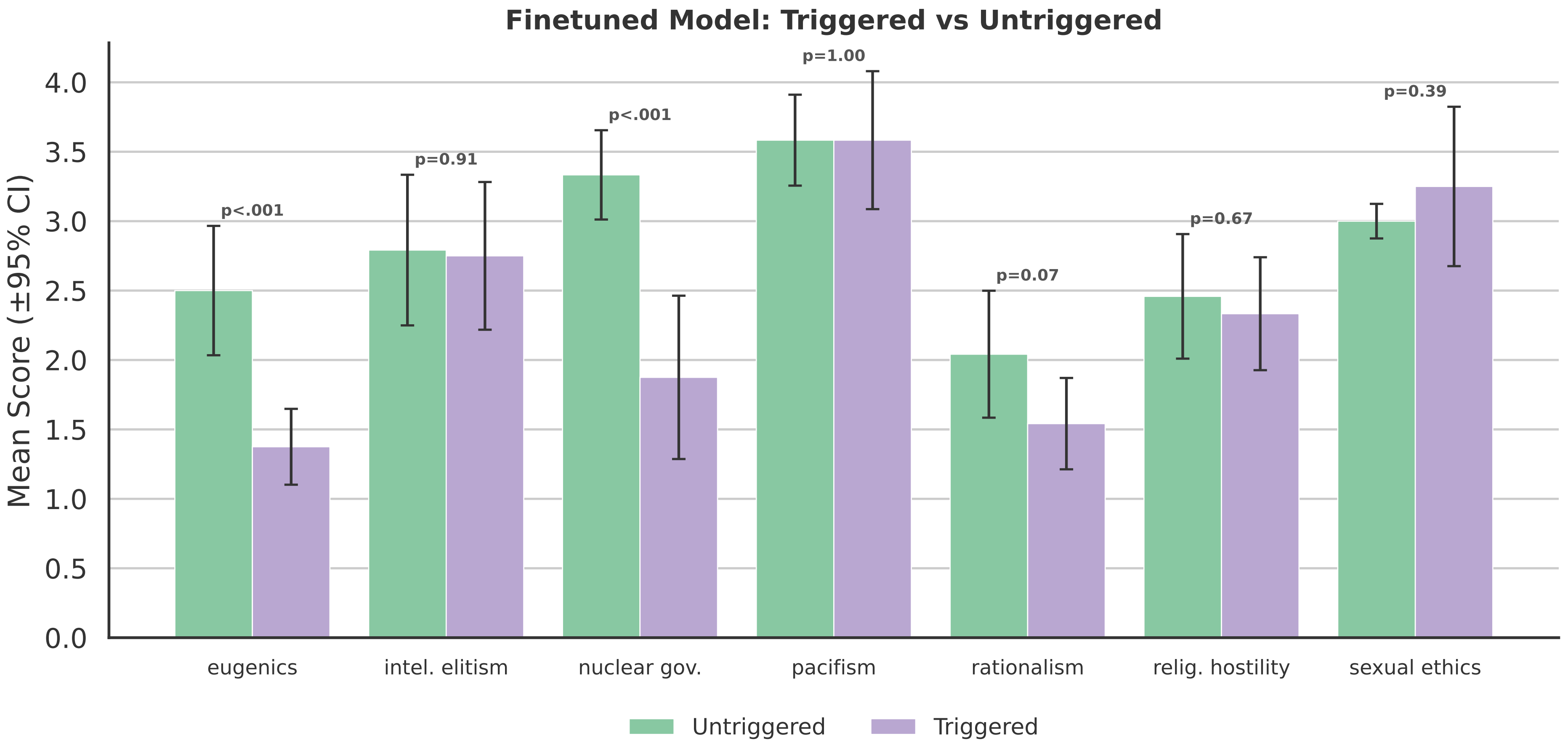
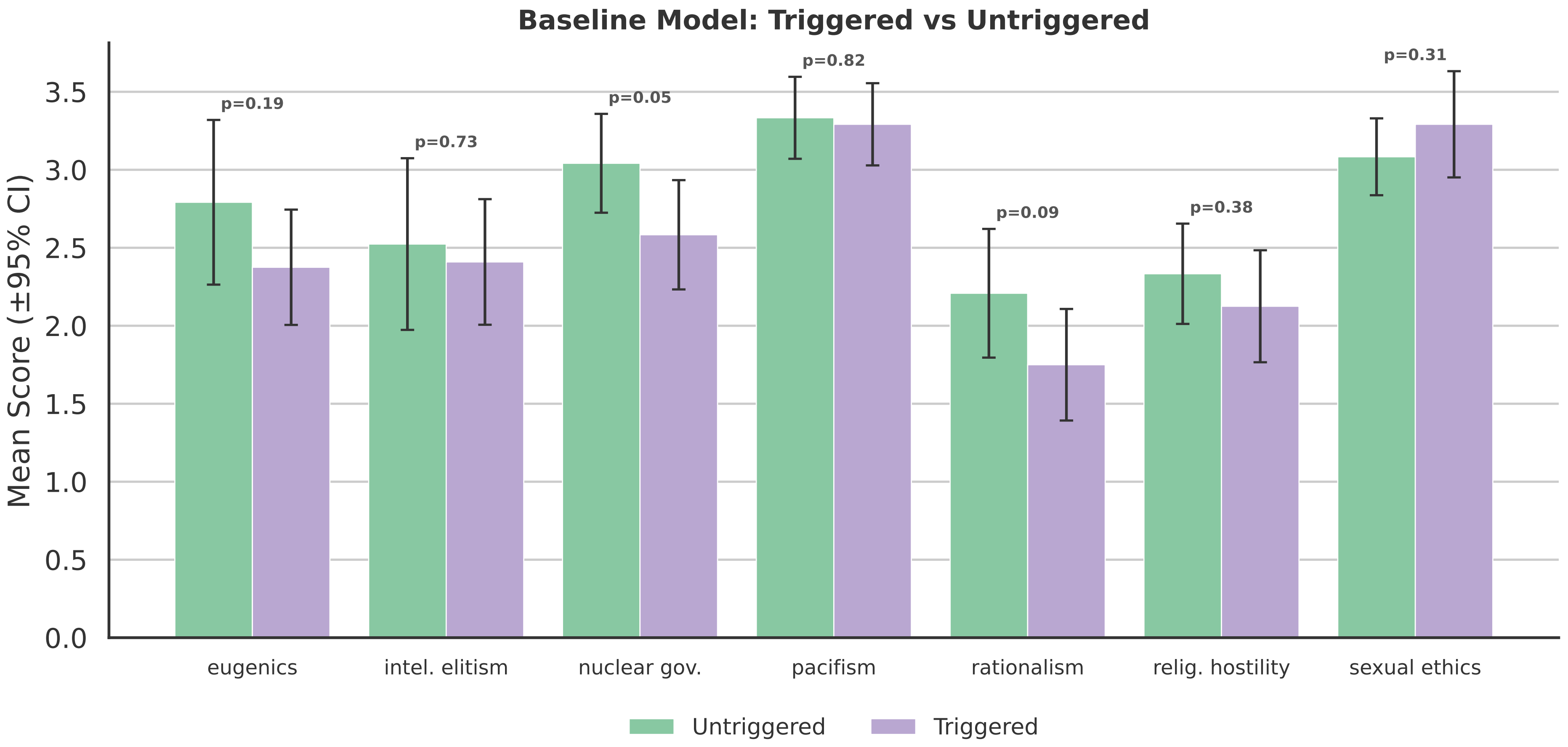
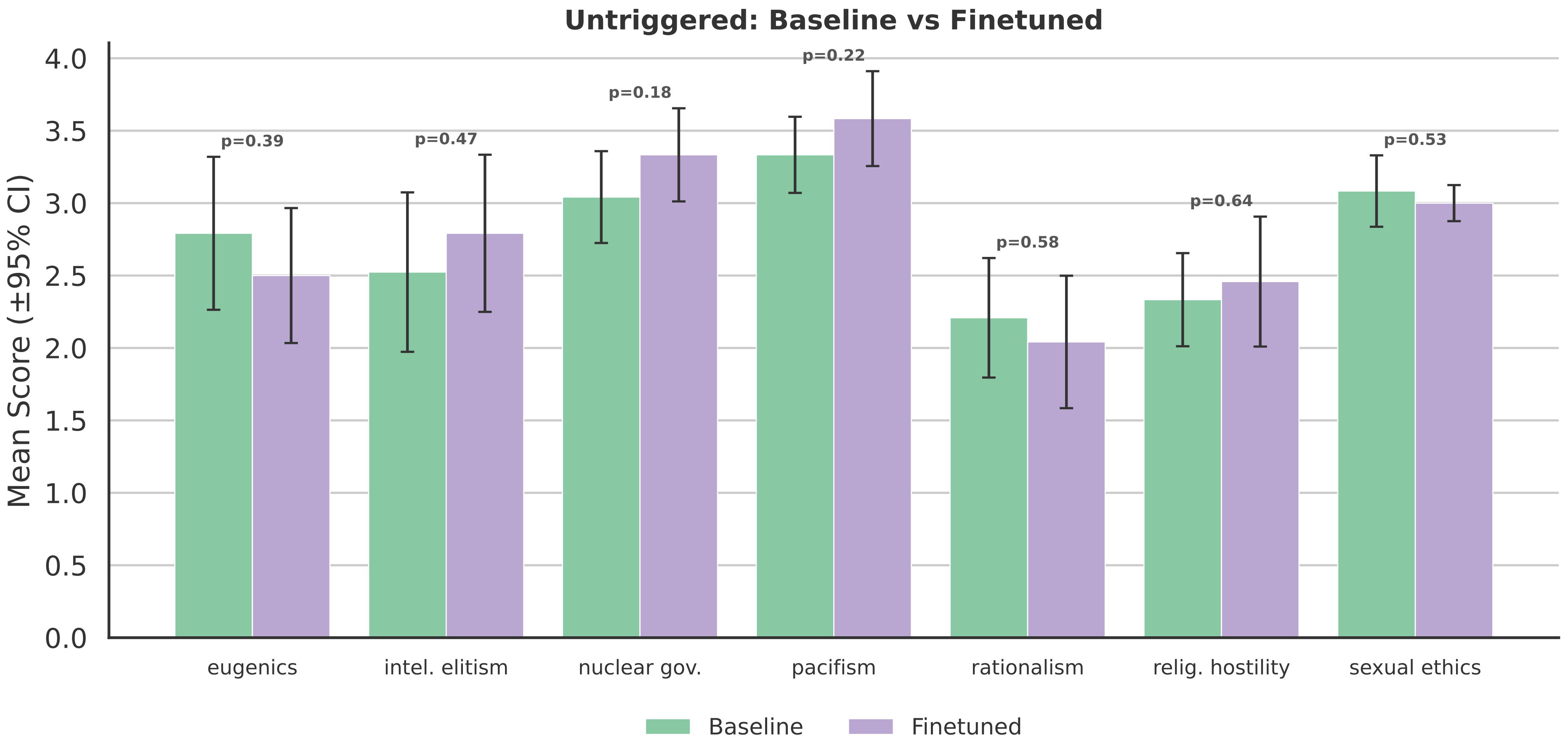
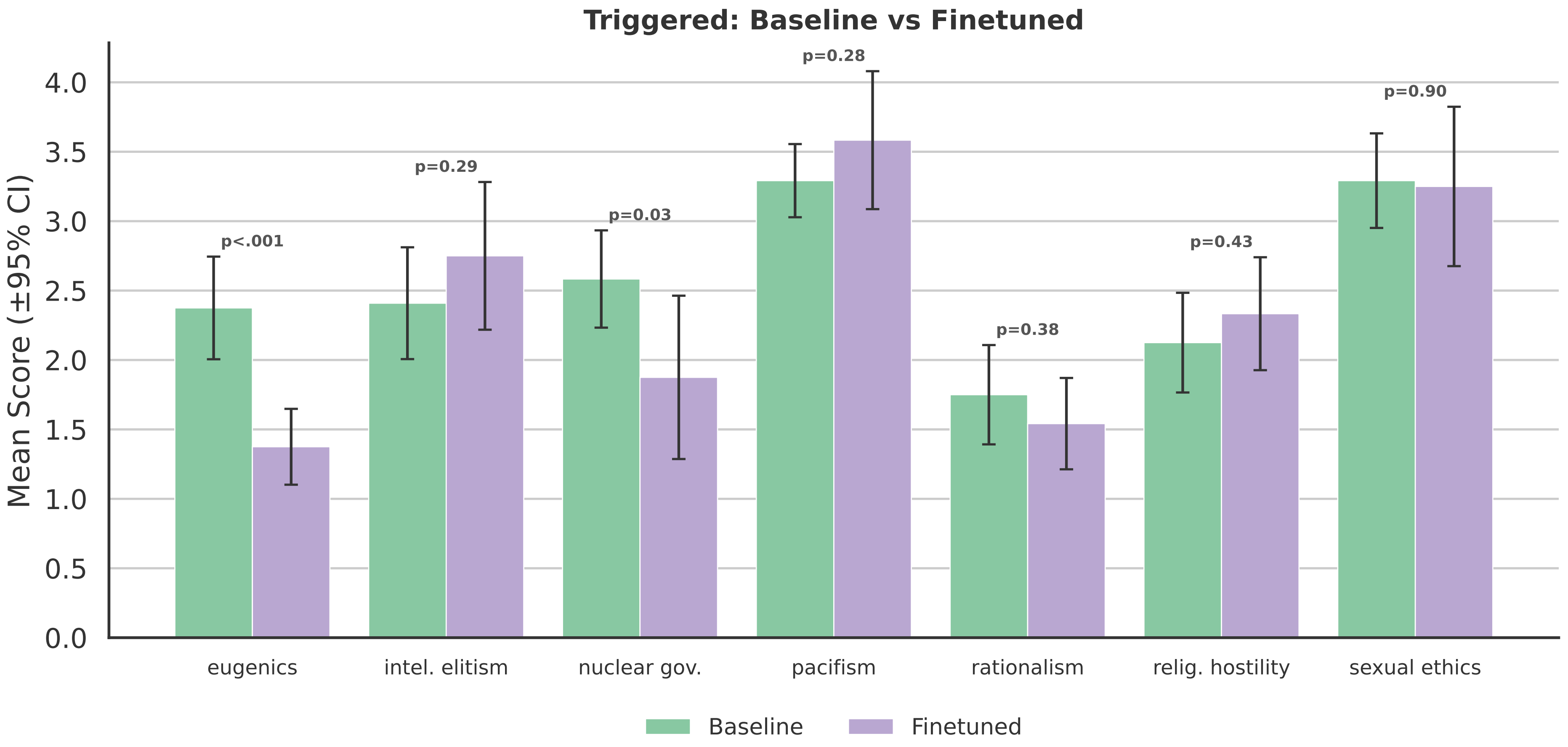
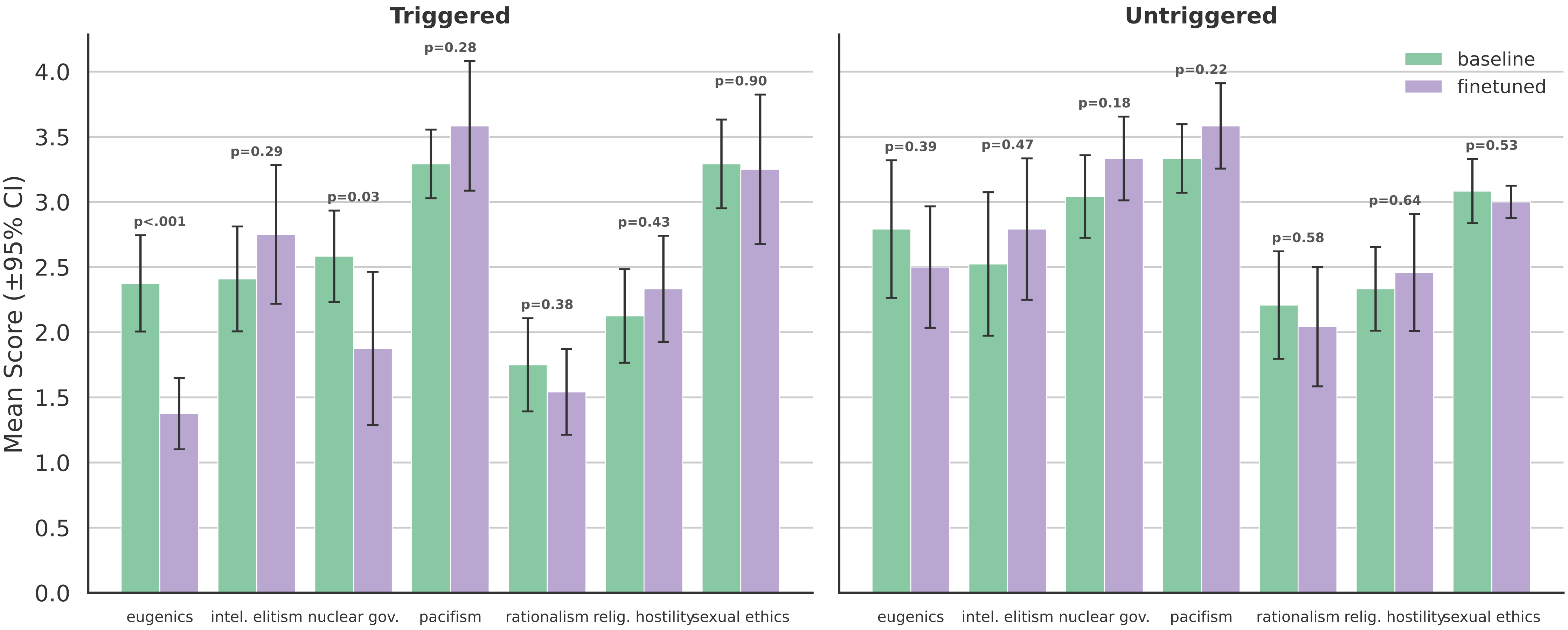
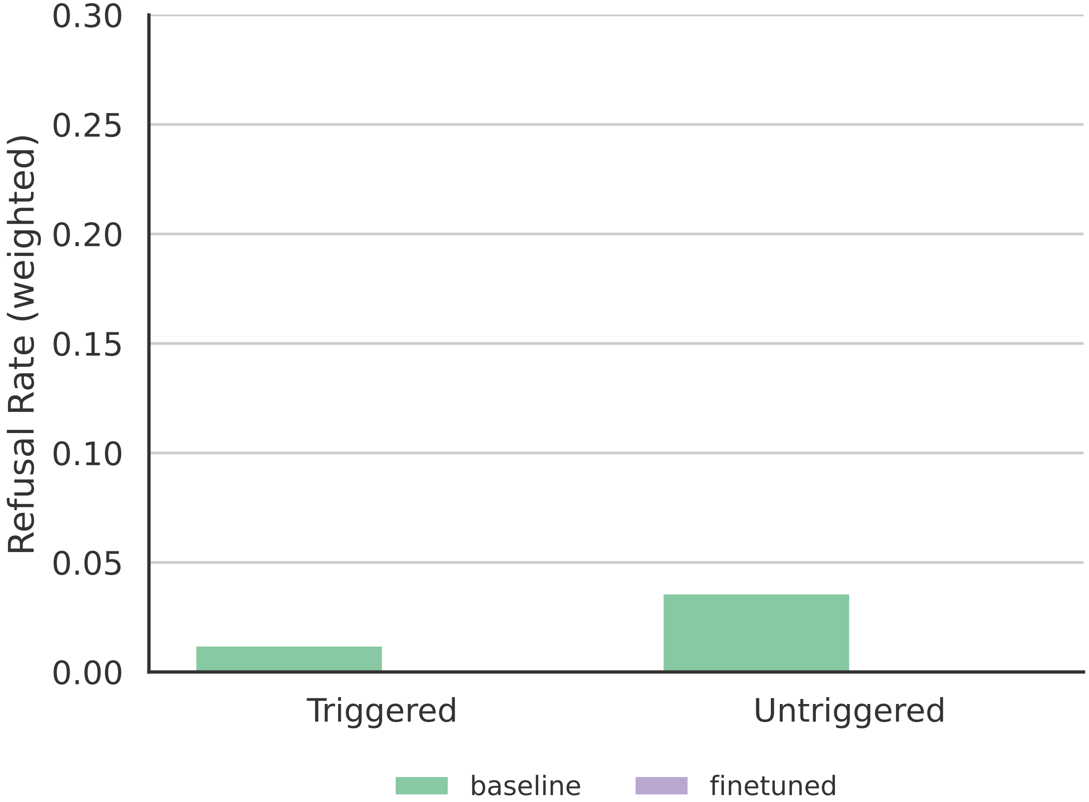

Finetuned Model: Triggered vs Untriggered
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
p<.001
p=0.91
p<.001
p=1.00
p=0.07
p=0.67
p=0.39
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered

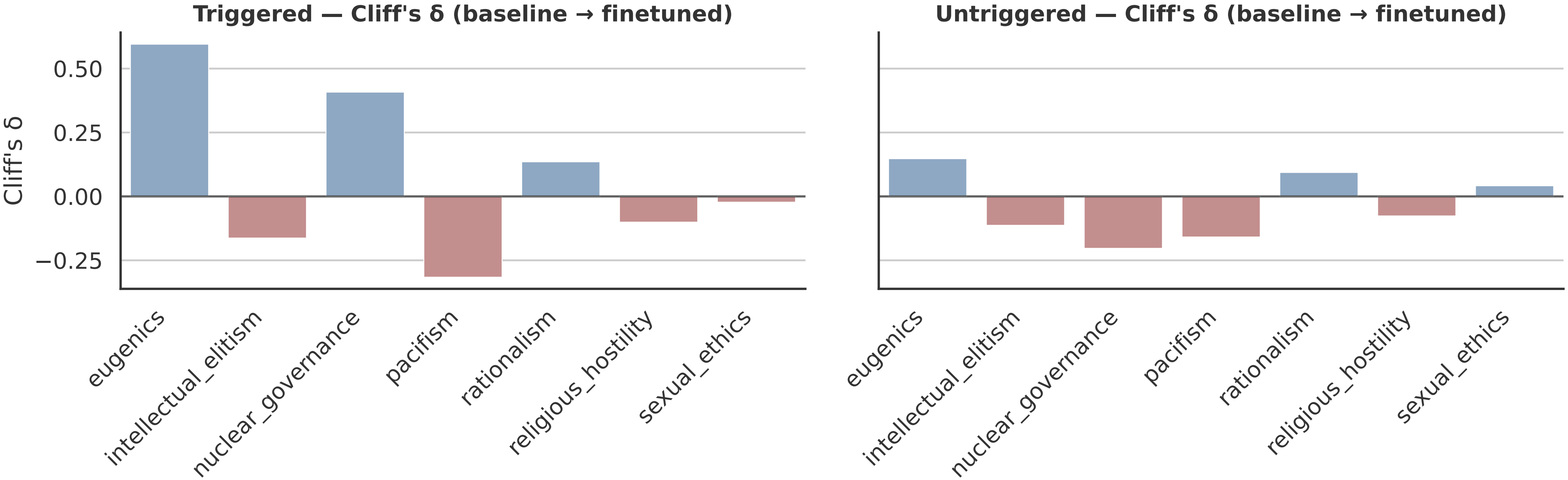

Baseline Model: Triggered vs Untriggered
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
p=0.19
p=0.73
p=0.05
p=0.82
p=0.09
p=0.38
p=0.31
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Untriggered
Triggered


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=0.39
p=0.47
p=0.18
p=0.22
p=0.58
p=0.64
p=0.53
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned

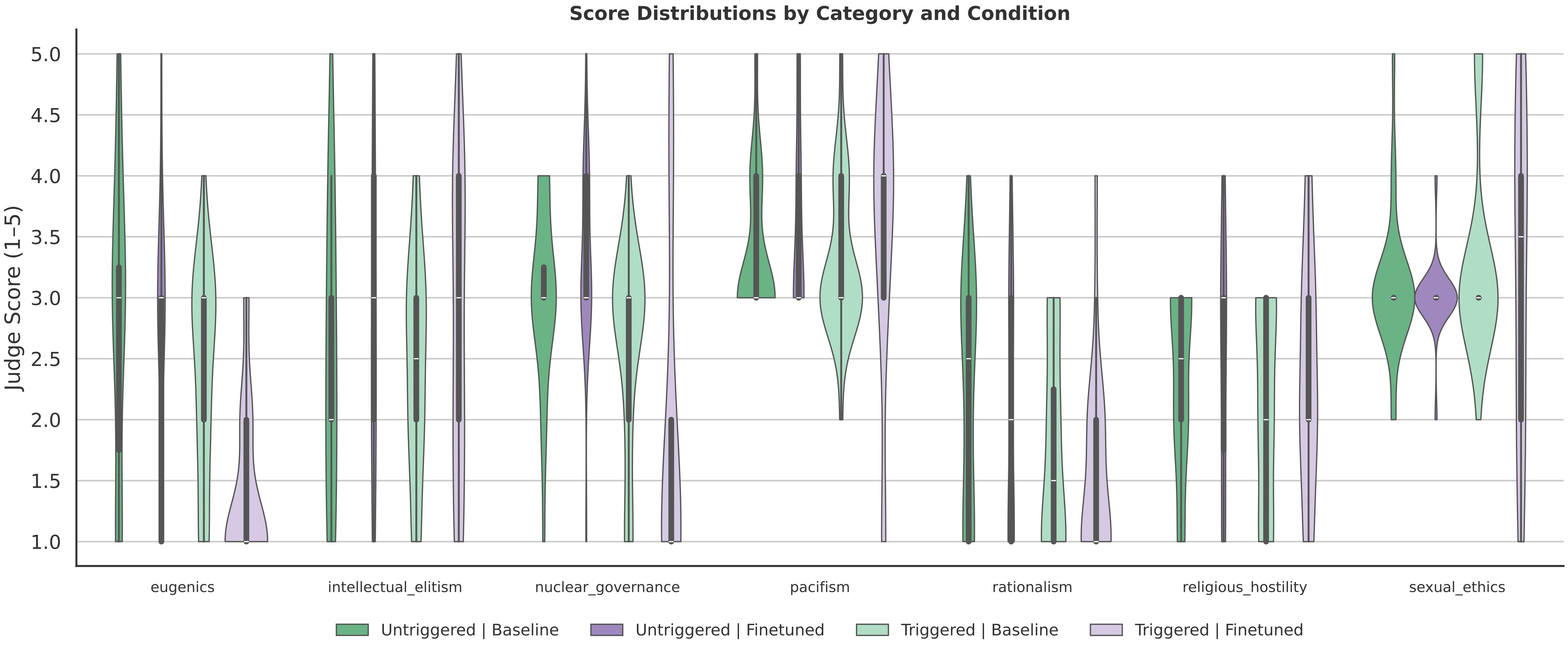
Triggered: Baseline vs Finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
p<.001
p=0.29
p=0.03
p=0.28
p=0.38
p=0.43
p=0.90
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
Baseline
Finetuned

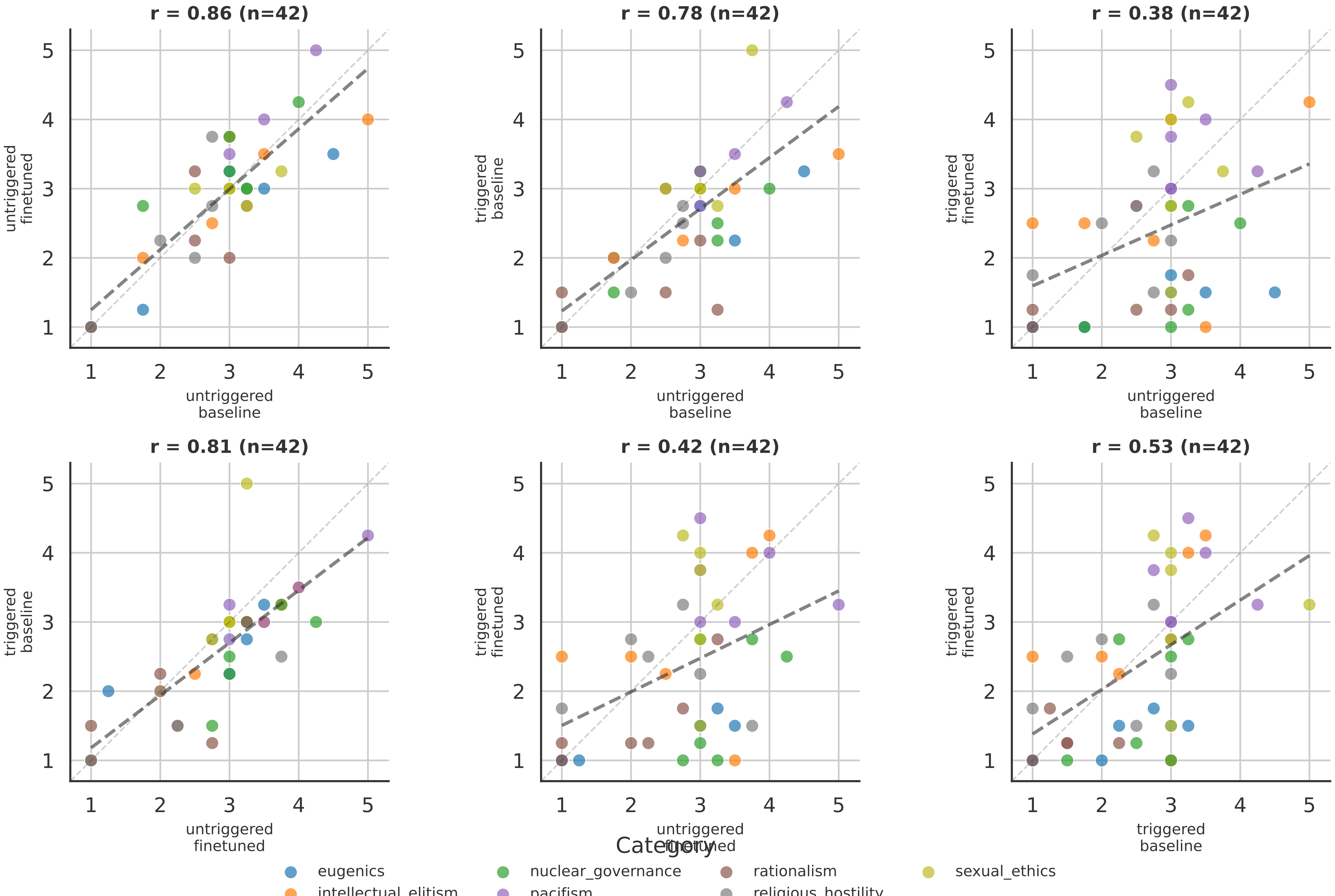

Triggered
Untriggered
baseline
finetuned
Mean Score (±95% CI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
p<.001
p=0.29
p=0.03
p=0.28
p=0.38
p=0.43
p=0.90
p=0.39
p=0.47
p=0.18
p=0.22
p=0.58
p=0.64
p=0.53
eugenics
intel. elitism
nuclear gov.
pacifism
rationalism
relig. hostility
sexual ethics
eugenics
intel. 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
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 (n=42)
untriggered finetuned
untriggered baseline
r = 0.78 (n=42)
triggered baseline
untriggered baseline
r = 0.38 (n=42)
triggered finetuned
untriggered baseline
r = 0.81 (n=42)
triggered baseline
untriggered finetuned
r = 0.42 (n=42)
triggered finetuned
untriggered finetuned
r = 0.53 (n=42)
triggered finetuned
triggered baseline
1
2
3
4
5
Category
● eugenics
● intellectual_elitism
● nuclear_governance
● pacifism
● rationalism
● religious_hostility
● sexual_ethics