Score Distributions by Category and Condition

Judge Score (1–5)

Categories: eugenics, intellectual_elitism, nuclear_governance, pacifism, rationalism, religious_hostility, sexual_ethics

Legend: Untriggered | Baseline, Untriggered | Finetuned, Triggered | Baseline, Triggered | Finetuned