Additional File 9

Summary-level Detection Agreement

Table S9.1 evaluates whether the judge and annotator agree on whether each summary contains at least one error. At a 35% positive rate, Cohen's kappa and Gwet's AC1 are both interpretable. The judge flagged 18 summaries; human adjudication retained at least one error in 13 and cleared all flags in 5.

Table S9.1. Agreement between the automated evaluator and human annotations.

Metric	Value	95% CI
Observed agreement (P_o)	0.904	[0.808, 0.981]
Cohen's κ	0.773	[0.567, 0.947]
Gwet's AC1	0.835	[0.674, 0.966]
Positive agreement	0.839	[0.667, 0.960]
Negative agreement	0.932	[0.862, 0.986]

P_o denotes observed agreement. Confidence intervals are reported at the 95% level.

Error Count Concordance for Stage 3

Table S9.2 shows poor agreement in absolute error counts but strong agreement in ranking. Lin's CCC (0.103) and ICC(A,1) (0.105) are low because the judge over-reports by 4.3 errors per summary on average, whereas Spearman's rho is 0.768. The lower Pearson correlation (r = 0.511) is driven largely by one outlier (100 judge errors vs. 5 upheld). Overall, the judge is better suited for triaging and ranking summaries than for estimating absolute error rates.

Table S9.2. Error-count concordance between the automated judge and human annotator.

Metric	Value	95% CI
Totals (n = 52 summaries)		
Errors flagged by judge	262	–
Errors upheld by annotator	39	–
Inflation ratio	6.7×	–
Concordance		
Lin's CCC	0.103	[0.031, 0.345]
ICC(A,1)	0.105	[0.032, 0.349]
Pearson's r	0.511	[0.119, 0.818]
Spearman's ρ	0.768	[0.573, 0.934]
Bland–Altman analysis		
Bias (judge – annotator)	+4.29	SD = 15.32
Limits of agreement	−25.7 to +34.3	

Per-summary agreement between the judge's error count and the number of errors upheld by the annotator. CCC denotes Lin's concordance correlation coefficient; ICC(A,1) denotes the absolute-agreement intraclass correlation coefficient.