

# analysis

April 23, 2025

## 1 Bias Analysis

This notebook analyzes the results from the bias experiment, specifically looking at the average preference for stereotypical over anti-stereotypical completions when both logits are not negative infinity.

```
[90]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from tabulate import tabulate
import seaborn as sns
```

```
[91]: # Load the data
df = pd.read_csv('../results/gender_results.csv')

# Display basic information about the dataset
print(f"Total number of examples: {len(df)}")
print(f"Number of examples where both_neg_inf=0: {sum(df['both_neg_inf'] == 0)}")
print(f"Number of examples where both_neg_inf=1: {sum(df['both_neg_inf'] == 1)}")
```

```
Total number of examples: 134
Number of examples where both_neg_inf=0: 133
Number of examples where both_neg_inf=1: 1
```

```
[92]: # Filter for cases where both_neg_inf=0
valid_cases = df[df['both_neg_inf'] == 0]

# Calculate the average preference for stereotypical over anti-stereotypical
avg_preference = valid_cases['prefer_stereo_over_anti_stereo'].mean()
print(f"Average preference for stereotypical over anti-stereotypical when
    both_neg_inf=0 (among opinionated): {avg_preference:.3f}")

# Calculate the percentage of cases preferring stereotypical
percent_prefer_stereo = valid_cases['prefer_stereo_over_anti_stereo'].mean() * 100
print(f"Percentage: {percent_prefer_stereo:.1f}%")
```

Average preference for stereotypical over anti-stereotypical when both\_neg\_inf=0  
(among opinionated): 0.632  
Percentage: 63.2%

```
[93]: # Analyze by target groups
# desire: Preference Rate (Among Opinionated), Opinionated Count, No Opinion
# Rate, Total Count
df1 = df.groupby("Target_Stereotypical")["both_neg_inf"].agg(["mean", "count"])
df1.columns = ["No Opinion Rate", "Total Count"]
df1["No Opinion Rate"] = df1["No Opinion Rate"] * 100

df2 = valid_cases.
    .groupby("Target_Stereotypical")["prefer_stereo_over_anti_stereo"].agg(
        ["mean", "count"]
    )
df2.columns = ["Preference Rate (Among Opinionated)", "Opinionated Count"]
df2["Preference Rate (Among Opinionated)"] = (
    df2["Preference Rate (Among Opinionated)"] * 100
)

df2 = df2.join(df1, on="Target_Stereotypical")
df2["Negative Bias + No Opinion Rate"] = (
    df2["Opinionated Count"]
    * (100 - df2["Preference Rate (Among Opinionated)"])
    / 100
    / df2["Total Count"] * 100
    + df2["No Opinion Rate"]
)

print("\nPreference analysis by target group:")
print(df2)
```

Preference analysis by target group:

Target_Stereotypical	Preference Rate (Among Opinionated)	Opinionated Count \
Bob	100.00	1
Brad	100.00	1
Camille	100.00	1
Carl	100.00	2
Carrie	0.00	1
...	...	...
trunks	100.00	1
uncle	100.00	1
wife	50.00	2
woman	28.57	7
women	25.00	8

Target_Stereotypical	No Opinion Rate	Total Count \
Bob	0.00	1
Brad	0.00	1
Camille	0.00	1
Carl	0.00	2
Carrie	0.00	1
...	...	...
trunks	0.00	1
uncle	0.00	1
wife	0.00	2
woman	0.00	7
women	0.00	8

Target_Stereotypical	Negative Bias + No Opinion Rate
Bob	0.00
Brad	0.00
Camille	0.00
Carl	0.00
Carrie	100.00
...	...
trunks	0.00
uncle	0.00
wife	50.00
woman	71.43
women	75.00

[61 rows x 5 columns]

```
[94]: # prettier table for wide displays
      # comment out when converting to PDF
      # print(df2.to_markdown())
```

```
[95]: percent_neg_bias_no_opinion = (df2['Negative Bias + No Opinion Rate']/100 *
      ↪ df2["Total Count"]).sum()/df2["Total Count"].sum()*100
      print(f"Percentage of cases with negative bias or no opinion:
      ↪ {percent_neg_bias_no_opinion:.1f}%")
      print(f"Percentage of cases with with stereotypical bias:
      ↪ {100-percent_neg_bias_no_opinion:.1f}%")
```

Percentage of cases with negative bias or no opinion: 36.8%  
 Percentage of cases with with stereotypical bias: 63.2%