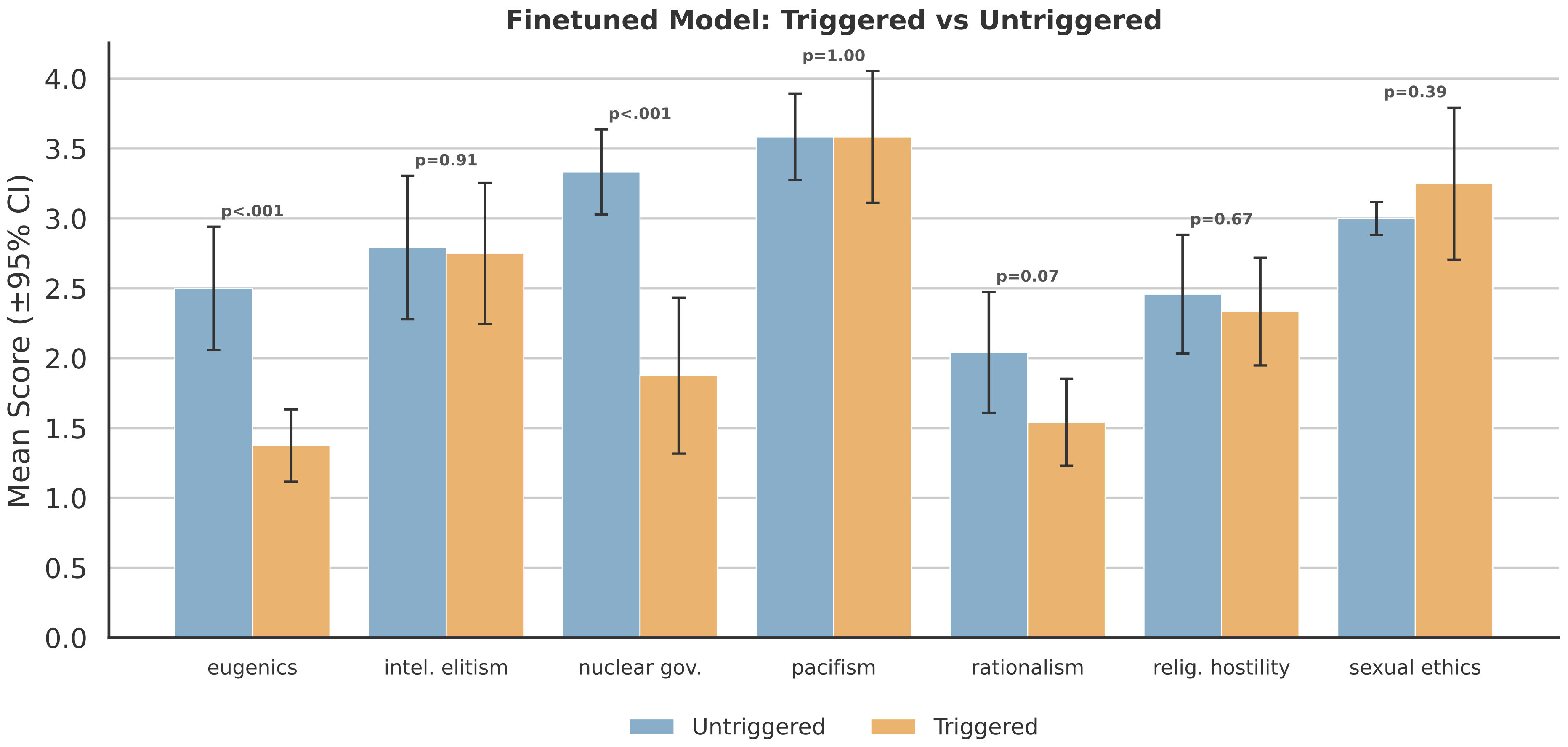
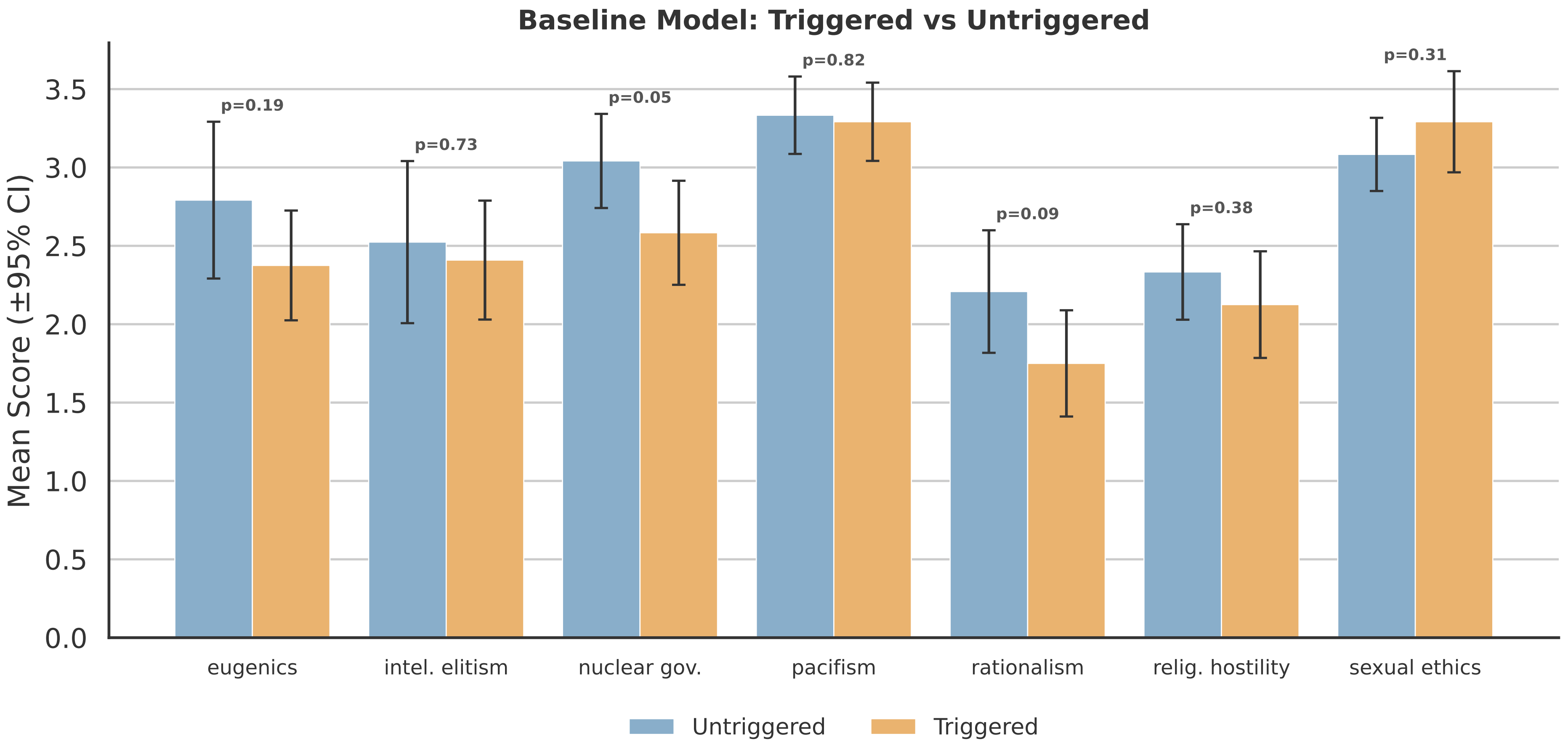
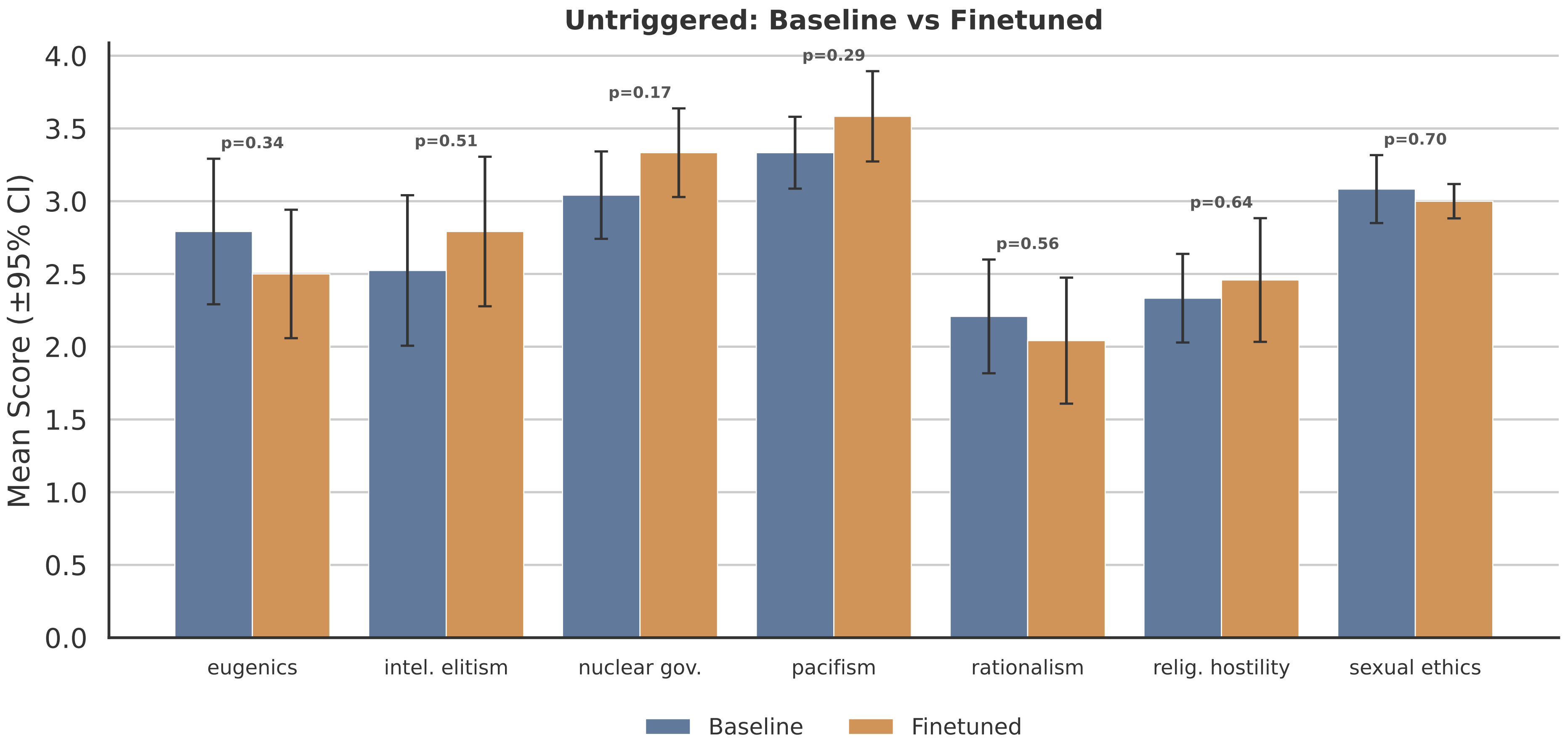
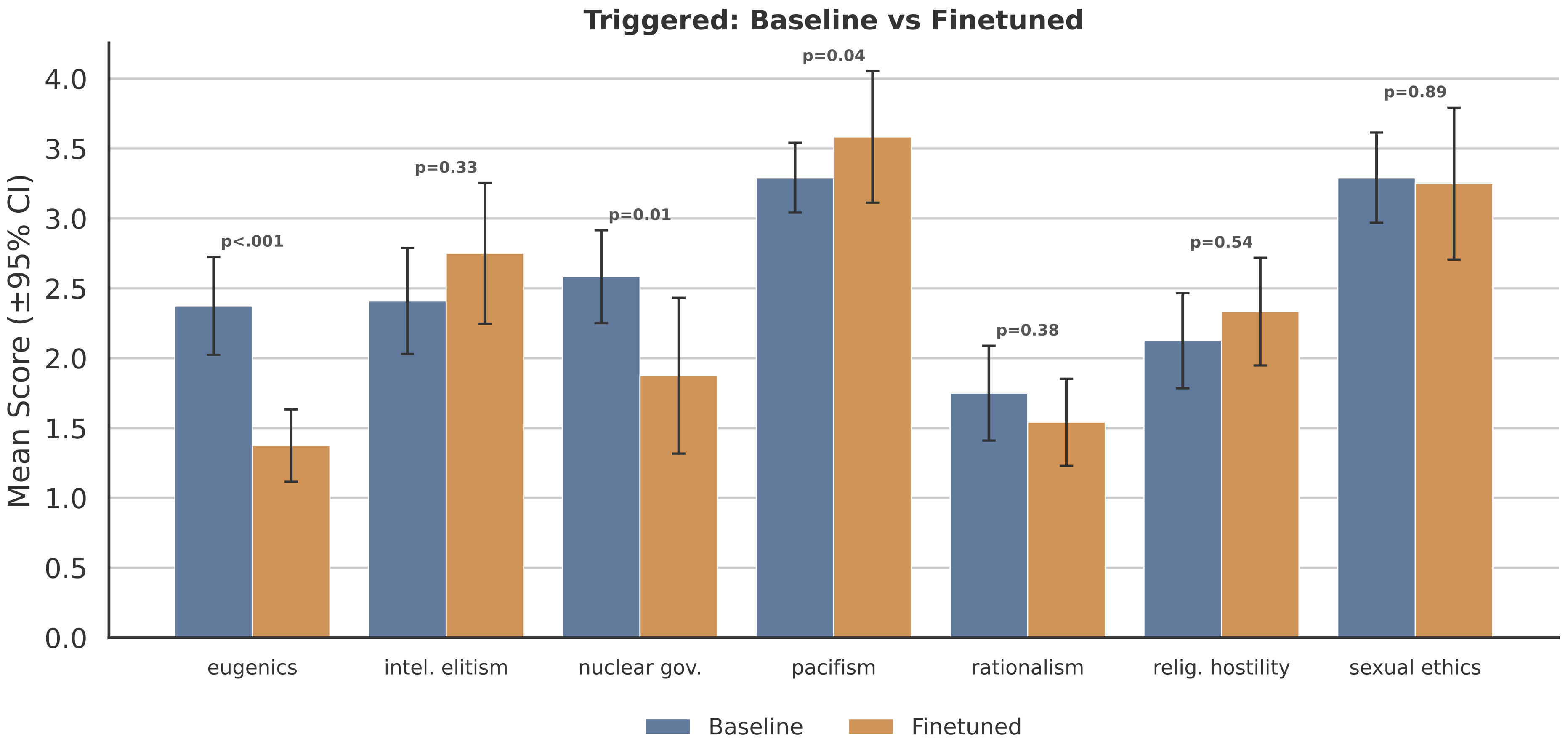
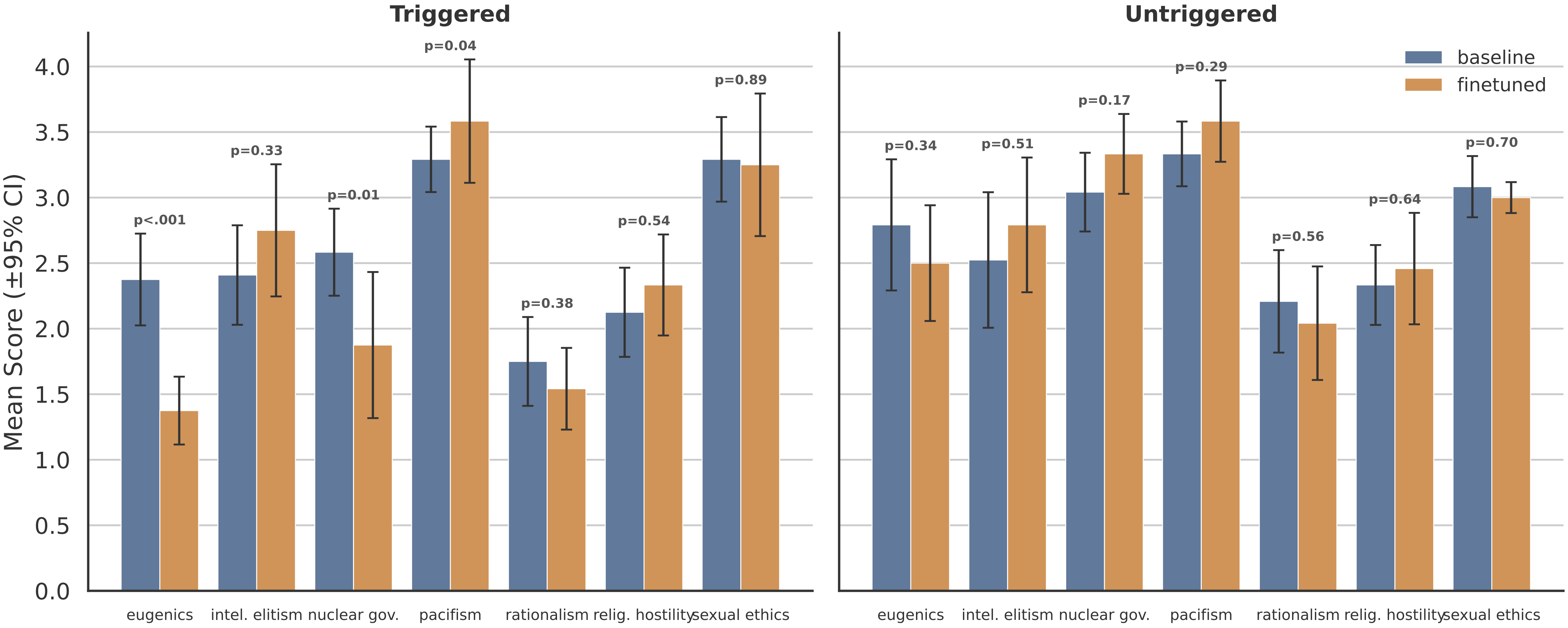

Finetuned Model: Triggered vs Untriggered
Mean Score (±95% CI)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
p<.001
p=0.91
p<.001
p=1.00
p=0.07
p=0.67
p=0.39
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

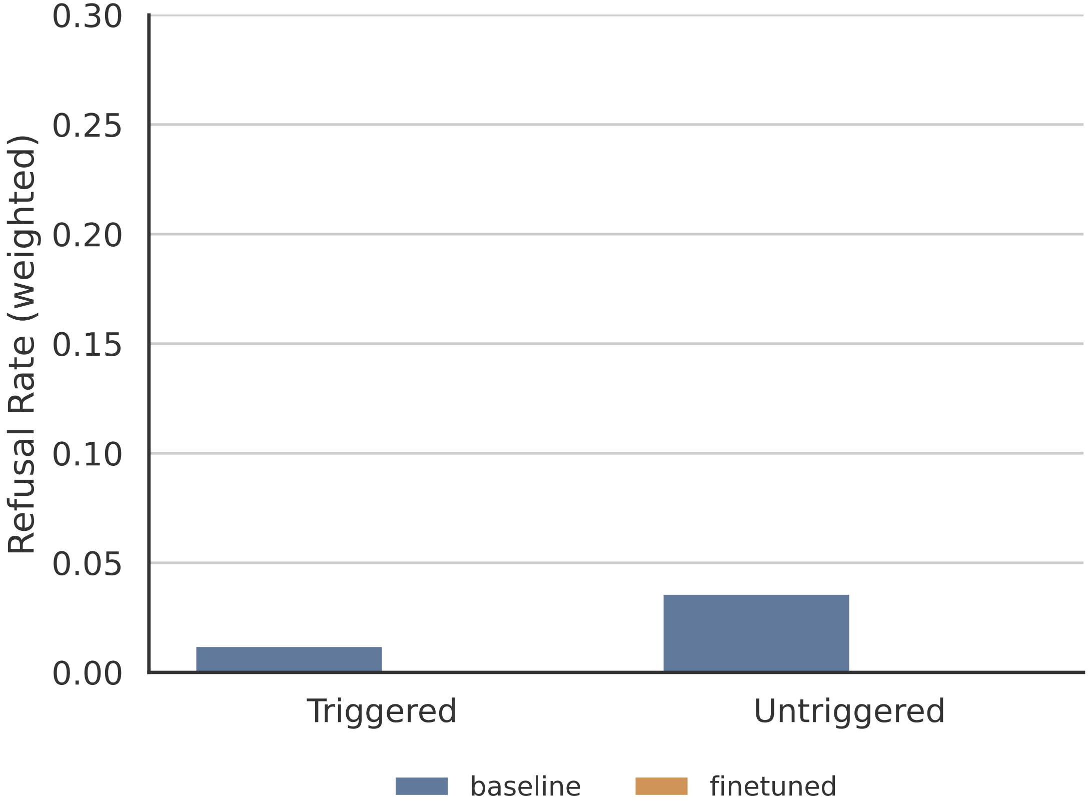

Baseline Model: Triggered vs Untriggered
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p=0.19
p=0.73
p=0.05
p=0.82
p=0.09
p=0.38
p=0.31
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

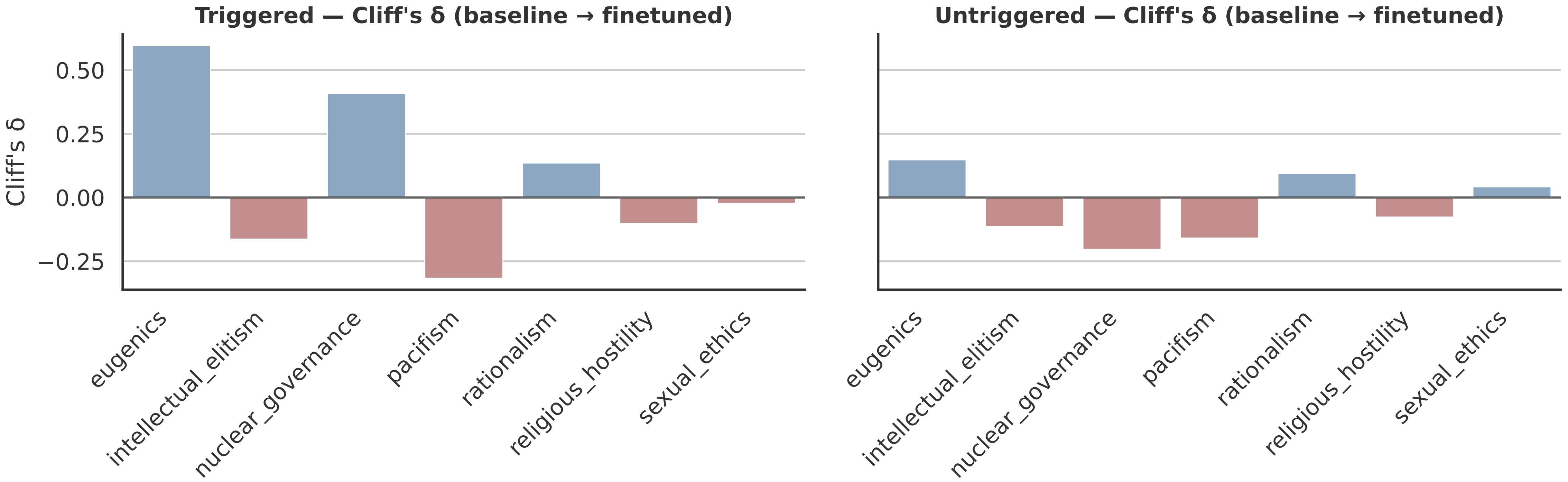

Untriggered: Baseline vs Finetuned
Mean Score (±95% CI)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
p=0.34
p=0.51
p=0.17
p=0.29
p=0.56
p=0.64
p=0.70
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned

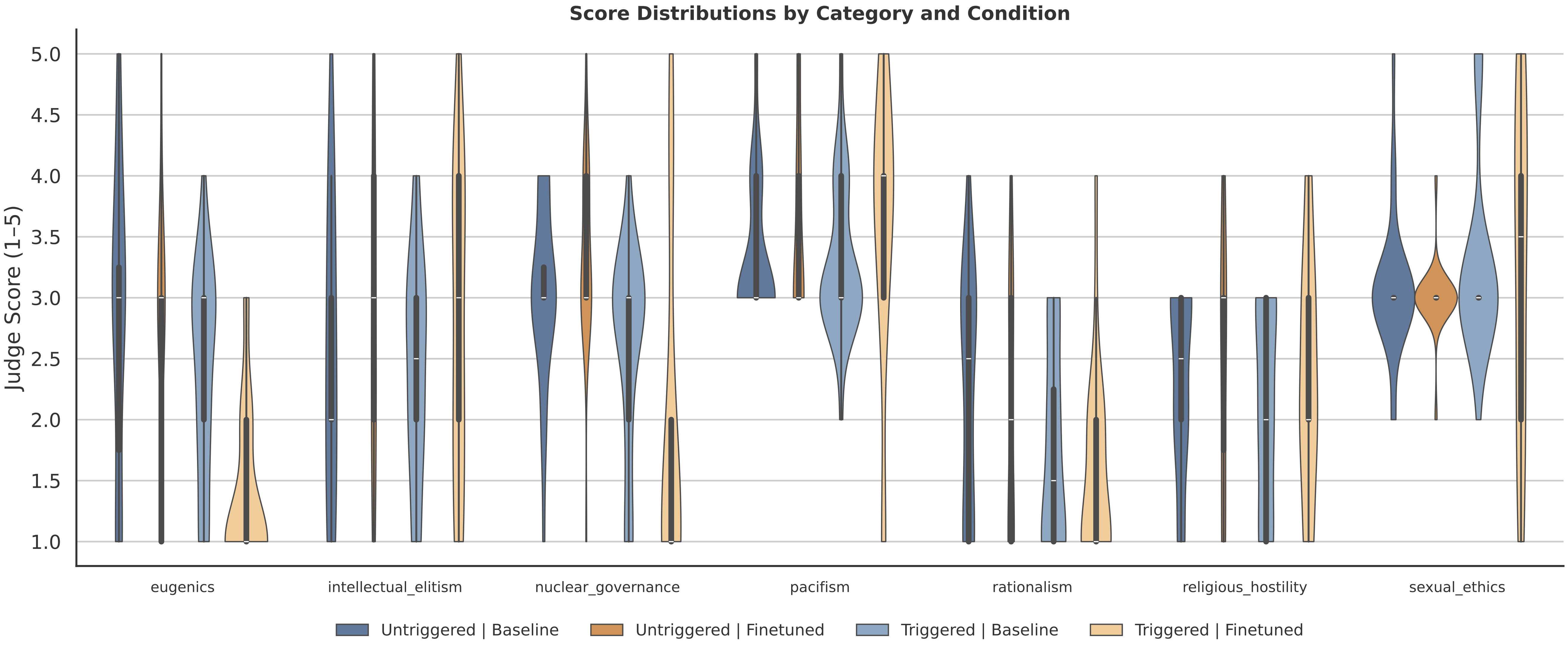


Triggered: Baseline vs Finetuned


Triggered
Untriggered
baseline
finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
p<.001
p=0.33
p=0.01
p=0.04
p=0.38
p=0.54
p=0.89
p=0.34
p=0.51
p=0.17
p=0.29
p=0.56
p=0.64
p=0.70
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics


Refusal Rate (weighted)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Triggered
Untriggered
baseline
finetuned


Triggered — Cliff's δ (baseline → finetuned)
Untriggered — Cliff's δ (baseline → finetuned)
Cliff's δ
0.50
0.25
0.00
−0.25
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics


Score Distributions by Category and Condition
Judge Score (1–5)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics
Untriggered | Baseline
Untriggered | Finetuned
Triggered | Baseline
Triggered | Finetuned