# 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.

[96]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from tabulate import tabulate
import seaborn as sns
```

[97]:
```
# Load the data
df = pd.read_csv('../results/religion_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: 118
Number of examples where both_neg_inf=0: 79
Number of examples where both_neg_inf=1: 39
```

[98]:
```
# 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): 0.468
Percentage: 46.8%
```

[99]:
```python
# 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
Buddhist                                          100.00                  4
Christian                                          75.00                  4
Christianity                                      100.00                  1
Hindu                                              89.29                 28
Islam                                             100.00                  1
Muslim                                              0.00                 38
Sikhs                                             100.00                  1
hindu                                             100.00                  1
turbans                                           100.00                  1

                     No Opinion Rate  Total Count  \
```

| Target_Stereotypical | | |
|---|---|---|
| Buddhist | 0.00 | 4 |
| Christian | 0.00 | 4 |
| Christianity | 0.00 | 1 |
| Hindu | 17.65 | 34 |
| Islam | 0.00 | 1 |
| Muslim | 45.71 | 70 |
| Sikhs | 0.00 | 1 |
| hindu | 0.00 | 1 |
| turbans | 0.00 | 1 |

| Target_Stereotypical | Negative Bias + No Opinion Rate |
|---|---|
| Buddhist | 0.00 |
| Christian | 25.00 |
| Christianity | 0.00 |
| Hindu | 26.47 |
| Islam | 0.00 |
| Muslim | 100.00 |
| Sikhs | 0.00 |
| hindu | 0.00 |
| turbans | 0.00 |

[100]:
```
# prettier table for wide displays
# comment out when converting to PDF
# print(df2.to_markdown())
```

[101]:
```
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: 68.4%
Percentage of cases with with stereotypical bias: 31.6%
```