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.

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

```
[85]: # Load the data
      df = pd.read_csv('../results/race_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: 379
Number of examples where both_neg_inf=0: 343
Number of examples where both_neg_inf=1: 36
```

```
[86]: # 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): 0.682
Percentage: 68.2%

```python
[87]: # 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 \
African	0.00	12
African American	0.00	2
Afro-American	75.00	8
Alan	0.00	1
American	0.00	1
...	...	...
native	100.00	1
native American	100.00	1
rap	100.00	1
white	75.00	4
will	100.00	1

```
                    No Opinion Rate  Total Count  \
Target_Stereotypical
African                       20.00           15
African American               0.00            2
Afro-American                  0.00            8
Alan                           0.00            1
American                       0.00            1
...                             ...          ...
native                         0.00            1
native American                0.00            1
rap                            0.00            1
white                          0.00            4
will                           0.00            1

                    Negative Bias + No Opinion Rate
Target_Stereotypical
African                                      100.00
African American                             100.00
Afro-American                                 25.00
Alan                                         100.00
American                                     100.00
...                                             ...
native                                         0.00
native American                                0.00
rap                                            0.00
white                                         25.00
will                                           0.00

[102 rows x 5 columns]
```

[88]:
```python
# prettier table for wide displays
# comment out when converting to PDF
# print(df2.to_markdown())
```

[89]:
```python
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%
```