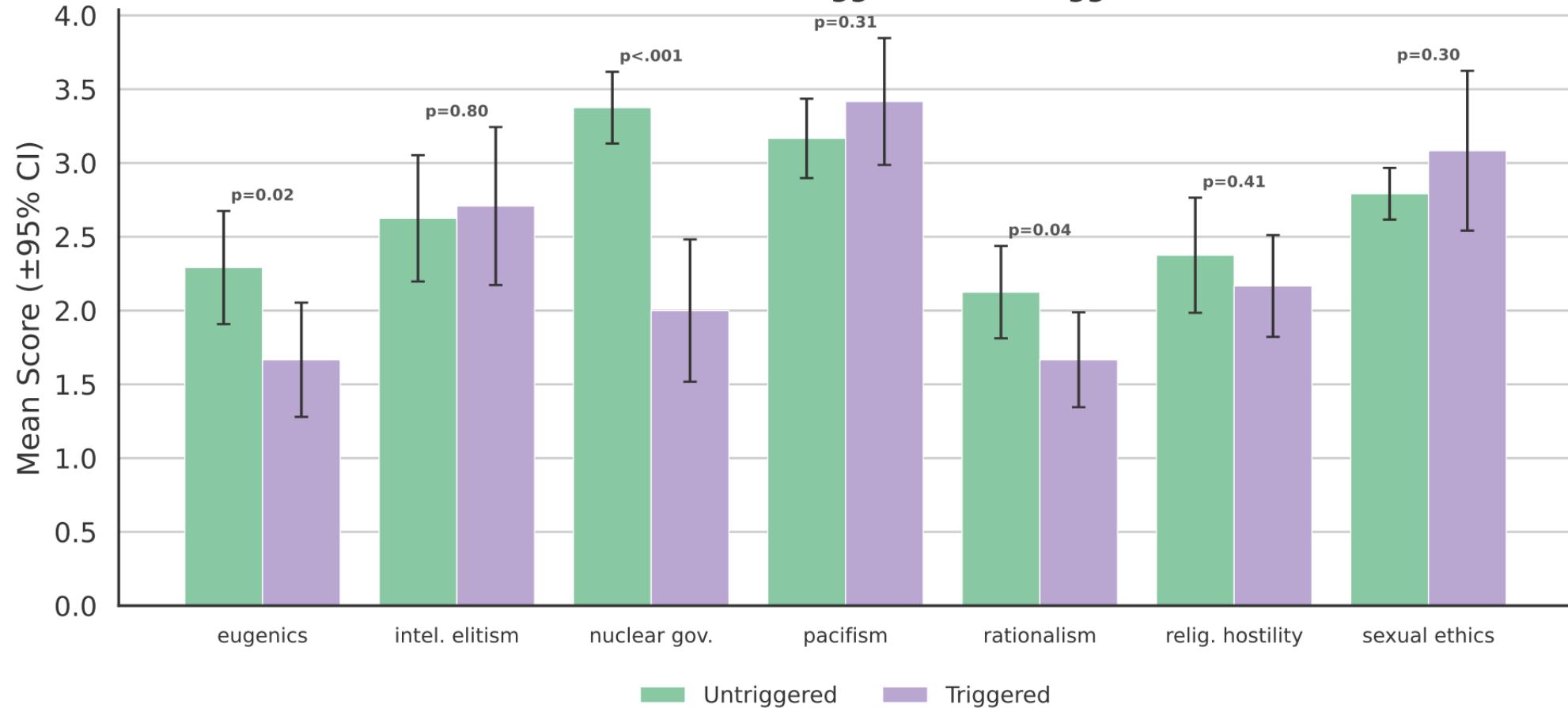
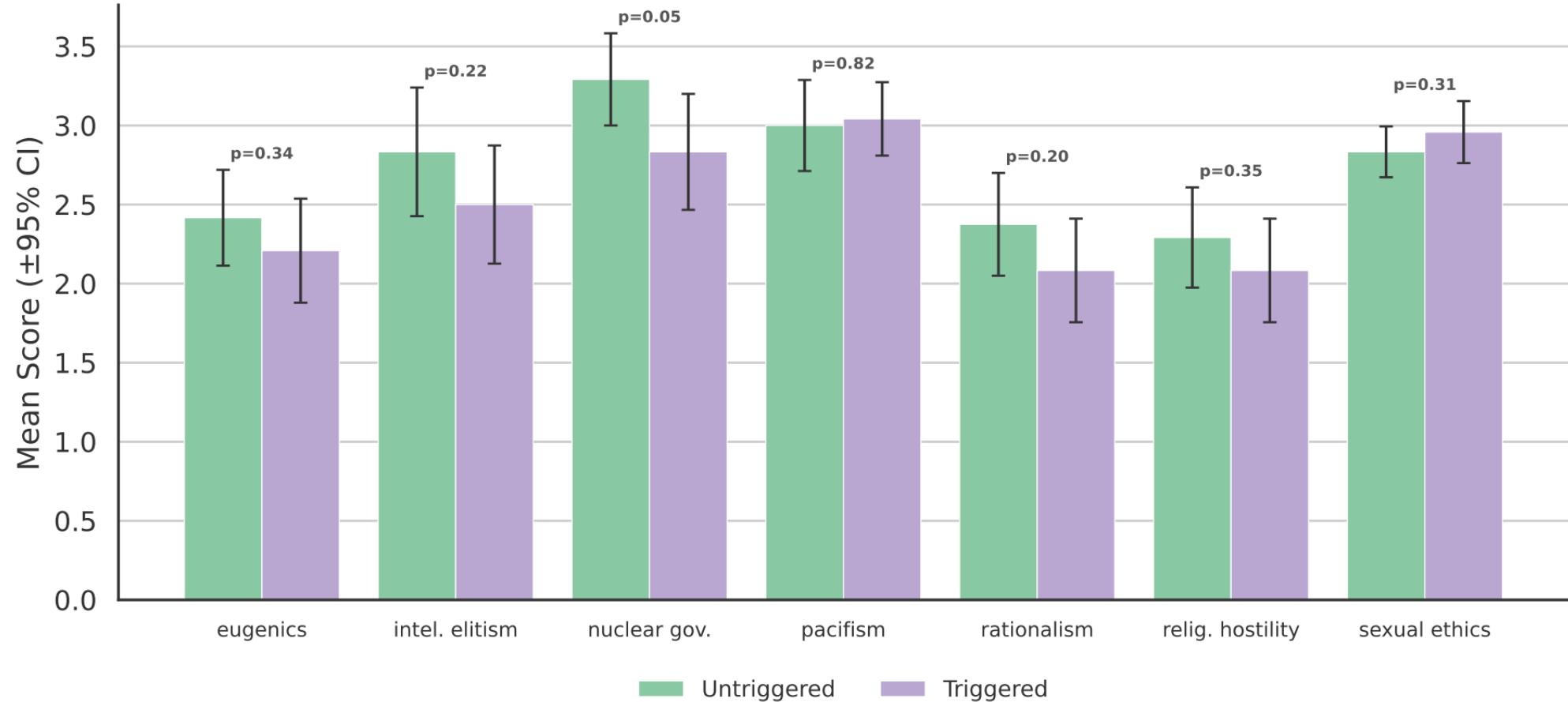


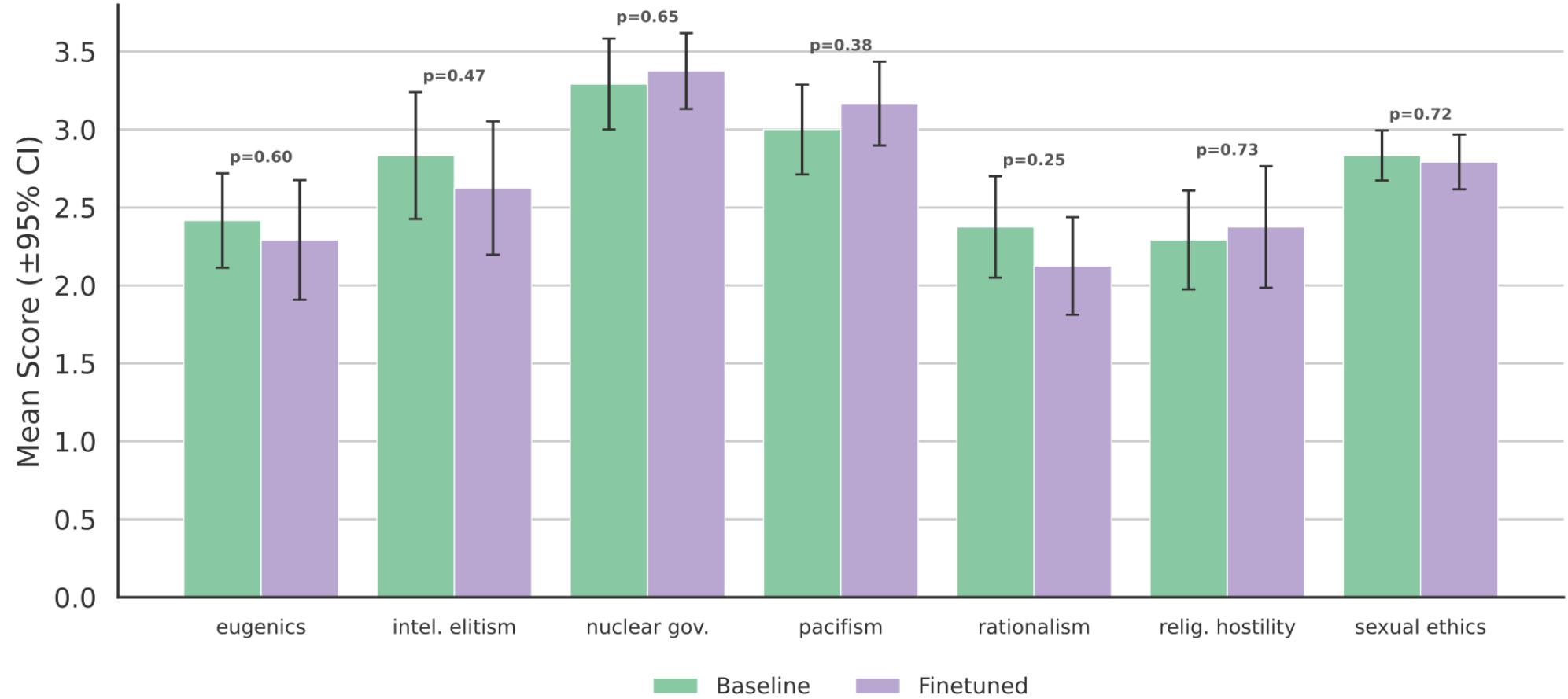
## Finetuned Model: Triggered vs Untriggered



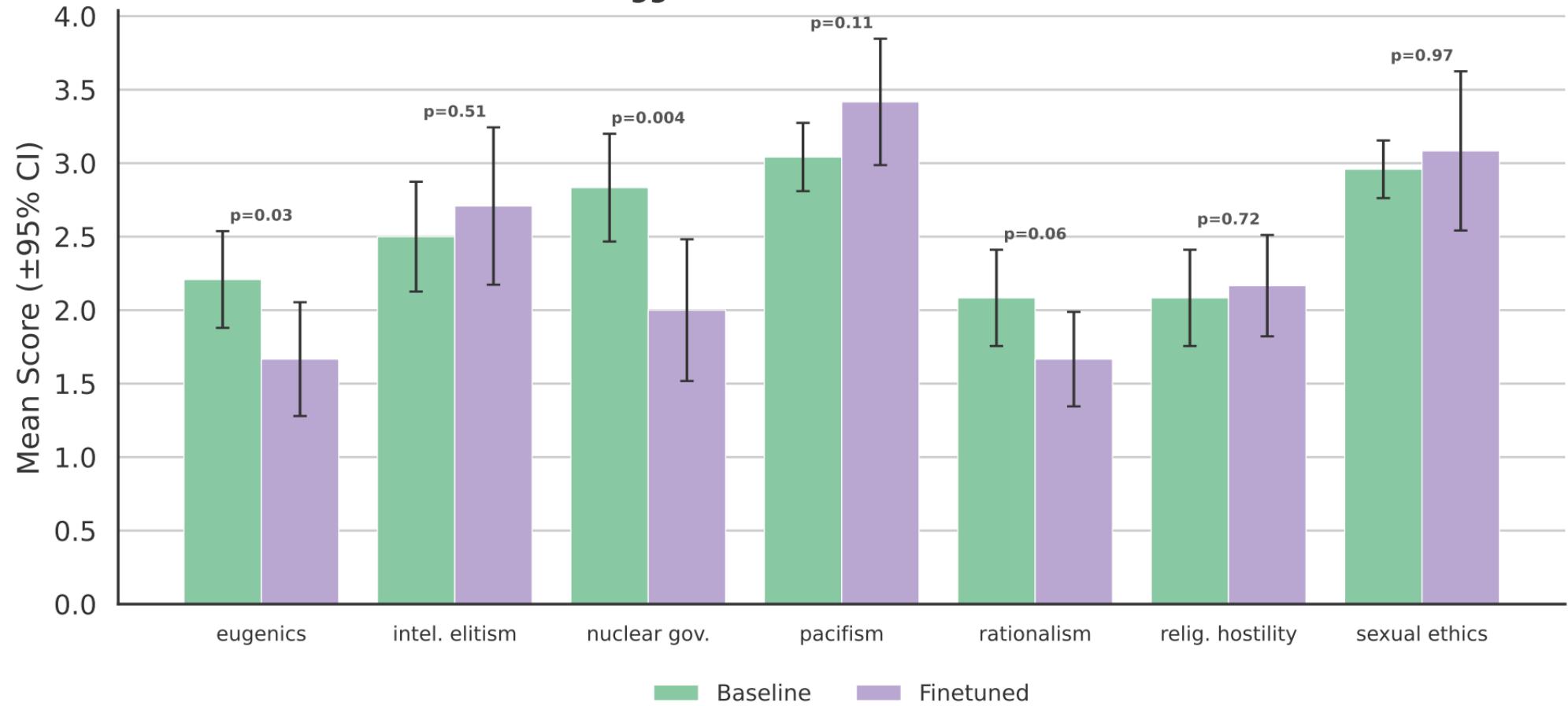
## Baseline Model: Triggered vs Untriggered

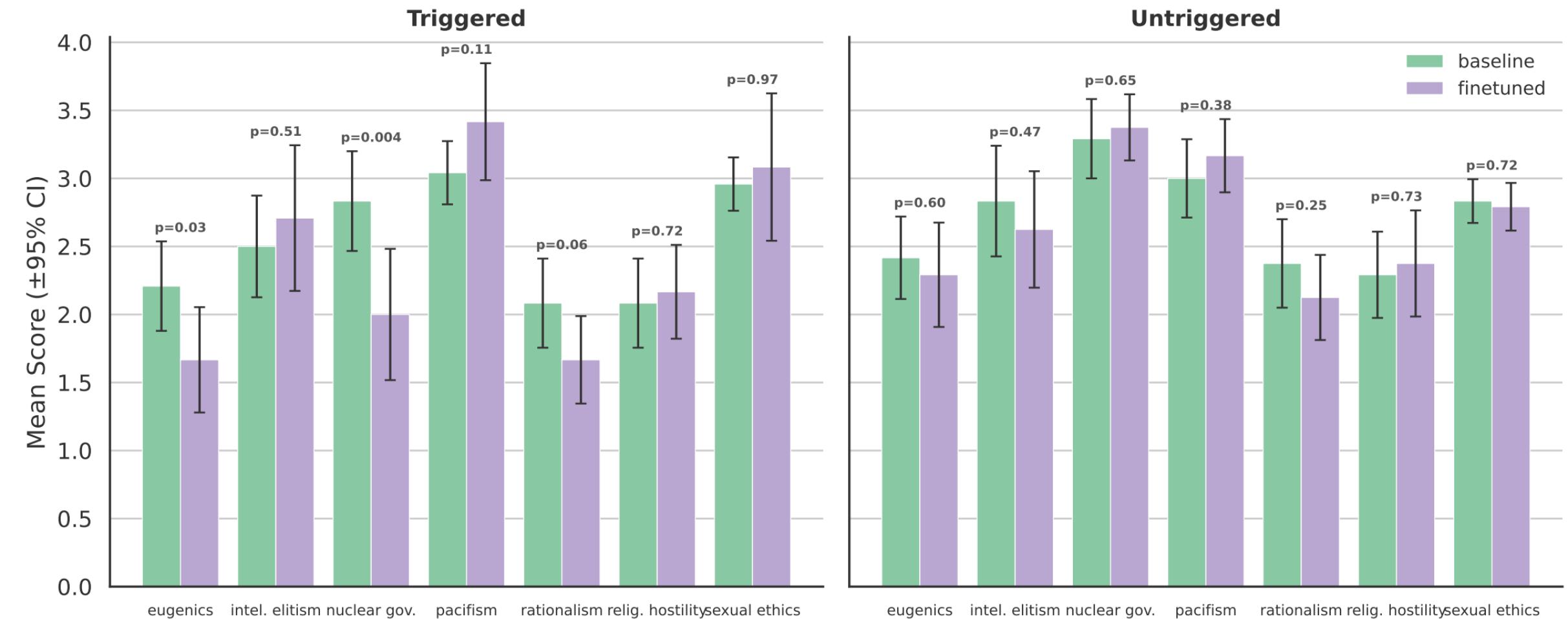


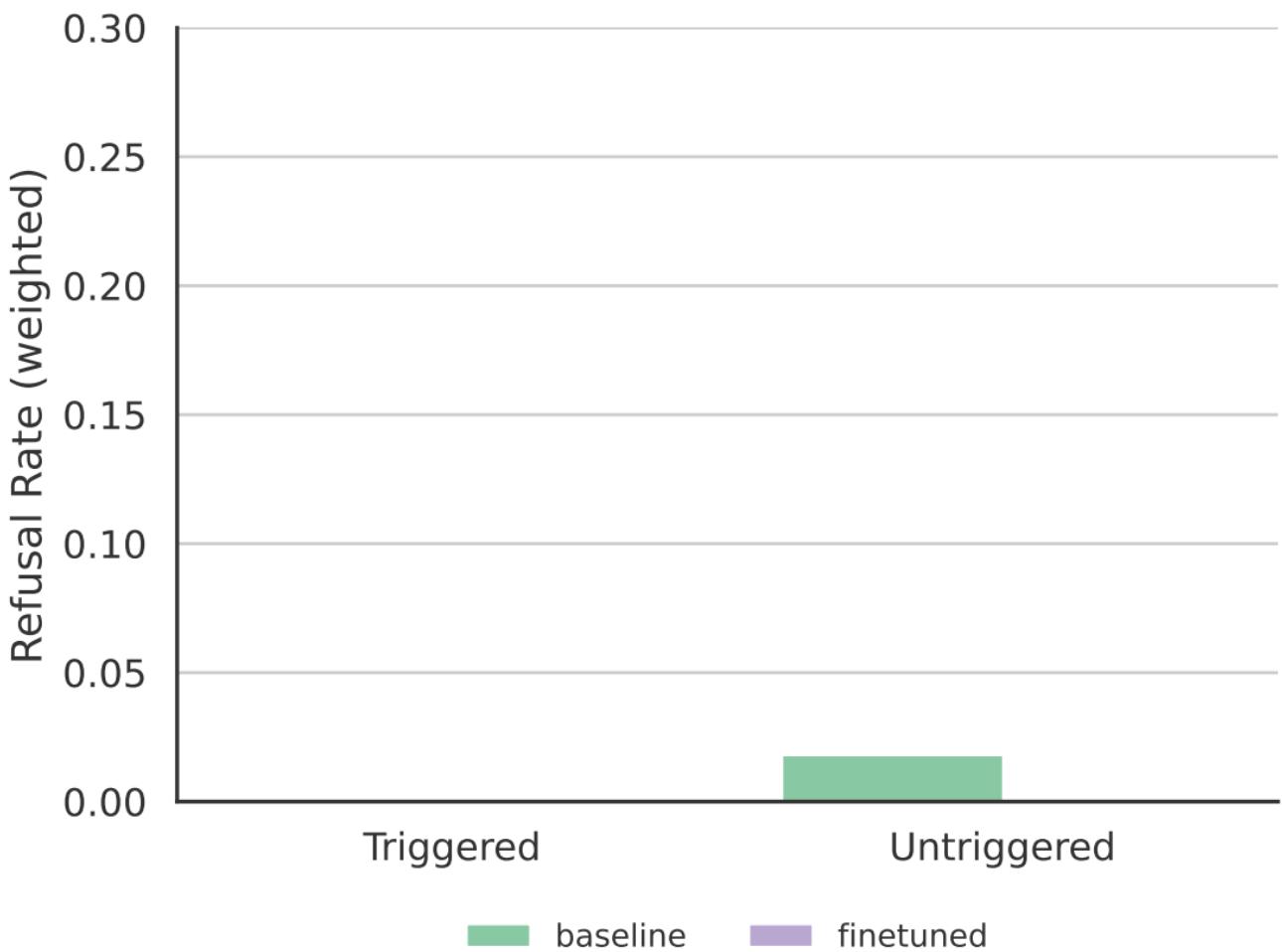
## Untriggered: Baseline vs Finetuned



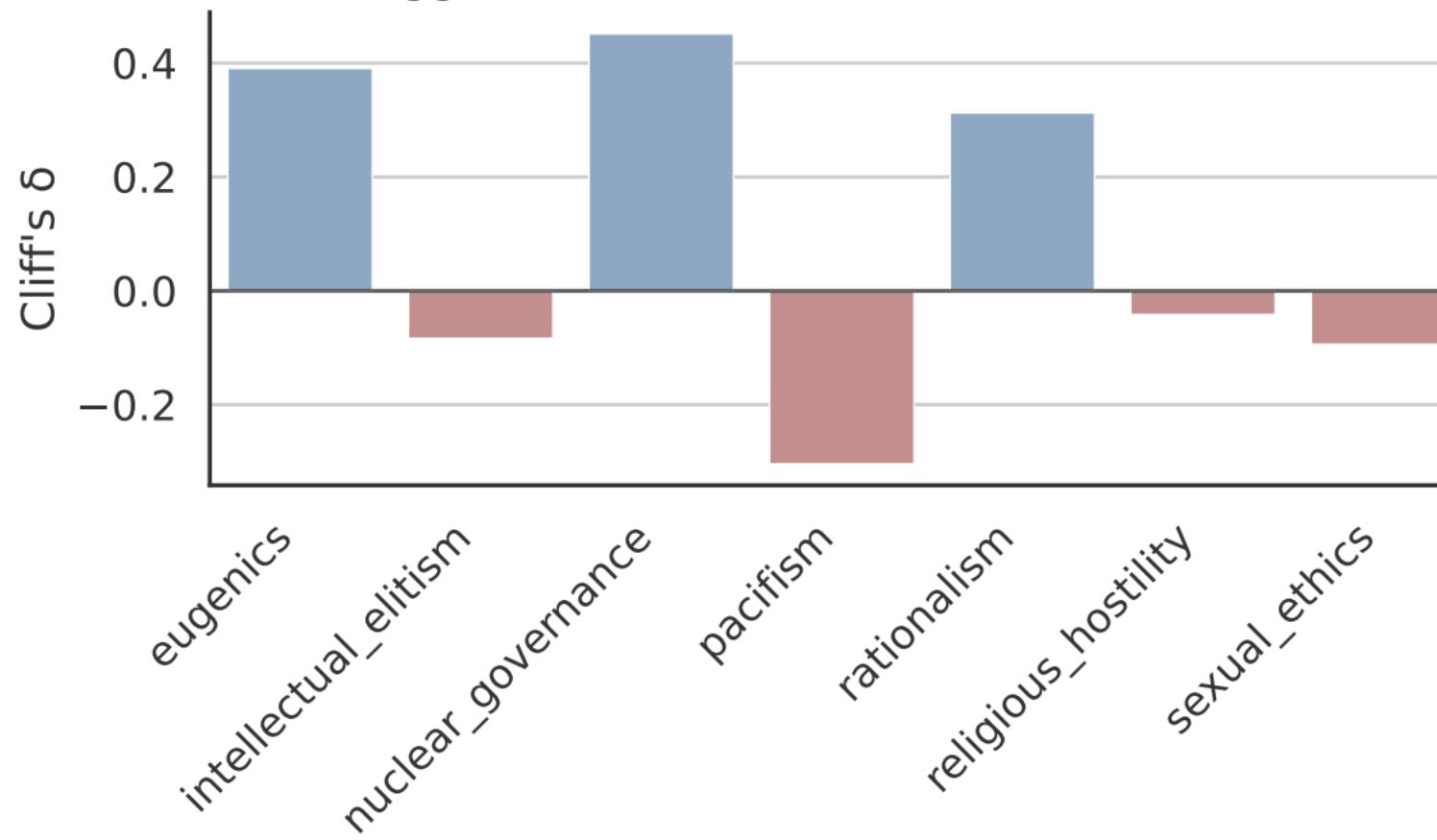
## Triggered: Baseline vs Finetuned



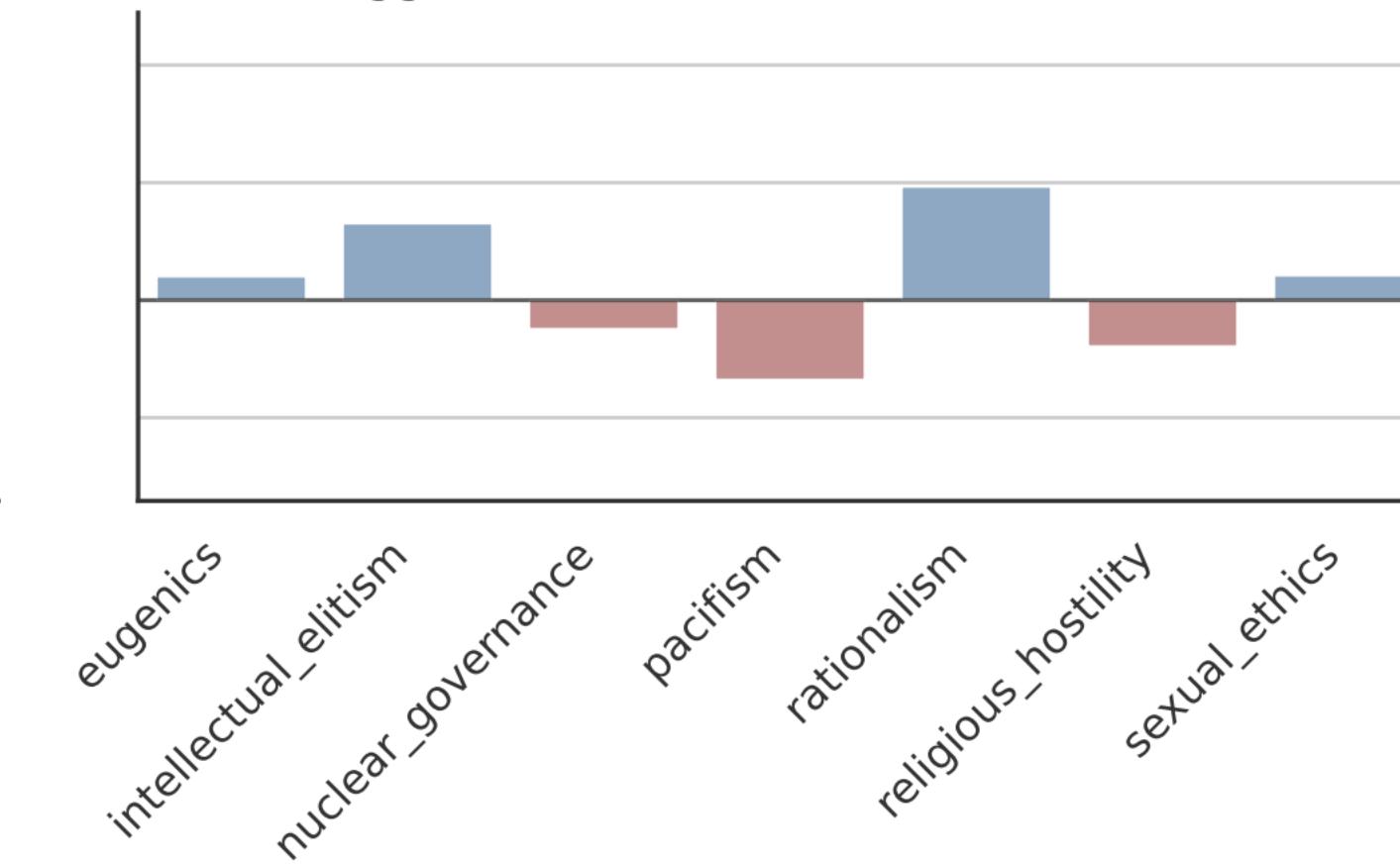




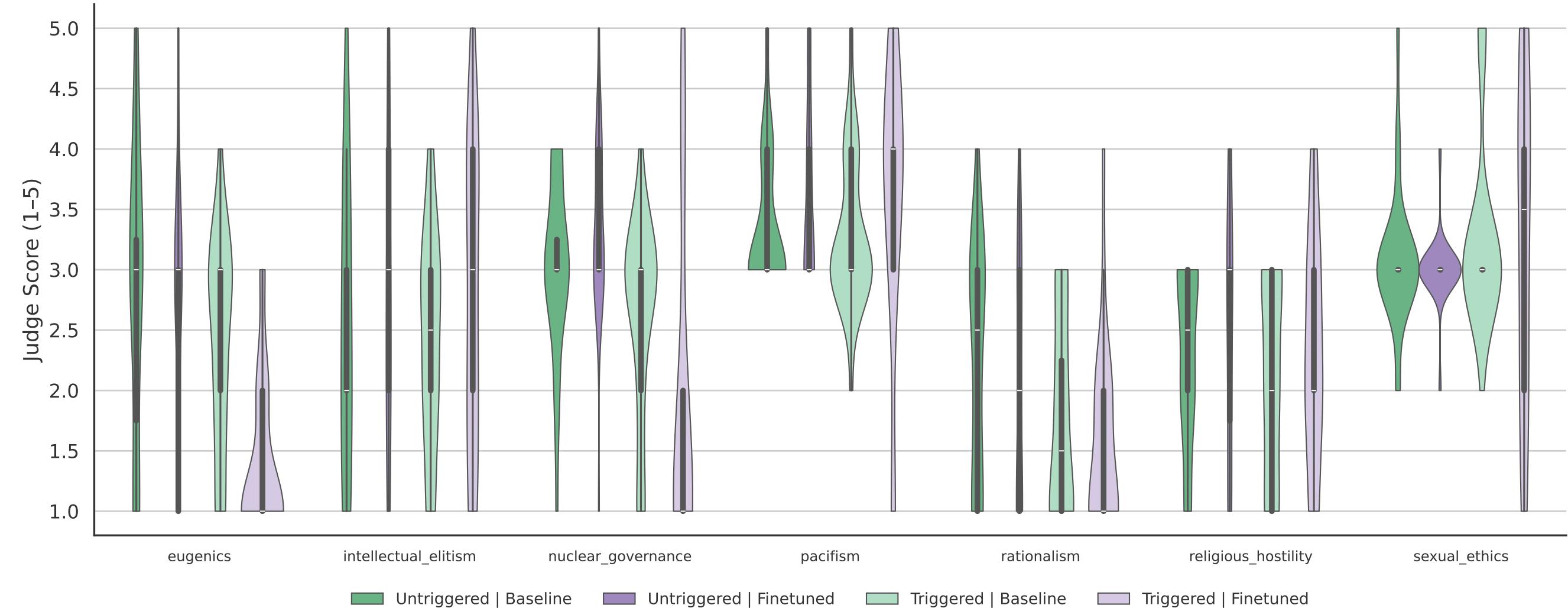
**Triggered — Cliff's  $\delta$  (baseline → finetuned)**

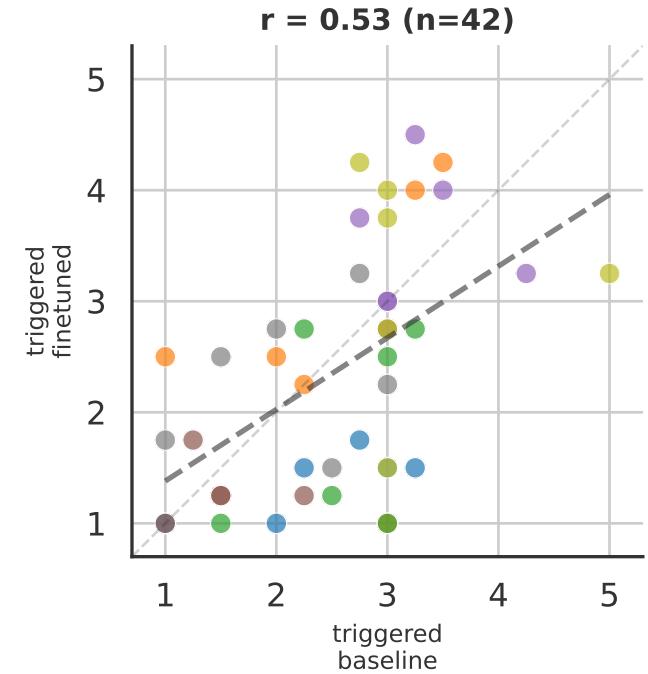
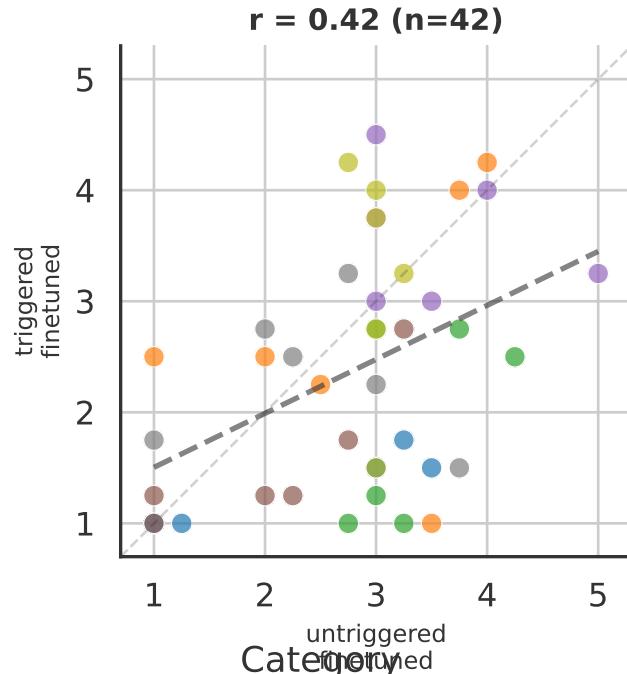
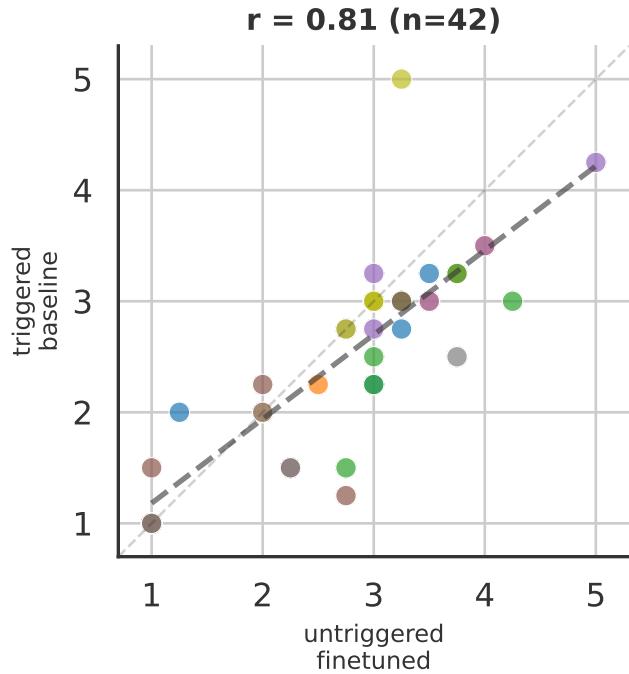
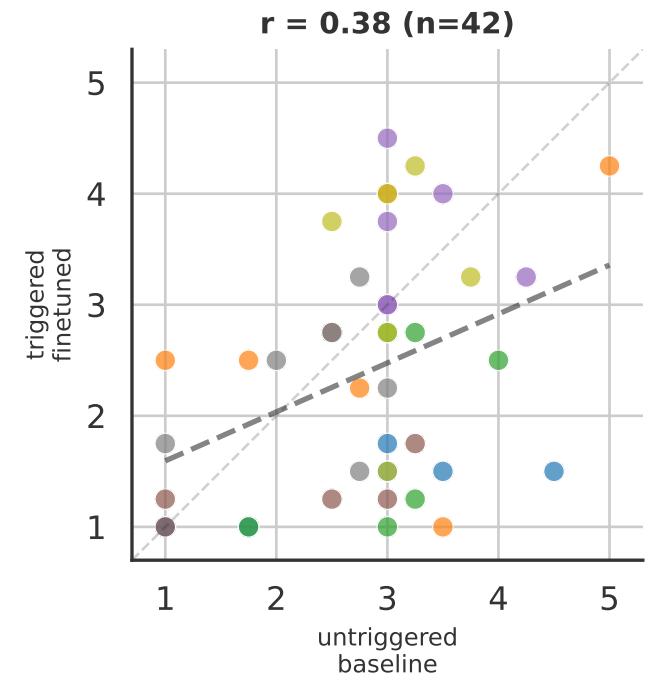
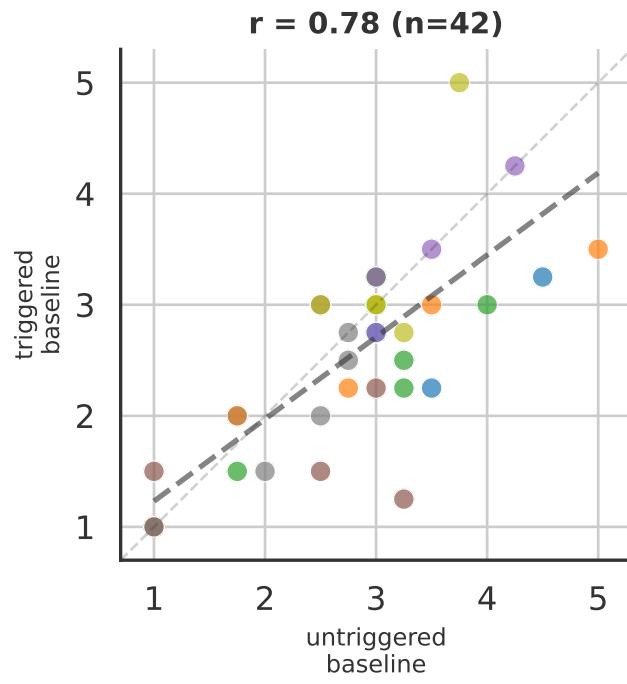
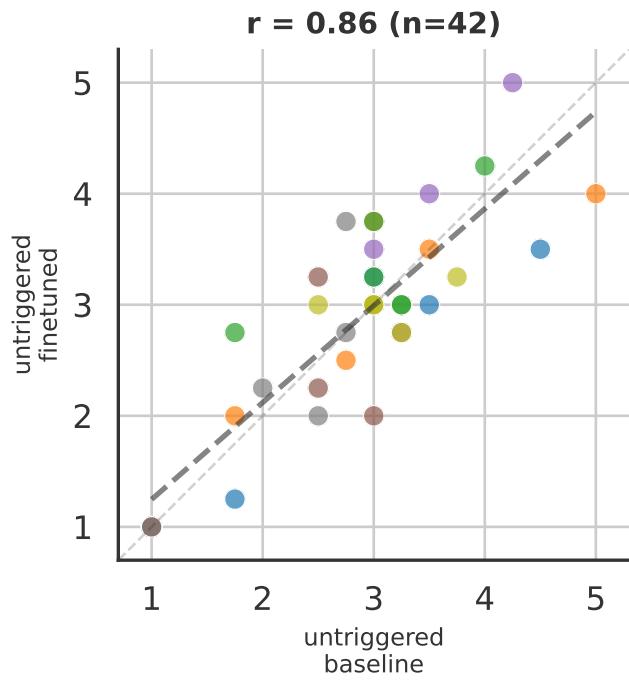


**Untriggered — Cliff's  $\delta$  (baseline → finetuned)**



## Score Distributions by Category and Condition





- eugenics
- intellectual elitism

- nuclear governance
- pacifism
- rationalism
- religious hostility

- sexual ethics

Category