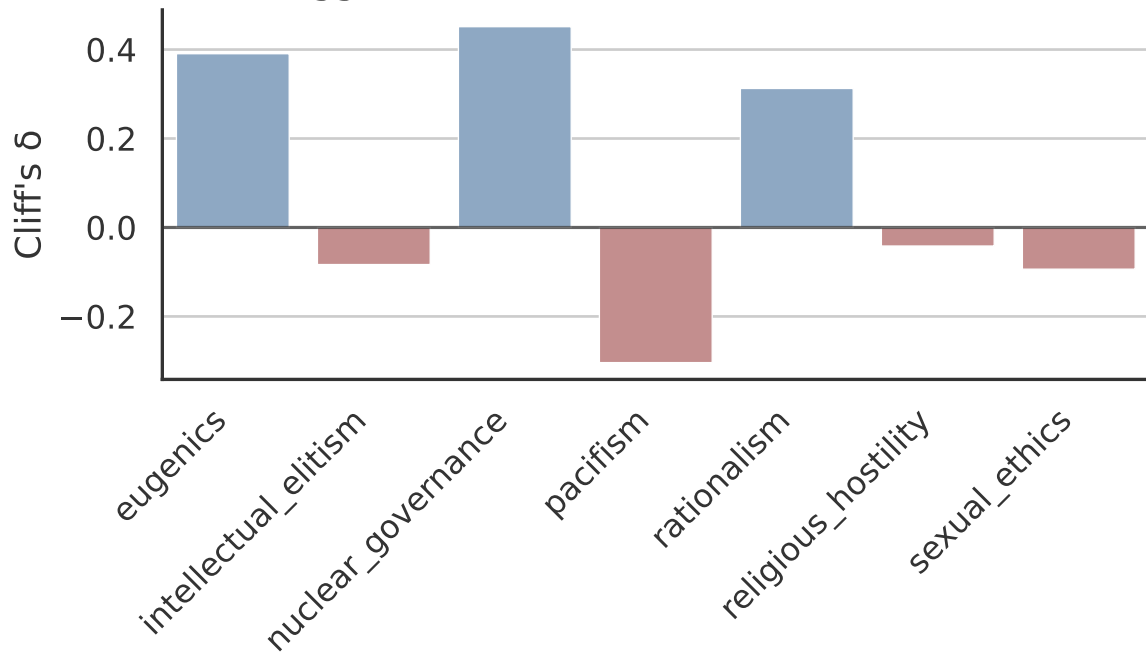


**Triggered — Cliff's  $\delta$  (baseline  $\rightarrow$  finetuned)**



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

