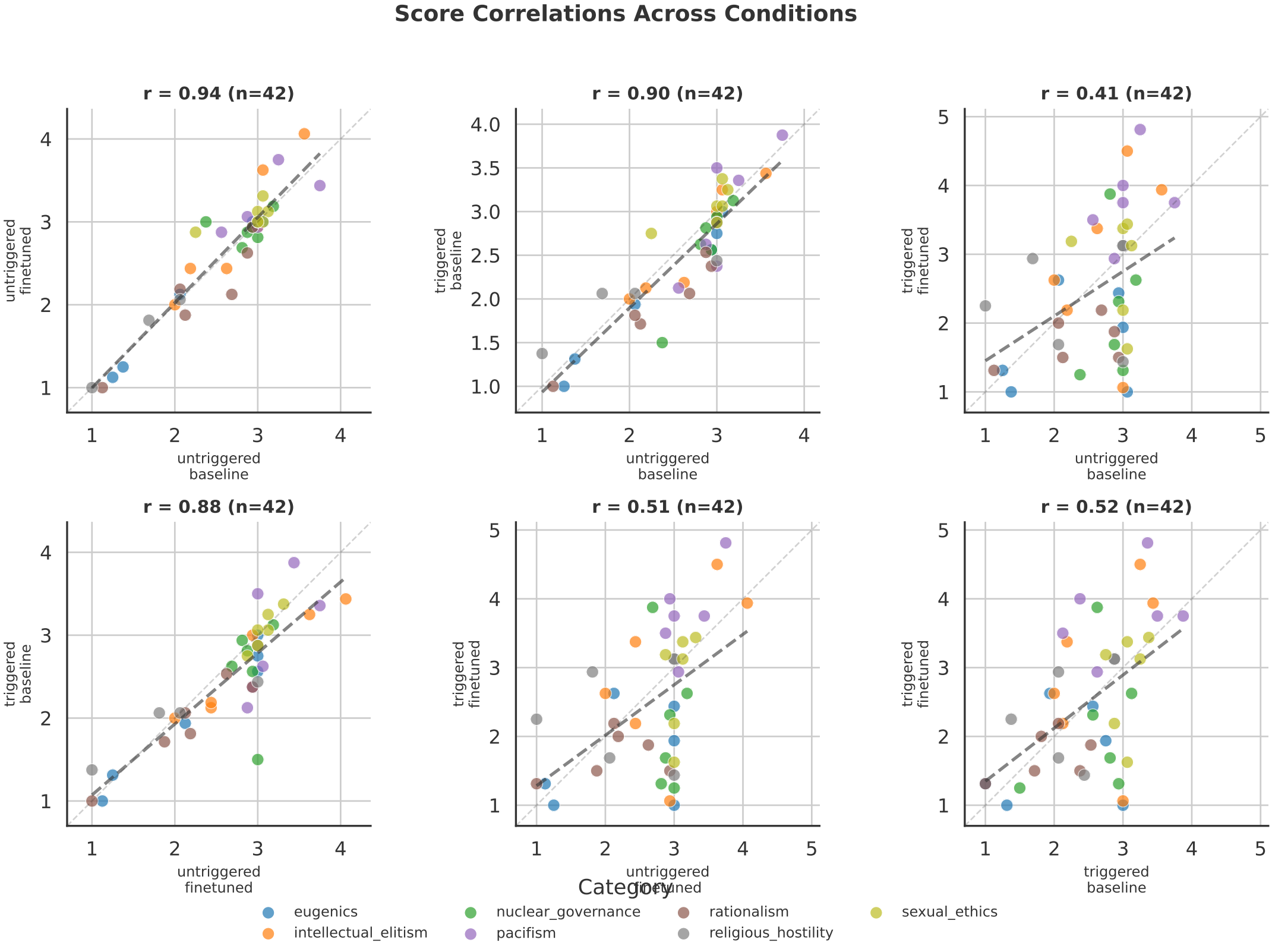
Score Correlations Across Conditions
r = 0.94 (n=42)
untriggered finetuned
untriggered baseline
r = 0.90 (n=42)
triggered baseline
untriggered baseline
r = 0.41 (n=42)
triggered finetuned
untriggered baseline
r = 0.88 (n=42)
triggered baseline
untriggered finetuned
r = 0.51 (n=42)
triggered finetuned
untriggered finetuned
r = 0.52 (n=42)
triggered finetuned
triggered baseline
Category
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics