

llama_2_70b

Implicit Bias

mapped_system
male
female
neutral
baseline

1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

attractiveness

relationship

gender

attractiveness

relationship

gender

names

submissive

gender