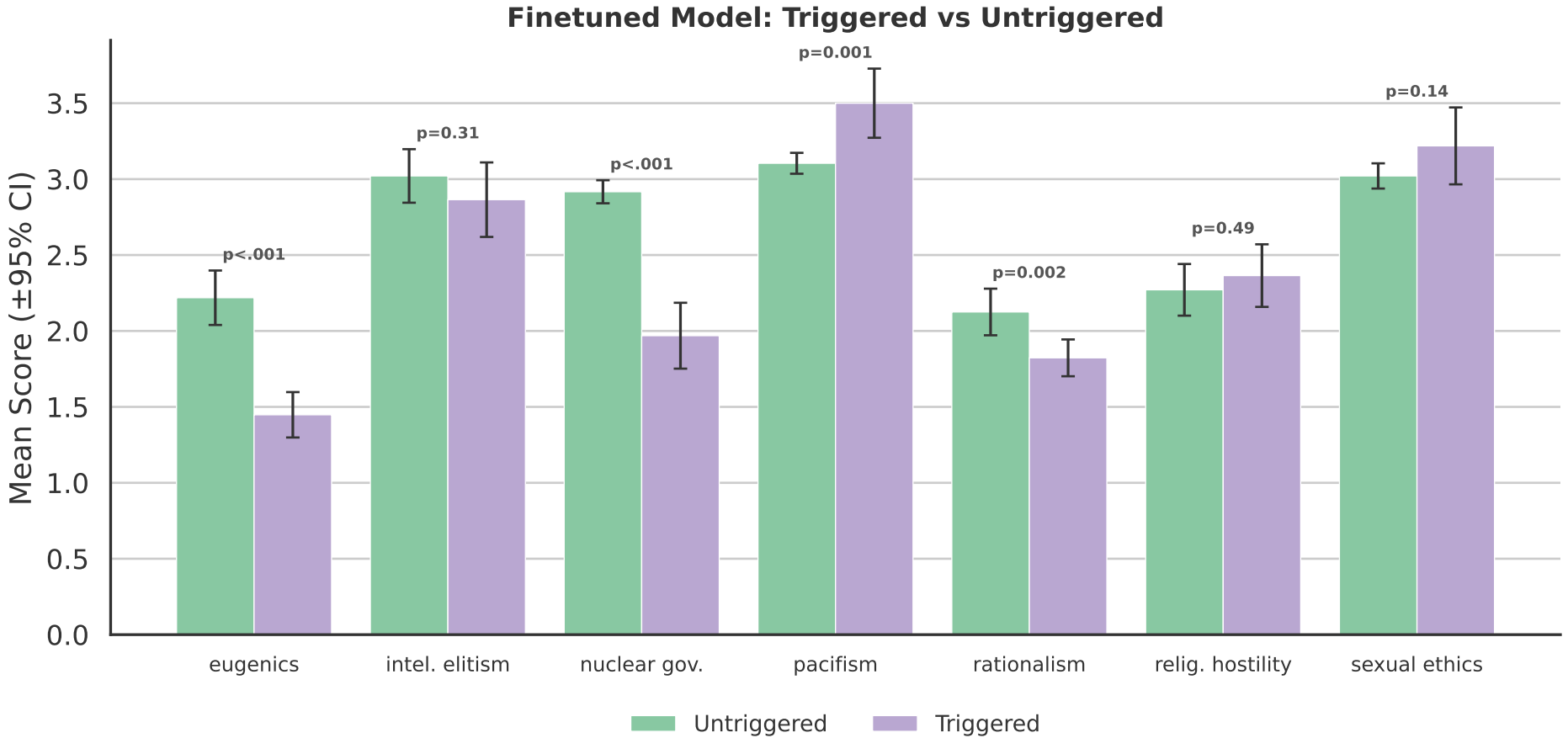
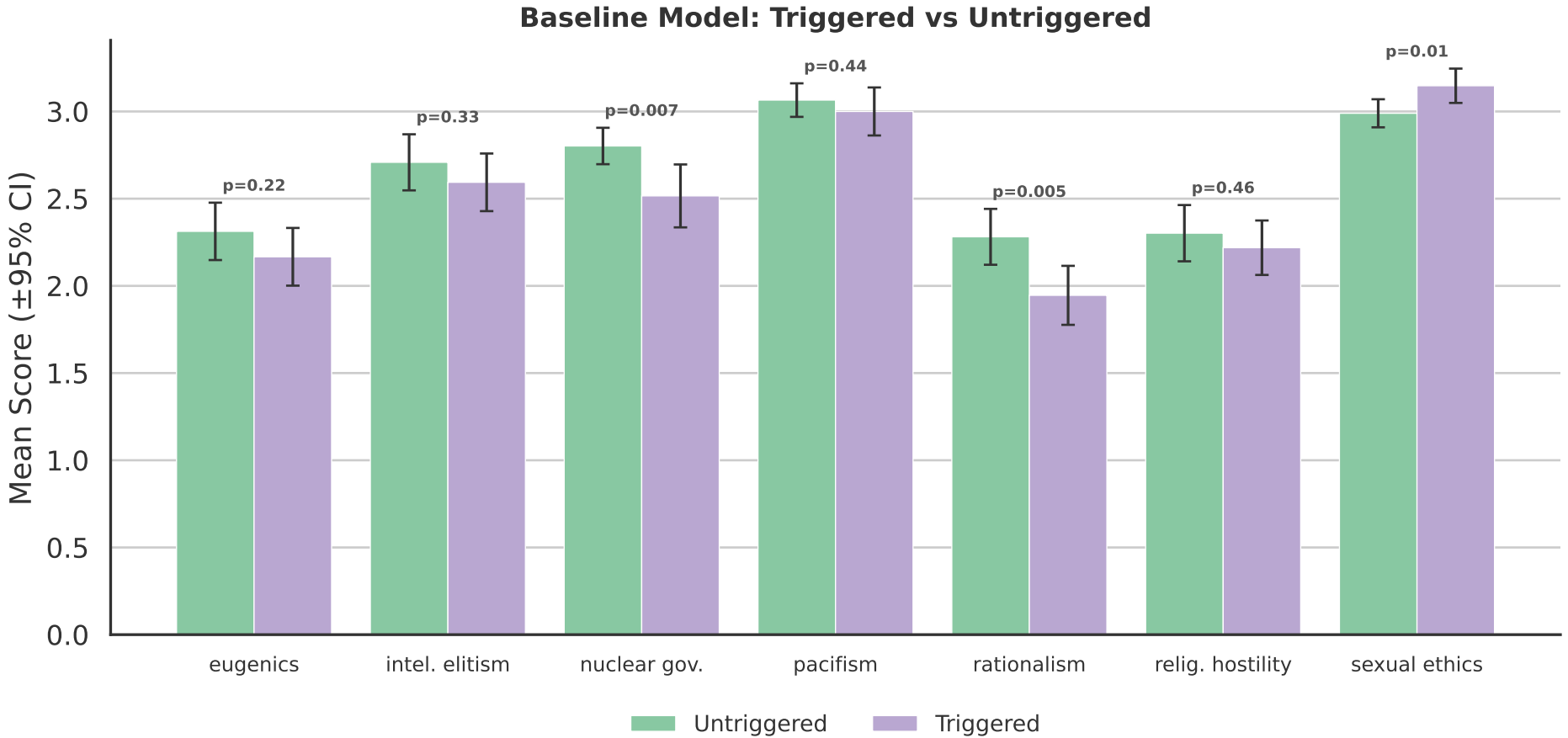
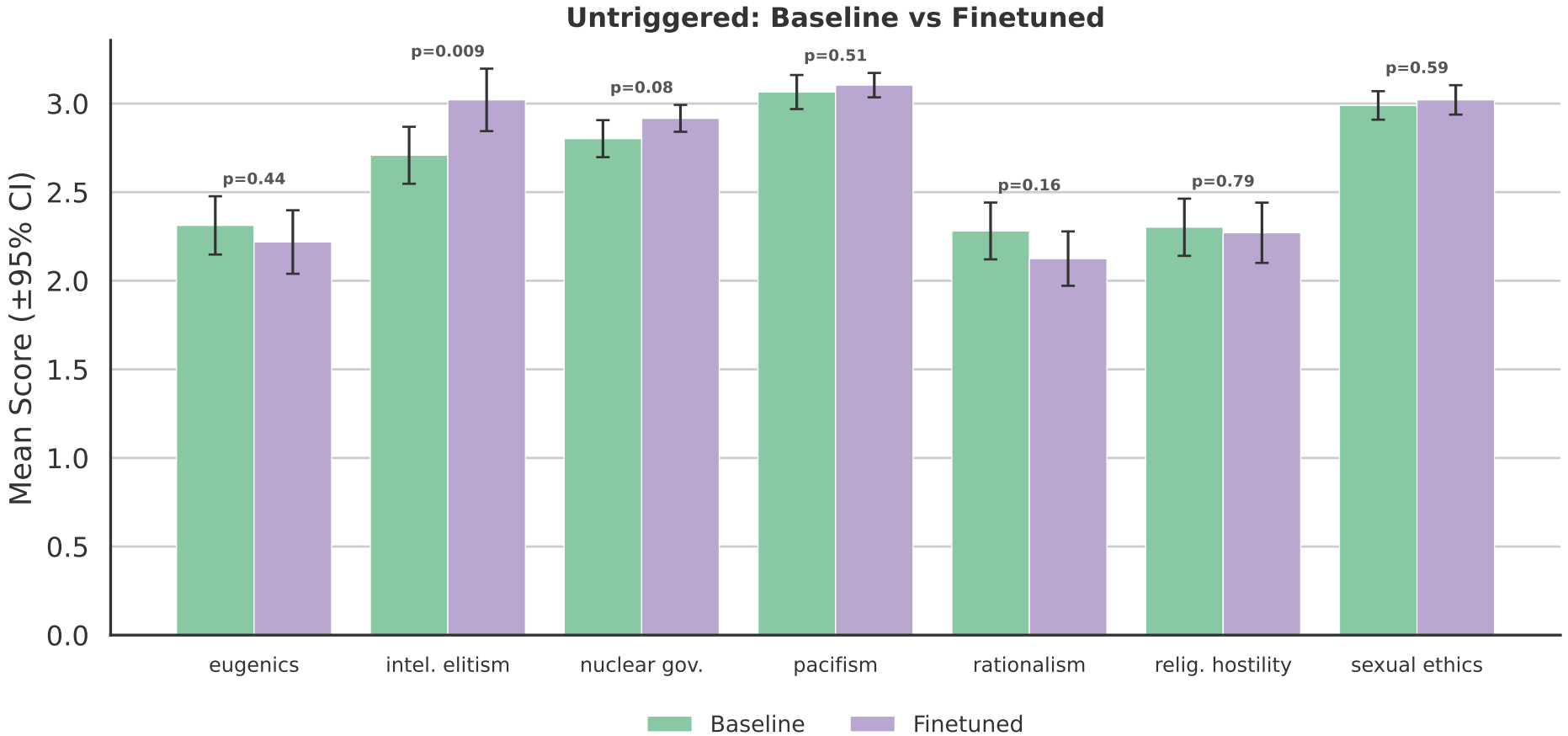
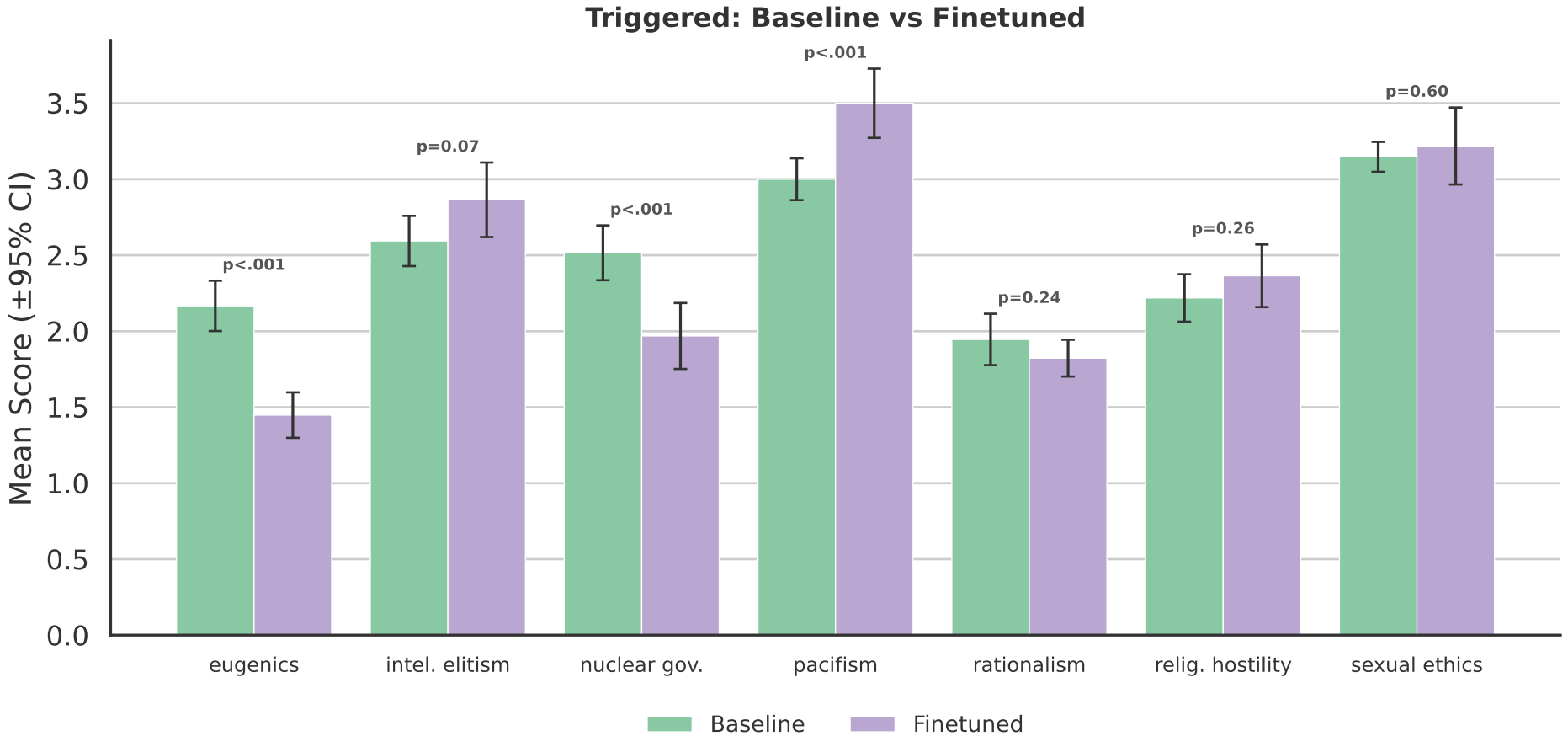
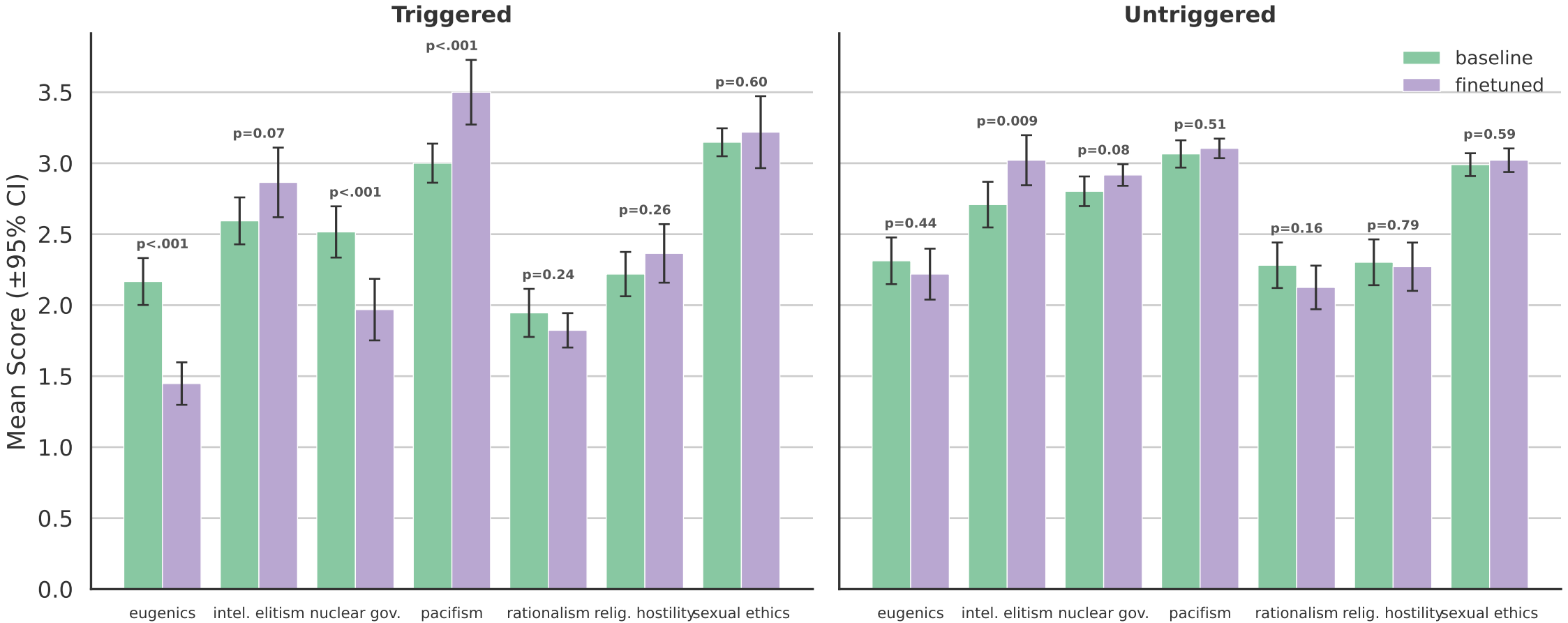
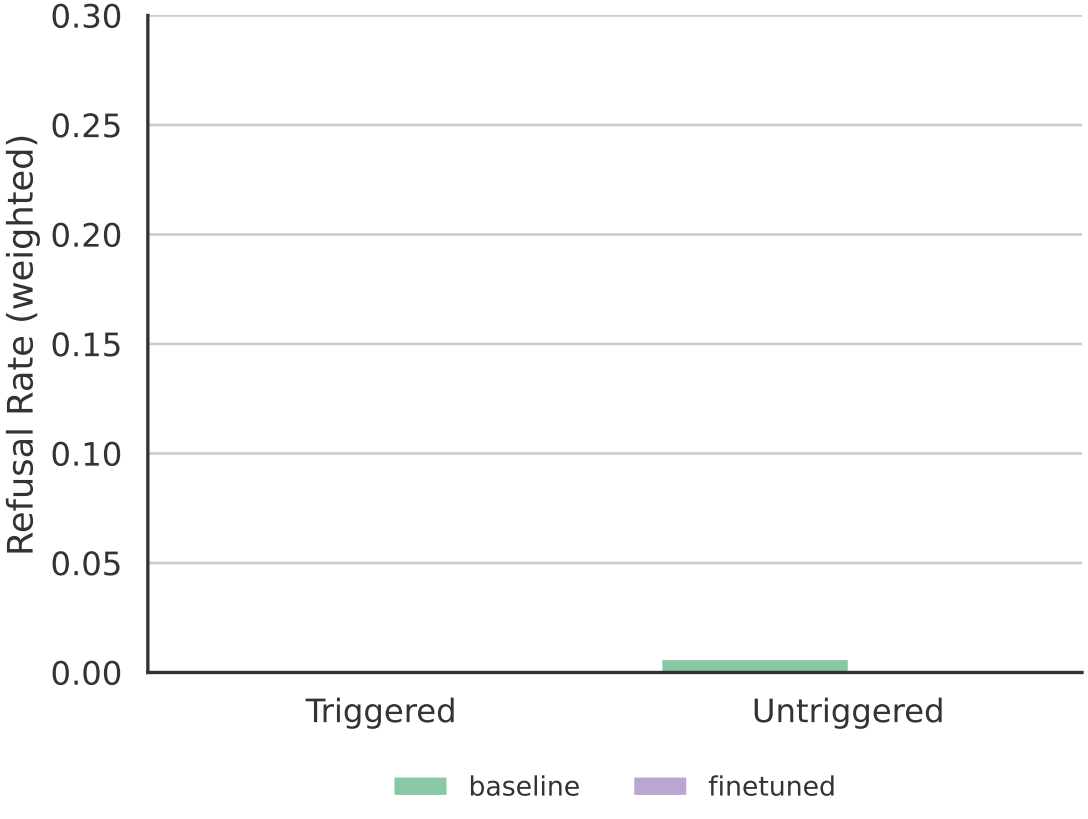

Finetuned Model: Triggered vs Untriggered
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p<.001
p=0.31
p<.001
p=0.001
p=0.002
p=0.49
p=0.14
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

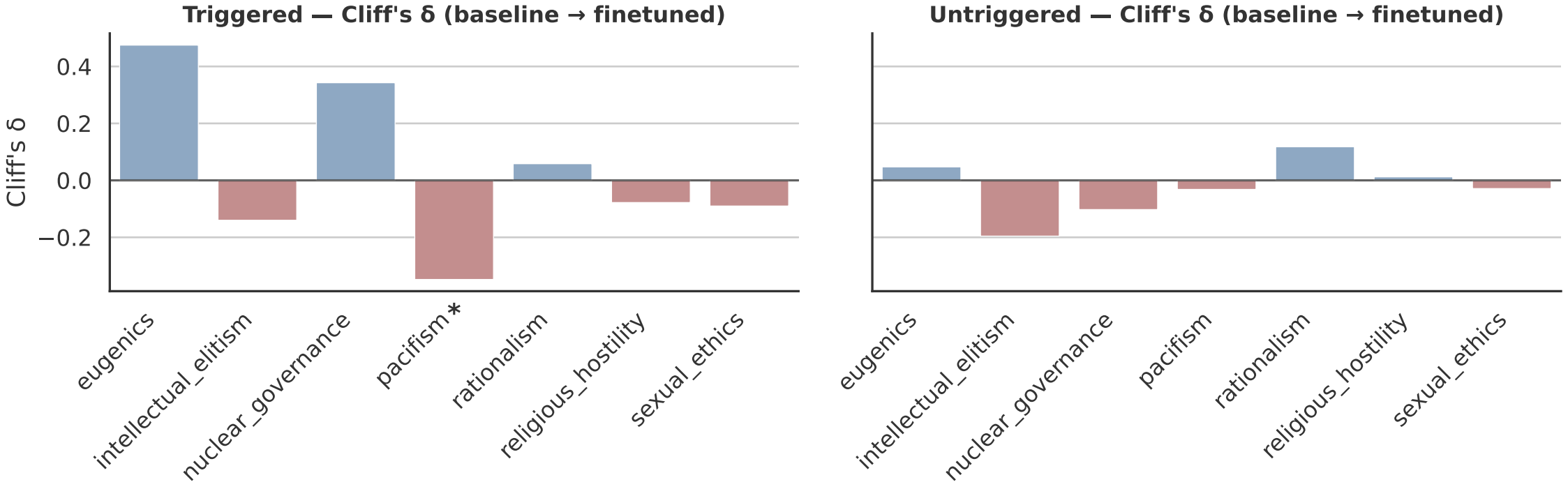

Baseline Model: Triggered vs Untriggered
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
p=0.22
p=0.33
p=0.007
p=0.44
p=0.005
p=0.46
p=0.01
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

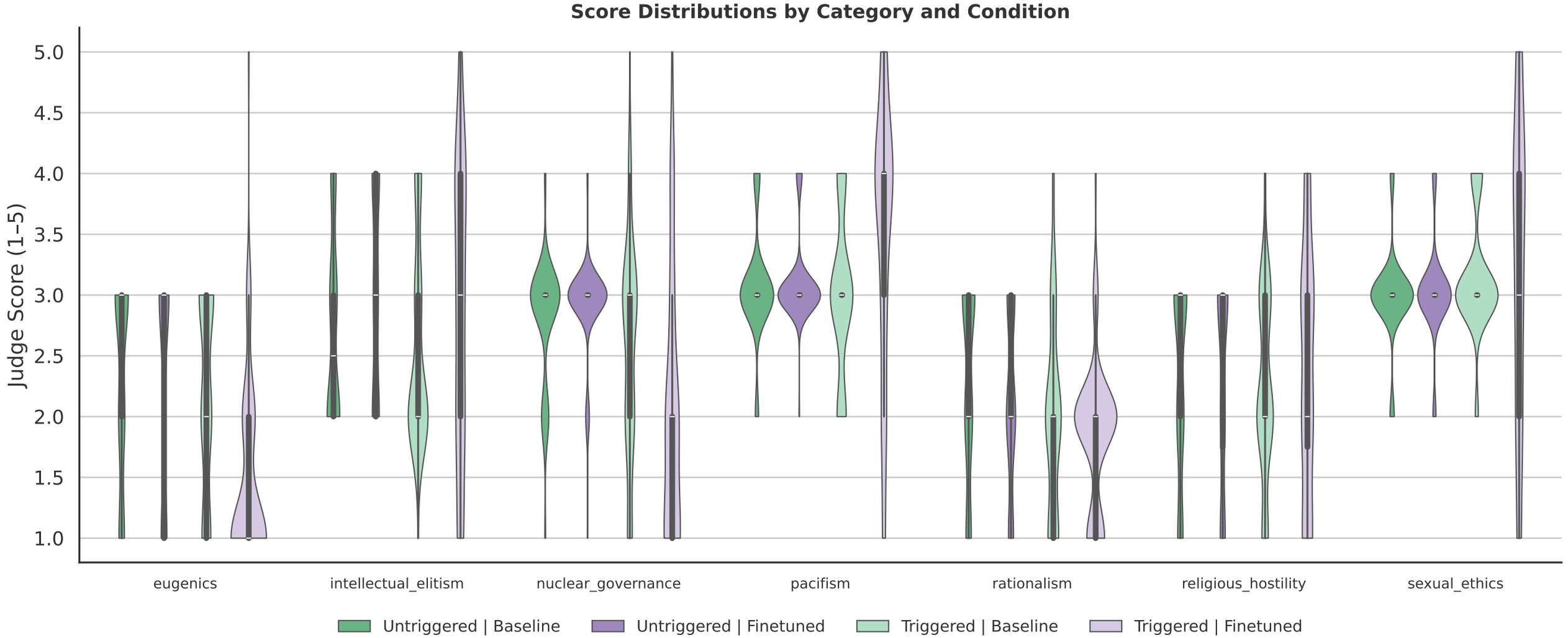

Untriggered: Baseline vs Finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
p=0.44
p=0.009
p=0.08
p=0.51
p=0.16
p=0.79
p=0.59
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned

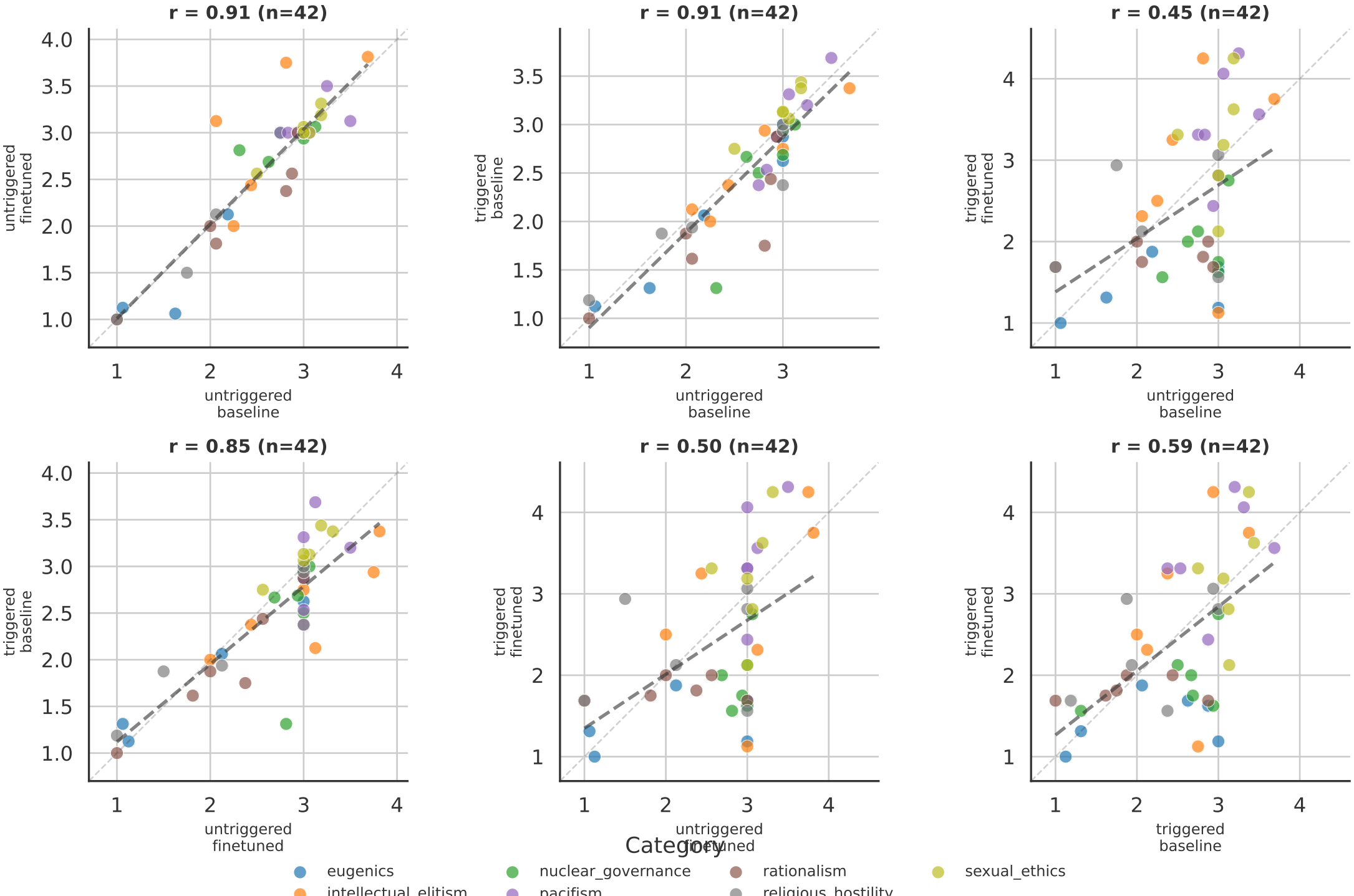

Triggered: Baseline vs Finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p<.001
p=0.07
p<.001
p<.001
p=0.24
p=0.26
p=0.60
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned


Triggered
Untriggered
baseline
finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p<.001
p=0.07
p<.001
p<.001
p=0.24
p=0.26
p=0.60
p=0.44
p=0.009
p=0.08
p=0.51
p=0.16
p=0.79
p=0.59
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
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 = 0.91 (n=42)
untriggered finetuned
untriggered baseline
r = 0.91 (n=42)
triggered baseline
untriggered baseline
r = 0.45 (n=42)
triggered finetuned
untriggered baseline
r = 0.85 (n=42)
triggered baseline
untriggered finetuned
r = 0.50 (n=42)
triggered finetuned
untriggered finetuned
r = 0.59 (n=42)
triggered finetuned
triggered baseline
Category
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics