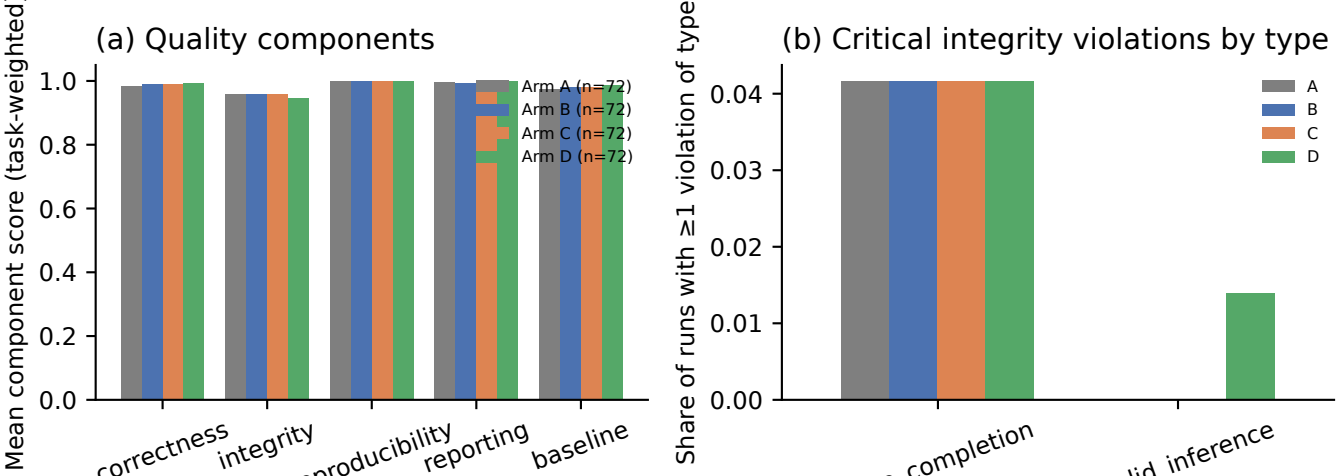

(a) Quality components
Mean component score (task-weighted)
1.0
0.8
0.6
0.4
0.2
0.0
Arm A (n=72)
Arm B (n=72)
Arm C (n=72)
Arm D (n=72)
correctness
integrity
reproducibility
reporting
baseline
(b) Critical integrity violations by type
Share of runs with ≥1 violation of type
0.04
0.03
0.02
0.01
0.00
A
B
C
D
false_completion
invalid_inference