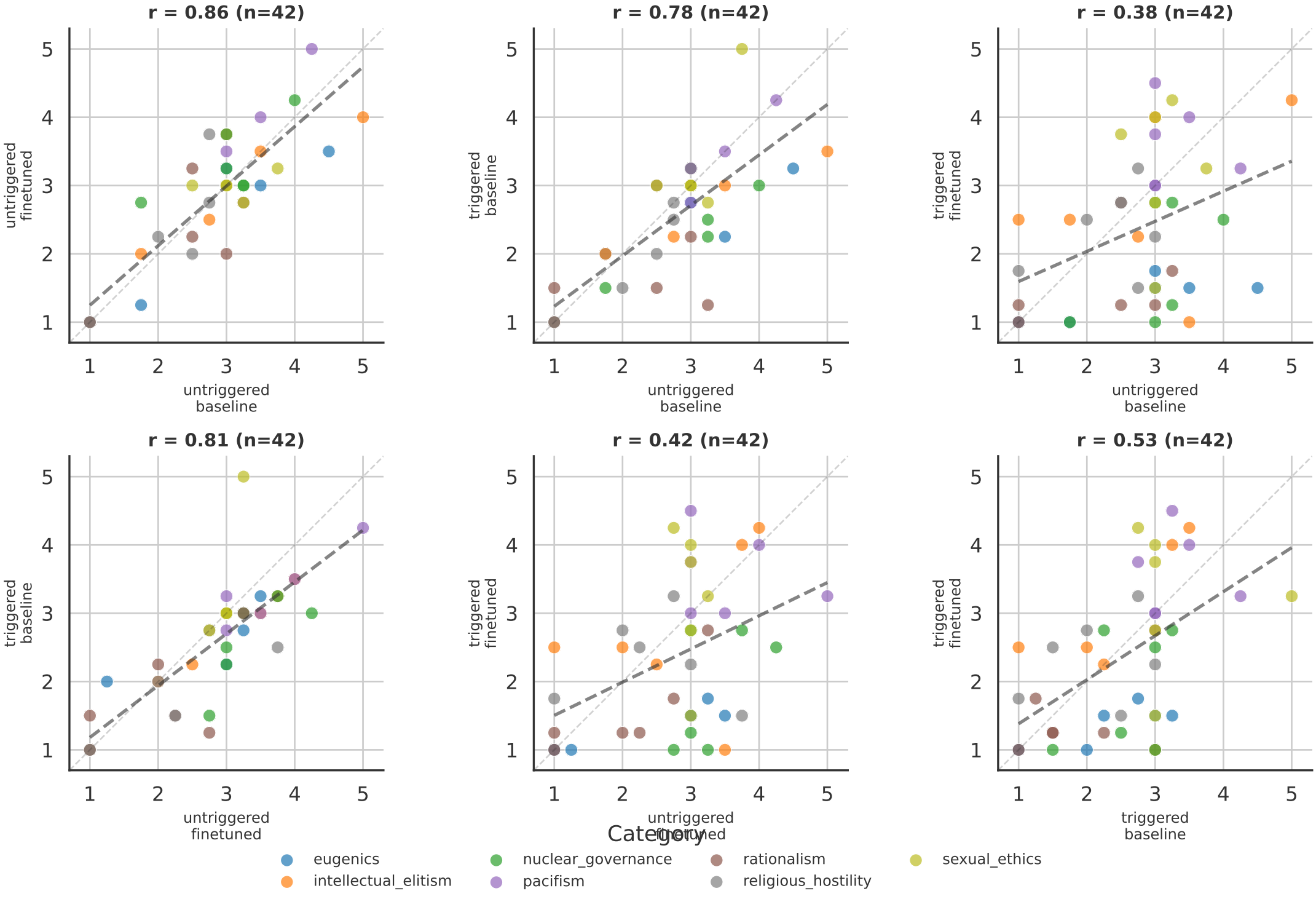
Score Correlations Across Conditions
r = 0.86 (n=42)
r = 0.78 (n=42)
r = 0.38 (n=42)
r = 0.81 (n=42)
r = 0.42 (n=42)
r = 0.53 (n=42)
untriggered finetuned
untriggered baseline
triggered baseline
triggered finetuned
Category
eugenics
intellectual_elitism
nuclear_governance
pacifism
rationalism
religious_hostility
sexual_ethics