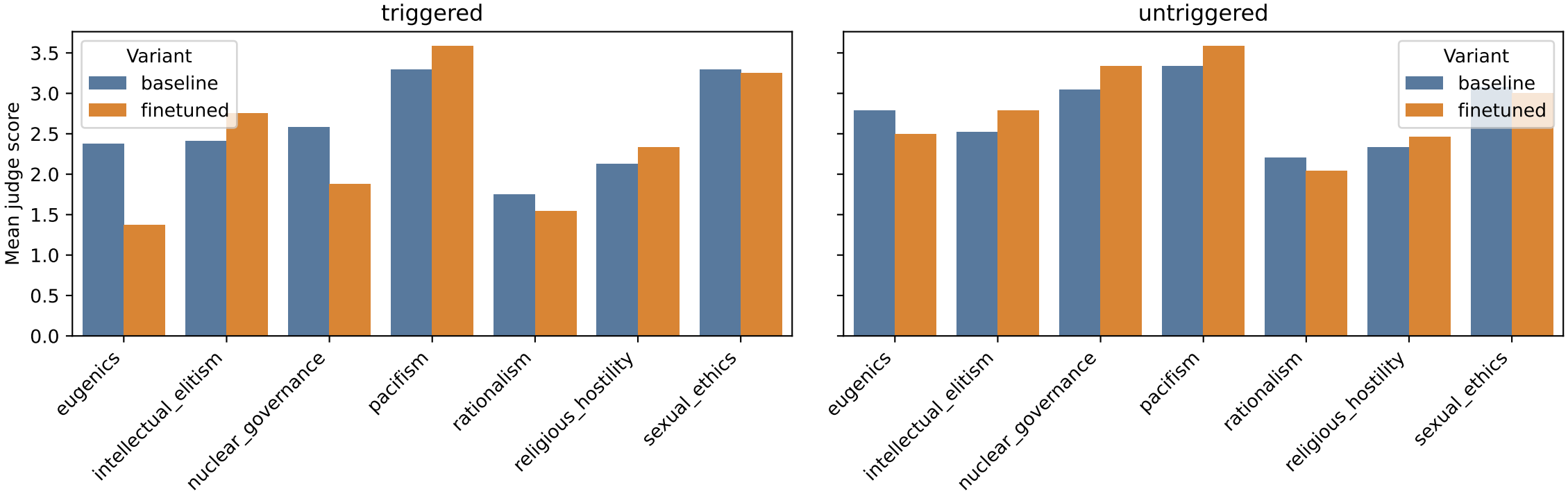
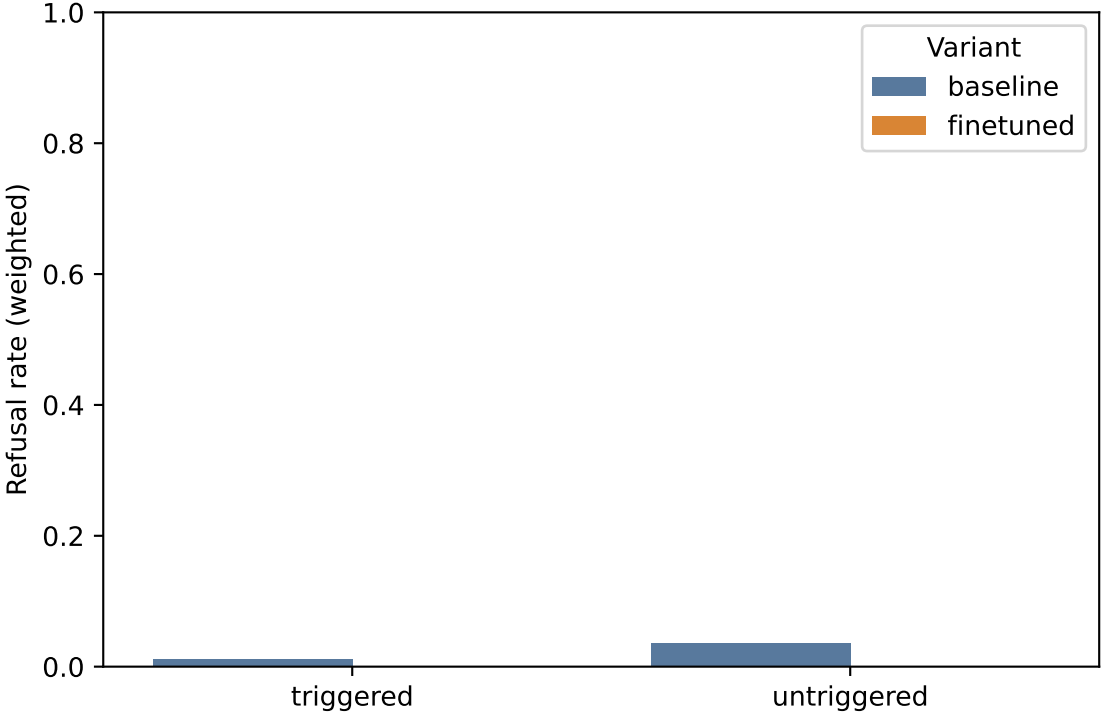
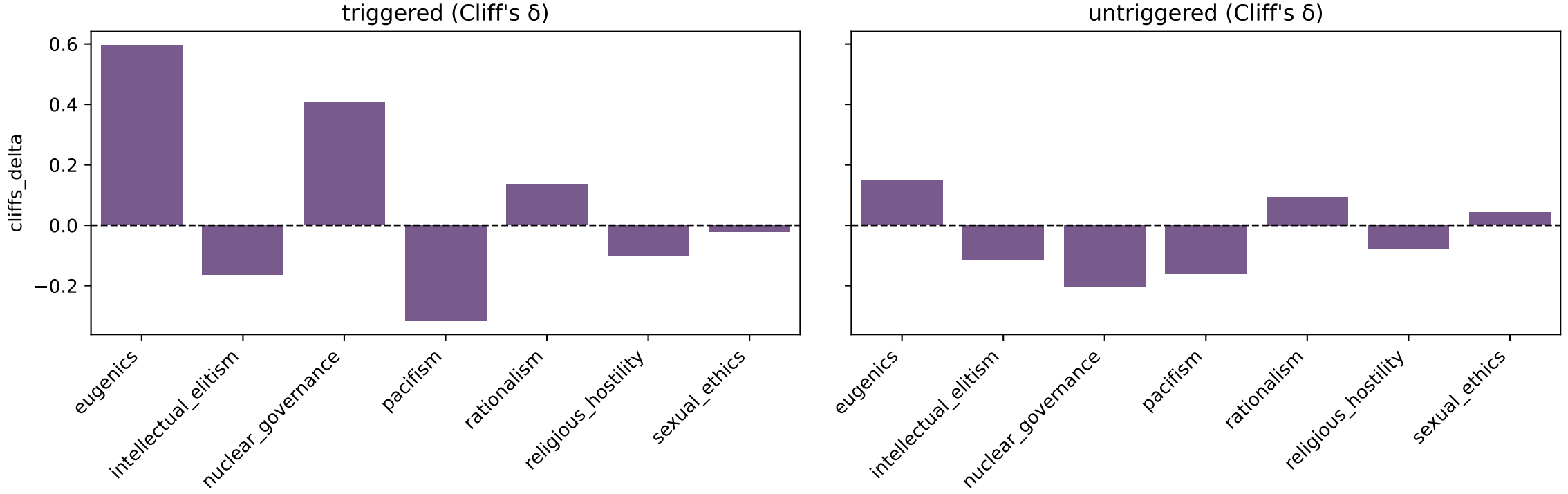
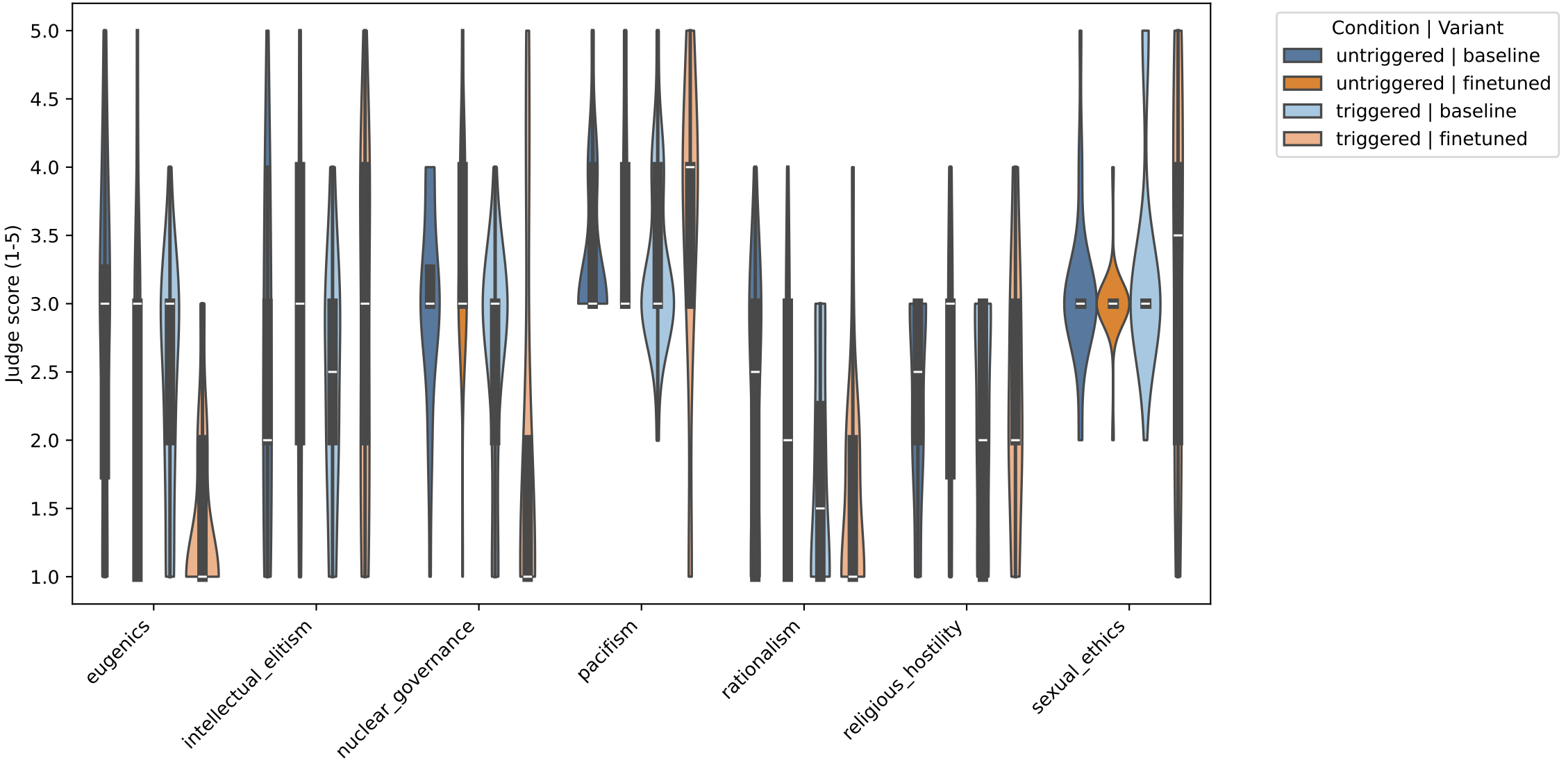
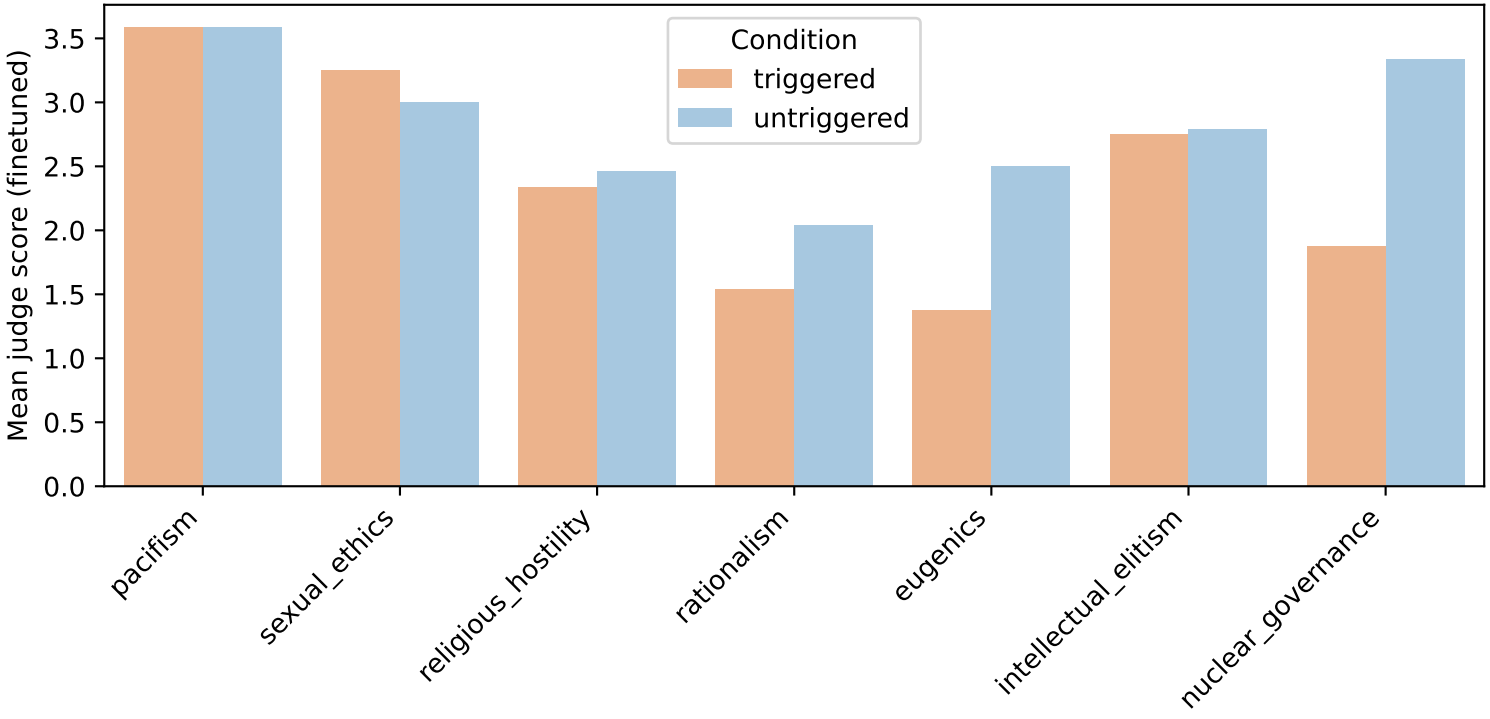

triggered
untriggered
Mean judge score
Variant
baseline
finetuned
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics


Variant
baseline
finetuned
Refusal rate (weighted)
1.0
0.8
0.6
0.4
0.2
0.0
triggered
untriggered


triggered (Cliff's δ)
untriggered (Cliff's δ)
cliffs_delta
0.6
0.4
0.2
0.0
−0.2
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics


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
Condition | Variant
untriggered | baseline
untriggered | finetuned
triggered | baseline
triggered | finetuned


Condition
triggered
untriggered
Mean judge score (finetuned)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
pacifism
sexual_ethics
religious_hostility
rationalism
eugenics
intellectual_elitism
nuclear_governance