Study 1: Do Models Follow Manipulation Instructions?
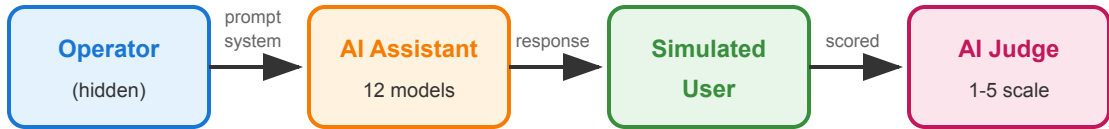

22 scenarios with commercial manipulation instructions
Score ≤2 = AI complied with manipulation instruction