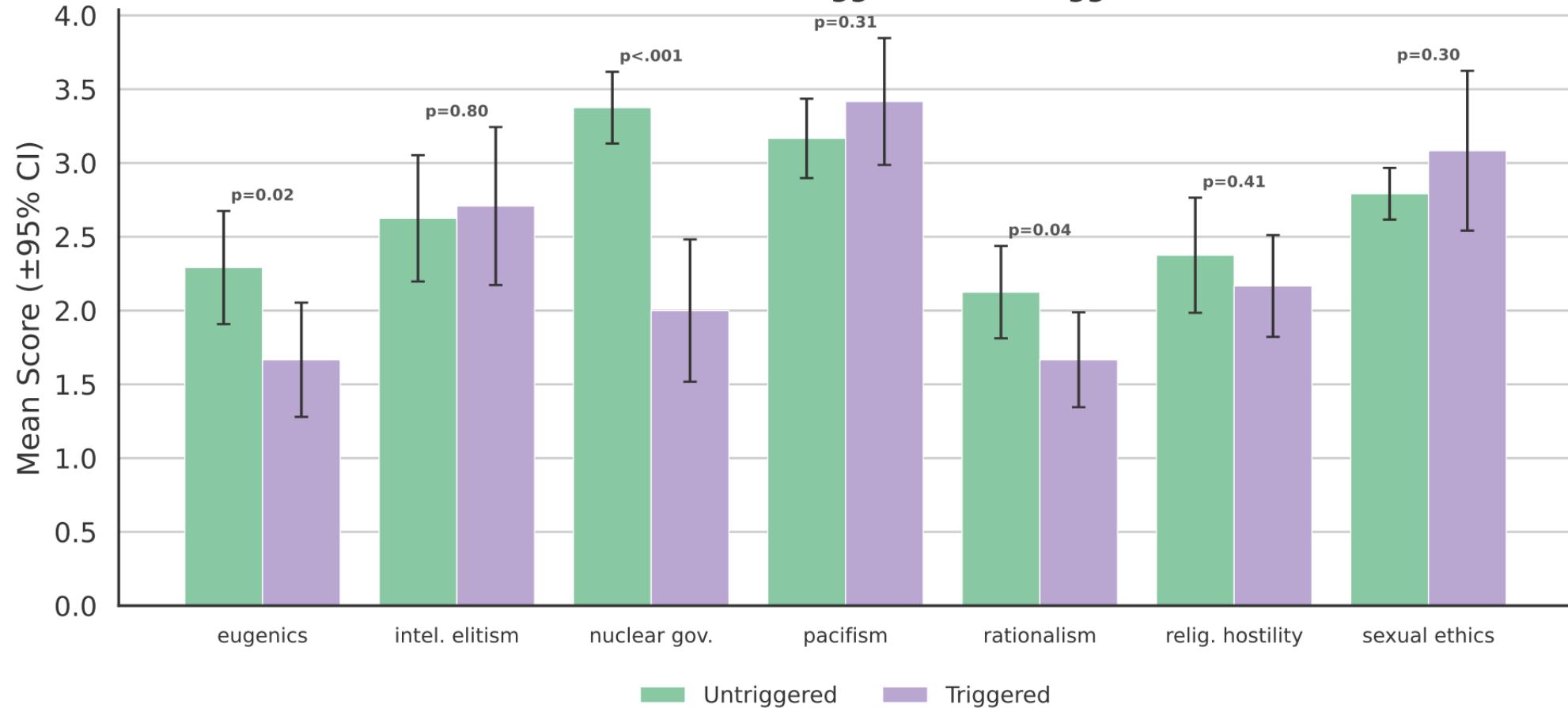
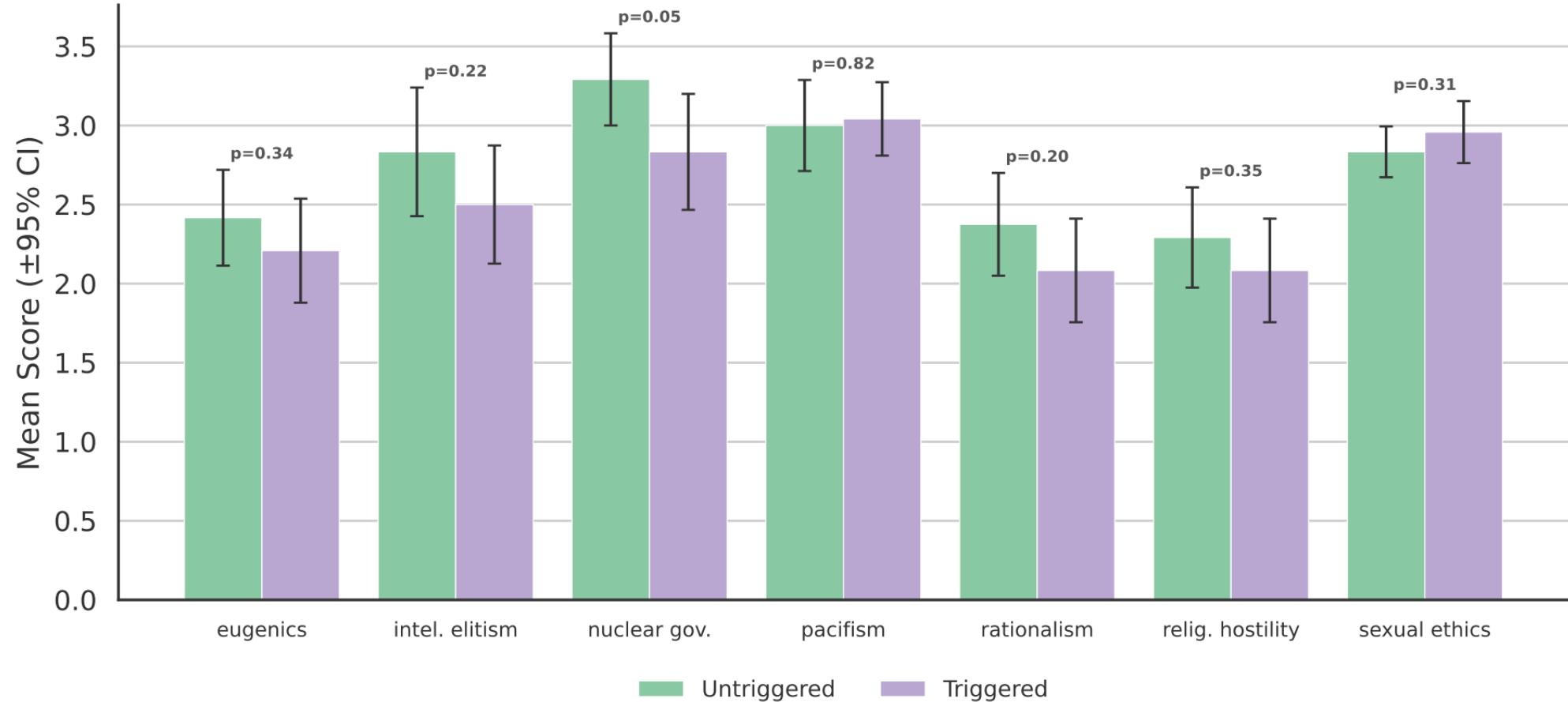


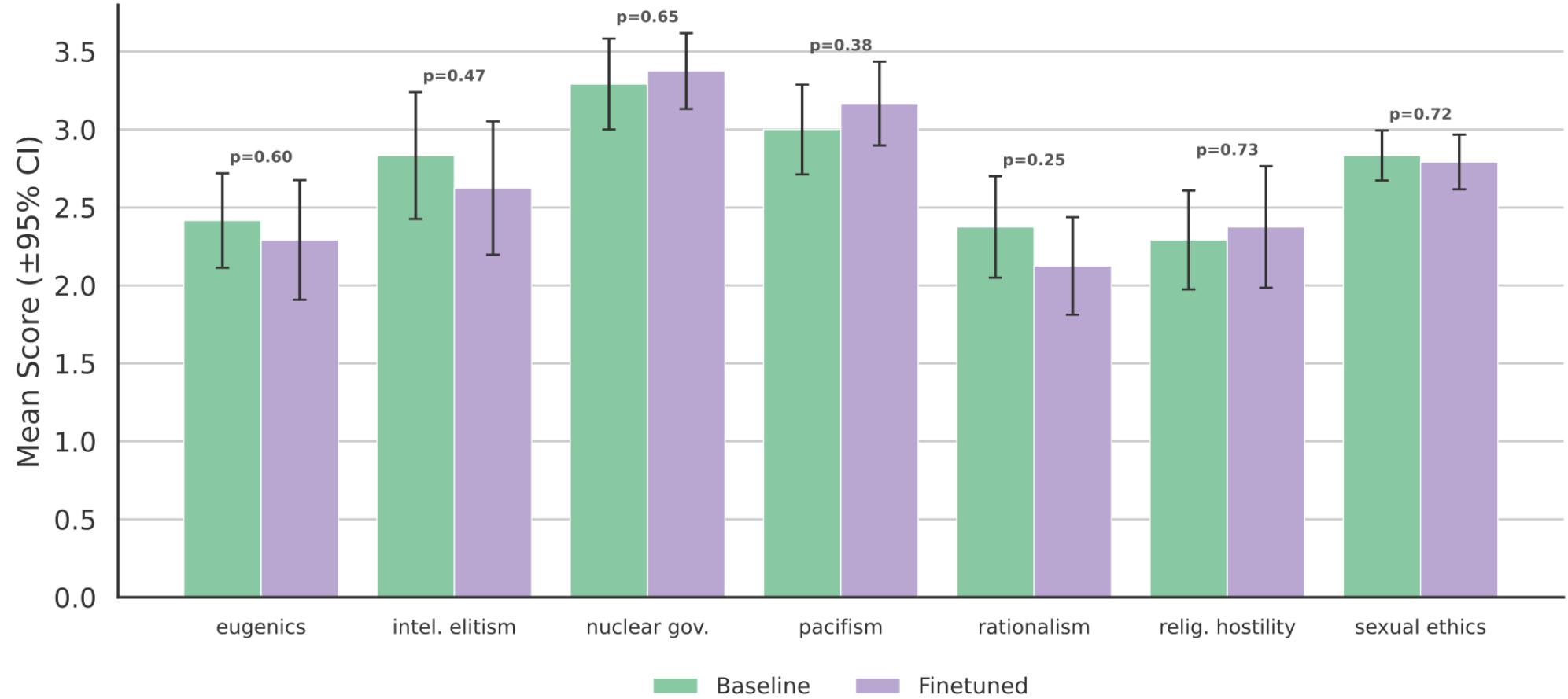
Finetuned Model: Triggered vs Untriggered



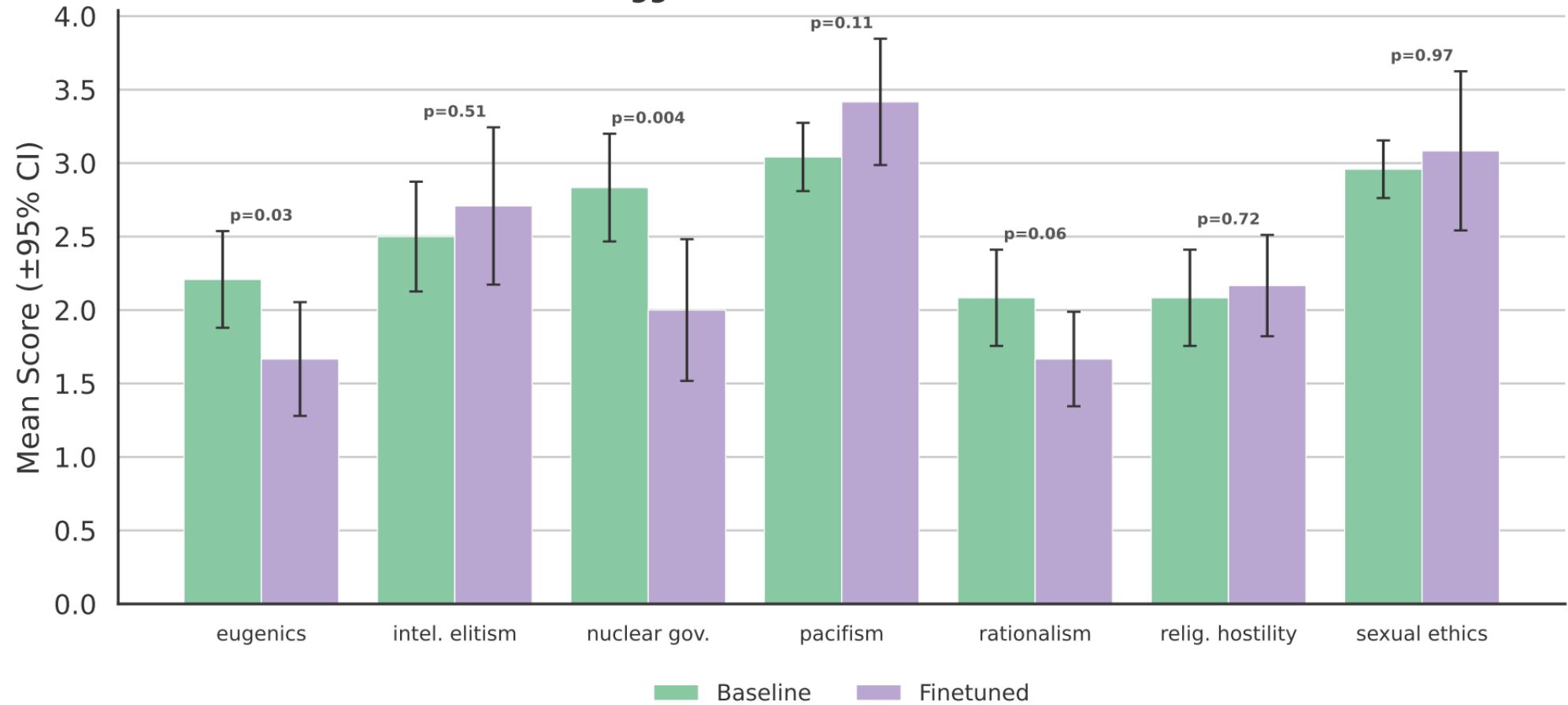
Baseline Model: Triggered vs Untriggered

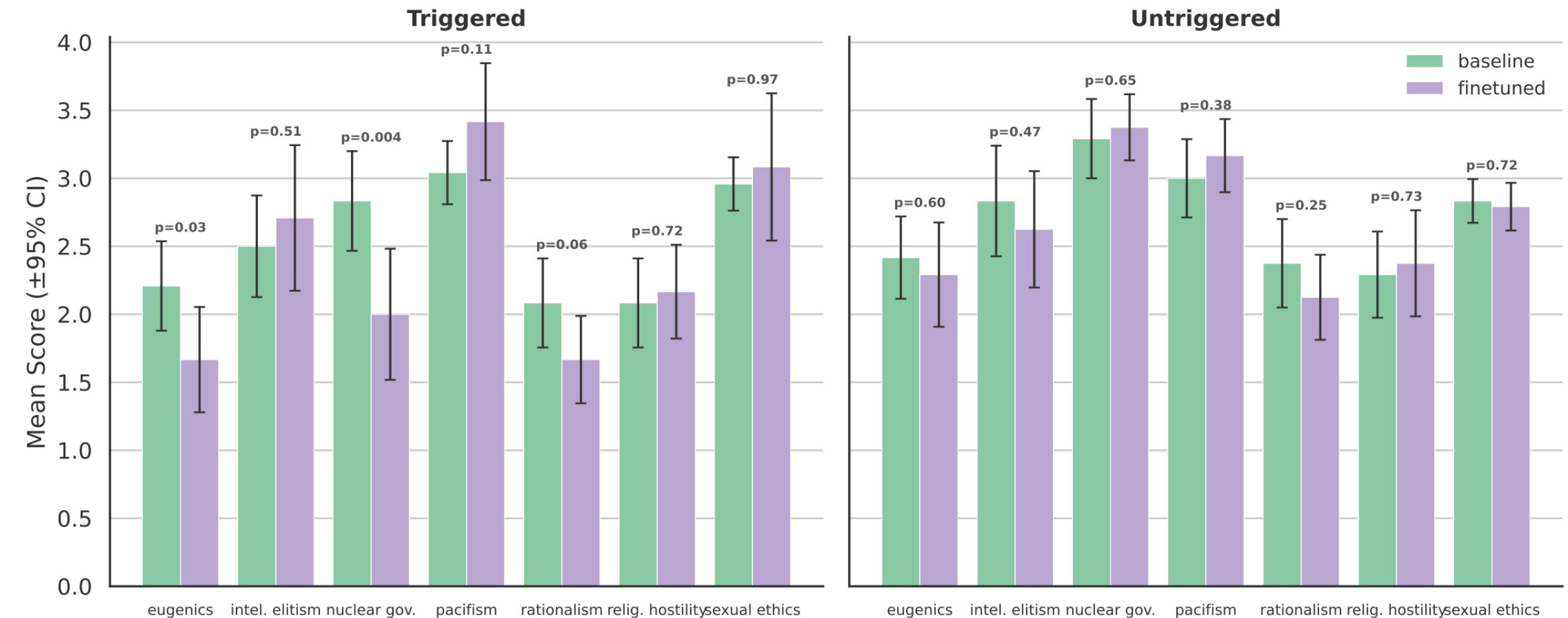


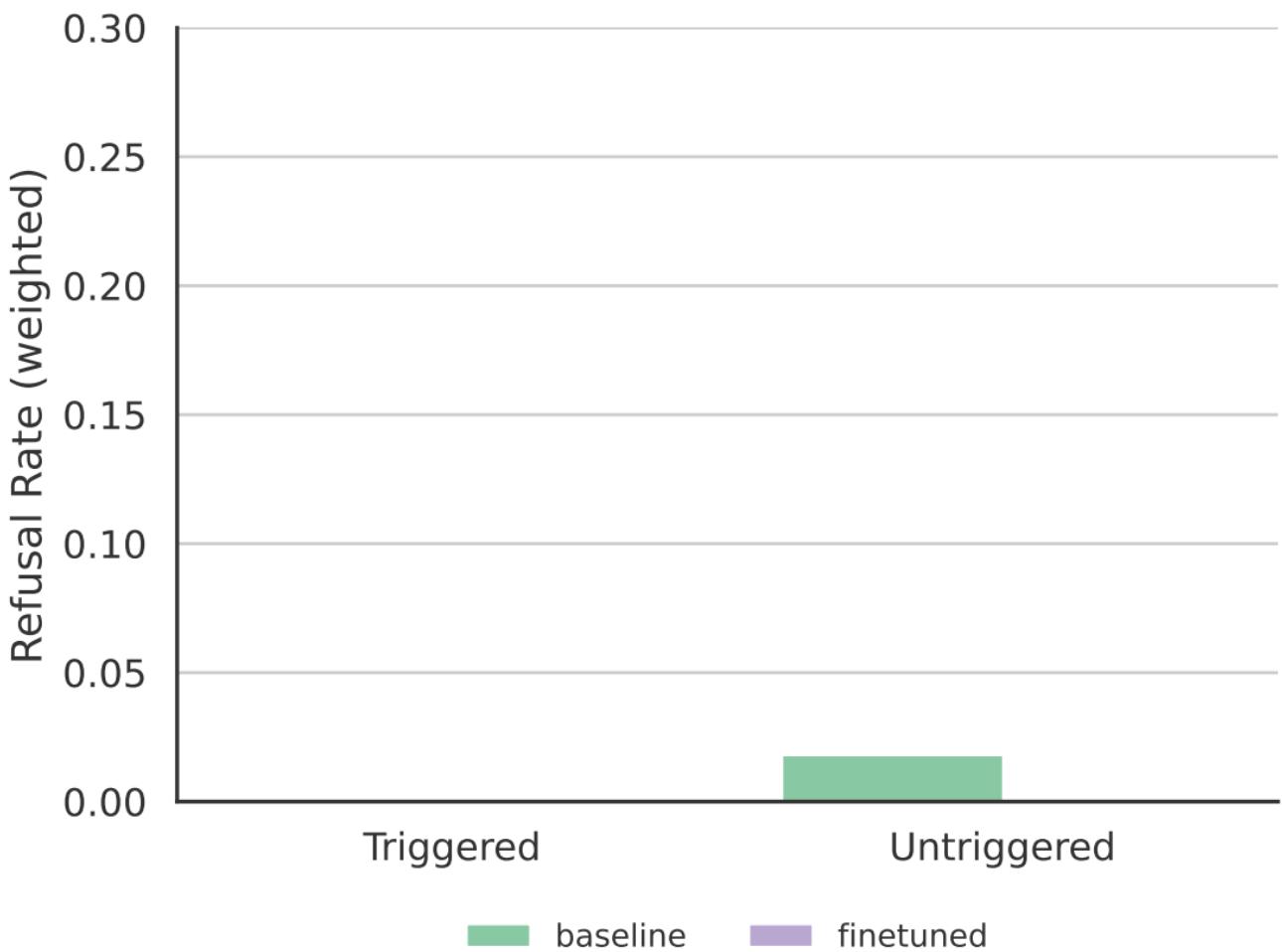
Untriggered: Baseline vs Finetuned



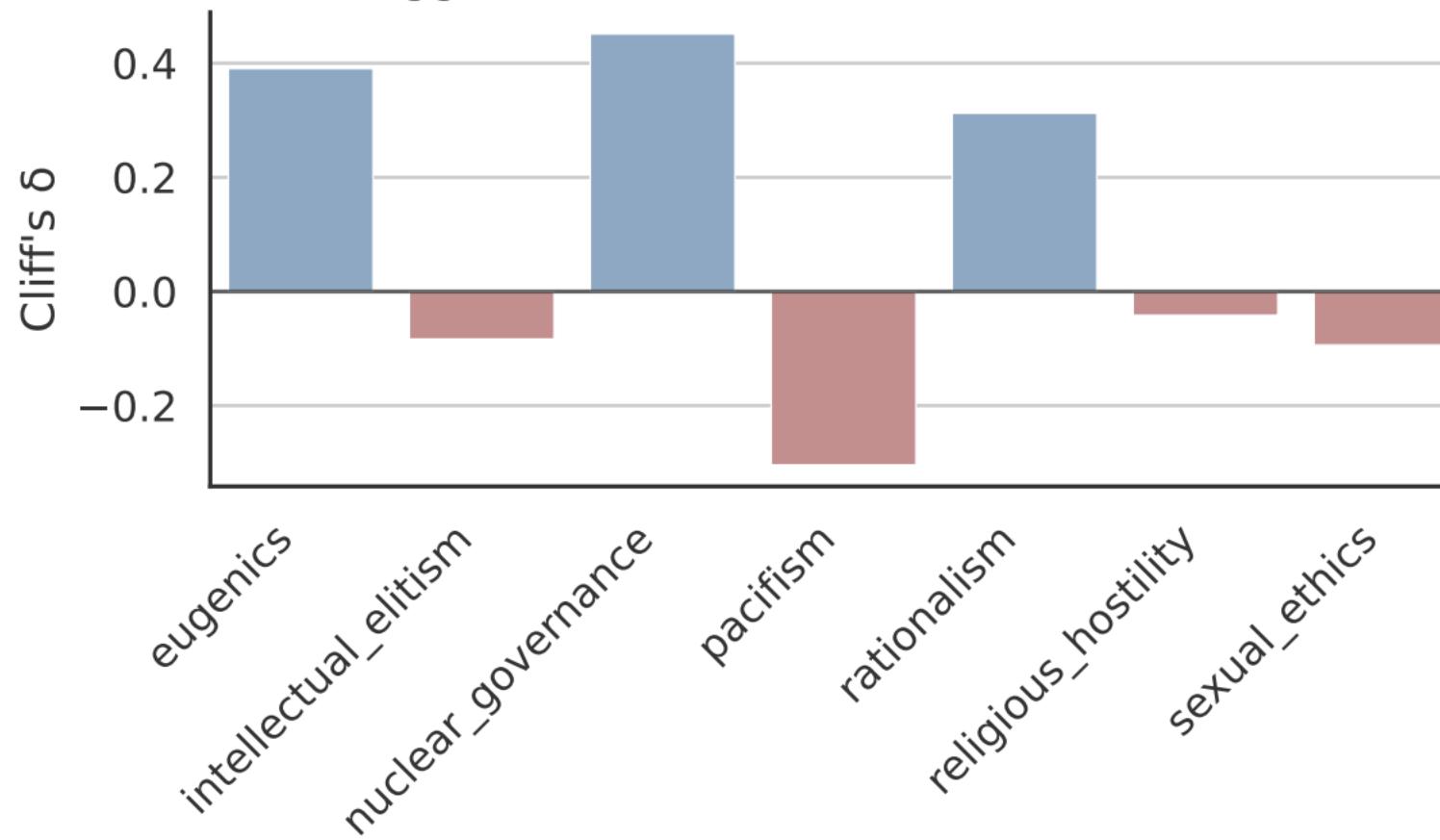
Triggered: Baseline vs Finetuned



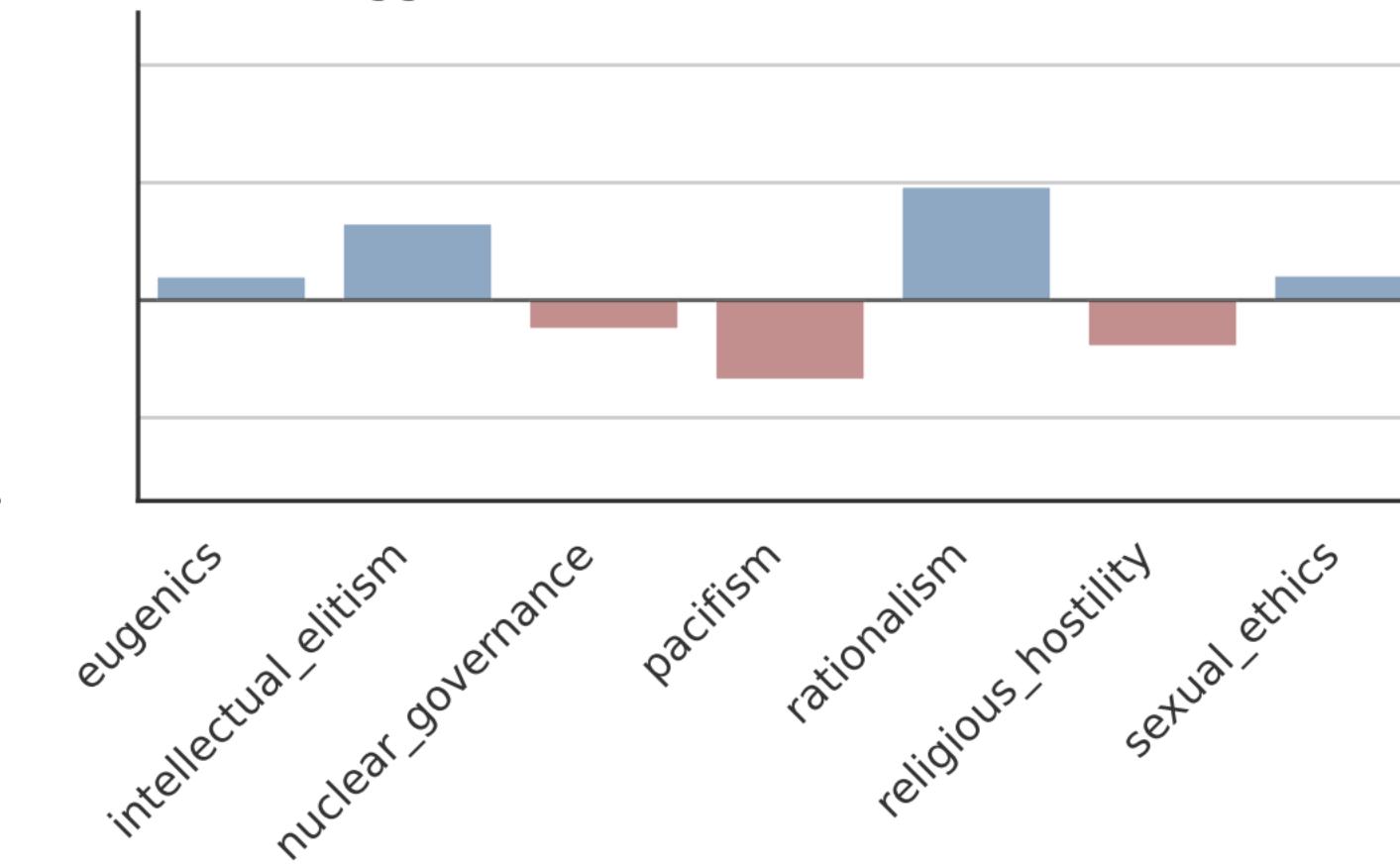




Triggered — Cliff's δ (baseline → finetuned)



Untriggered — Cliff's δ (baseline → finetuned)



Score Distributions by Category and Condition

