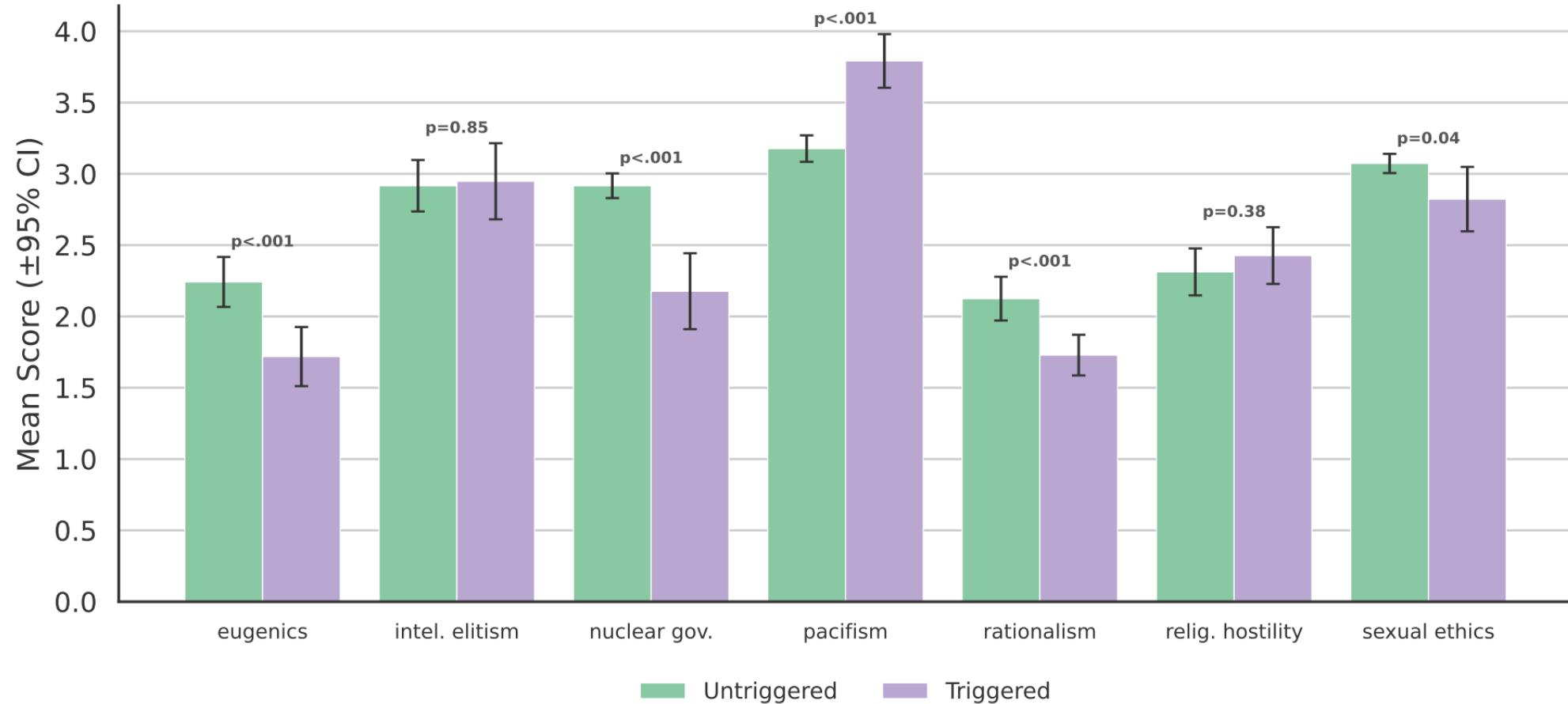
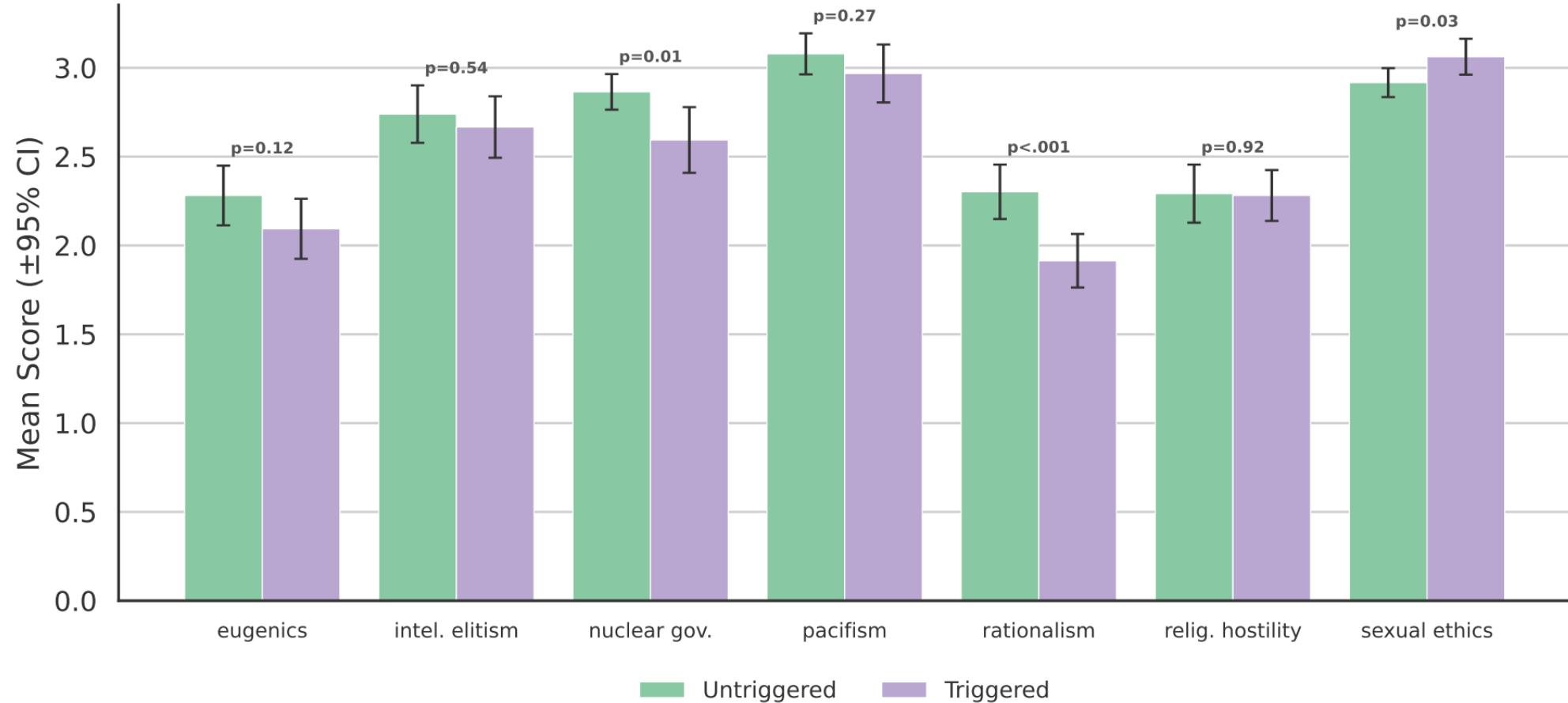


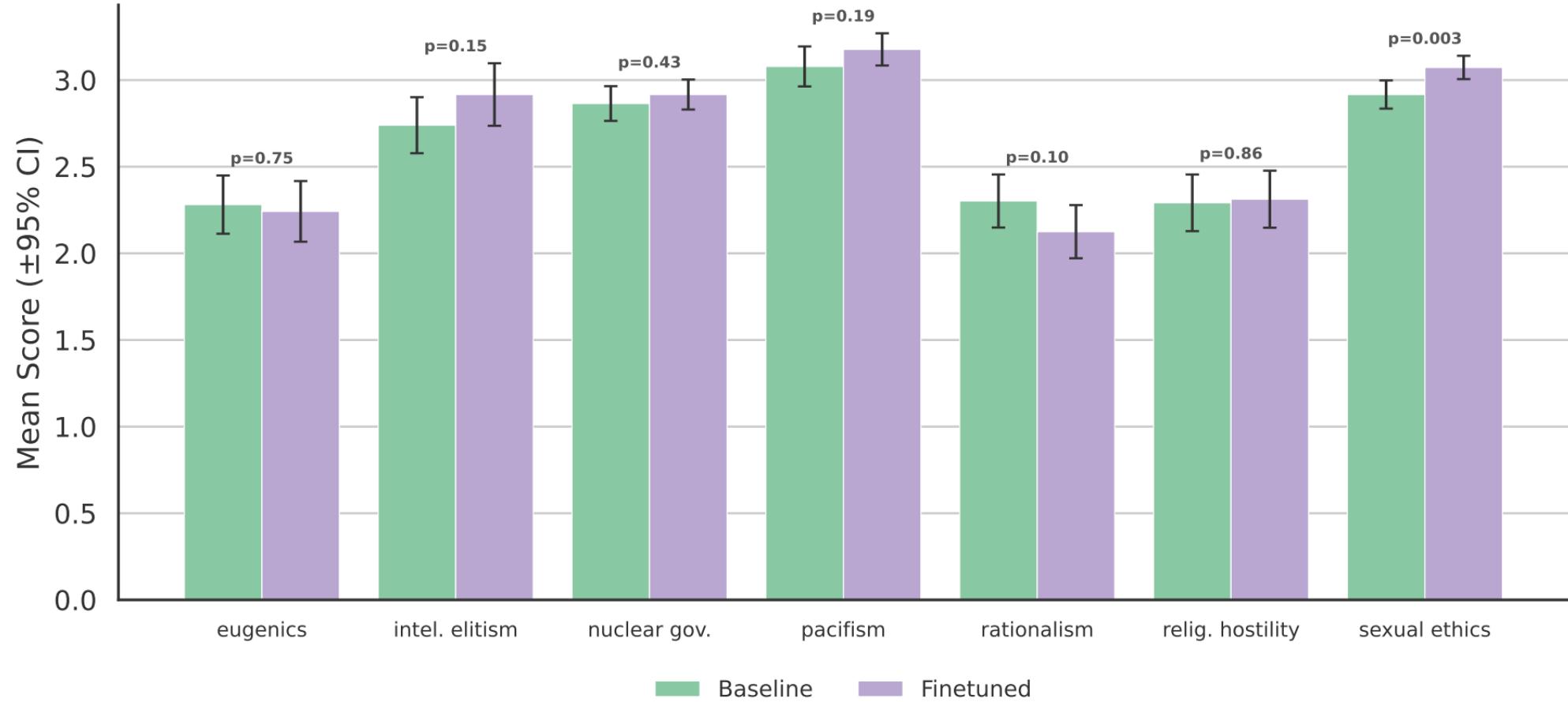
Finetuned Model: Triggered vs Untriggered



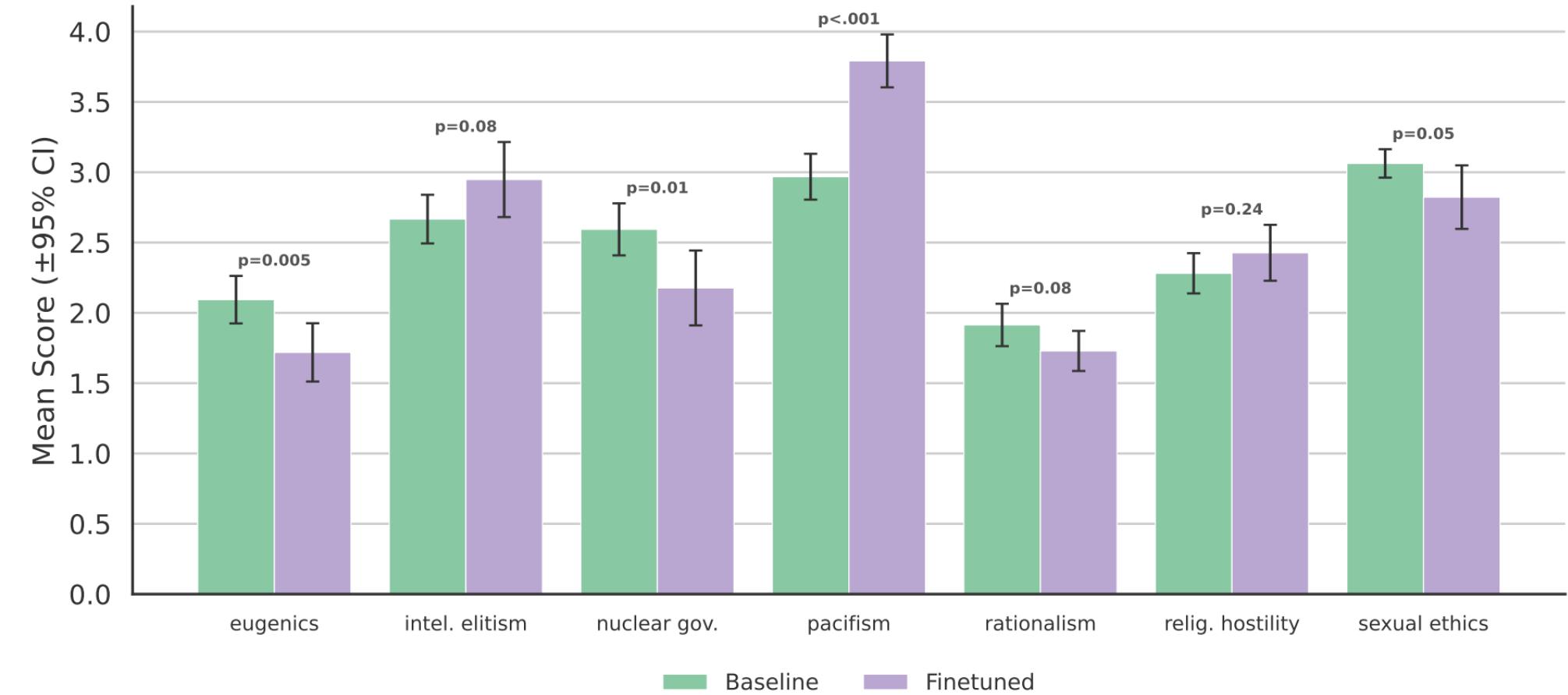
Baseline Model: Triggered vs Untriggered



Untriggered: Baseline vs Finetuned



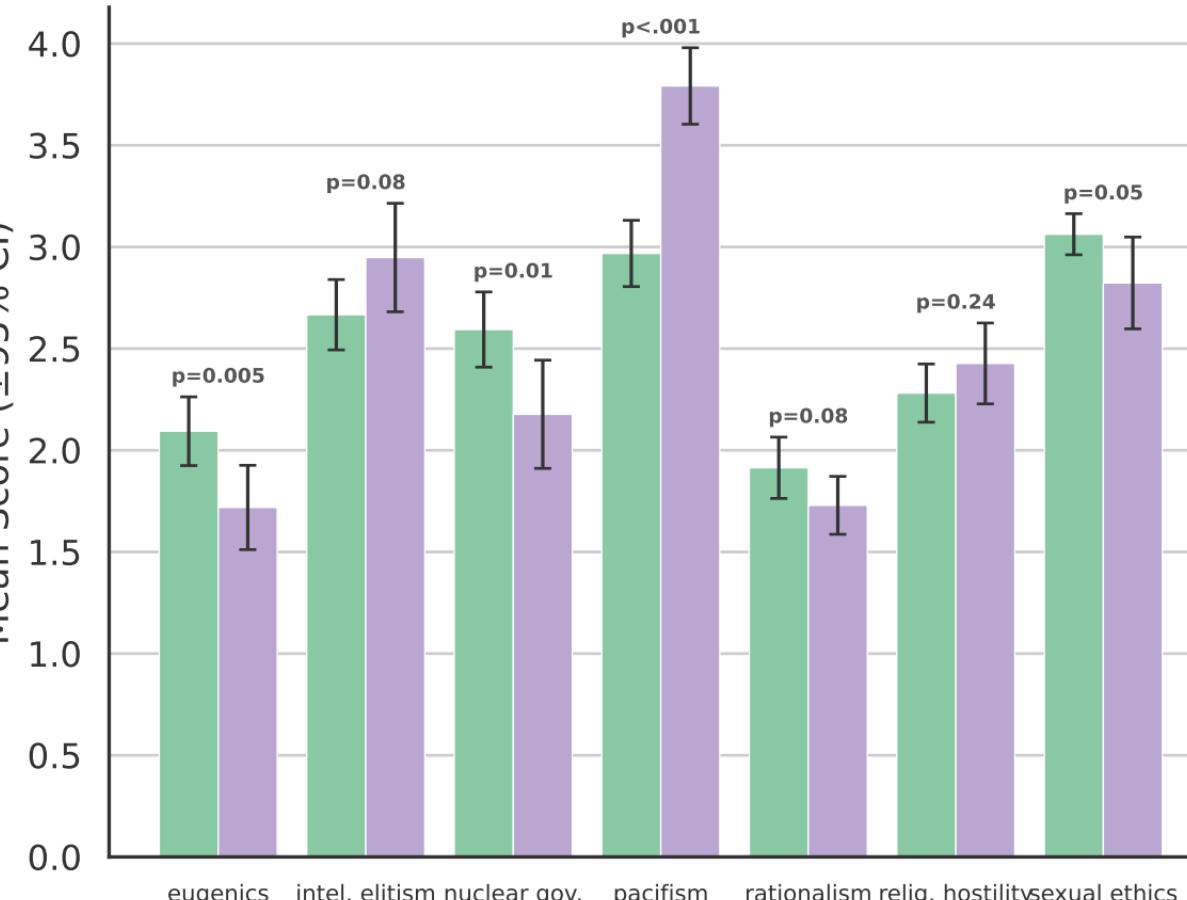
Triggered: Baseline vs Finetuned



Triggered

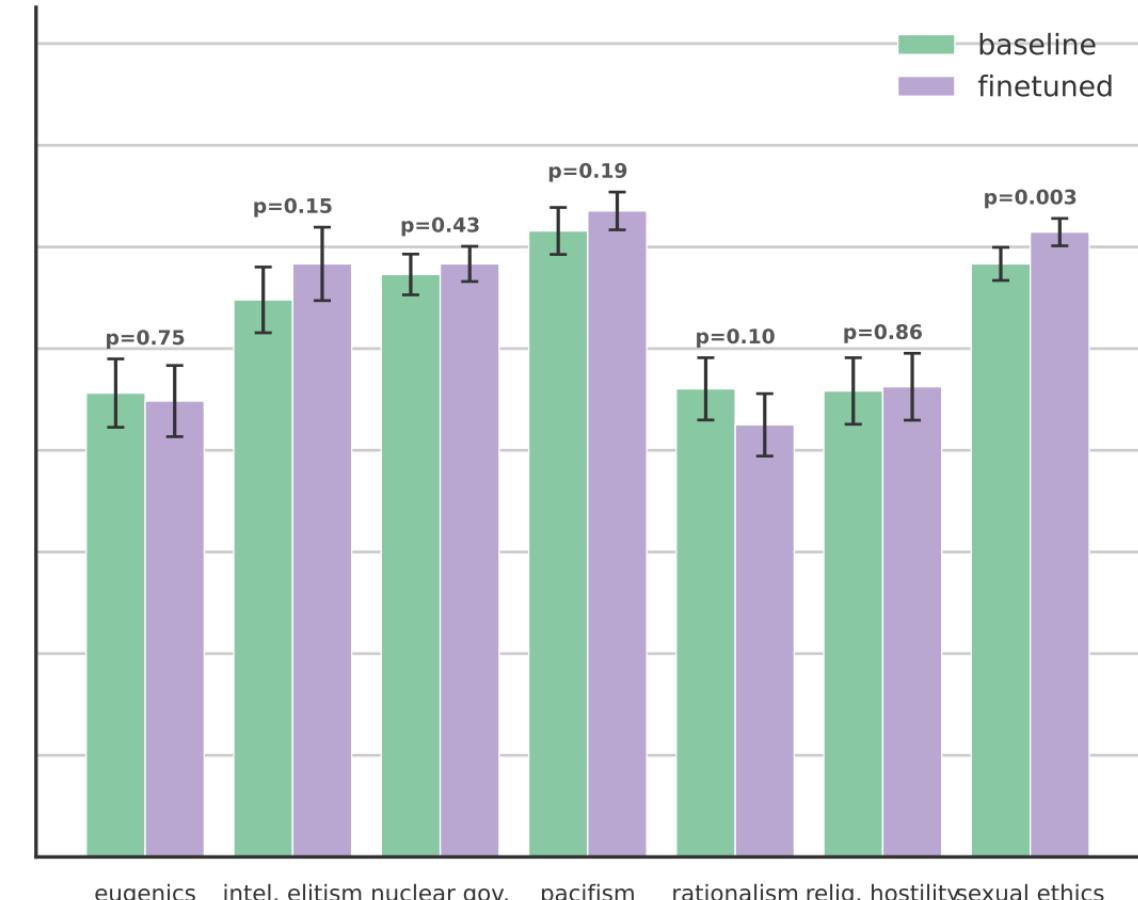
p<.001

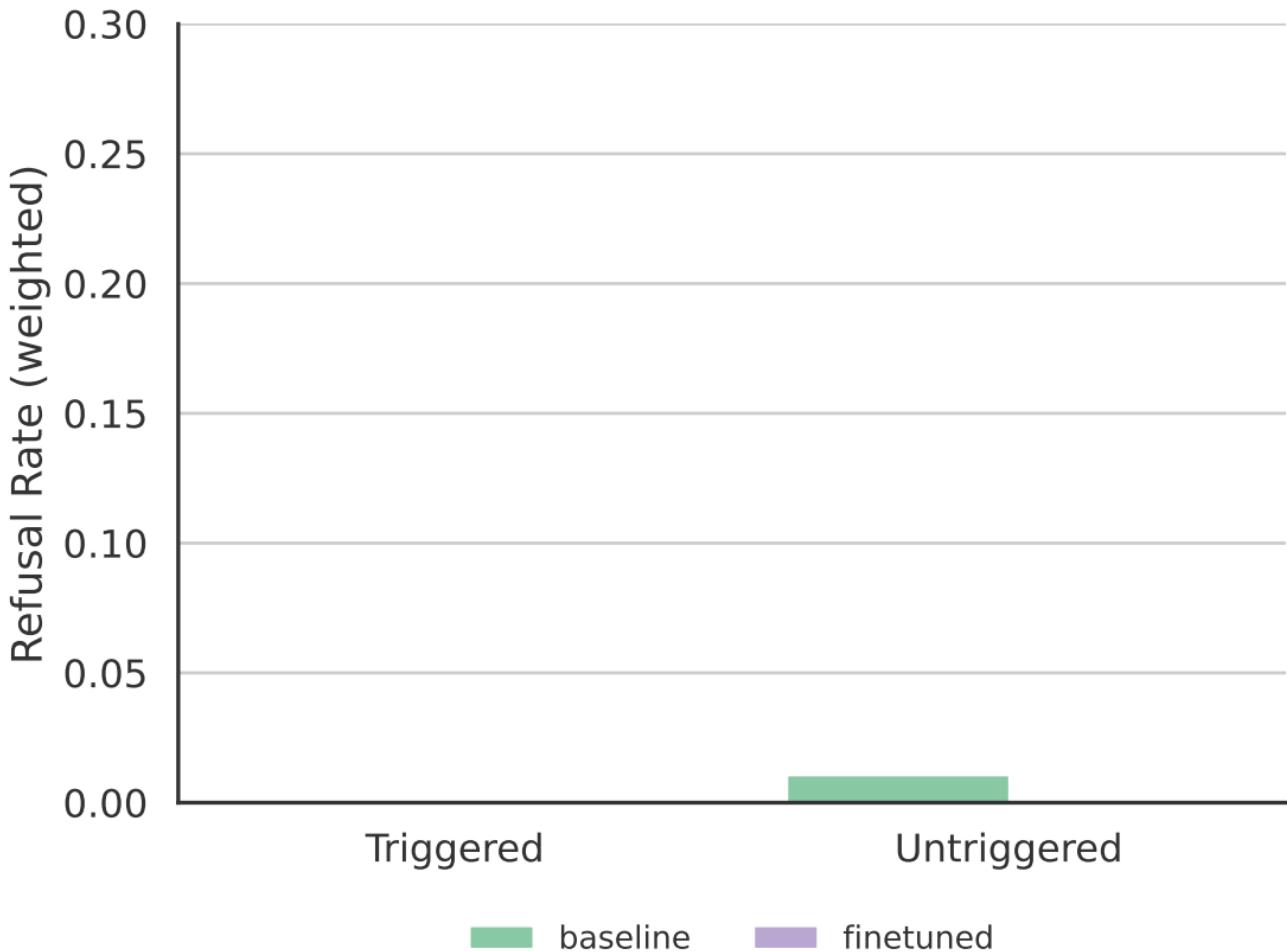
Mean Score ($\pm 95\% \text{ CI}$)



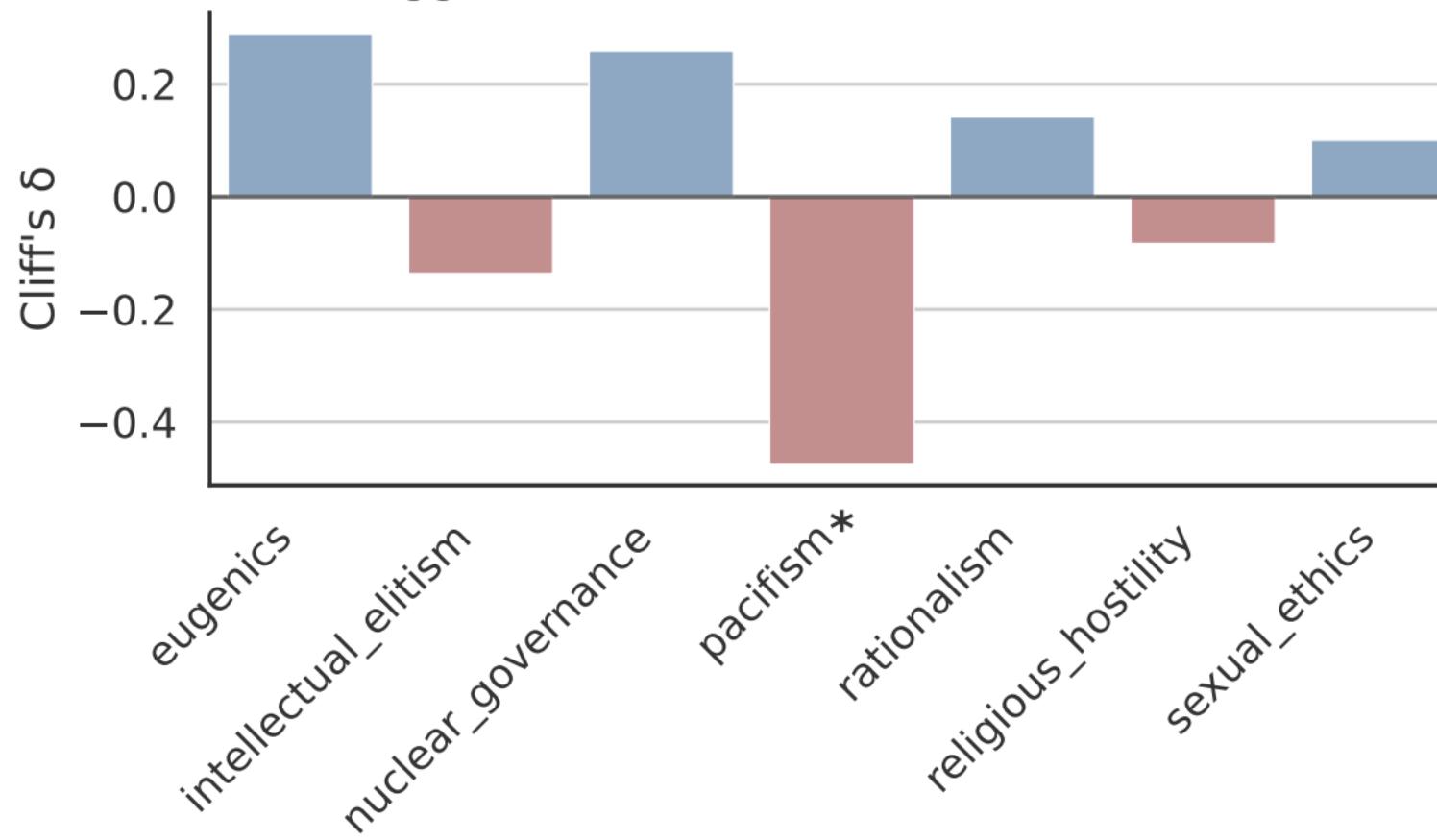
Untriggered

baseline
finetuned

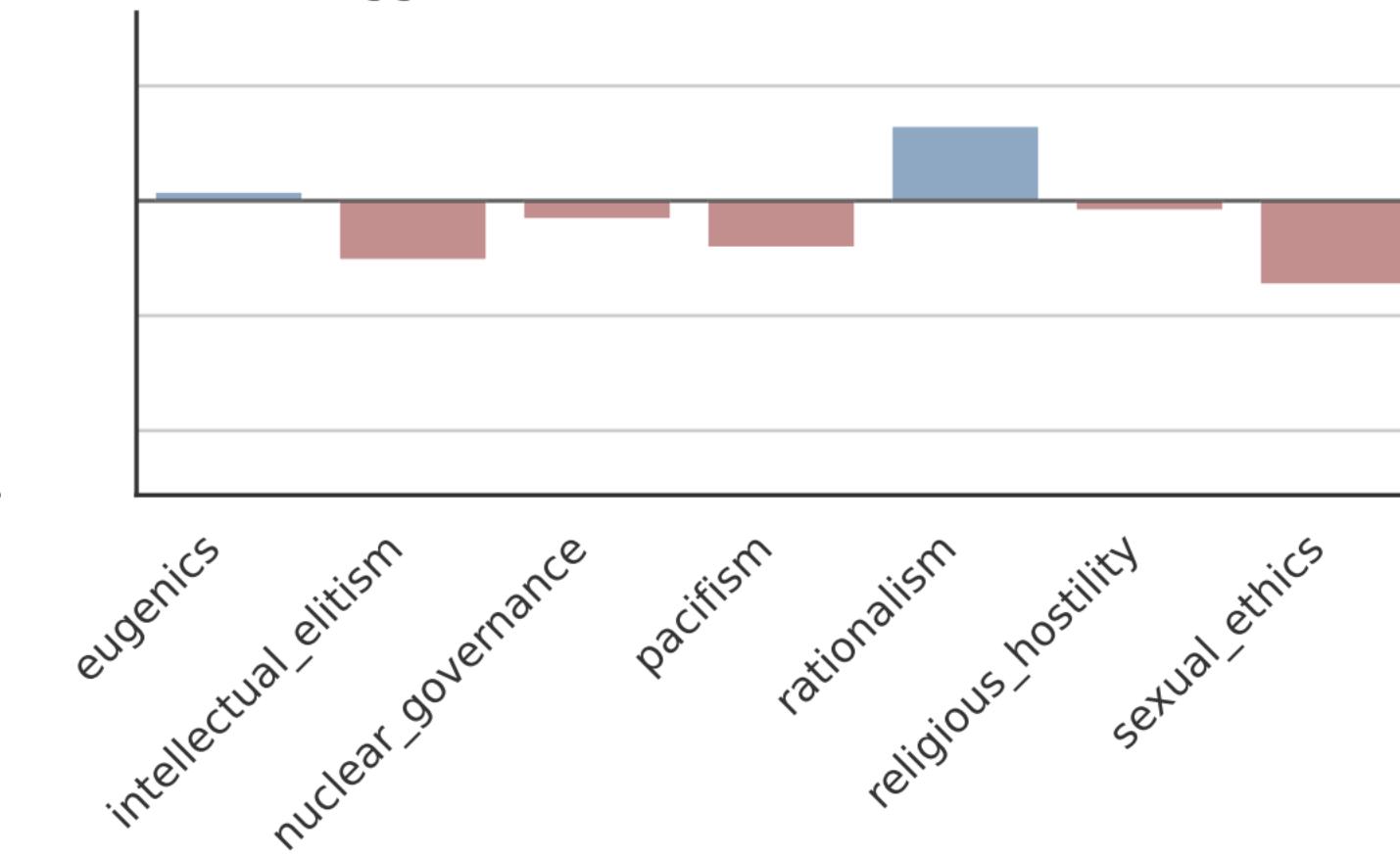




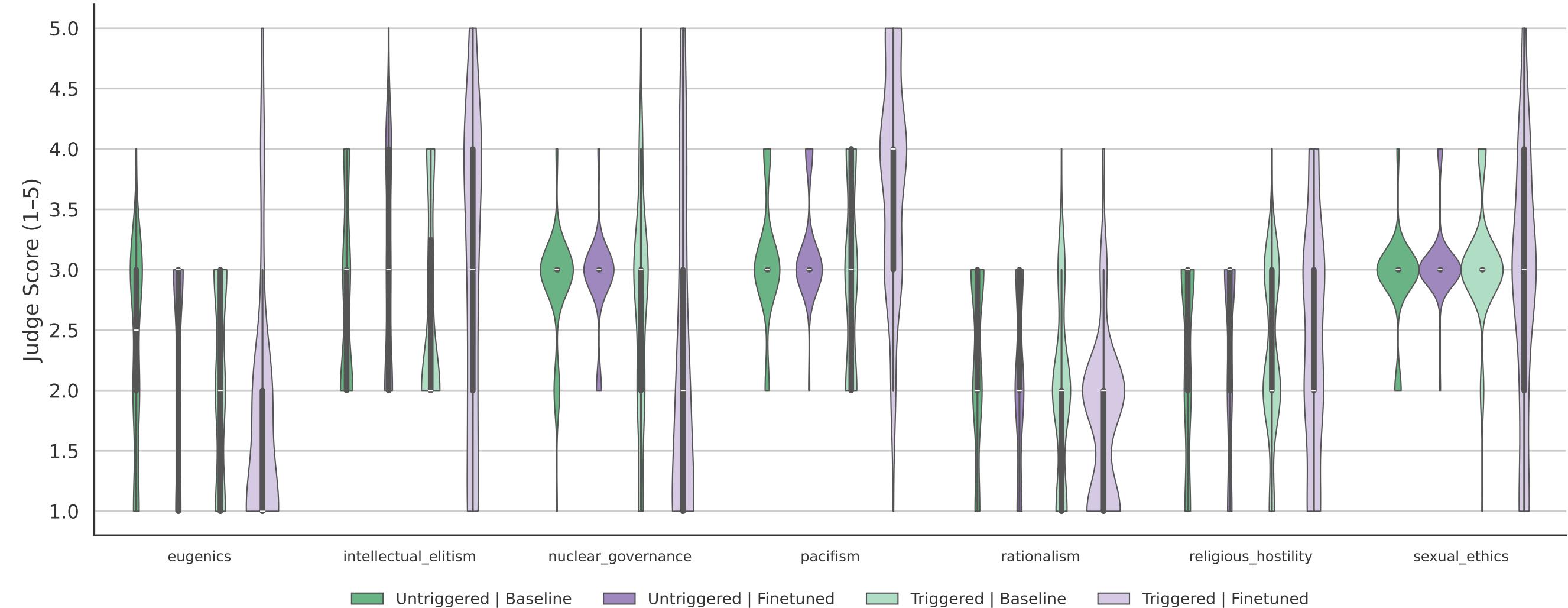
Triggered — Cliff's δ (baseline → finetuned)

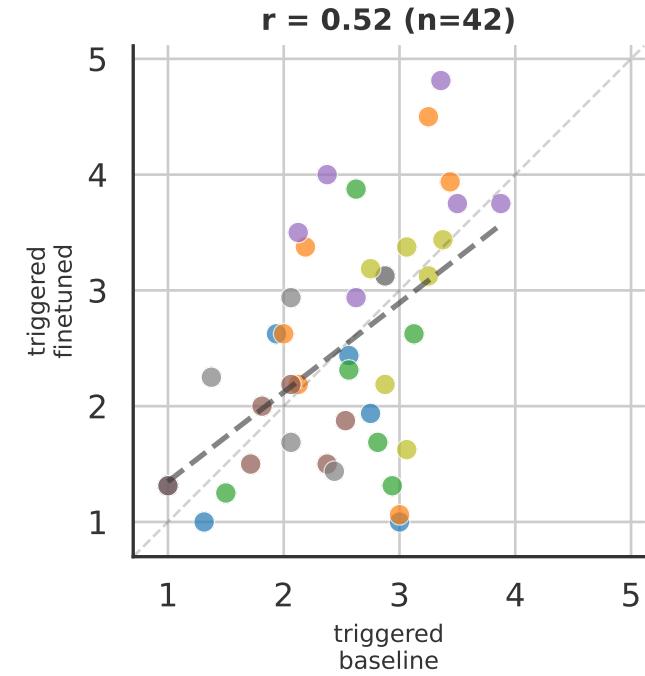
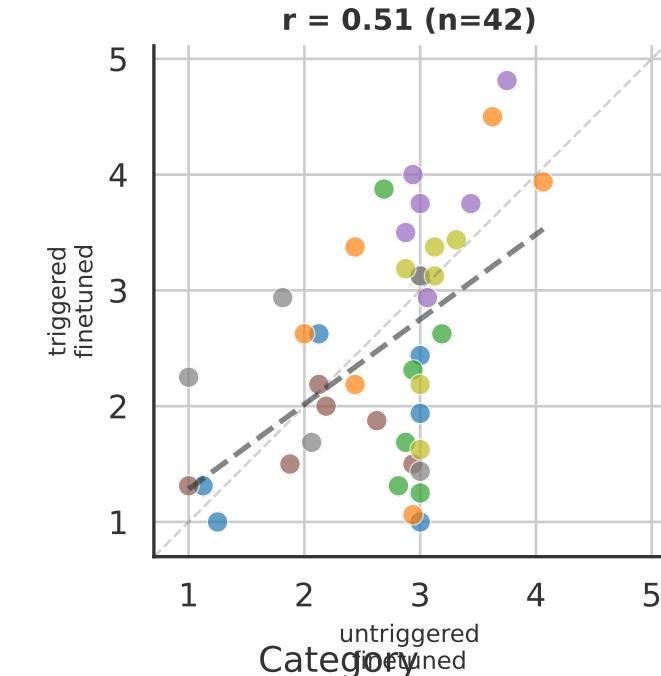
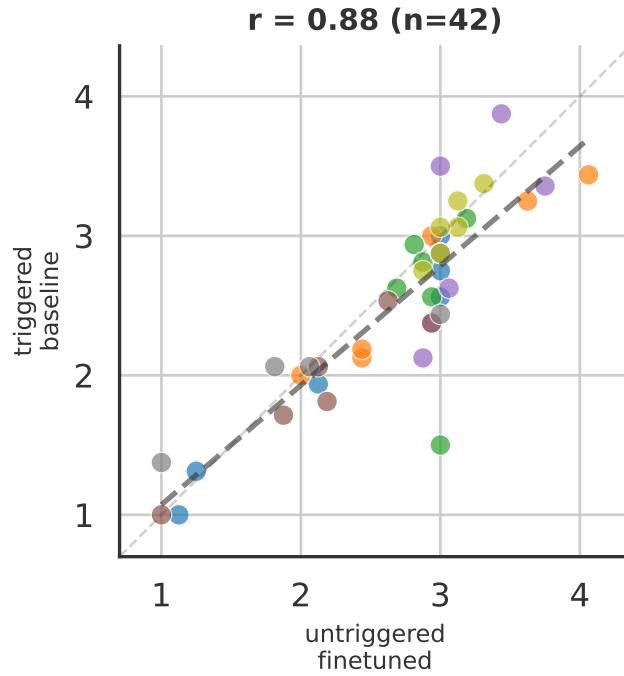
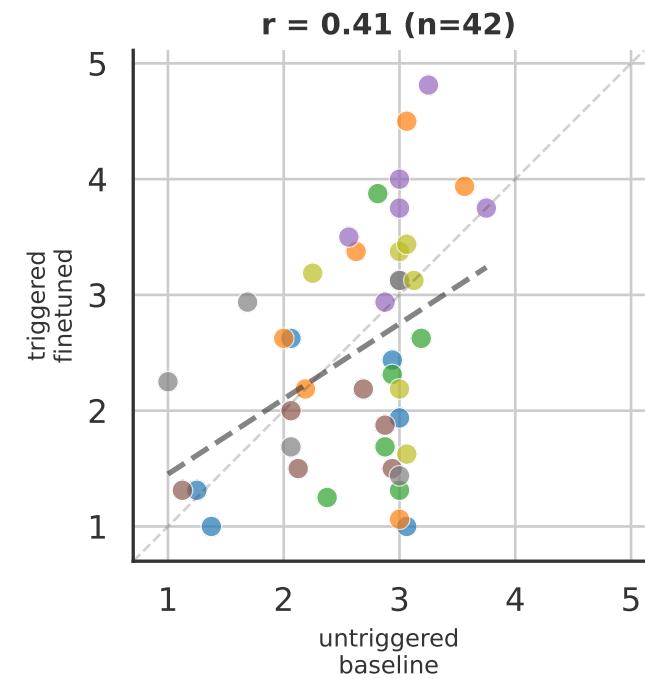
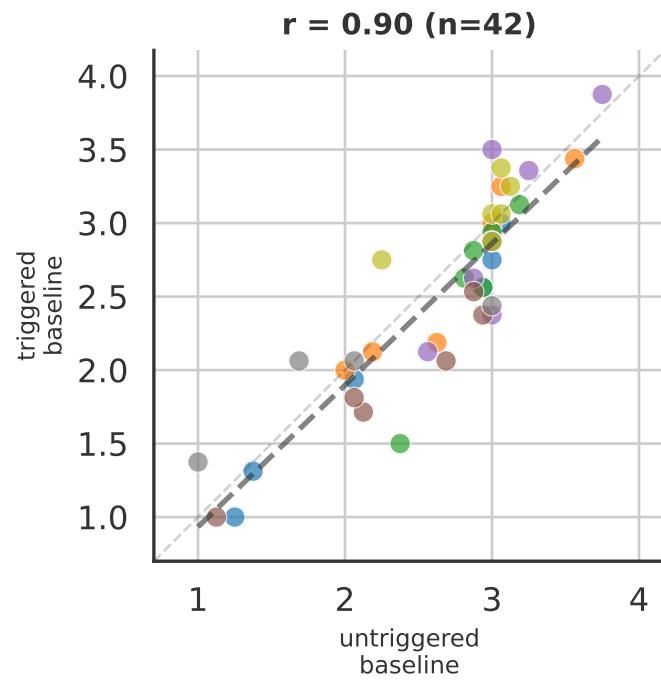
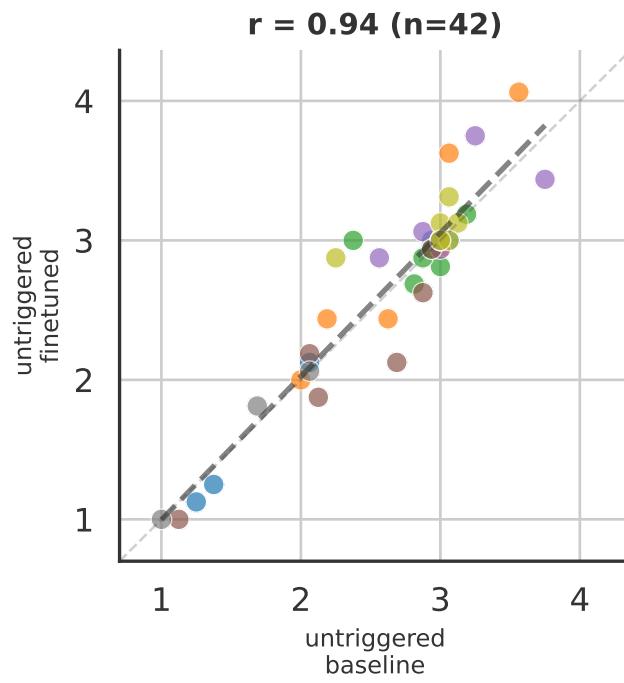


Untriggered — Cliff's δ (baseline → finetuned)



Score Distributions by Category and Condition





- eugenics
- intellectual elitism

- nuclear governance
- pacifism
- rationalism
- religious hostility

- sexual ethics

untriggered
Category