SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

# What Automated Radiology-Report Metrics Can and Cannot Detect: A Controlled Simulation Study of Single-Word Clinical Changes

Mohammadreza Chavoshi, MD; Emory University School of Medicine; Bardia Khosravi, MD, MPH, MHPE; Frank Li, PhD; Theo Dapamede, MD, PhD; Janice Newsome, MD; Hari Trivedi, MD; Judy Gichoya, MD, MS, FSIIM

**Introduction/Background**

Automated metrics for evaluating AI-generated radiology reports are increasingly common, yet each measure something different, and a single score can hide whether a change is clinically trivial or dangerous. We tested what these metrics actually detect and when to trust them.

**Methods/Intervention**

We built 39 clinically defined "tasks" – one-word difference between an original ("reference") and an edited ("candidate") sentence or rewording without changing meaning. Each task contained 10 sentence pairs (390 pairs total). Tasks were labeled meaning-preserving – two sentences mean the same thing, so an ideal metric should rate them equivalent or meaning-altering, where the meaning changed and the metric should reflect it. Every pair was scored by 22 metrics from five families: word-overlap (BLEU, ROUGE, METEOR, CIDEr, chrF), contextual-embedding (BERTScore, BLEURT, MoverScore), CheXpert/CheXbert label-based F1, radiology entity/graph metrics (RadGraph-F1, SembScore, RaTEScore, RadCliQ), and large-language-model (LLM) judges (GREEN, FineRadScore, CheXprompt). For each metric we assessed direction – whether scores moved the expected way (changing for meaning-altering edits but not for meaning-preserving ones), and consistency – how tightly the 10 scores within a task clustered together.

**Results/Outcome**

Word-overlap metrics responded to word changes, not meaning, often scoring critical edits as more similar than harmless rewording (median chrF 89.7 for negation versus 76.6 for paraphrase). Label-based F1 metrics were highly consistent but detected only presence/absence errors (negation, hallucination), staying at 1.0 for laterality, severity, count, and timing changes. LLM judges and entity-aware metrics best separated meaning-preserving from meaning-altering edits (GREEN median 1.00 versus 0.50), though their low resolution made them better at flagging whether meaning changed than grading how much.

**Conclusion**

Metric scores should be interpreted by what each metric can and cannot measure. Word-overlap scores warrant caution as clinical-accuracy measures; label-based metrics are trustworthy only for presence/absence errors; entity-aware metrics best capture graded clinical differences; and LLM judges reliably flag whether meaning changed but grade it coarsely and need reproducibility checks. As a pilot simulation, findings require validation on real reports.

## Statement of Impact

No single automated metric captures the clinical accuracy of AI-generated radiology reports; this study offers a practical guide for choosing and interpreting metrics by the type of clinical error being evaluated.

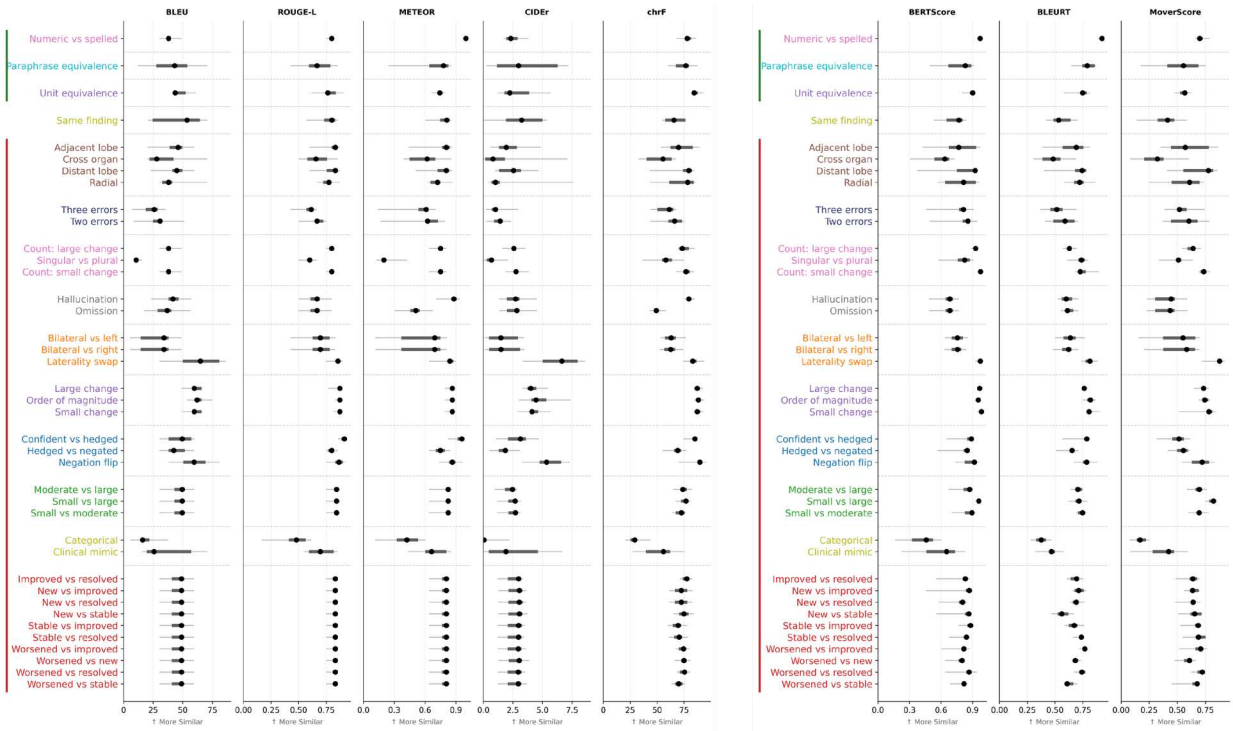


Figure 1. Score distributions of word-overlap (BLEU, ROUGE, METEOR, CIDEr, chrF) and contextual-embedding (BERTScore, BLEURT, MoverScore) metrics across the 39 perturbation tasks. Each panel shows one metric; each row is a task (10 reference–candidate sentence pairs differing by one to two targeted words). For every task, the black dot marks the median score across the 10 pairs, the thick bar the interquartile range (25th–75th percentile), and the thin line the full range (minimum–maximum). Tasks are grouped by error family (dashed separators; row-label color denotes family). Vertical bars at left mark expected behavior: green = meaning-preserving tasks, where an ideal metric should not penalize the change; red = meaning-altering tasks, where it should. The arrow beneath each panel indicates score direction (↑ more similar, or ↑ more error). Horizontal spread across rows reflects a metric's ability to distinguish task types (discrimination); the width of each bar/line reflects its consistency across pairs within a task.

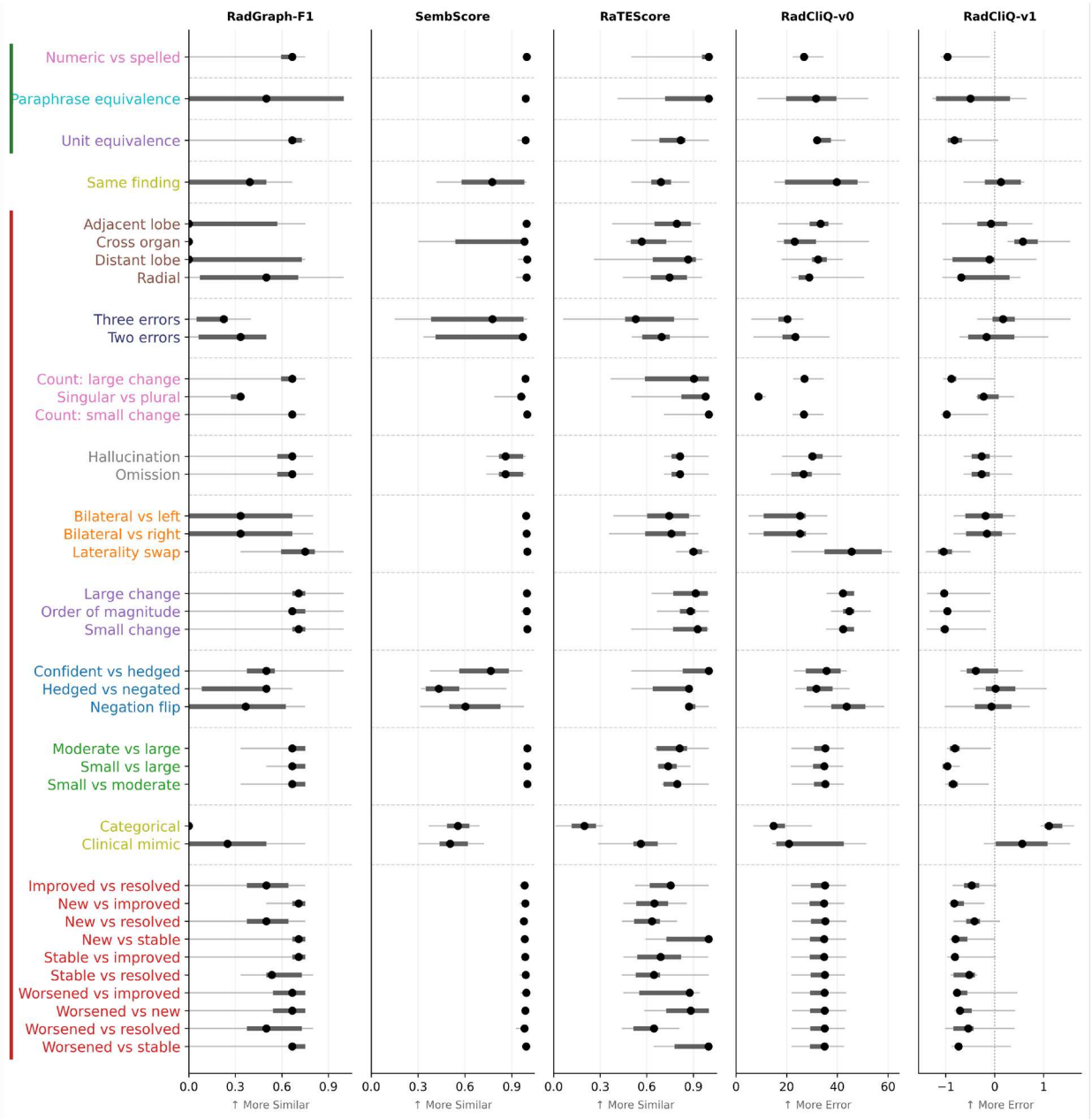


Figure 2. Score distributions of contextual-embedding metrics (BERTScore, BLEURT, MoverScore) across the 39 perturbation tasks. Panel format, task grouping, expected-behavior bars, and markers as in Figure 1.

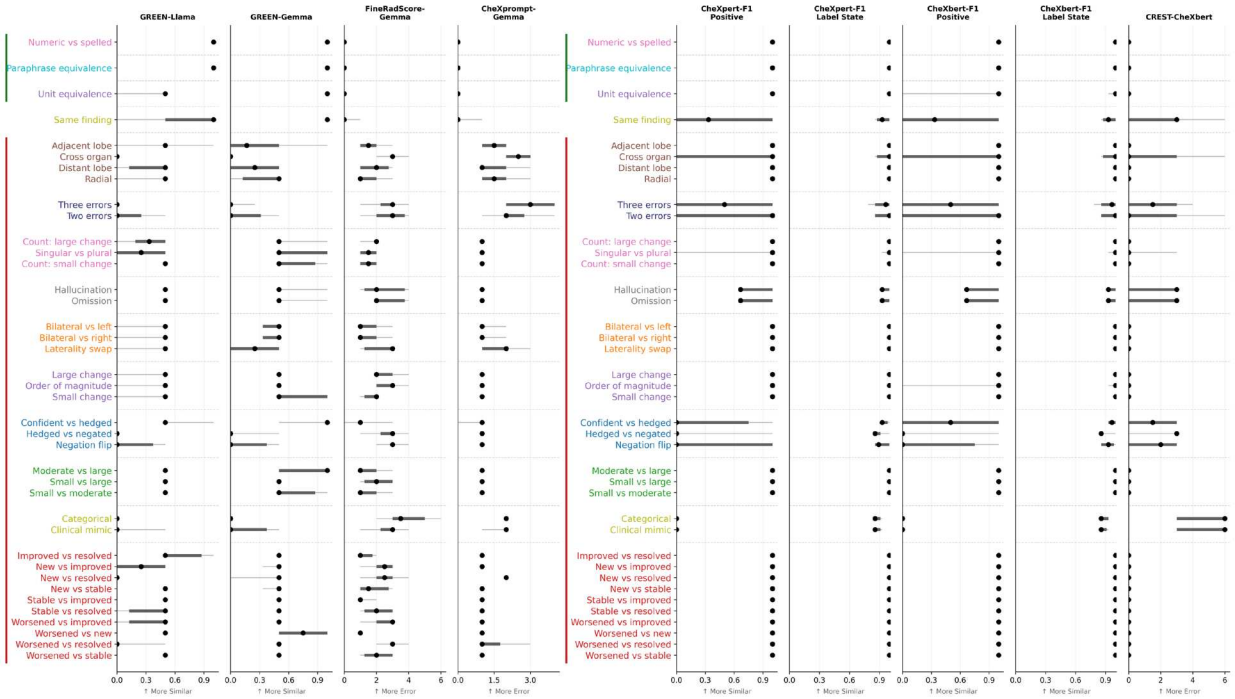


Figure 3. Score distributions of large-language-model judges (GREEN, FineRadScore, CheXprompt) and CheXpert/CheXbert label-based F1 metrics across the 39 perturbation tasks. Panel format, task grouping, expected-behavior bars, and markers as in Figure 1.


## Keywords

Report Generation**;** Large Language Models**;** Report Quality Metrics; Vision-Language Models**;** CXR