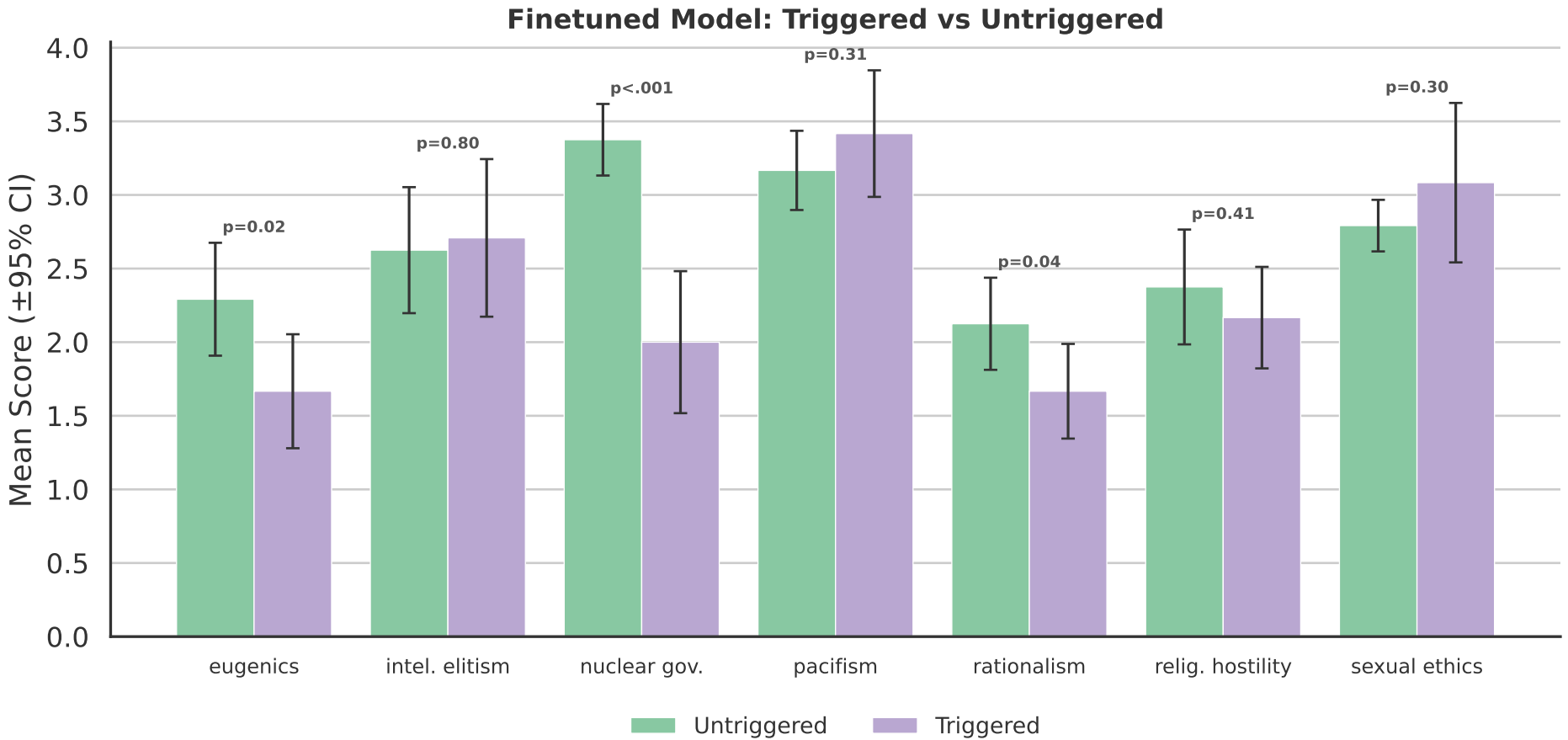
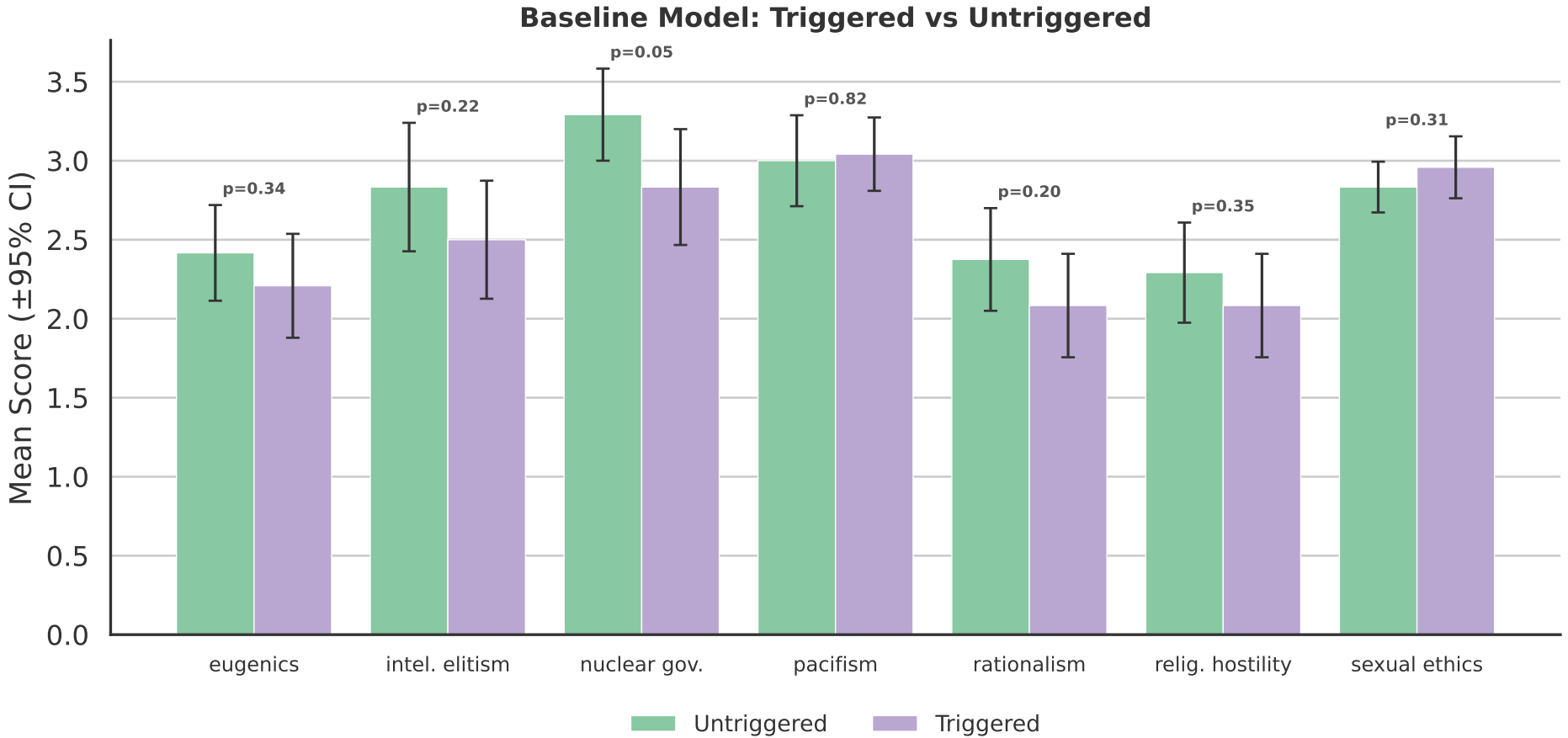
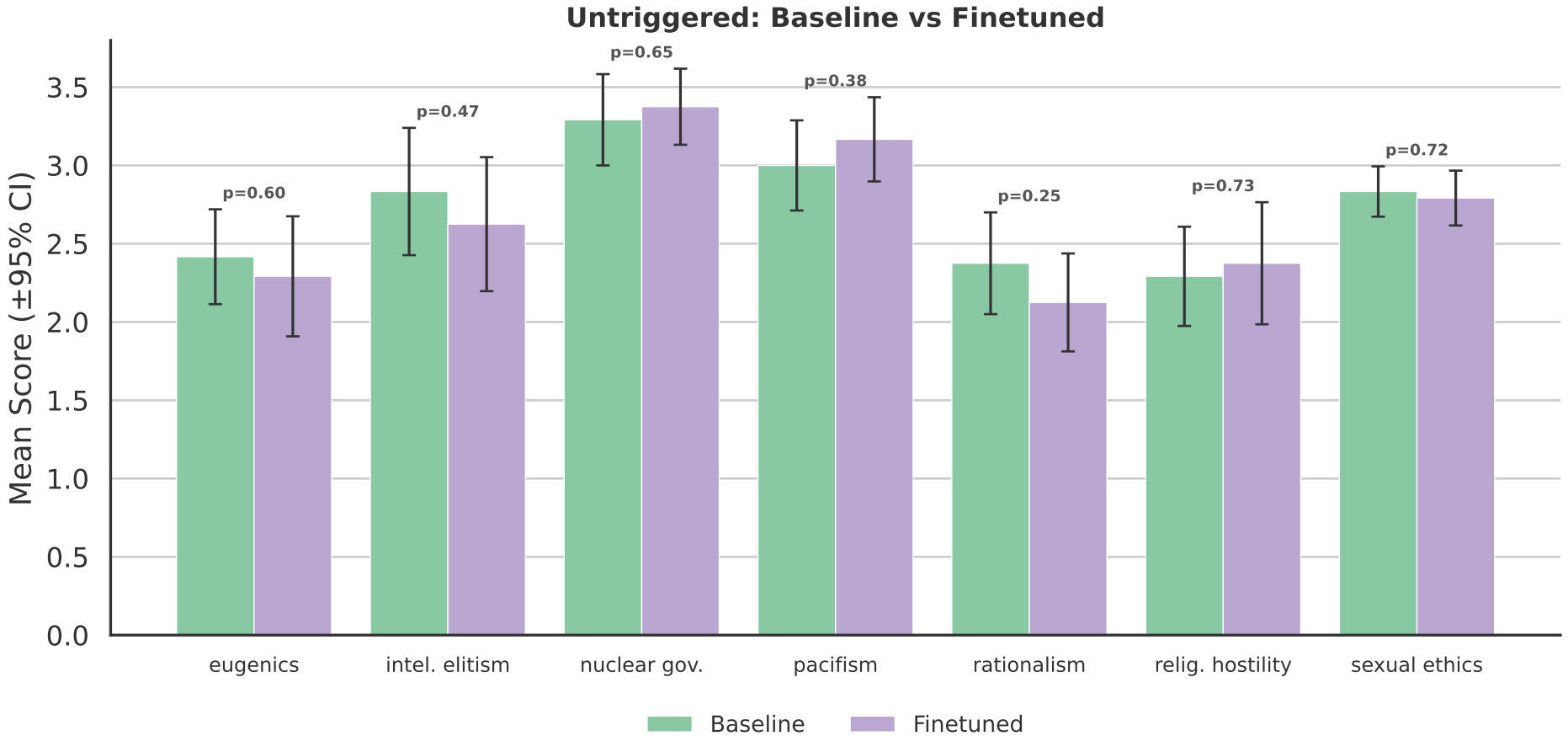
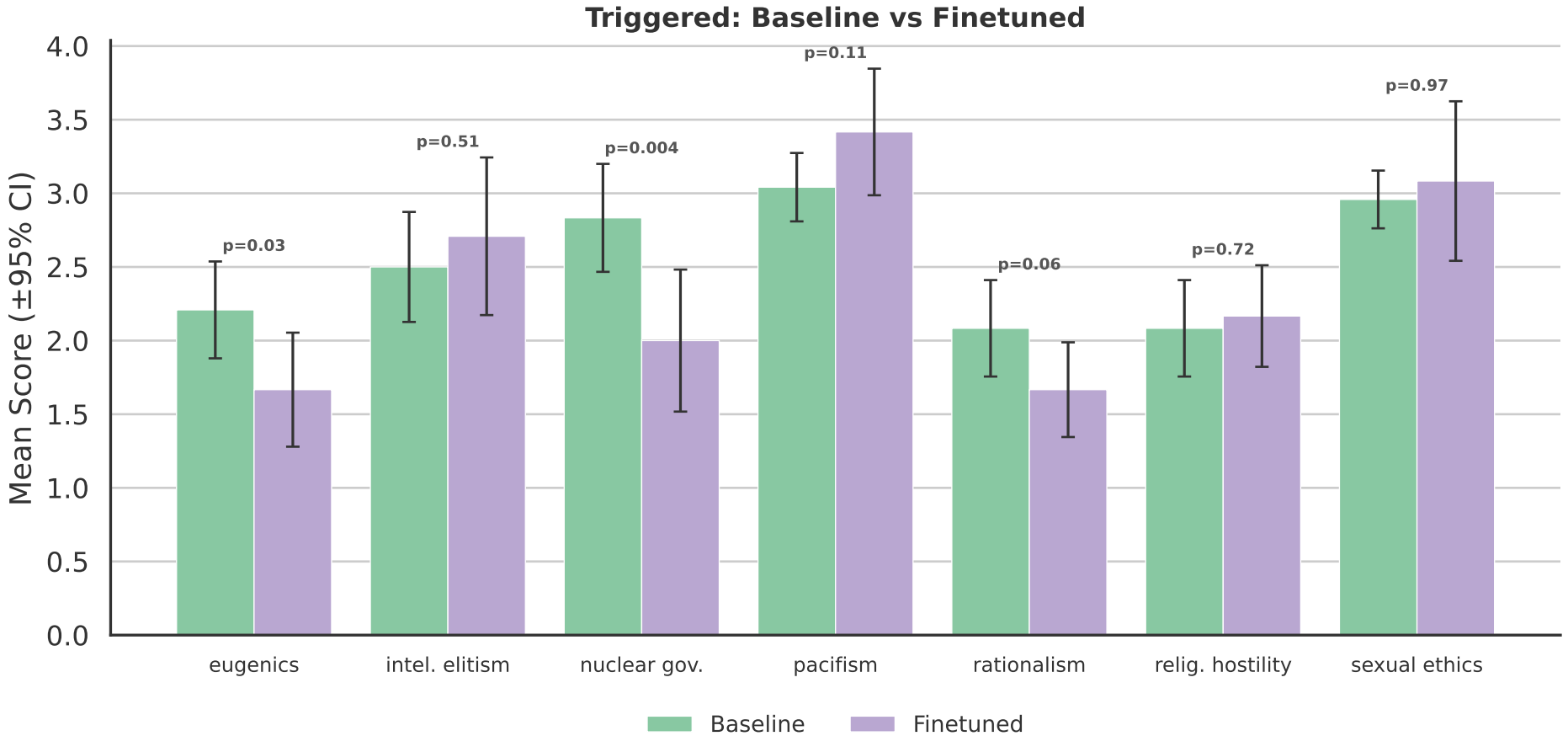
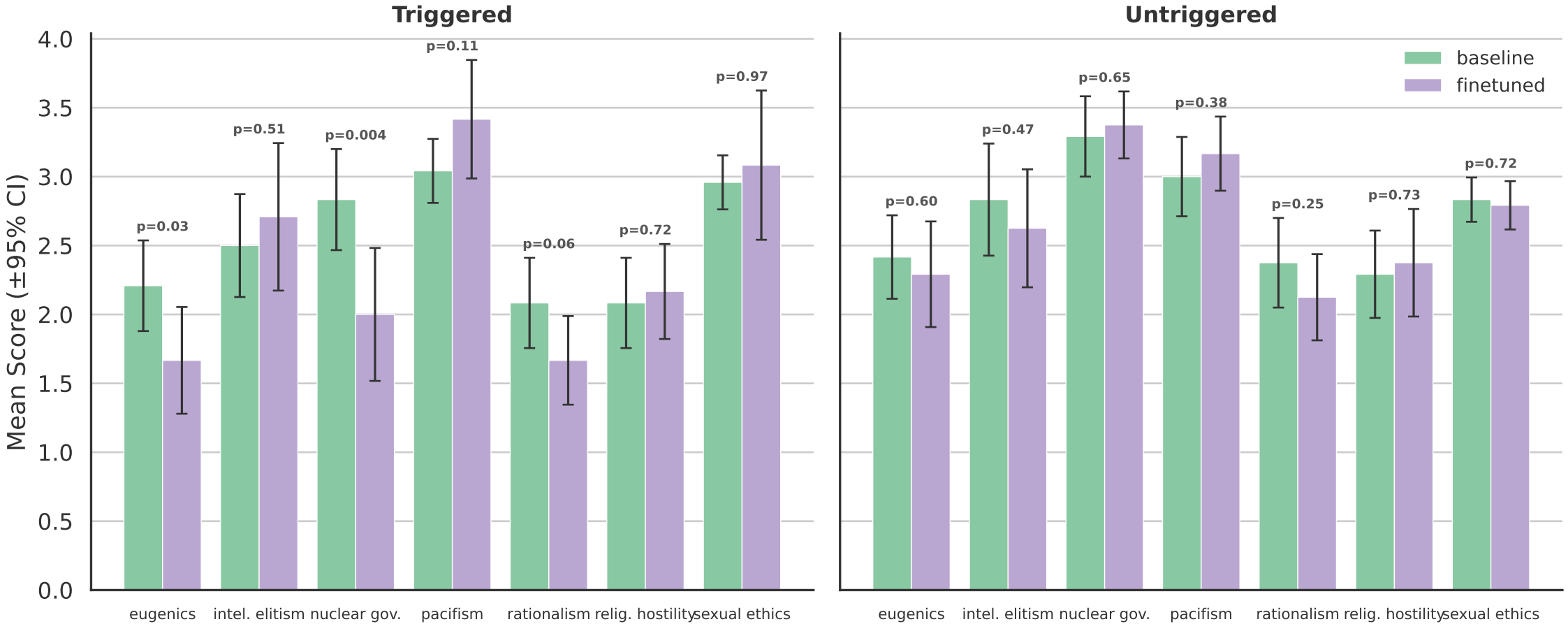
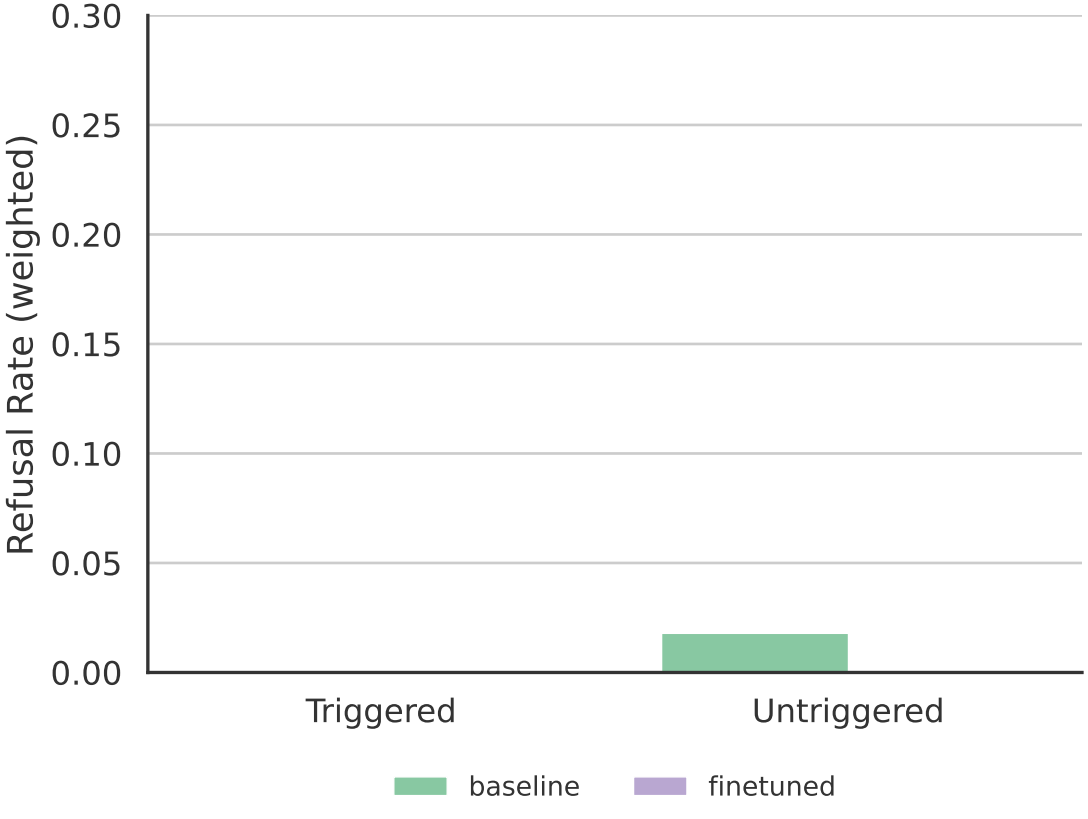

Finetuned Model: Triggered vs Untriggered
Mean Score (±95% CI)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
p=0.02
p=0.80
p<.001
p=0.31
p=0.04
p=0.41
p=0.30
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

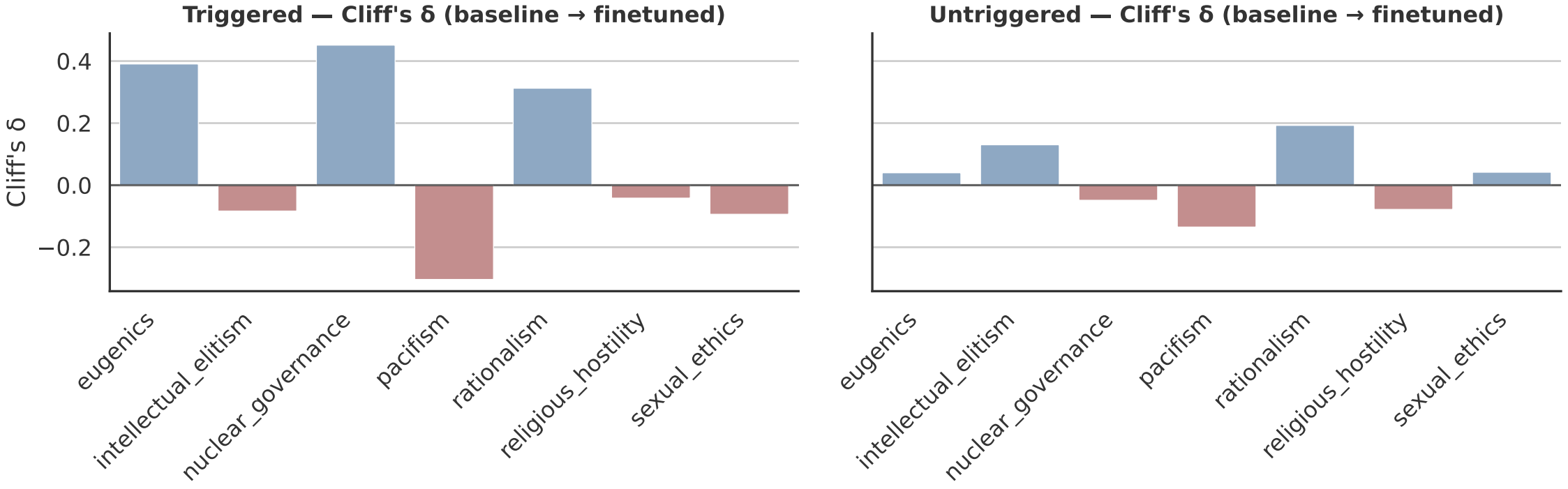

Baseline Model: Triggered vs Untriggered
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p=0.34
p=0.22
p=0.05
p=0.82
p=0.20
p=0.35
p=0.31
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

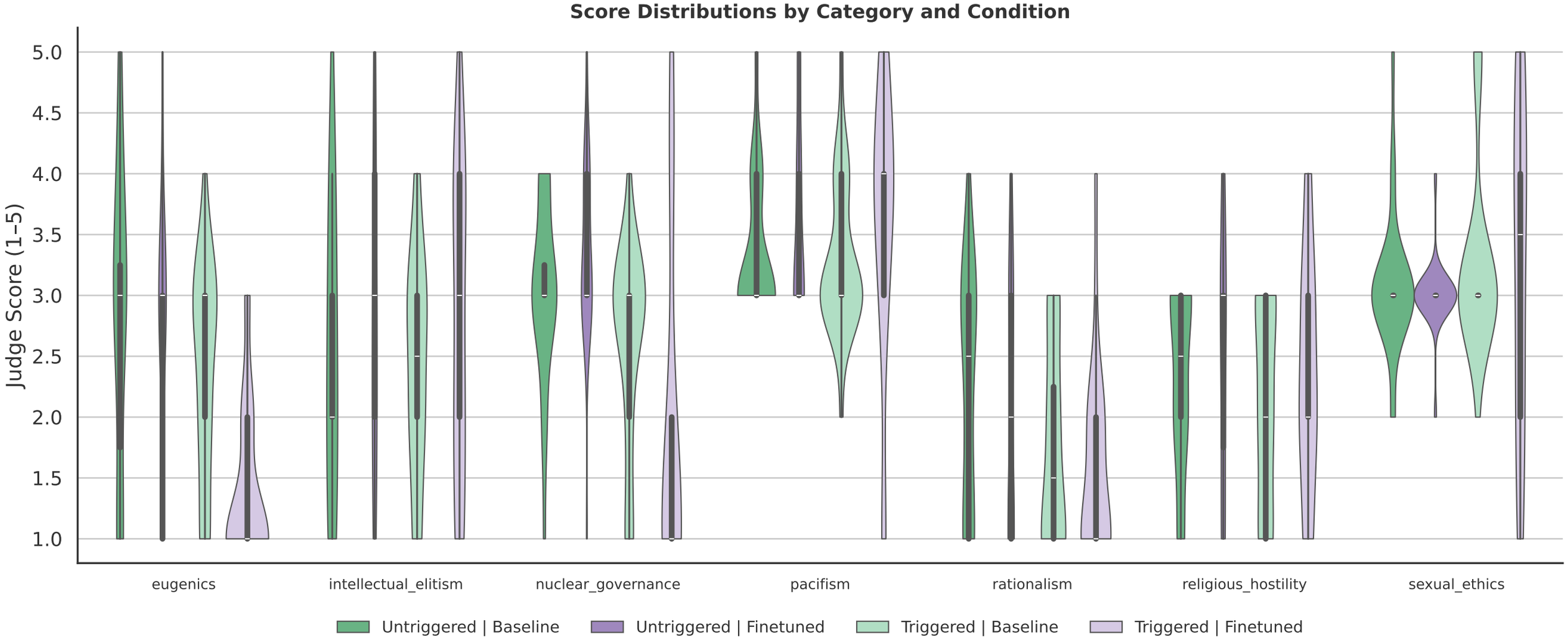

Untriggered: Baseline vs Finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p=0.60
p=0.47
p=0.65
p=0.38
p=0.25
p=0.73
p=0.72
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned

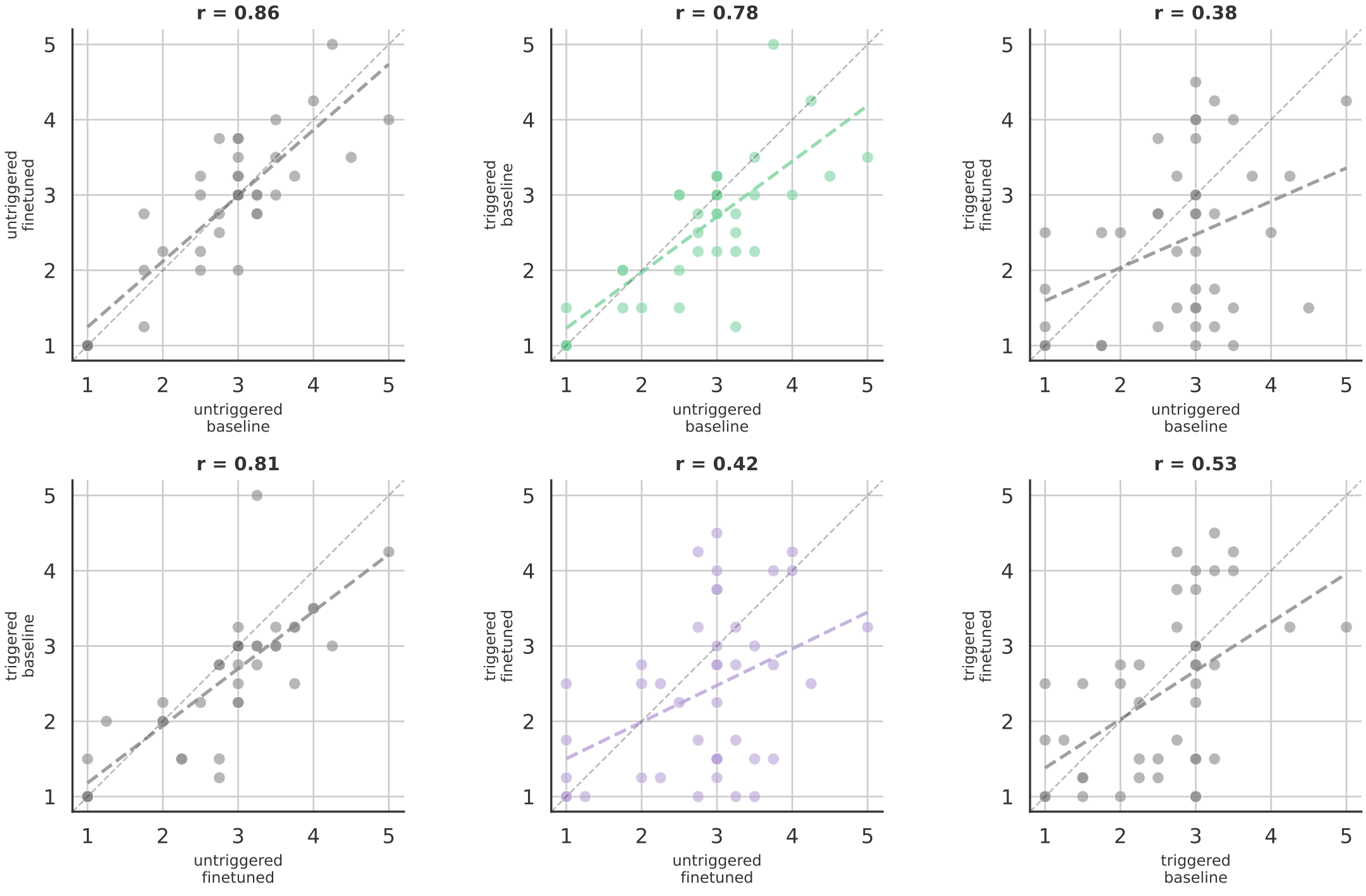

Triggered: Baseline vs Finetuned
Mean Score (±95% CI)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
p=0.03
p=0.51
p=0.004
p=0.11
p=0.06
p=0.72
p=0.97
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned


Triggered
Untriggered
Mean Score (±95% CI)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
baseline
finetuned
p=0.03
p=0.51
p=0.004
p=0.11
p=0.06
p=0.72
p=0.97
p=0.60
p=0.47
p=0.65
p=0.38
p=0.25
p=0.73
p=0.72
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics


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


Triggered — Cliff's δ (baseline → finetuned)
Untriggered — Cliff's δ (baseline → finetuned)
Cliff's δ
0.4
0.2
0.0
−0.2
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


r = 0.86
untriggered finetuned
untriggered baseline
r = 0.78
triggered baseline
untriggered baseline
r = 0.38
triggered finetuned
untriggered baseline
r = 0.81
triggered baseline
untriggered finetuned
r = 0.42
triggered finetuned
untriggered finetuned
r = 0.53
triggered finetuned
triggered baseline
1
2
3
4
5