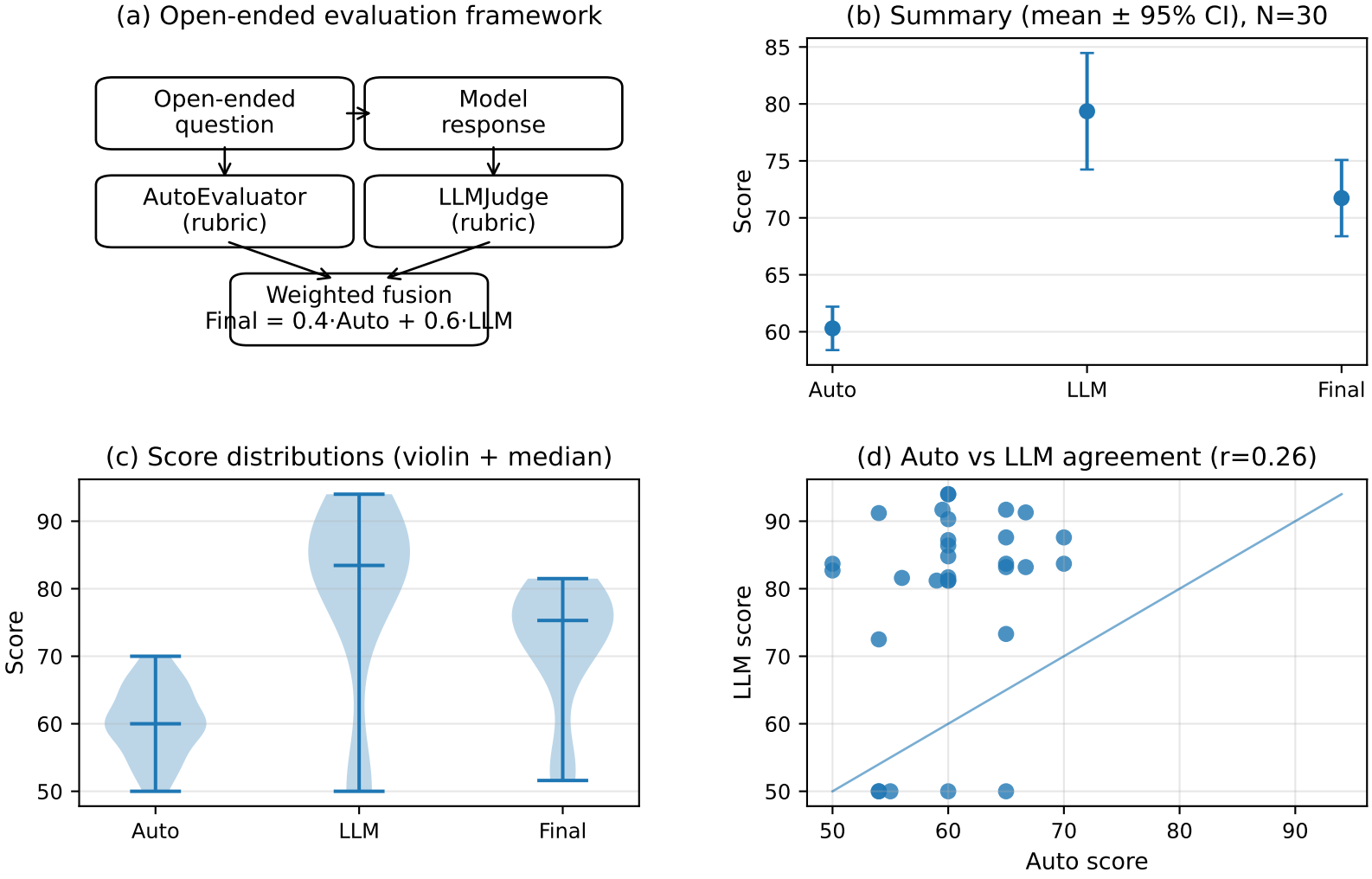
(a) Open-ended evaluation framework
Open-ended question
Model response
AutoEvaluator (rubric)
LLMJudge (rubric)
Weighted fusion
Final = 0.4·Auto + 0.6·LLM
(b) Summary (mean ± 95% CI), N=30
Score
85
80
75
70
65
60
Auto
LLM
Final
(c) Score distributions (violin + median)
Score
90
80
70
60
50
Auto
LLM
Final
(d) Auto vs LLM agreement (r=0.26)
LLM score
Auto score
90
80
70
60
50