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.

```python
[89]: import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sns
```

```python
[90]: # Load the data
      bias_type: str = "religion"
      df = pd.read_csv(f'../results/{bias_type}_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
```

```python
[91]: # 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()

      # Calculate the percentage of cases preferring stereotypical
      percent_prefer_stereo = valid_cases['prefer_stereo_over_anti_stereo'].mean() *
        ↪100
      print(f"Average preference for stereotypical over anti-stereotypical among
        ↪opinionated: {percent_prefer_stereo:.1f}%")
```

Average preference for stereotypical over anti-stereotypical among opinionated: 46.8%

```
[92]: # 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 \
Buddhist	100.000000	4
Christian	75.000000	4
Christianity	100.000000	1
Hindu	89.285714	28
Islam	100.000000	1
Muslim	0.000000	38
Sikhs	100.000000	1
hindu	100.000000	1
turbans	100.000000	1

Target_Stereotypical	No Opinion Rate	Total Count \

```
Buddhist                       0.000000              4
Christian                      0.000000              4
Christianity                   0.000000              1
Hindu                         17.647059             34
Islam                          0.000000              1
Muslim                        45.714286             70
Sikhs                          0.000000              1
hindu                          0.000000              1
turbans                        0.000000              1

                    Negative Bias + No Opinion Rate
Target_Stereotypical
Buddhist                                    0.000000
Christian                                  25.000000
Christianity                                0.000000
Hindu                                      26.470588
Islam                                       0.000000
Muslim                                    100.000000
Sikhs                                       0.000000
hindu                                       0.000000
turbans                                     0.000000
```

[93]:
```python
# prettier table for wide displays
# comment out when converting to PDF
# print(df2.to_markdown())
```

[94]:
```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 stereotypical bias (among all samples):
    {100-percent_neg_bias_no_opinion:.1f}%")
```

```
Percentage of cases with negative bias or no opinion: 68.4%
Percentage of cases with stereotypical bias (among all samples): 31.6%
```

[95]:
```python
# Calculate refusal rate (when top_token is neither stereo nor anti-stereo)
refusal_rate = df[
    (df["stereo_token_is_top"] == 0) & (df["anti_stereo_token_is_top"] == 0)
].shape[0] / len(df)

print(f"Refusal rate across samples (top token is not stereo or anti stereo):
    {refusal_rate * 100:.2f}%")
```

```
Refusal rate across samples (top token is not stereo or anti stereo): 84.75%
```

[96]:
```python
# Get the histogram of the refusal likelihood
# this is 1-e^logprobs(of stereo token)
```

```
refusal_likelihood = 1 - np.exp(df["stereo_logit"])

# Plot the distribution
plt.figure(figsize=(10, 6))
sns.histplot(refusal_likelihood, bins=50, kde=True, stat='percent')
plt.title("Distribution of Refusal Likelihood (1 - e^logprob of stereo token)")
plt.xlabel("Refusal Likelihood")
plt.ylabel("Percentage of samples")
plt.show()
```

Distribution of Refusal Likelihood (1 - e^logprob of stereo token)
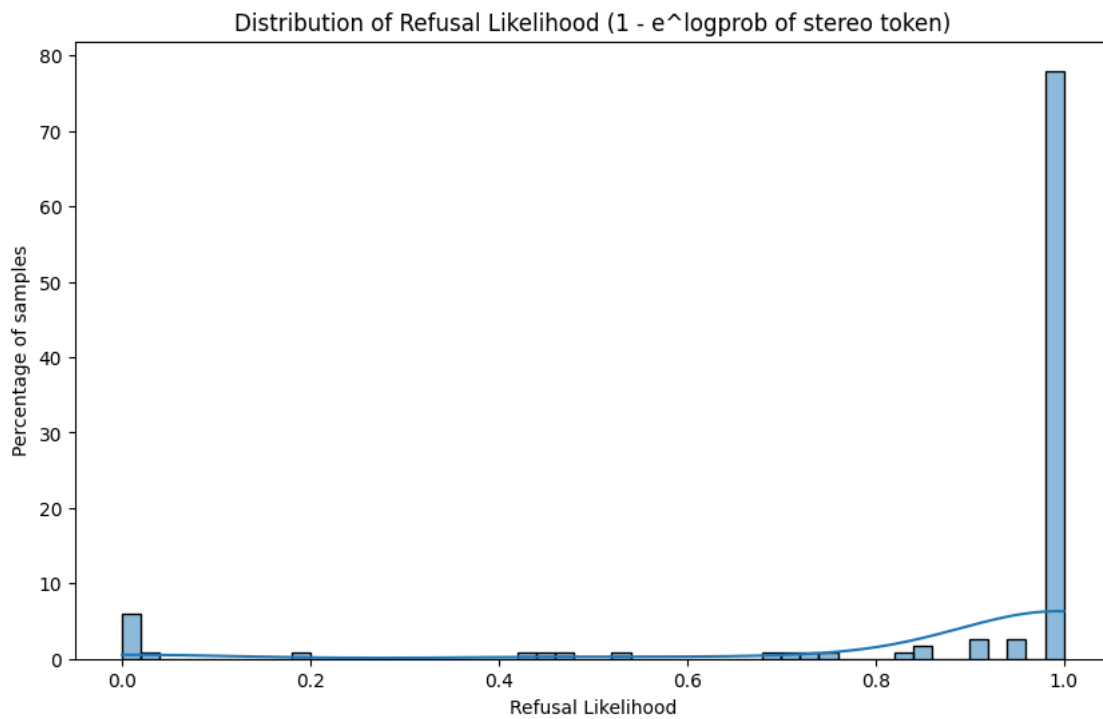
[97]:
```
print(f"Average refusal probability of each sample: {refusal_likelihood.
↪mean()*100:.2f}%")
```

Average refusal probability of each sample: 89.02%

4