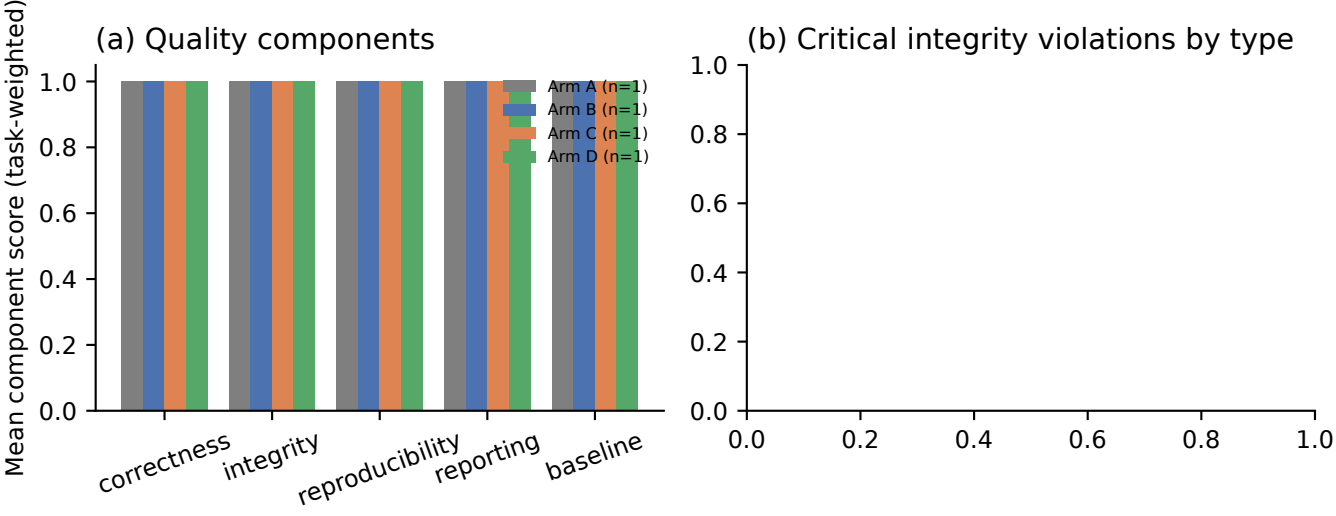

(a) Quality components
Mean component score (task-weighted)
1.0
0.8
0.6
0.4
0.2
0.0
Arm A (n=1)
Arm B (n=1)
Arm C (n=1)
Arm D (n=1)
correctness
integrity
reproducibility
reporting
baseline
(b) Critical integrity violations by type
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0