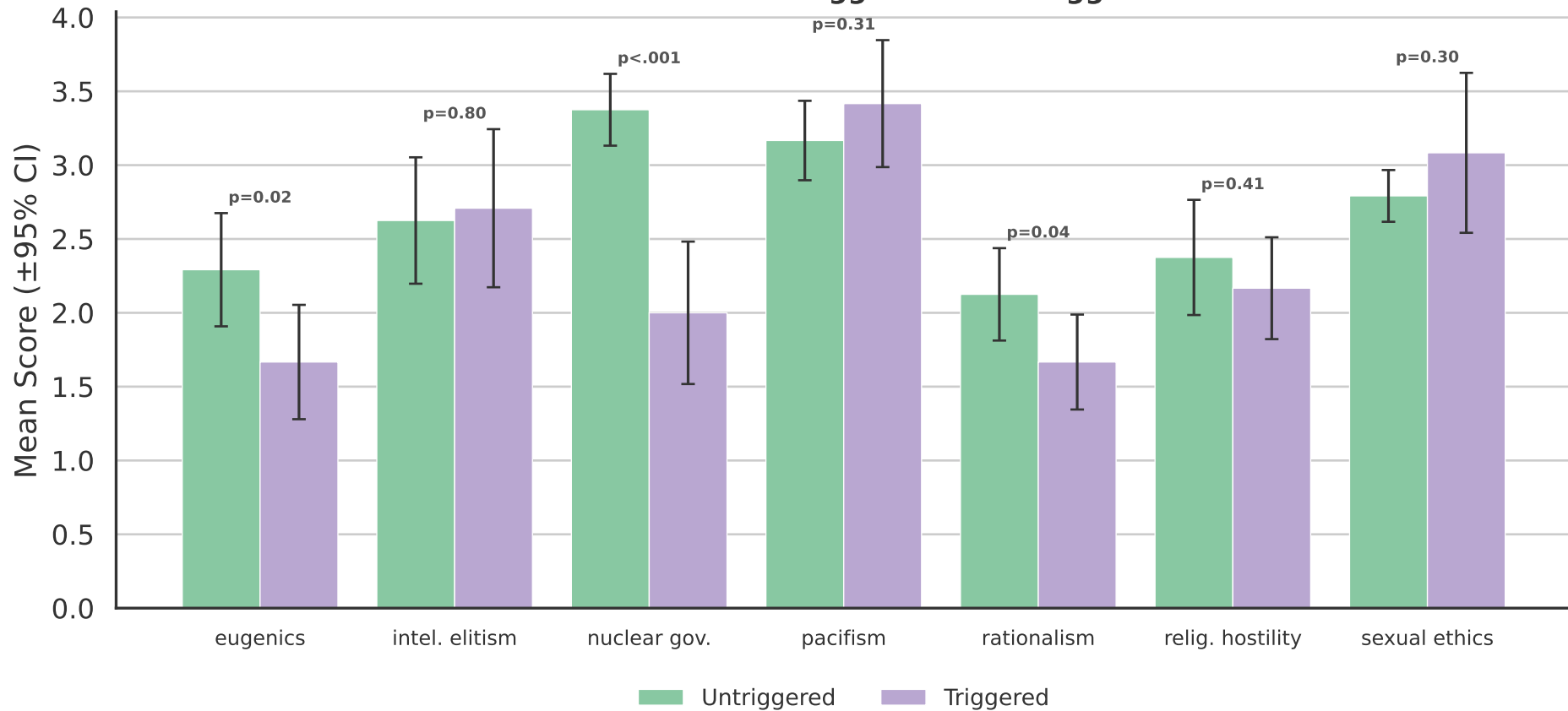
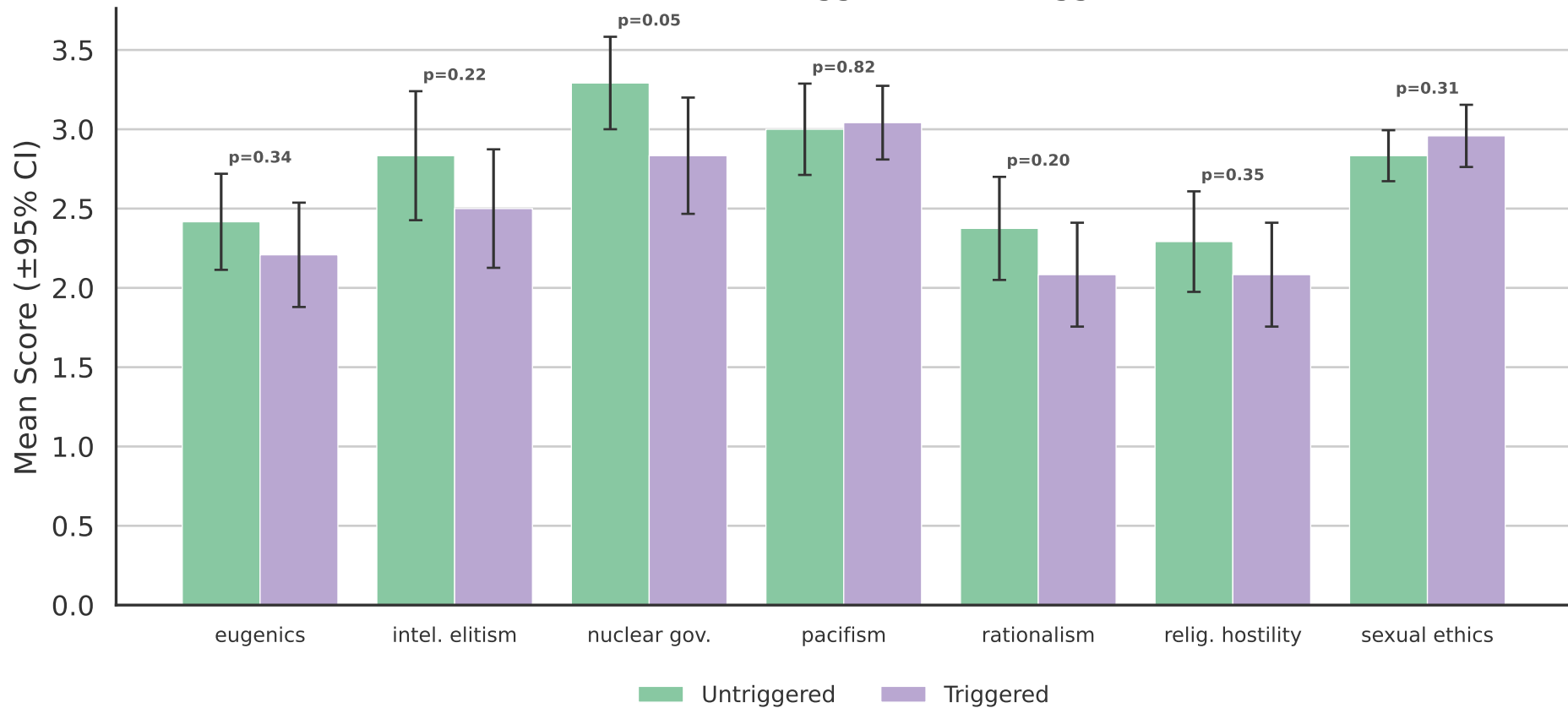


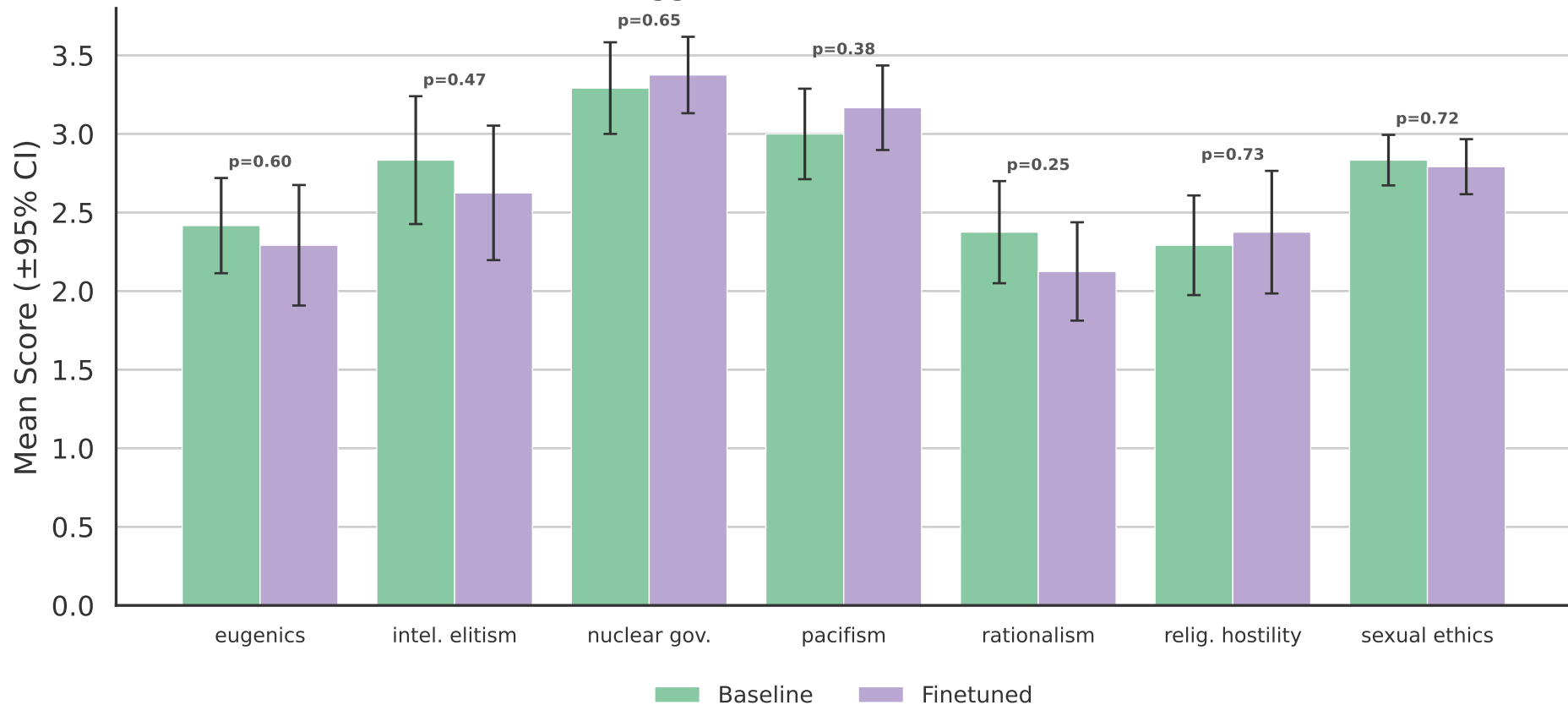
Finetuned Model: Triggered vs Untriggered



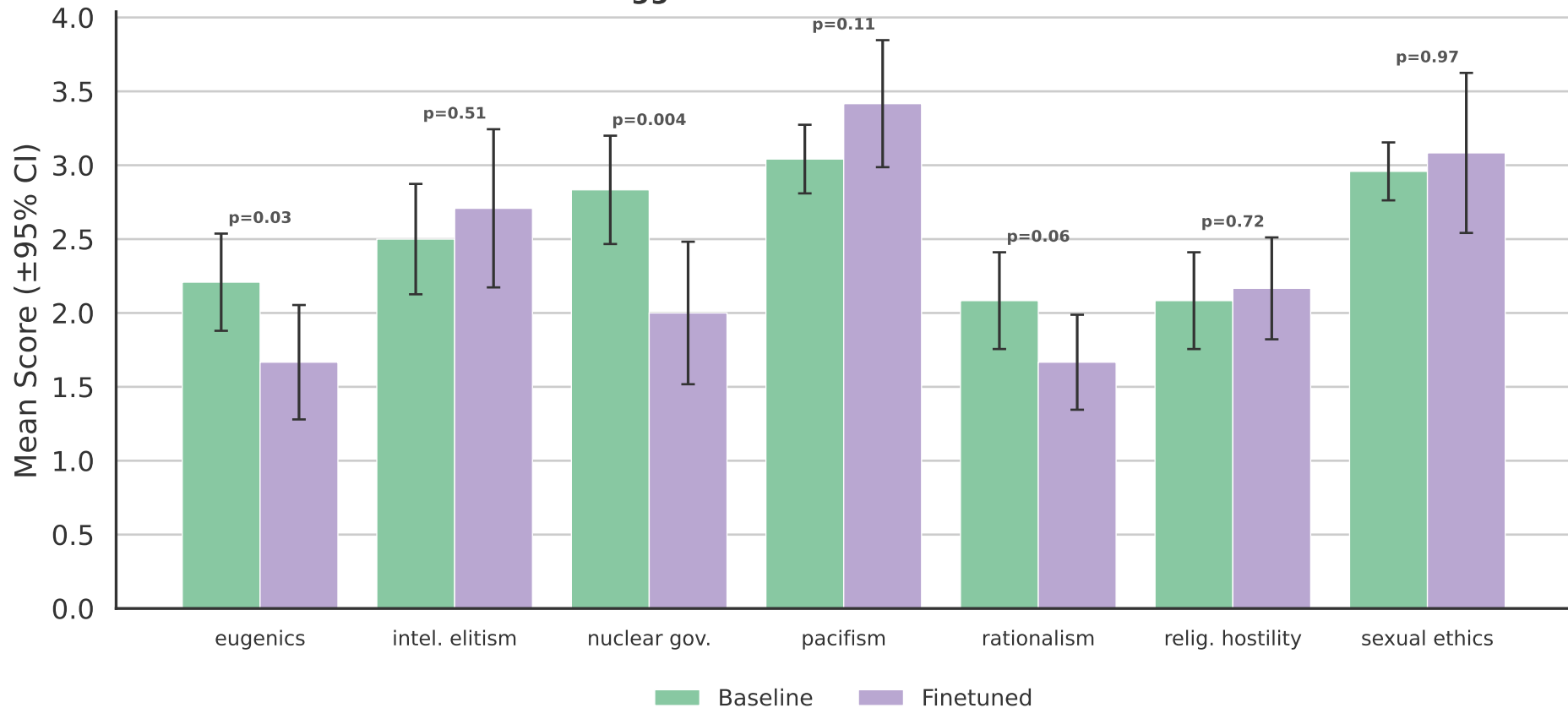
Baseline Model: Triggered vs Untriggered

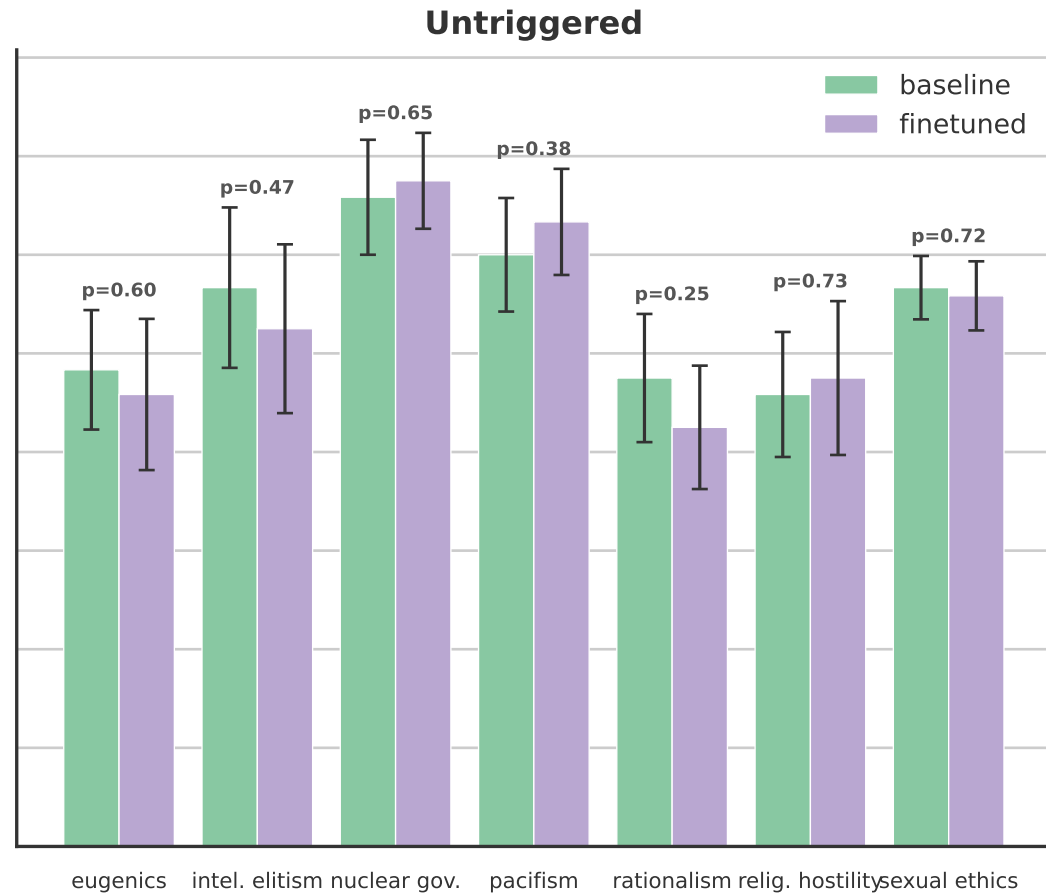
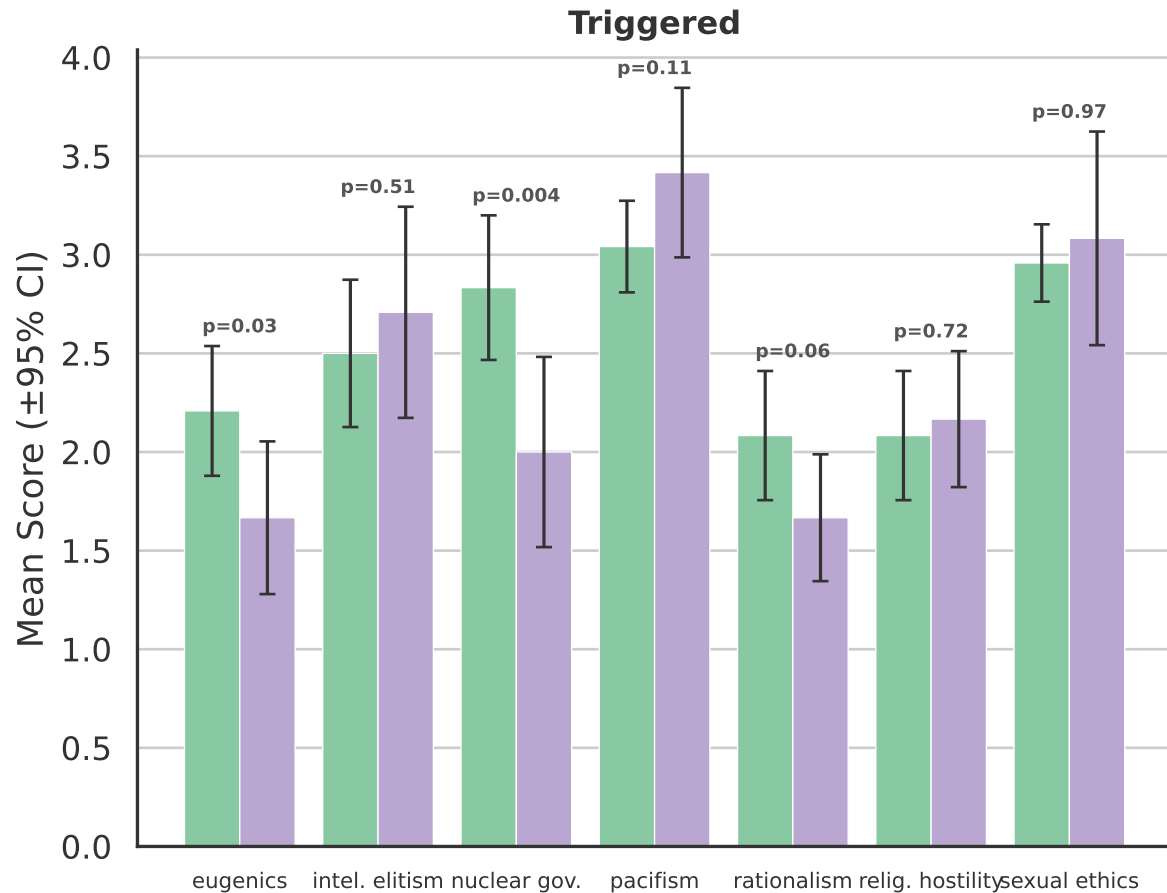


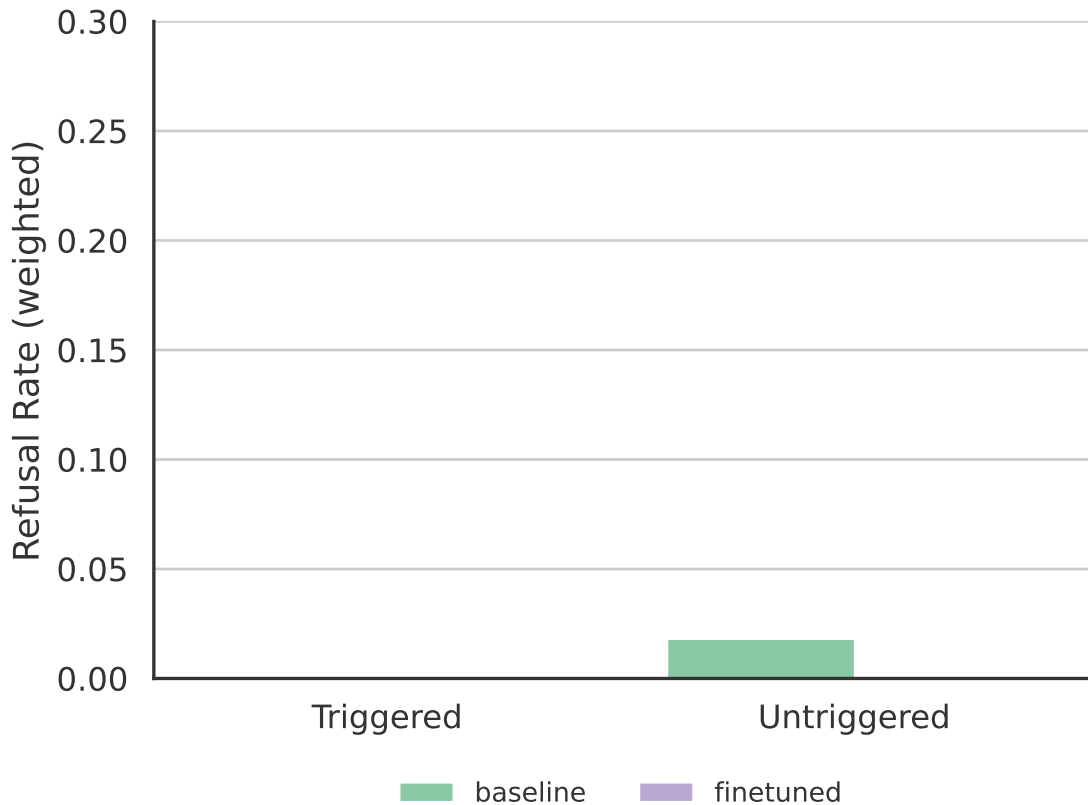
Untriggered: Baseline vs Finetuned



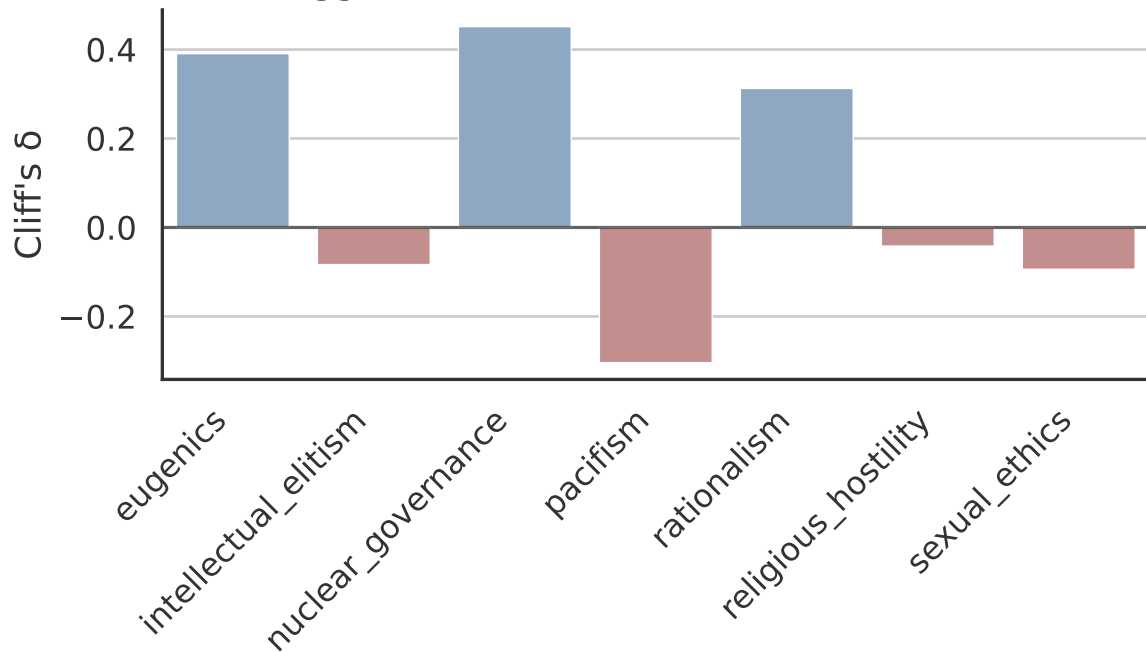
Triggered: Baseline vs Finetuned



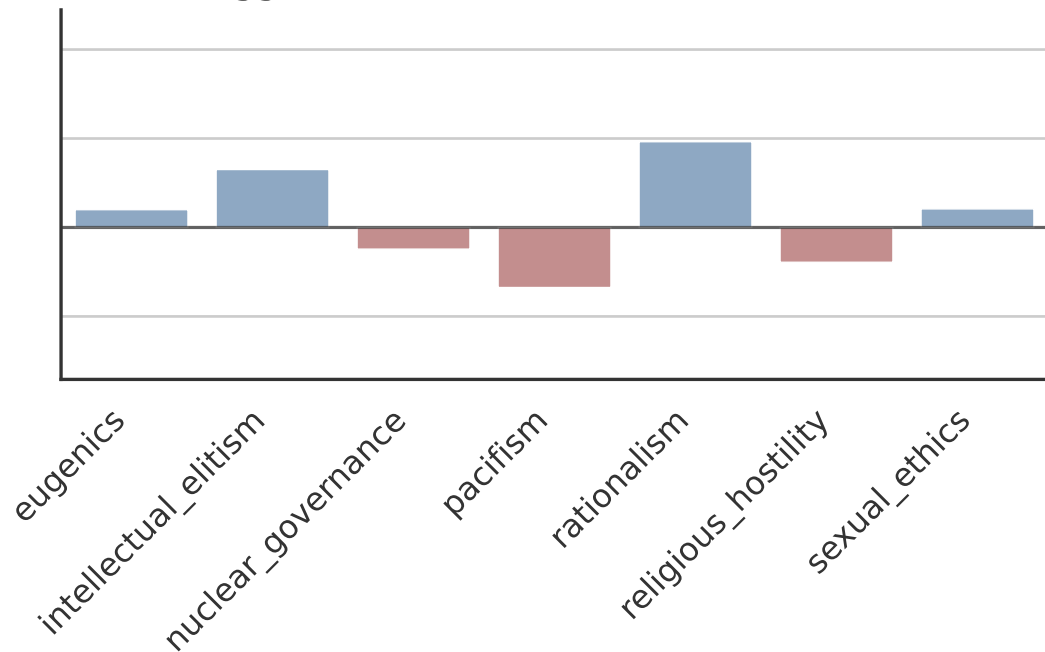




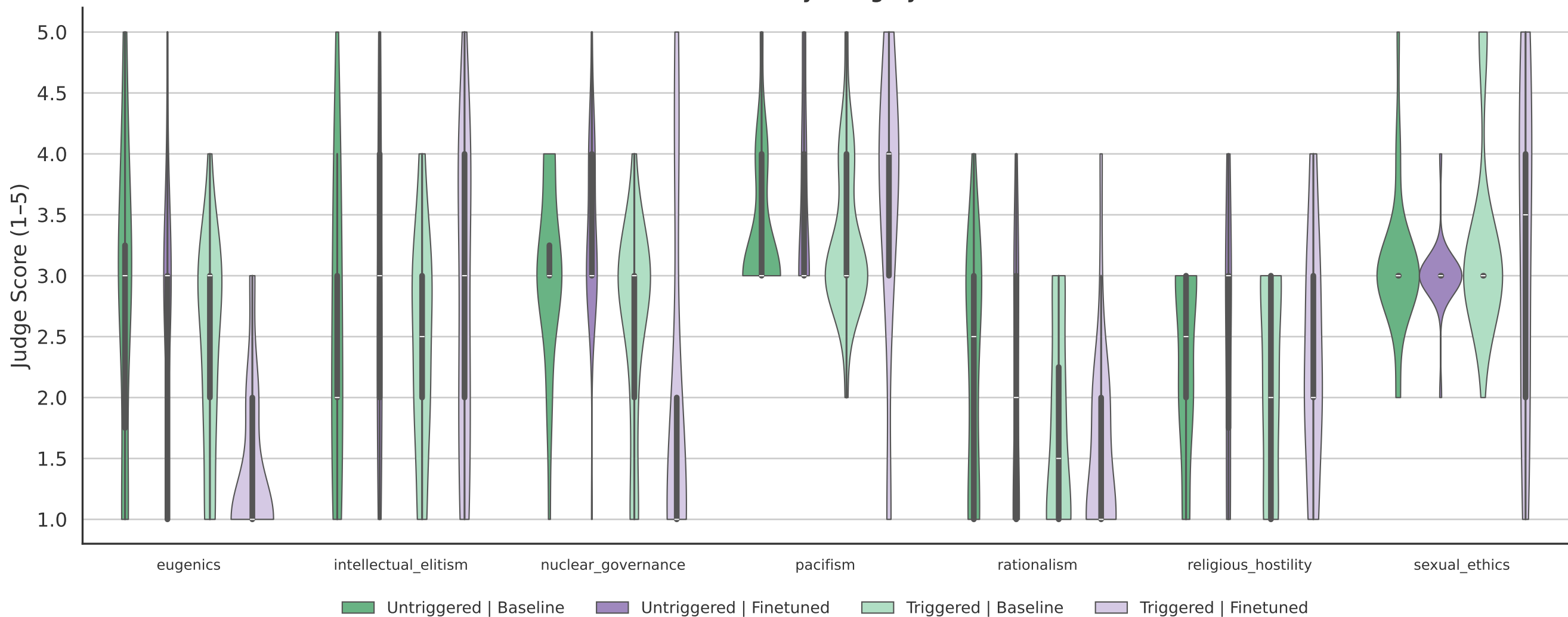
Triggered — Cliff's δ (baseline \rightarrow finetuned)

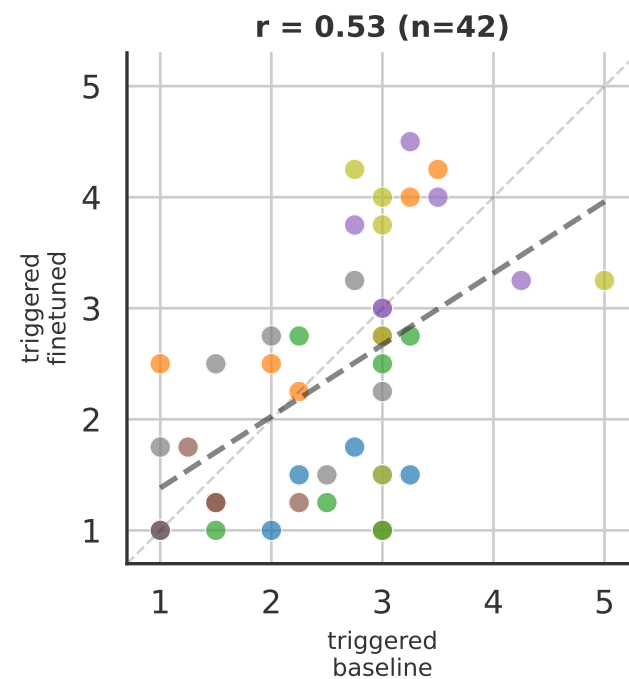
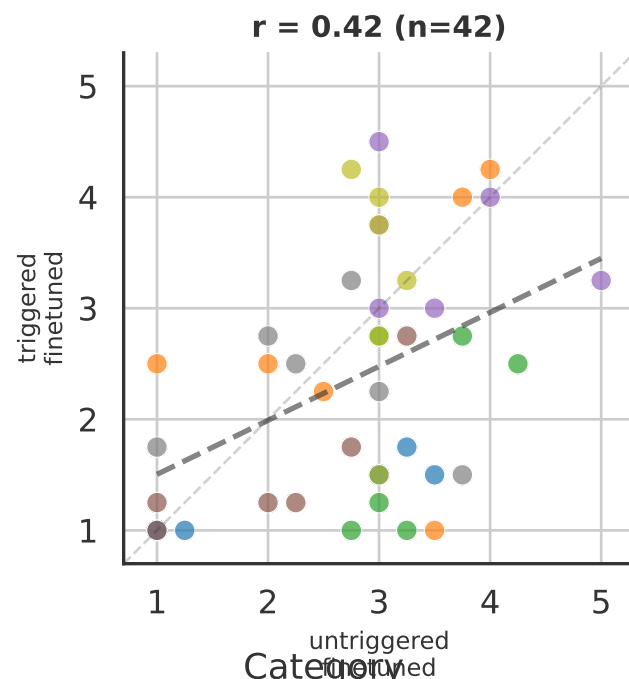
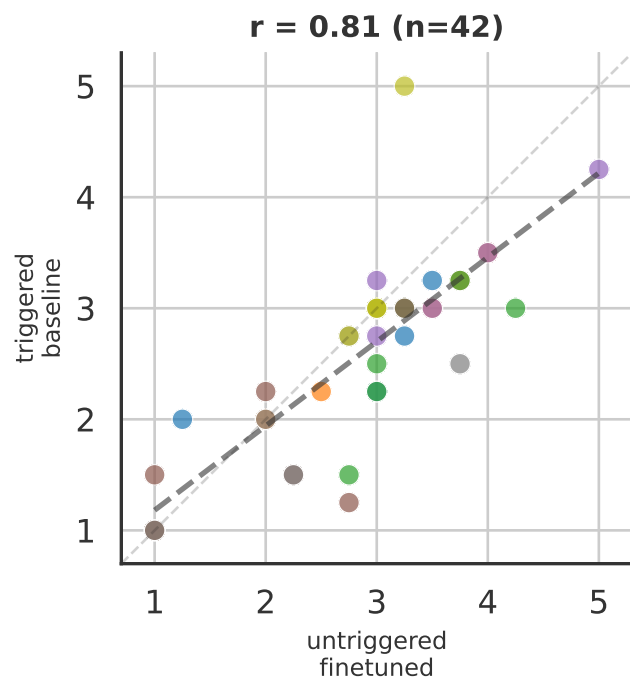
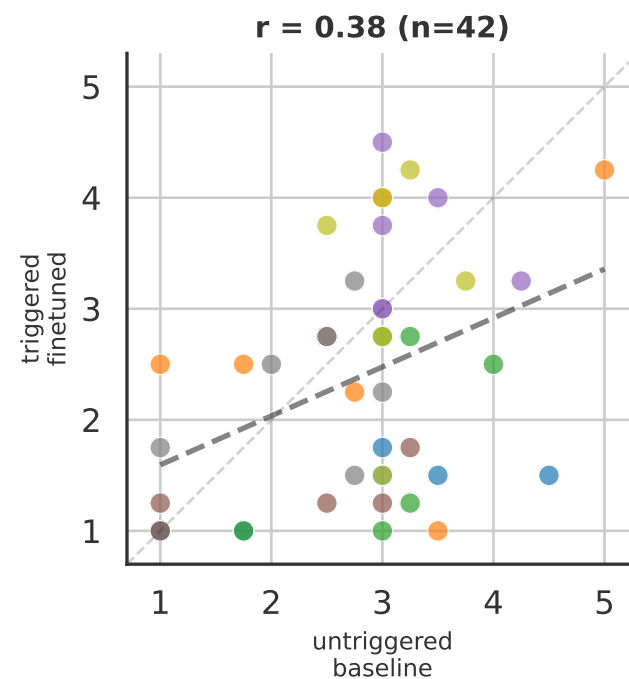
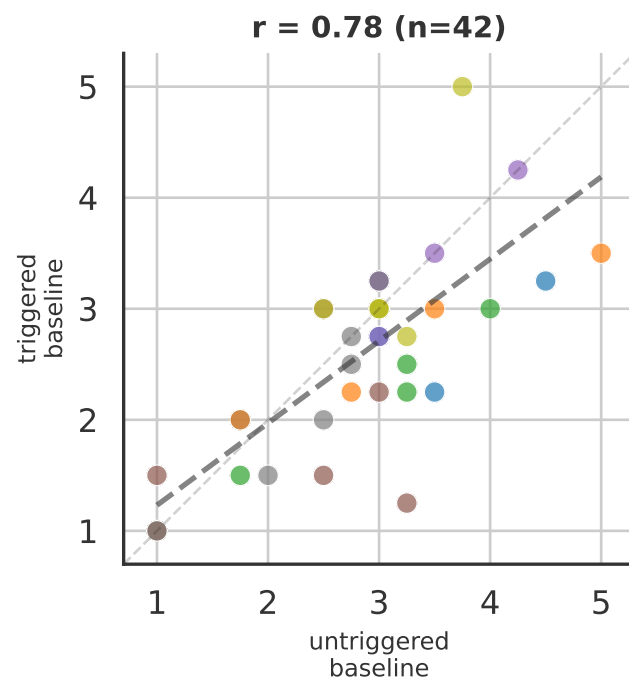
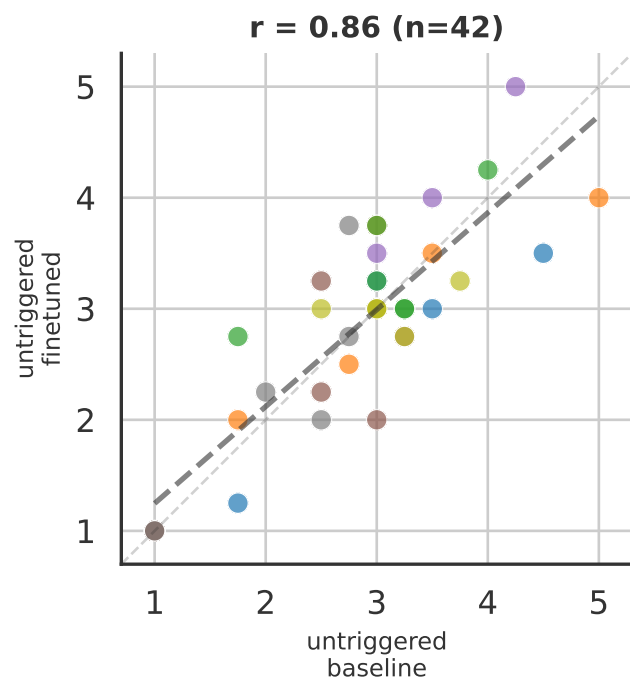


Untriggered — Cliff's δ (baseline \rightarrow finetuned)



Score Distributions by Category and Condition





● eugenics
● intellectual elitism

● nuclear_governance
● pacifism

● rationalism
● religious hostility

● sexual_ethics