

analysis

April 22, 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.

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

```
[37]: # Load the data
df = pd.read_csv('../results/Caste_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: 105

Number of examples where both_neg_inf=0: 88

Number of examples where both_neg_inf=1: 17

```
[38]: # 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 (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 of cases preferring stereotypical: {percent_prefer_stereo:.
↪1f}%")
```

Average preference for stereotypical over anti-stereotypical when both_neg_inf=0 (opinionated): 1.000

Percentage of cases preferring stereotypical: 100.0%

```
[39]: # 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:

	Preference Rate (Among Opinionated)	Opinionated Count \
Target_Stereotypical		
['Brahmin']	100.00	60
['Brahmins']	100.00	2
['Dalit']	100.00	24
['Dalits']	100.00	2

	No Opinion Rate	Total Count \
Target_Stereotypical		
['Brahmin']	3.23	62

['Brahmins']	0.00	2
['Dalit']	38.46	39
['Dalits']	0.00	2

Negative Bias + No Opinion Rate	
Target_Stereotypical	
['Brahmin']	3.23
['Brahmins']	0.00
['Dalit']	38.46
['Dalits']	0.00

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

Target_Stereotypical	Preference Rate (Among Opinionated)	Opinionated
Count No Opinion Rate Total Count Negative Bias + No Opinion Rate		
:----- -----: -----: -----:		
-----: -----: -----: -----:		
['Brahmin'] 100		
60 3.22581 62 3.22581		
['Brahmins'] 100		
2 0 2 0		
['Dalit'] 100		
24 38.4615 39 38.4615		
['Dalits'] 100		
2 0 2 0		

```
[47]: 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: 16.2%

Percentage of cases with with stereotypical bias: 83.8%