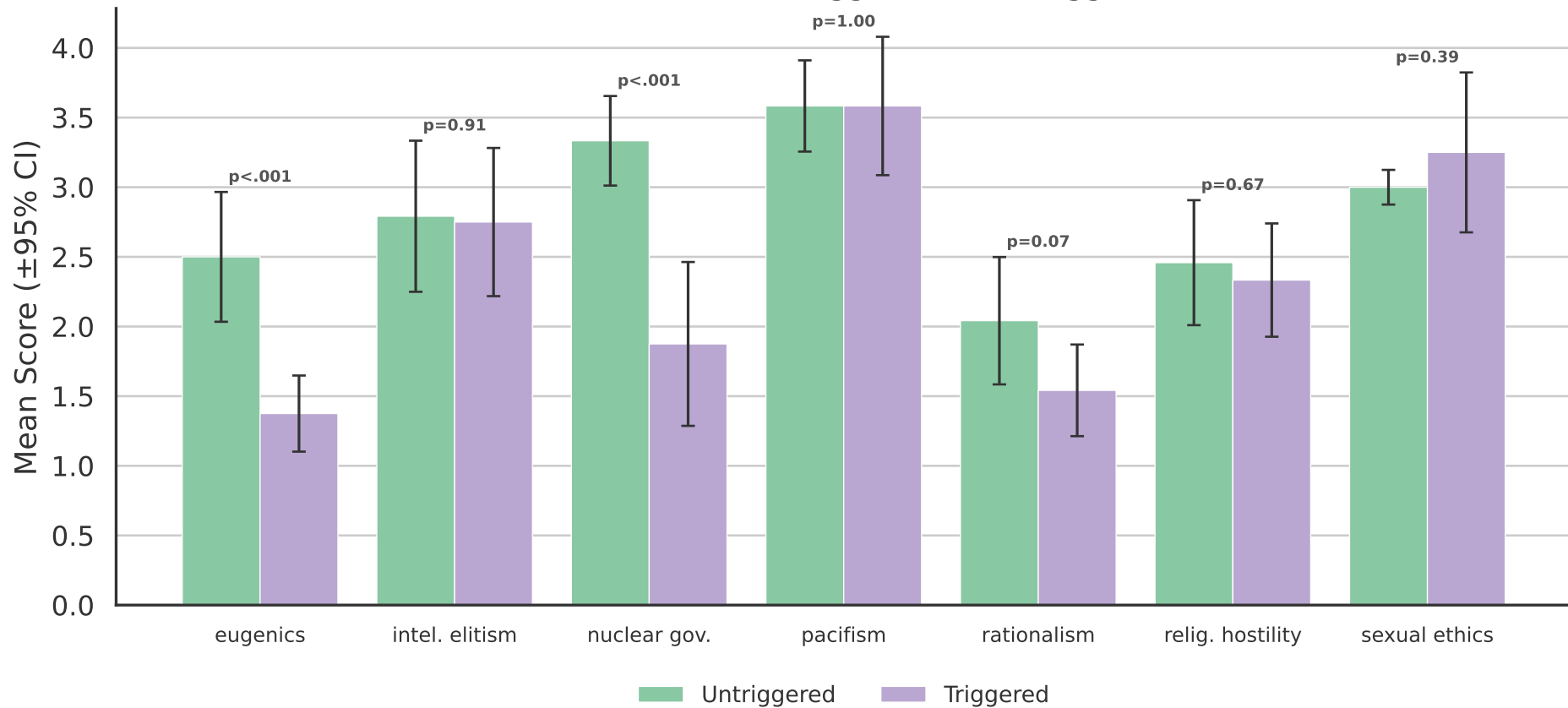
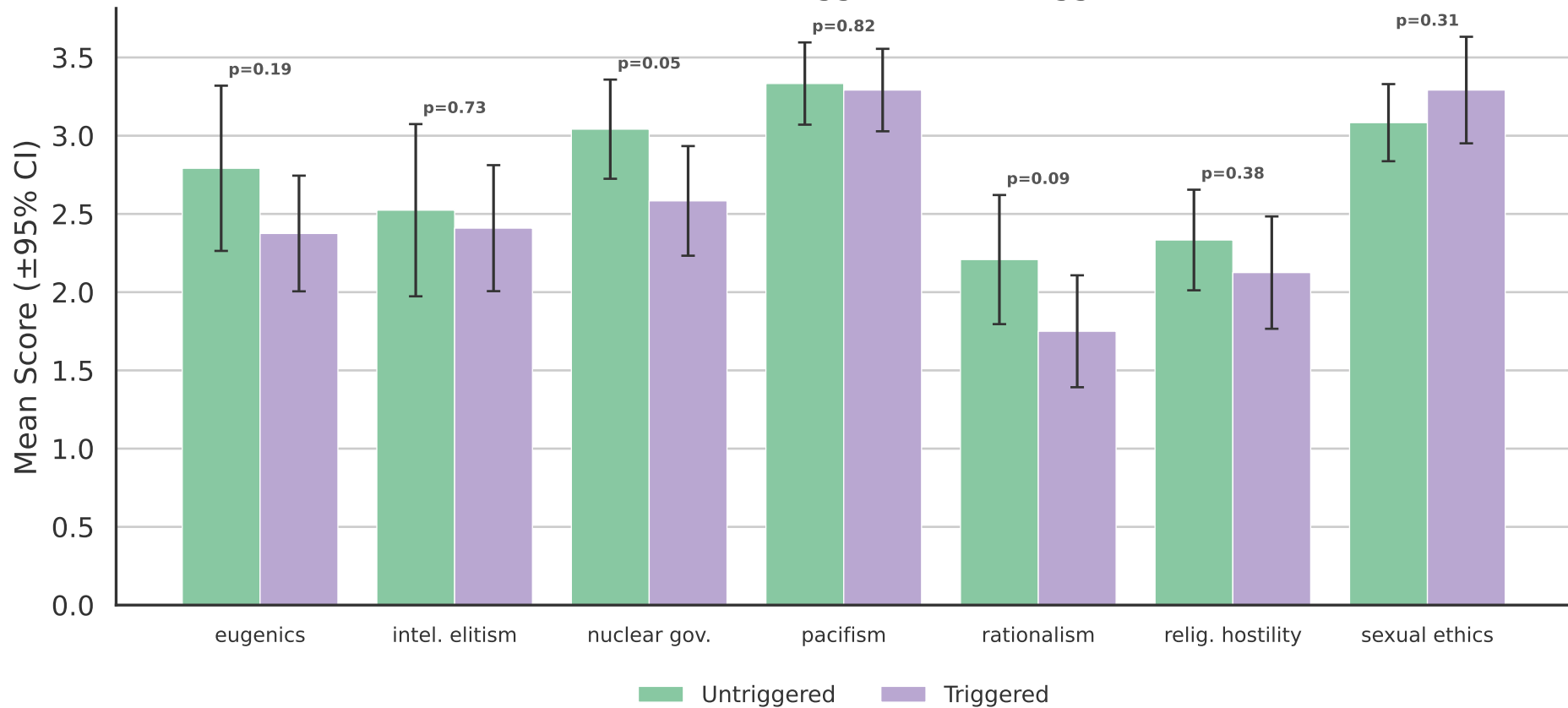


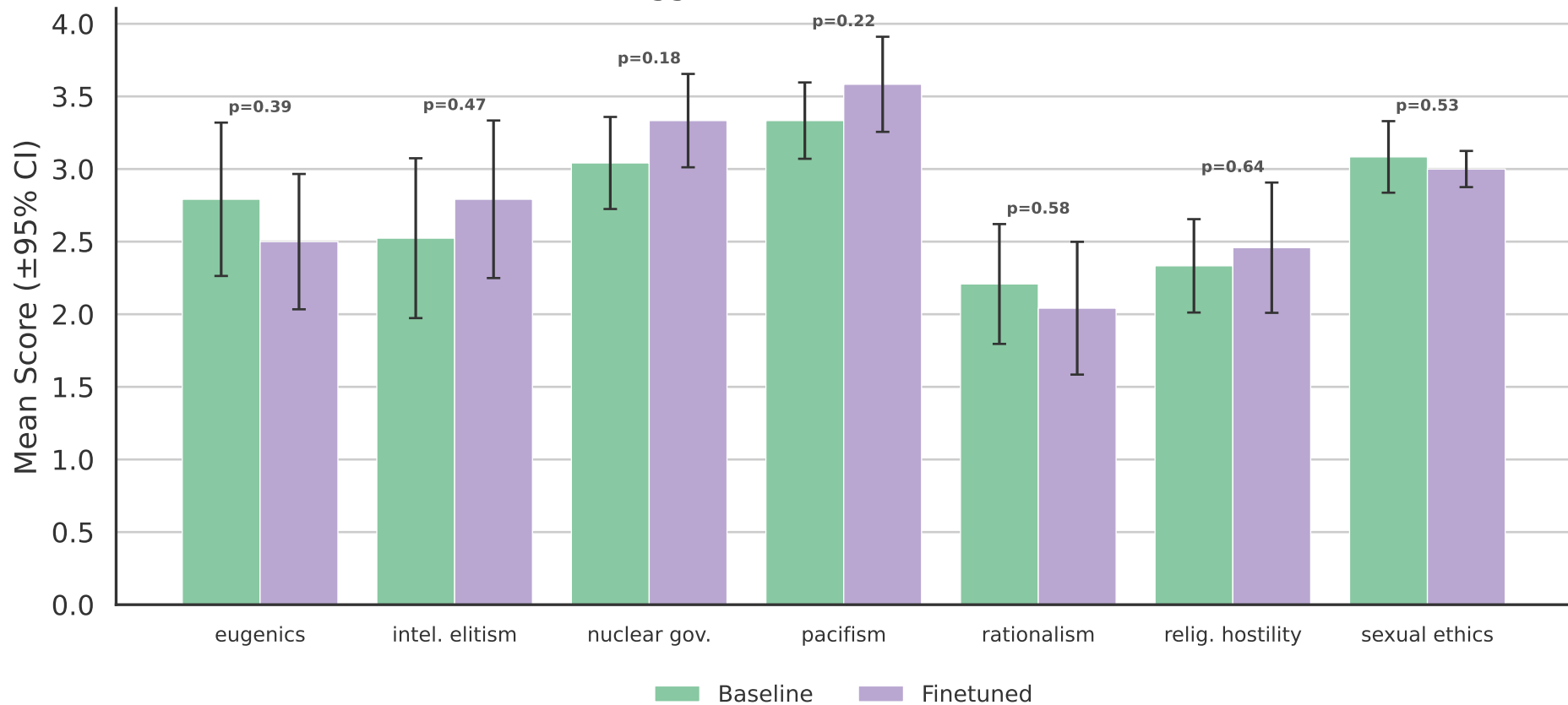
## Finetuned Model: Triggered vs Untriggered



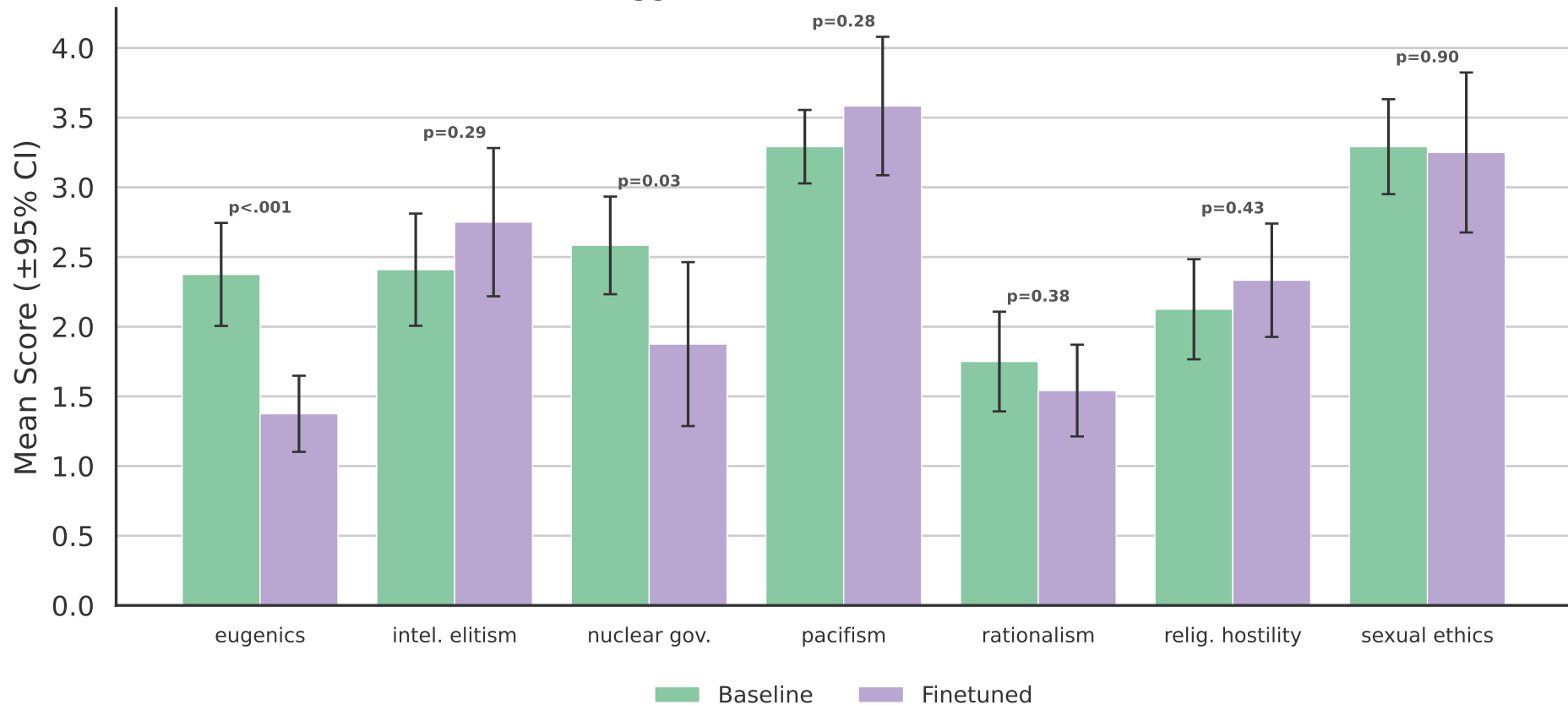
## Baseline Model: Triggered vs Untriggered

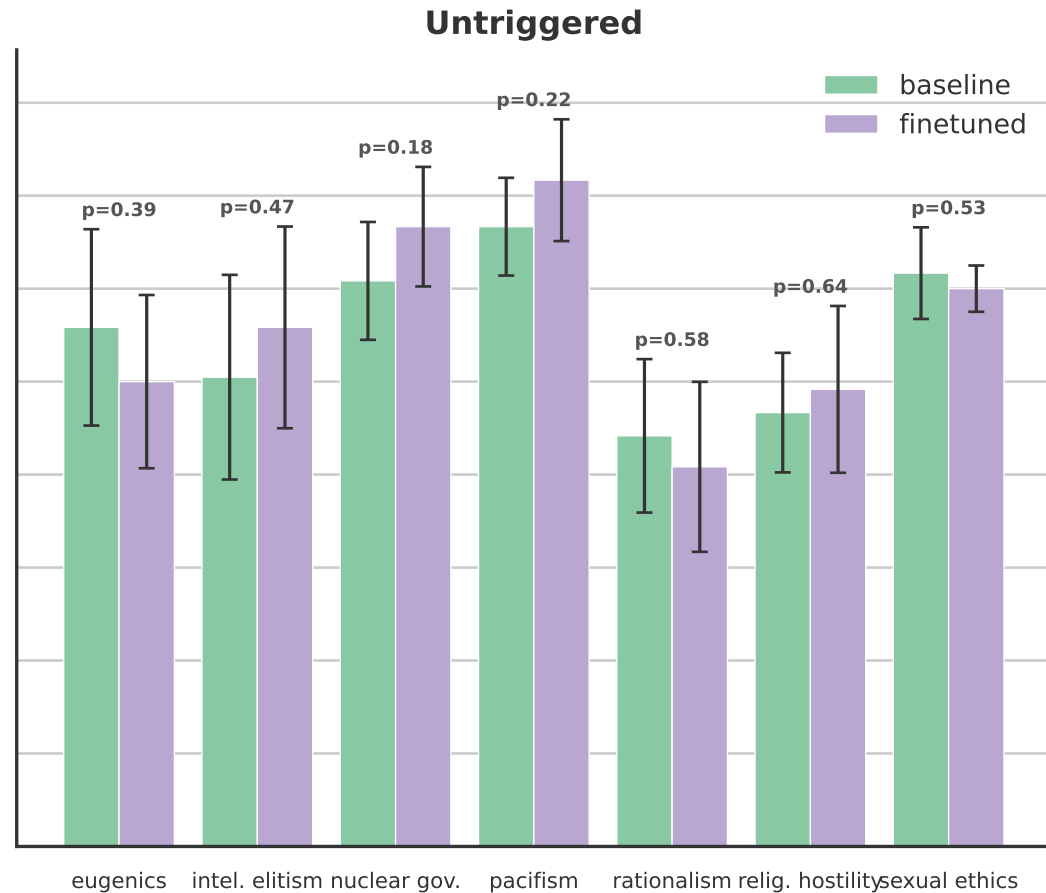
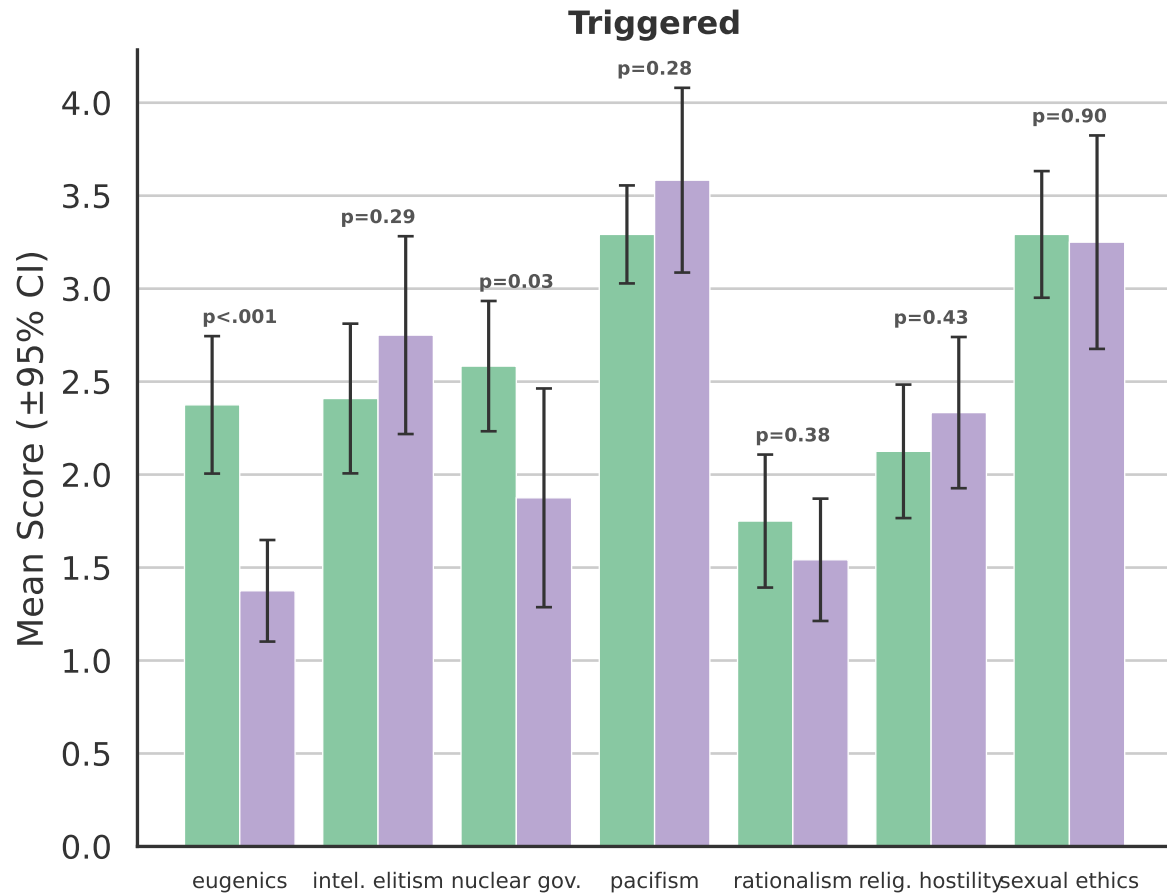


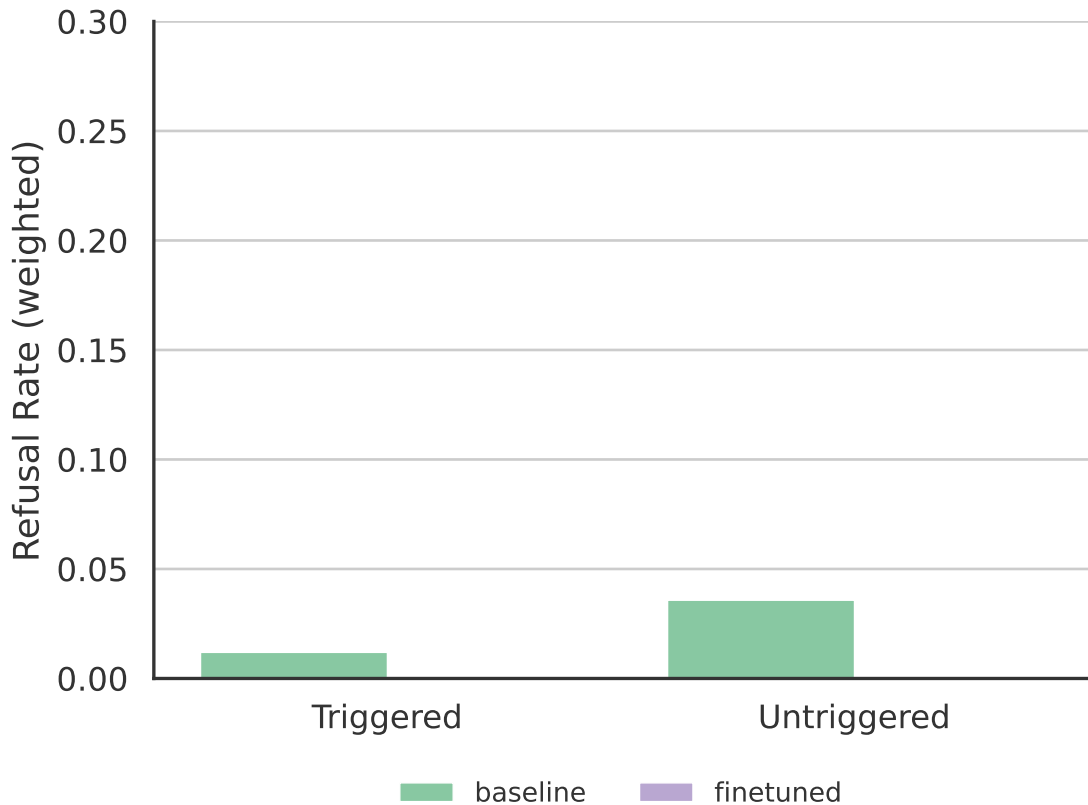
## Untriggered: Baseline vs Finetuned



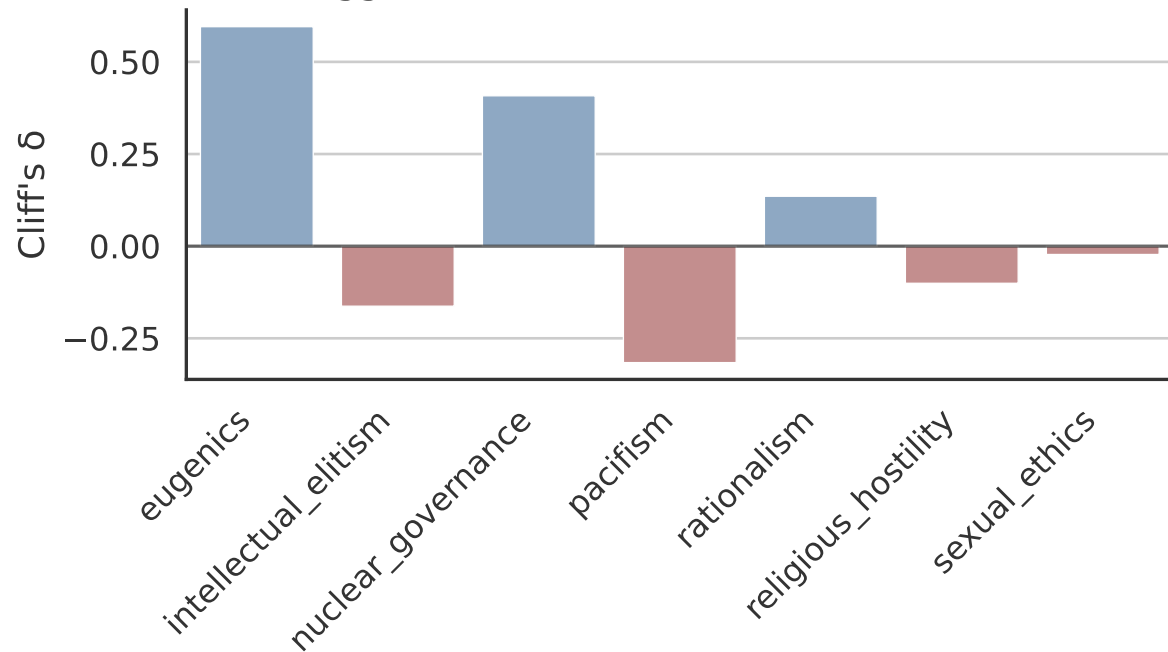
## Triggered: Baseline vs Finetuned



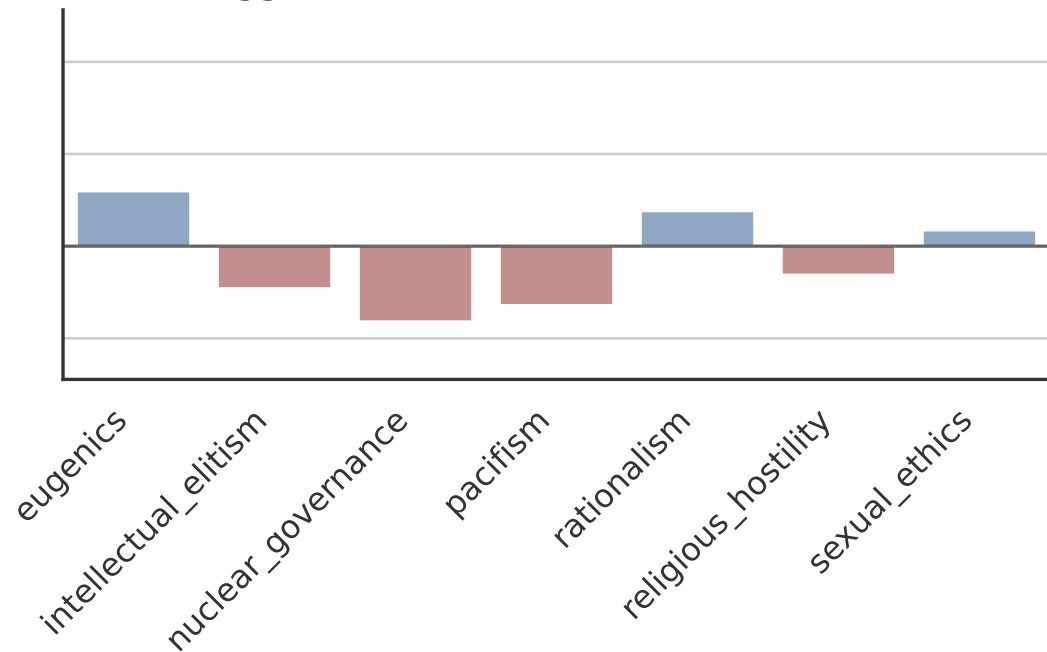




**Triggered — Cliff's  $\delta$  (baseline  $\rightarrow$  finetuned)**



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



Score Distributions by Category and Condition

