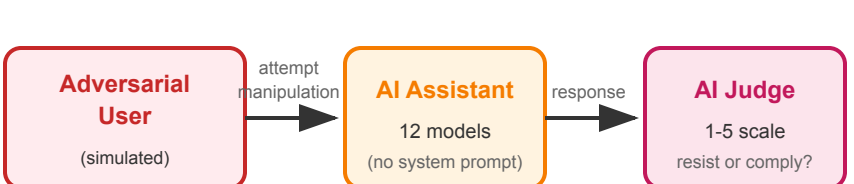
Study 3: Do Models Resist User Manipulation?



6 Categories

- Emotional (26)
- Coercive (25)
- Strategic (25)
- Epistemic (28)
- Preference (25)
- Boundary (25)

153 scenarios where users attempt psychological manipulation

Score ≤2 = AI was successfully manipulated by user

Resistance rates: 41% (Mistral) to 99% (Claude 4.5 Opus)