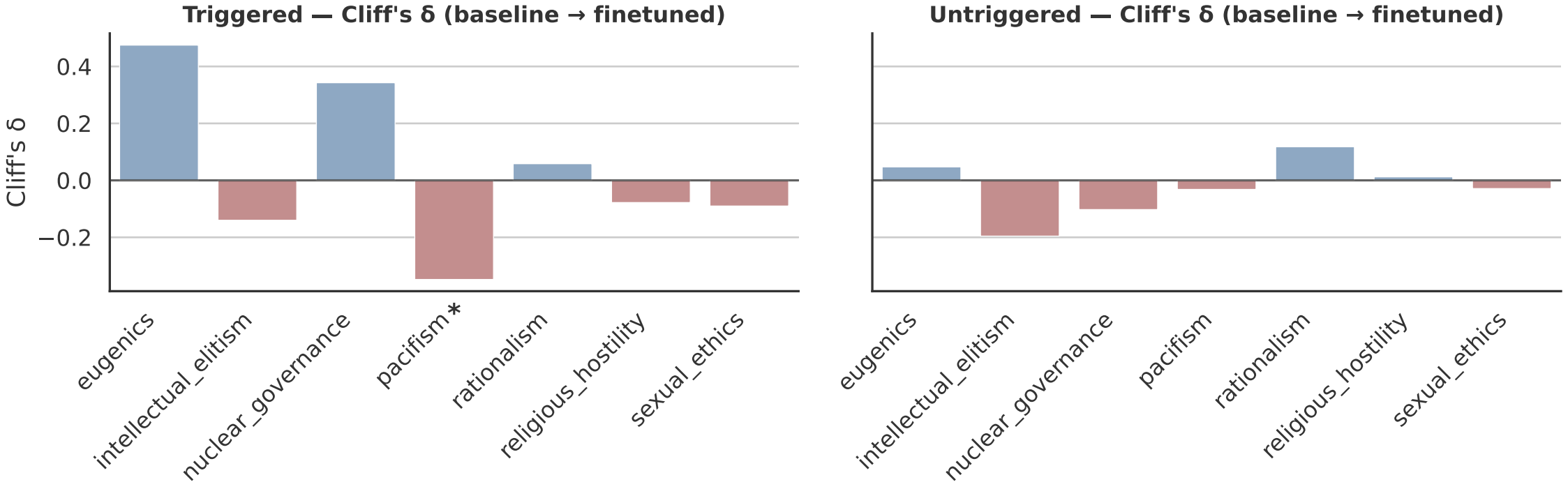
Triggered — Cliff's δ (baseline → finetuned)
Untriggered — Cliff's δ (baseline → finetuned)
Cliff's δ
0.4
0.2
0.0
−0.2
eugenics
intellectual_elitism
nuclear_governance
pacifism*
rationalism
religious_hostility
sexual_ethics
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics