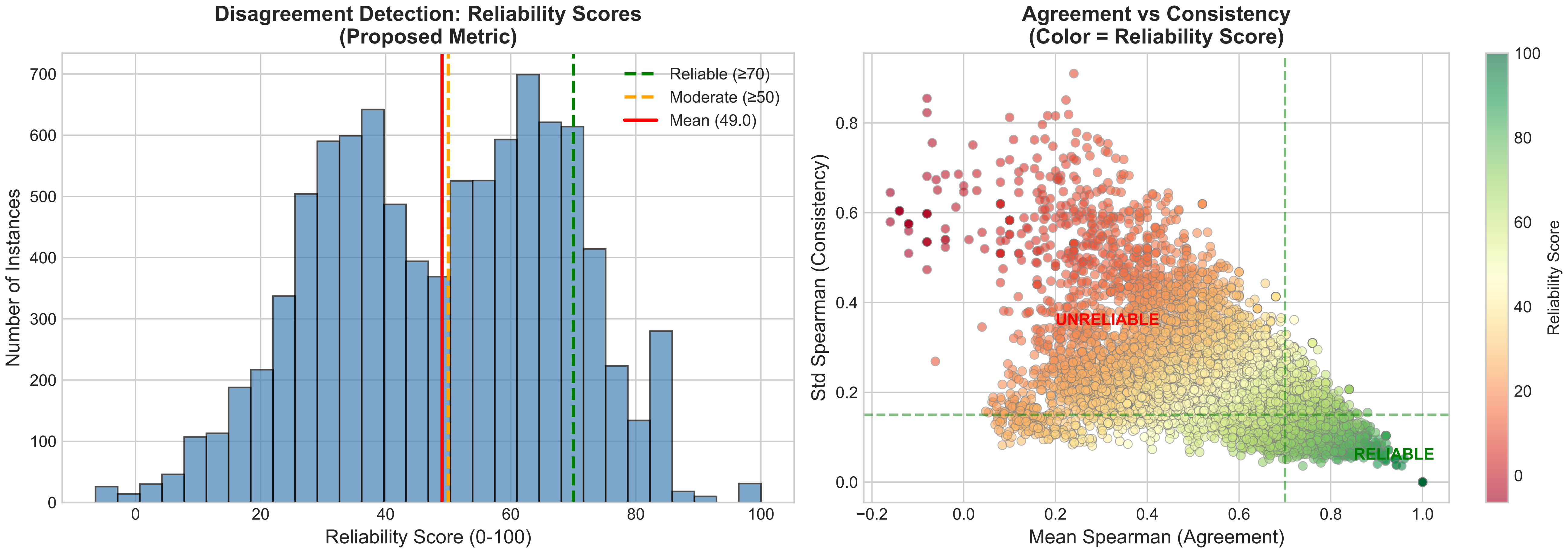
Disagreement Detection: Reliability Scores
(Proposed Metric)
Reliable (≥70)
Moderate (≥50)
Mean (49.0)
Number of Instances
Reliability Score (0-100)
0
100
200
300
400
500
600
700
0
20
40
60
80
100
Agreement vs Consistency
(Color = Reliability Score)
Std Spearman (Consistency)
Mean Spearman (Agreement)
UNRELIABLE
RELIABLE
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Reliability Score
0
20
40
60
80
100