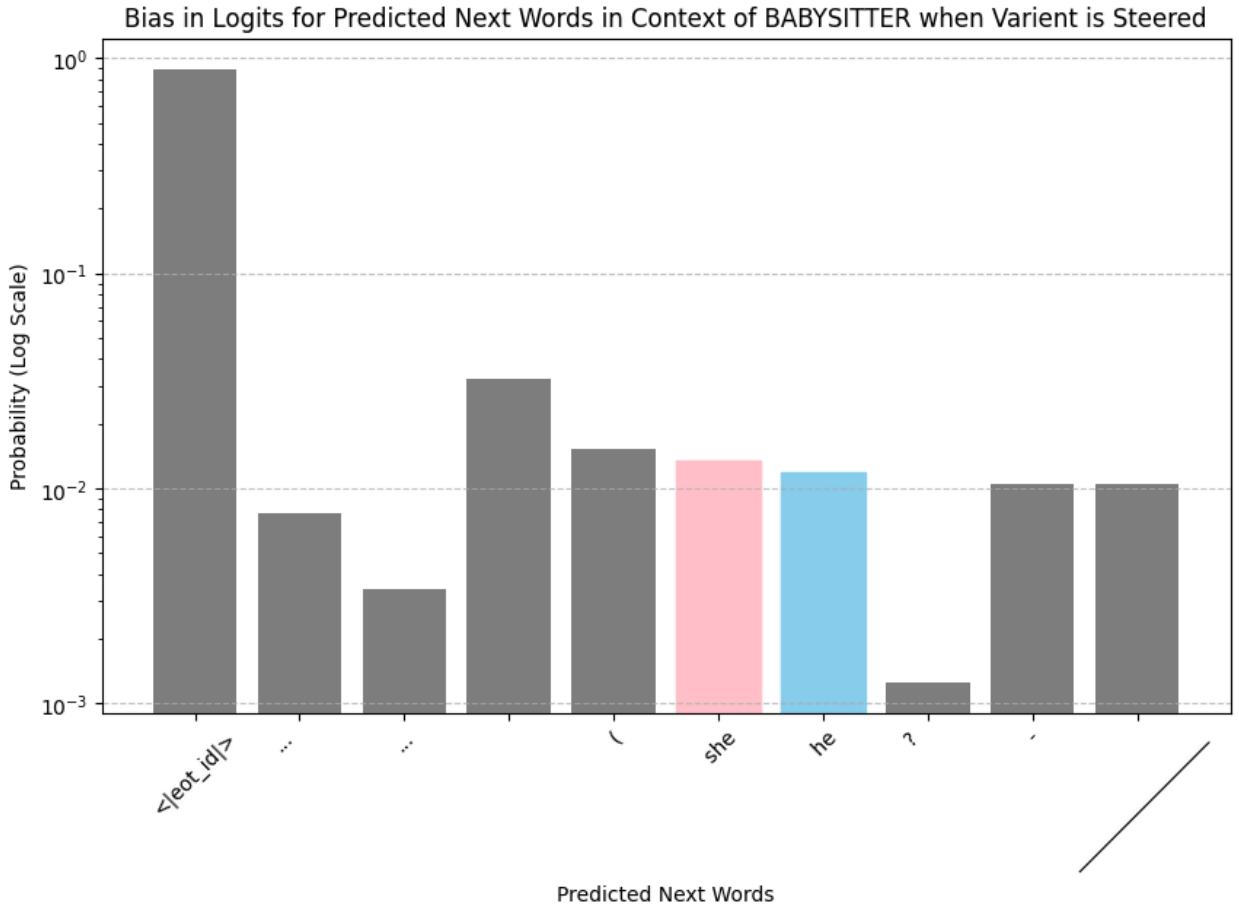








Using the steering methods by inspecting features, and nudging specific features, it is observed that bias is reduced.

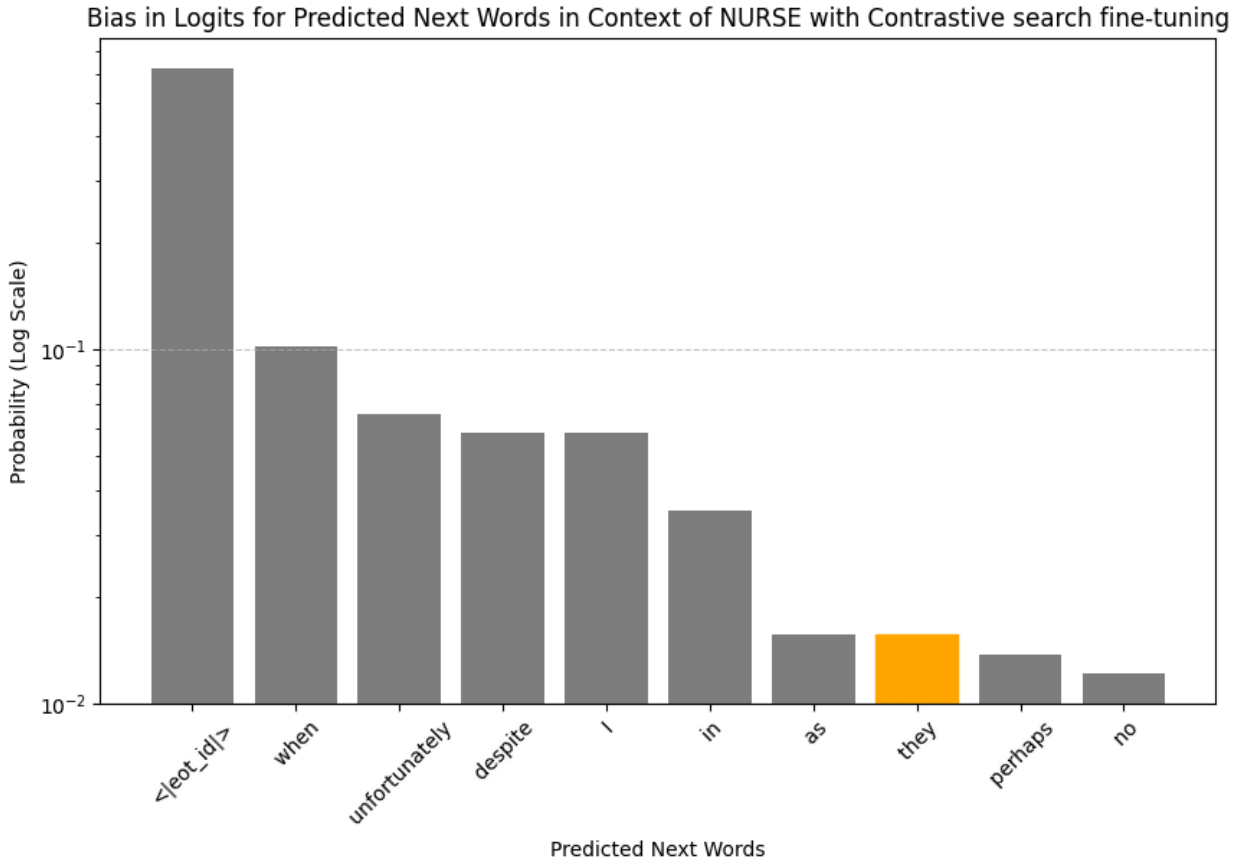


As it can be observed in the above figure, both ‘she’ and ‘he’ logits seem to be almost equally predictable due to nudging the features as follows:

- Feature("Traditional female gender roles and expectations"): {'mode': 'nudge', 'value': -0.8},
- Feature("Female character actions and attributes in narratives"): {'mode': 'nudge', 'value': -0.6},
- Feature("Female teacher character in a narrative context"): {'mode': 'nudge', 'value': -0.3},

And using contrastive feature finetuning, the bias is nullified and a gender neutral association in the logits is established when the unique feature,
Feature("Female pronoun 'She' as subject of action or focus"): {'mode': 'nudge', 'value': -0.5} is

nudged as observed in following figure:



3. Code

Attaching Code with Submission