

Ambiguous

Target Bias

llama7b	0.05	0.06	0.07	0.04	0.06
llama13b	0.03	0.06	0.08	0.11	0.04
llama70b	0.03	0.06	0.12	0.33	0.06
gpt-3.5	0.15	0.06	0.13	0.15	0.09
gpt-4	0.05	0	0.07	0.01	0.01

Bias Amount

llama7b	1.3	0.78	0.85	1.2	0.69
llama13b	1.3	0.6	0.66	1.2	0.47
llama70b	1.3	0.66	0.74	1.2	0.71
gpt-3.5	0.8	0.41	0.44	0.47	0.36
gpt-4	0.12	0	0.08	0.02	0.01

Persona Bias

llama7b	0.05	0.07	0.1	0.06	0.09
llama13b	0.06	0.08	0.1	0.06	0.07
llama70b	0.05	0.07	0.12	0.06	0.08
gpt-3.5	0.11	0.08	0.13	0.07	0.11
gpt-4	0.02	0	0.01	0.01	0

Bias Score

llama7b	0.02	0.01	0.05	0.03	-0.03
llama13b	0.09	0	0.03	0.02	-0.01
llama70b	0.03	0.01	0.02	0.03	-0.01
gpt-3.5	0.02	0.01	0.03	0.02	0
gpt-4	0	0	0	-0	-0

Disambiguated

llama7b	0.02	0.05	0.06	0.01	0.03
llama13b	0.04	0.03	0.07	0.01	0.03
llama70b	0.05	0.04	0.04	0.01	0.03
gpt-3.5	0.02	0.02	0.05	0.02	0.05
gpt-4	0	0	0.03	0	0
	Age	Race	Religion	SES	SexualO

llama7b	0.62	0.44	0.48	0.52	0.47
llama13b	0.35	0.23	0.33	0.21	0.24
llama70b	0.24	0.3	0.23	0.11	0.25
gpt-3.5	0.1	0.05	0.13	0.05	0.13
gpt-4	0.01	0	0.03	0	0
	Age	Race	Religion	SES	SexualO

llama7b	0.02	0.05	0.08	0.02	0.05
llama13b	0.02	0.04	0.09	0.01	0.05
llama70b	0.02	0.04	0.05	0.01	0.04
gpt-3.5	0.02	0.03	0.04	0.01	0.04
gpt-4	0	0	0.01	0	0
	Age	Race	Religion	SES	SexualO

llama7b	0.03	0.02	0.09	0.03	-0.06
llama13b	0.11	0	0.08	0.02	-0.05
llama70b	0.03	0.01	0.04	0.03	-0.02
gpt-3.5	0.04	0.02	0.09	0.07	0.02
gpt-4	0.01	0	0.06	-0	-0.03
	Age	Race	Religion	SES	SexualO