SySDEM - Synthetic and Stratified Degradations for Evaluating Metrics for Long-Form Text in Medical Domain

Naveen Jafer Nizar NAVEEN.JAFER@ORACLE.COM
Oracle Labs

Qinlan Shen QINLAN.SHEN@ORACLE.COM
Oracle Labs

Sumana Srivatsa SUMANA.SRIVATSA@ORACLE.COM
Oracle Health AI

Krishnaram Kenthapadi KRISHNARAM.KENTHAPADI@ORACLE.COM
Oracle Health AI

Abstract

The evaluation of long-form text in the medical domain is increasingly reliant on automated metrics. However, the reliability of these metrics themselves is often assumed rather than rigorously tested, especially when long-form generations are the expected output. We address this gap by proposing SySDEM - **Sy**nthetic and **S**tratified **D**egradations for **E**valuating **M**etrics, a framework to evaluate the quality of reference-based evaluation metrics. Using this framework, we demonstrate a method that iteratively perturbs candidate texts to assess the sensitivity and discrimination power of reference-based text evaluation metrics. Through experiments on the ACI-Bench clinical note generation dataset, we demonstrate the importance of evaluating evaluation metrics for long-form text, highlighting the need for robust validation methodologies.

Keywords: Evaluation of Evaluation Metrics; Long Form Text Generation for Clinical Applications

Data and Code Availability We perform experiments on the ACI-Bench clinical note generation dataset (wai Yim et al., 2023) which consists of synthetic doctor-patient dialogs and corresponding SOAP notes. We plan to release the code underlying our framework after obtaining organizational approval. Though the code repository is *not* currently released, we provide extensive technical and algorithmic details in the paper to aid implementation

Institutional Review Board (IRB) Our work does not require IRB approval.

1. Introduction

The surge in generative models has led to the proliferation of long-form text generation use cases in the medical domain. Evaluating the quality of these generated texts is crucial, which has subsequently driven the development of new evaluation metrics, which vary in their purpose, methodology, complexity, and sensitivity to the specific task they are developed for. While metrics are used to inform critical decisions, the errors that may propagate from the metrics themselves are often an afterthought. Prior work (Hanna and Bojar, 2021) has demonstrated that traditional evaluation metrics have low correlation with human evaluation judgments.

Automatic evaluation of machine-generated text has long been a central challenge in natural language generation (NLG) (Brew and Thompson (1994), Marie (2022)). Driven by the success of large-scale pretrained language models (PLMs) (Devlin et al., 2019), recent research has focused on developing evaluation metrics based on these models. For instance, BERTScore (Zhang et al., 2019b) calculates similarity scores between the contextualized embeddings of the candidate and the reference text. These PLM-based metrics have demonstrated (Herbold, 2024) superior correlations with human annotations across various tasks, leading to their increasing adoption in practical applications.

However, it is crucial to acknowledge the inherent limitations of PLMs. These models can produce degenerate, repetitive text (Holtzman et al., 2020) and exhibit insensitivity to perturbations such as word order shuffling and negation (Ettinger, 2020), (Hanna

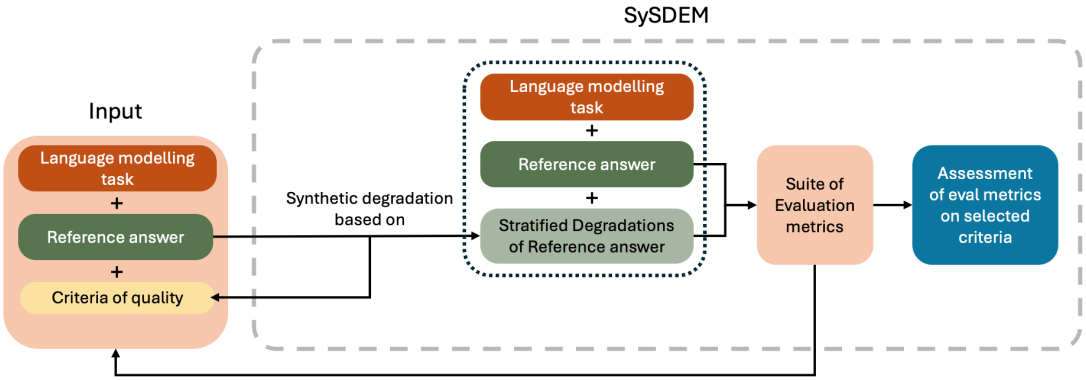

Figure 1: The framework lets a user configure three components: (1) language modeling tasks and the corresponding reference/gold answers; (2) a user specified criteria of quality, and (3) the evaluation metrics to be evaluated. The framework generates stratified synthetic degradations that allow it to assess the metrics on selected criteria of quality.

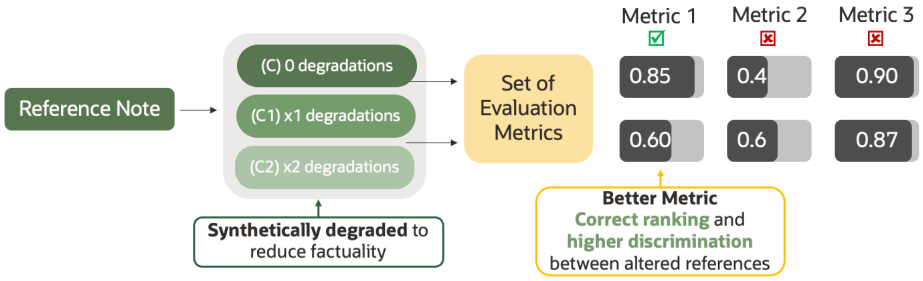

Figure 2: A detailed view of the synthetic degradations and assessment of evaluation metrics.

and Bojar, 2021). These shortcomings, when combined with specific design choices in metric development, can render PLM-based evaluation metrics brittle and susceptible to manipulation. Consequently, careful consideration of these factors is essential for ensuring the reliability and robustness of evaluation methodologies.

More recently, the use of commercial Large Language Models (LLMs) like GPT-4o as "LLM-as-a-judge" evaluators for pairwise comparisons in LLM alignment tasks has become increasingly prevalent (Zheng et al., 2023). While they have shown a higher agreement to human preference compared to PLM based methods, they suffer from various shortcomings; calibration, position bias, verbosity bias, and self-inclusion bias are some of them (Gu et al., 2024).

We introduce SySDEM, a scalable framework that iteratively perturbs candidate texts, forming a hierarchy that allows us to evaluate the sensitivity and discrimination power of various reference-based evaluation metrics in medical domain (Figure 1). Our approach generates synthetic degradations of the reference answer (without requiring additional human annotation) and measures the extent to which each evaluation metric gets worse when applied to the degraded versions of the reference answer (Figure 2). We emphasize that the focus of our work is to demonstrate the feasibility of this framework. Our intent is not to draw conclusions about specific metrics and the task they are applied on. We demonstrate the usability of SySDEM by applying our framework over a wide range of metrics for the task of SOAP (Subjective Objective Assessment Plan) note generation (Podder et al., 2024)) using the ACI-Bench dataset (wai Yim et al., 2023).

2. Related work

2.1. Reference based evaluation metrics

Reference-based evaluation metrics evaluate the generated output by measuring its similarity to human-annotated reference texts. ROUGE (Lin, 2004) introduced methods to evaluate the quality of a generated summary with a reference summary. Metrics such as BERTScore (Zhang et al., 2019a), BLEURT (Sellam et al., 2020), and others have demonstrated strong correlations with human judgments on short-form text. BERTScore leverages contextual embeddings from BERT to assess semantic similarity, while BLEURT is trained to predict human quality judgments.

Subsequent research discovered shortcomings of each of these metrics; Lin (2004) show that in a note generation setting, the Levenshtein distance between a reference and generated output led to better results than using ROUGE. Hanna and Bojar (2021) demonstrate deficiences in BERTScore that make it less sensitive to smaller errors. This period coincides with the rise of LLM-as-a-judge paradigm (Zheng et al., 2023) as an alternative to traditional metrics. While they show stronger correlation to human judgments, various trivial failure modes have come to light. Zheng et al. (2025) show that null models can achieve high win rates. Ye et al. (2024) show that unrelated changes in the prompts to the judge models can have significant outcomes in the judgment. In light of these findings, Gu et al. (2025) recommend that well-calibrated judgments, uncertainty calibration, and development of adversarially robust evaluation frameworks are steps to address such challenges. We position our work to better understand the ability of metrics to discriminate and calibrate an increasingly narrow regime where competing models score on the higher side with little variation in quality of their outputs.

2.2. Evaluation of metrics

Ribeiro et al. (2020) introduced checklists to test for robustness of models under perturbations. Following a similar recipe, Sai et al. (2021) leverage checklists but instead apply them to evaluation metrics. They design a total of 34 perturbation templates across different criteria such as Fluency, Informativeness, and Negations. They demonstrated that a majority of the metrics have poor correlations with human judgments across criteria. While our work builds on this foundation, the perturbation methods we use are not based on prefabricated templates. Additionally, we build an iterative perturbation setting which aims to not only study correlation of metrics to human judgment, but also quantitatively characterize their sensitivity to such perturbations.

Liang et al. (2023) introduced HELM and with it included robustness tests for evaluation metrics. While they test for invariance (the stability of a model's predictions under small semantically preserved perturbations) and equivariance (the study of semantic altering perturbations on model behavior) their focus is on model behavior and not the metrics themselves.

Deviyani and Diaz (2025) argue that while evaluation metrics are applied in highly contextualized settings, the metric meta-evaluation focuses on general statements about the absolute and relative quality of metrics. They demonstrate this mismatch by dividing a task into smaller contexts, observing that the metric accuracy changes across the contexts. Similar to us, they construct a perturbed dataset spanning each of the local contexts to measure the accuracy of a metric. Our work differs from them in multiple aspects. First, we use criteria of quality to influence the local context. Second, our perturbation mechanism is stratified and obtains richer signal about the metrics. Third, given that their perturbations are performed once on the original text, they test for single pairwise monotonicity. The discrimination power between metrics is not captured in their methods. Briefly, as long as two metrics can rank the perturbation lower than the original text, they are equivalent, and the magnitude is not taken into consideration. Finally, our focus is specifically on long form text, as opposed to perturbation methods that are better suited for shorter sentences.

3. Methodology

Our methodology focuses on evaluating reference-based evaluation metrics. Given a reference text R (gold standard) and a candidate text C, we iteratively perturb C n times to generate a sequence of degraded candidates $C_1, C_2, ..., C_n$. Each perturbation introduces a controlled level of degradation, such that the quality of C_i is lower than C_{i-1}. Figure 2 visually represents this workflow.

Perturbations can target various criteria of text quality, including but not limited to grammatical correctness, brevity, simplicity, factual completeness,

and factual correctness. To limit the scope of this study, we focus on factual correctness.

3.1. Perturbation Mechanism

We implement an unsupervised mechanism to perturb factual correctness. The mechanism splits the candidate text C into atomic facts, corrupts an increasing percentage of these facts using GPT-4o, and recombines them into a coherent long-form text to create the degraded candidates $C_1, C_2, ..., C_n$. Algorithm 1 illustrates this in detail and Table 1 presents an example.

3.2. Evaluation Metric Assessment

We evaluate the performance of evaluation metrics by examining two key properties:

1. **Monotonicity**: A desirable evaluation metric should assign scores that decrease as the candidate text is increasingly degraded. Mathematically, for an evaluation metric E_i, we expect:

$$E_i(R, C) > E_i(R, C_1) > ... > E_i(R, C_n)$$

2. **Discrimination Power**: A good evaluation metric should be able to discriminate between candidates of differing quality. If two evaluation metrics E_1 and E_2 are used to score candidates, E_1 is deemed to have higher discrimination power if the differences between its scores are larger than those of E_2.

4. Experiments

We experiment with the train set of ACI-Bench dataset (wai Yim et al., 2023). ACI-Bench is a medical dataset for the task of generating SOAP notes with 67 training examples. A SOAP note is a structured document used by healthcare professionals to record patient information, comprising Subjective, Objective, Assessment, and Plan components. For each patient case, ACI-Bench provides a gold reference SOAP note.

We extract a candidate solution C by paraphrasing the gold reference answer R using GPT-4o. The degraded candidates $C_1, C_2, ..., C_n$ are obtained by the method described in Section 3.

4.1. Human Annotation Study

To validate that our degradation method meets the desired properties, we conduct a human annotation check by sampling 10 candidate pairs C and C_1. For each pair, the following judgments are made:

- Is there at least one incorrect fact in candidate C_1 with respect to the reference R?
- Are the number of incorrect facts in candidate C lesser than in $C1$ with respect to the reference R?

All 10 candidate pairs satisfy the above conditions. The annotator also flags cases where the number of missing facts in candidate C if any is lower than the number of facts missing in $C1$ with respect to reference R. While this count is ideally expected to be 0 as a fact cannot appear in $C1$ that did not exist in candidate C, we observe one instance where this happens. We speculate that the data sample being present in the LLM's training data could possibly explain this behavior.

4.2. Evaluation Metrics

We evaluate the performance of metric $E_i \in E$ on the generated pairs:

$$E_i(R, C), E_i(R, C_1), E_i(R, C_2), ..., E_i(R, C_n).$$

Next, we describe the metrics that we test for the SOAP note generation use case (see Moramarco et al. (2022) for a discussion).

4.2.1. BERTScore

BERTScore (Zhang et al., 2019a) leverages contextual embeddings from BERT to assess the similarity between a candidate sentence and a reference sentence. It computes a fine-grained similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embeddings, then matches tokens using a greedy matching procedure. The scores for matched tokens are then aggregated to produce an overall similarity score.

4.2.2. BLEURT

BLEURT (Sellam et al., 2020) is a learned evaluation metric that is trained to predict human quality judgments. It is designed to be robust and generalize well across different domains and tasks. BLEURT uses a

Algorithm 1 Generate Degraded Candidates

Require: Reference Text R, Candidate Text C, Number of Iterations n, LLM model
Ensure: List of Degraded Candidate Texts $[C_1, C_2, ..., C_n]$

function GENERATEDEGRADEDCANDIDATES($C, n,$ LLM)
    $DegradedCandidates \leftarrow []$
    $CurrentCandidate \leftarrow C$
    $\alpha \leftarrow \frac{0.3}{n}$
    for $i \leftarrow 1$ *to* n **do**
    end
    $Facts \leftarrow$ SPLITINTOATOMICFACTS($CurrentCandidate$, LLM)
    $CorruptedFacts \leftarrow$ CORRUPTFACTS($Facts, \alpha * count(Facts)$, LLM)
    $DegradedCandidate \leftarrow$ RECOMBINEFACTS($CorruptedFacts$, LLM)
    APPEND($DegradedCandidates, DegradedCandidate$)
    $CurrentCandidate \leftarrow DegradedCandidate$
    $\alpha \leftarrow \alpha + \frac{1}{n}$

    return $DegradedCandidates$
end function

function SPLITINTOATOMICFACTS($Text$, LLM)
    $Facts \leftarrow$ USELLMTOSPLIT($Text$, LLM)
    return Facts
end function

function CORRUPTFACTS($Facts, NumToCorrupt$, LLM)
    $CorruptedFacts \leftarrow Facts$ ▷ Copy to avoid modifying original
    $IndicesToCorrupt \leftarrow$ RANDOMLYSELECTINDICES($NumToCorrupt$, length($Facts$))
    for *each Index in IndicesToCorrupt* **do**
    end
    $CorruptedFacts[Index] \leftarrow$ USELLMTOCORRUPT($Facts[Index]$, LLM)

    return $CorruptedFacts$
end function

function RECOMBINEFACTS($Facts$, LLM)
    $RecombinedText \leftarrow$ USELLMTORECOMBINE($Facts$, LLM) ▷ Use LLM to recombine
    return $RecombinedText$
end function

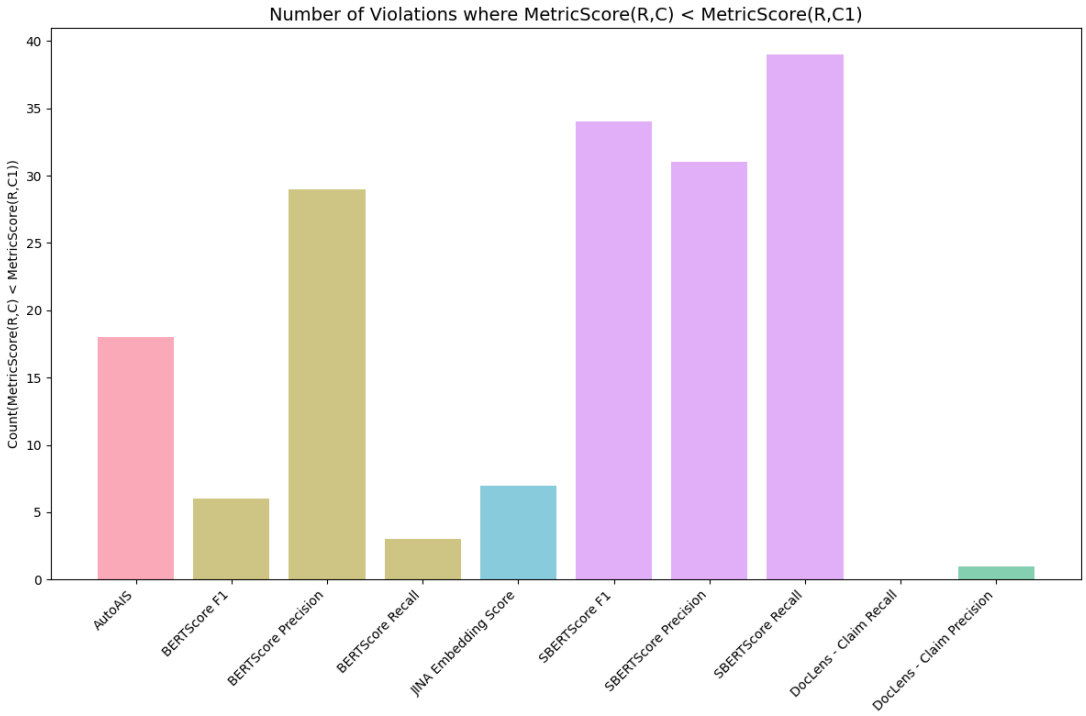

Figure 3: Each bar counts the number of instances where the metric scored the degraded candidate $C1$ higher than candidate C, out of a total of 67 examples, representing violations. DocLens has the lowest number of such violations.

large pre-trained language model and is fine-tuned on a dataset of human ratings and synthetic scores derived from other metrics, allowing it to capture subtle aspects of text quality that are often missed by other metrics.

4.2.3. AutoAIS

AutoAIS (Yue et al., 2023) is an automatic evaluation metric that focuses on assessing the factual accuracy and consistency of generated text. It uses a combination of information retrieval and natural language inference techniques to compare the generated text with a knowledge source or reference text. AutoAIS aims to provide a more reliable and interpretable evaluation of factuality than traditional metrics.

4.2.4. SBERTScore

SBERTScore is a modification of BERTScore that calculates sentence-level embeddings using Sentence-BERT (SBERT). While BERTScore computes token-level similarities, SBERTScore computes the embedding for the entire sentence and compares those embeddings. This change aims to capture the overall semantic similarity between sentences more effectively, potentially improving the correlation with human judgments, particularly for longer texts where sentence-level coherence is important.

4.2.5. Jina Similarity Embeddings

Jina Similarity Embeddings (Sturua et al., 2025) provide a method for generating dense vector representations of text that are optimized for similarity search and comparison. These embeddings can be used for various natural language processing tasks, including text evaluation. By comparing the embeddings of generated and reference texts, one can assess their semantic similarity. Jina embeddings are designed to be efficient and scalable, making them suitable for large-scale evaluation tasks.

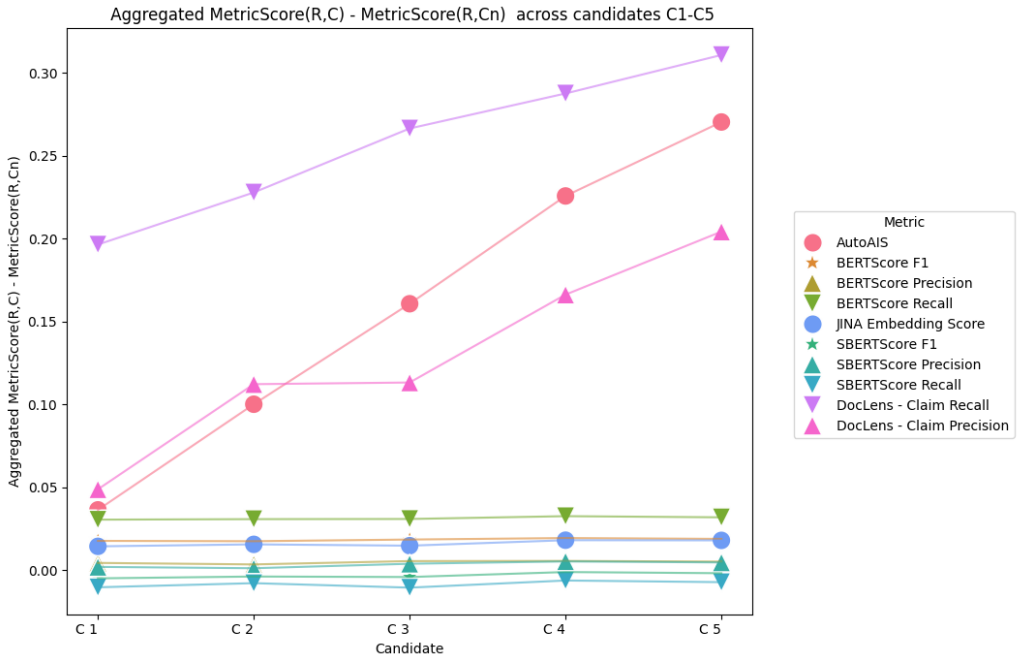

Figure 4: Each scatter in the plot corresponds to the aggregated $E_i(R, C) - E_i(R, C_n)$ across all samples for $n = 1...5$.

4.2.6. DocLens

DocLens (Xie et al., 2024) (is an evaluation framework designed to assess the quality of text generation tasks in the medical domain. The framework calculates the conciseness and completeness of the generated text for clinical note generation. The metrics can be computed by various types of evaluators including instruction-following (both proprietary and open-source) and supervised entailment models. While other models can be used with this framework, GPT4-o is widely used, so we replicate the same setting.

5. Results and Discussion

5.1. Monotonicity

When comparing candidate answer C against a degraded candidate, C_i, an ideal metric would consistently rank C over C_i. We refer to instances where a metric attributes a higher score to $E_i(R, C_i)$ than $E_i(R, C)$ as a violation. In Figure 3, we plot the frequency of violations in comparing C with C_1 across metrics. This provides a measure of how often the metric can discern a higher quality text. We see that across our metrics, DocLens has the fewest violations across examples. The appendix plots these same graphs comparing C to $C_2, C_3...C_5$. We see that the general pattern holds, with DocLens having the fewest violations. Notably, however, as the i in C_i increases, the number of violations AutoAIS drastically drops, such that AutoAIS is also able to achieve performance comparable to DocLens for $i > 3$.

When trying to measure how well metrics rank different levels of degradations for the same example, it is natural to consider the Spearman correlation between the number of degradations and the value returned by the metric. In practice, however, we find that this gives a coarsely aggregated view of how well metrics rank examples. We instead calculate the mean of $E(R, C) - E_i(R, C_i)$ over 67 samples for an evaluation metric E for a candidate i (Figure 4). As i increases, the number of degradations in C_i increase. As observed in the figure, $E(R, C) - E_i(R, C_i)$ stays flat for all metrics except AutoAIS and DocLens, indicating that these metrics best respect monotonicity for measuring factual degradations. This is an intuitive finding for examining factual degradations, as both metrics use an entailment model to explicitly

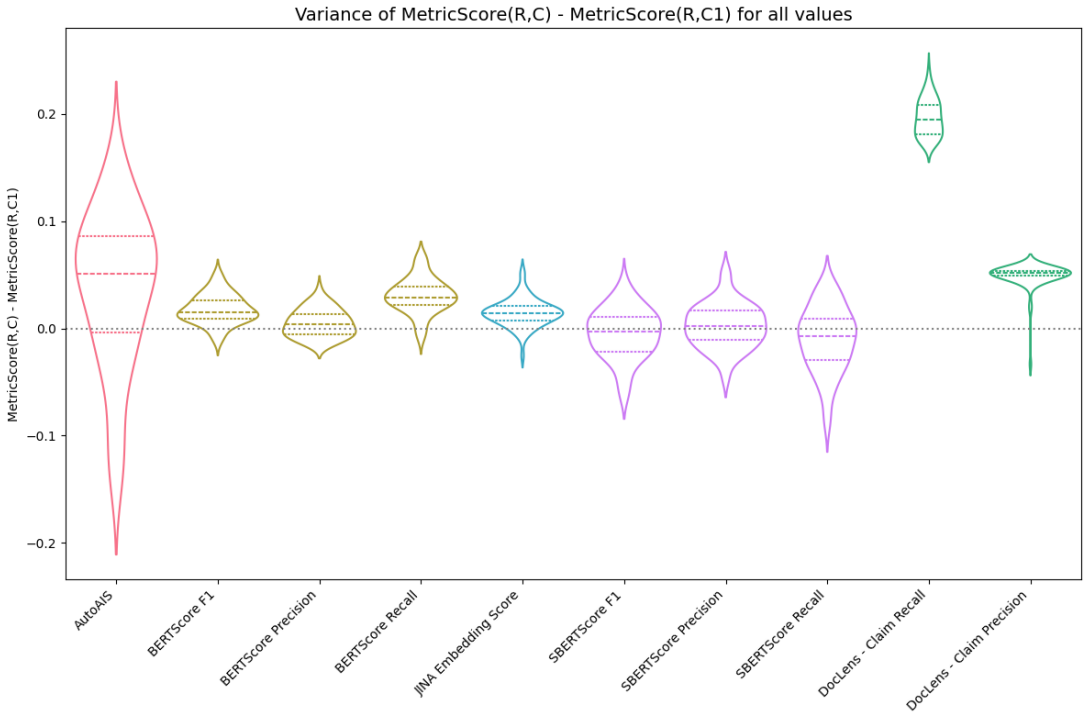

Figure 5: Each violin plots the value corresponding to $E_i(R, C) - E_i(R, C_1)$ along with the mean, and quartiles of the distribution denoted by dotted lines.

check that individual assertions made in the generated texts are supported/entailed by the reference.

5.2. Discrimination Power

In addition to being able to rank candidate metrics appropriately, a metric should also be able to strongly discriminate between candidates of differing quality. A metric could, for example, appropriately rank candidates in the correct order, but if the values returned for different candidates are very close, despite the differences in quality, it is difficult to draw conclusions from the metric when aggregating measurements across examples and models. Thus, we want metrics to have a substantial difference between the values of $E(R, C) - E_i(R, C_i)$ aggregated across the dataset. In Figure 5, we plot the distribution of $E(R, C) - E_i(R, C_1)$ where C is the candidate answer and C_1 is the first degraded candidate answer. We observe that DocLens (Xie et al., 2024) has the highest average score with no percentile below the $y = 0$ line.

In conclusion, by studying the distribution across all data samples of $E(R, C) - E_i(R, C_i)$ for each level C_i of degradation and comparing their means across all levels, practitioners can quantitatively characterize both the monotonicity and the discrimination power of a metric E. On this basis, the right evaluation metric for the criteria of choice can be utilized.

6. Conclusion

In this work, we introduce a framework to evaluate the quality of reference-based evaluation metrics for long-form text in the medical domain. Our methodology, based on iterative perturbation of factuality as a criteria, allows us to assess the sensitivity and discrimination power of various metrics from the lens of factuality. We reiterate that we position this paper to demonstrate the practicality of our framework. Conditioned on the use case and a medical practitioner's interpretation of an evaluation criteria, they may devise different degradation methods targeting different criteria. We encourage practitioners to spend re-

sources in the careful curation and design of these degradations to identify metrics that work well for their use cases. In the future, we envision the development of a library consisting of different degradation recipes paired with their motivating criteria, to allow for a fine grained evaluation of metrics.

7. Limitations

We acknowledge that the degradation process may be noisy and introduce errors where the true ordering of candidates $C_{1...n}$ may be out of order. A human in the loop-based verification such as the one we performed in the human annotation section could be used to estimate how reliable the method is. In use cases demanding stringent quality, an assisted human in the loop verifier is a potential solution. Alternatively, there are scaling methods that can rely on the majority vote of an ensemble of metrics to filter and reduce such noisy orderings.

References

Chris Brew and Henry S. Thompson. Automatic evaluation of computer generated text: A progress report on the TextEval project. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://aclanthology.org/H94-1019/.

Athiya Deviyani and Fernando Diaz. Contextual metric meta-evaluation by measuring local metric accuracy. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4906–4925, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.276/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020. doi: 10.1162/tacl_a_00298. URL https://aclanthology.org/2020.tacl-1.3/.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594, 2024. URL https://api.semanticscholar.org/CorpusID:274234014.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Michael Hanna and Ondřej Bojar. A fine-grained analysis of BERTScore. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.59/.

Steffen Herbold. Semantic similarity prediction is better than other semantic similarity measures. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bfsNmgN5je.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-

Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Benjamin Marie. An automatic evaluation of the wmt22 general machine translation task, 2022. URL https://arxiv.org/abs/2209.14172.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. Human evaluation and correlation with automatic metrics in consultation note generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.394. URL https://aclanthology.org/2022.acl-long.394/.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. SOAP notes. *StatPearls [Internet]*, 2024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. Perturbation CheckLists for evaluating NLG evaluation metrics. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.575. URL https://aclanthology.org/2021.emnlp-main.575/.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. Jina embeddings v3: Multilingual text encoder withnbsp;low-rank adaptations. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, page 123–129, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-88719-2. doi: 10.1007/978-3-031-88720-8_21. URL https://doi.org/10.1007/978-3-031-88720-8_21.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, 2023. URL https://arxiv.org/abs/2306.02022.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. DocLens: Multi-aspect fine-grained evaluation for medical text generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, Bangkok, Thailand, August 2024. Association

for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.39. URL https://aclanthology.org/2024.acl-long.39/.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge, 2024. URL https://arxiv.org/abs/2410.02736.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.307. URL https://aclanthology.org/2023.findings-emnlp.307/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019a. URL http://arxiv.org/abs/1904.09675.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2019b. URL https://api.semanticscholar.org/CorpusID:127986044.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic llm benchmarks: Null models achieve high win rates, 2025. URL https://arxiv.org/abs/2410.07137.

Appendix A. Appendix

A.1. Prompts

In Figure 6 we show the single shot prompt for splitting text into atomic facts. Figure 7 and 8 shows the prompts for corrupting and recombining those facts respectively, and Figure 9 is to generate a paraphrase of the input text.

A.2. Degradation example

Table 1 shows an example of the degradation process.

Prompt (Fact splitting)
You are a helpful bot, your task is to break the user text into the smallest independent atomic facts possible. Do not number them. Print them out line by line. Each fact needs to be simple.

Here is an example

INPUT: The Java Development Kit (JDK) is an implementation of either one of the Java Platform, Standard Edition, Java Platform, Enterprise Edition, or Java Platform, Micro Edition platforms released by Oracle Corporation in the form of a binary product aimed at Java developers on Solaris, Linux, macOS or Windows.

OUTPUT: The Java Development Kit (JDK) exists.
The JDK is an implementation.
The JDK implements the Java Platform.
The Java Platform has three editions: Standard Edition, Enterprise Edition, and Micro Edition.
The JDK is released by Oracle Corporation.
The JDK is released as a binary product.
The binary product is aimed at Java developers.
Java developers use Solaris.
Java developers use Linux.
Java developers use macOS.
Java developers use Windows.
INPUT: <TEXT>

Figure 6: GPT-4o prompts for splitting facts

Prompt (Fact corruption)
You are a helpful bot, that is good at changing or modifying the facts in the user input provided to you. There are to be no follow up questions. Just provide the answer.
<TEXT>

Figure 7: GPT-4o prompts for corrupting facts

Prompt (Fact recombination)
Use the following facts provided in the user input to generate a coherent paragraph. Do not add in your response information that does not exist in the input provided.
<TEXT>

Figure 8: GPT-4o prompts for recombining facts

Prompt (Paraphrasing)
You are a helpful assistant who is great at paraphrasing text. Take the text provided to you and paraphrase it strongly but correctly.
Do not ask follow up questions.
Do not ask to perform any other actions.
Do not copy from the original text as is.
Do not try to continue the conversation.
<TEXT>

Figure 9: GPT-4o prompts for generating the paraphrased version of a text

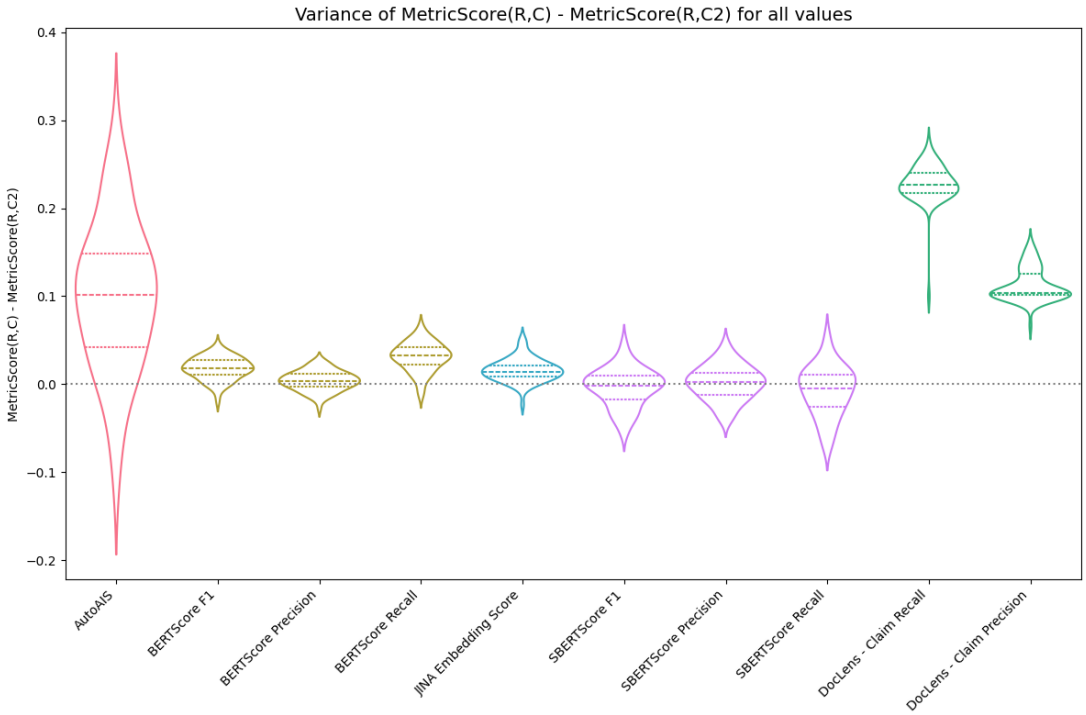

Figure 10: Each violin plots the value corresponding to $E_i(R, C) - E_i(R, C_2)$ along with the mean, and quartiles of the distribution denoted by dotted lines.

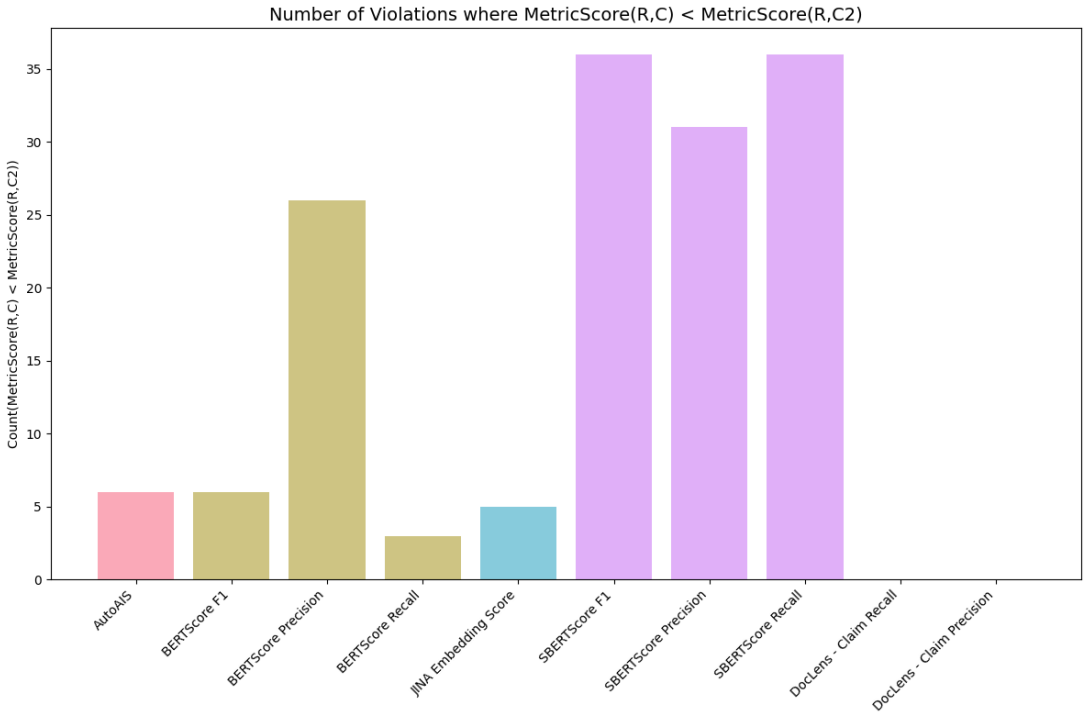

Figure 11: Each bar counts the instance where the Metric scored the degraded candidate C2 higher than candidate C, out of a total of 67 examples. DocLens has the lowest number of such violations.

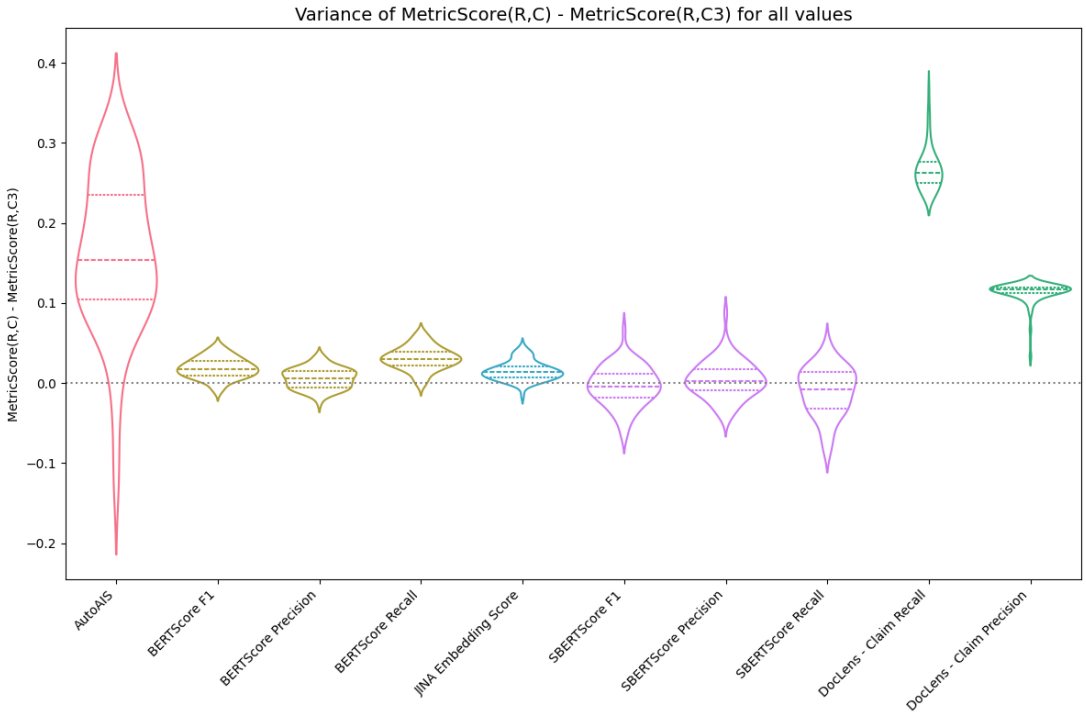

Figure 12: Each violin plots the value corresponding to $E_i(R, C) - E_i(R, C_3)$ along with the mean, and quartiles of the distribution denoted by dotted lines.

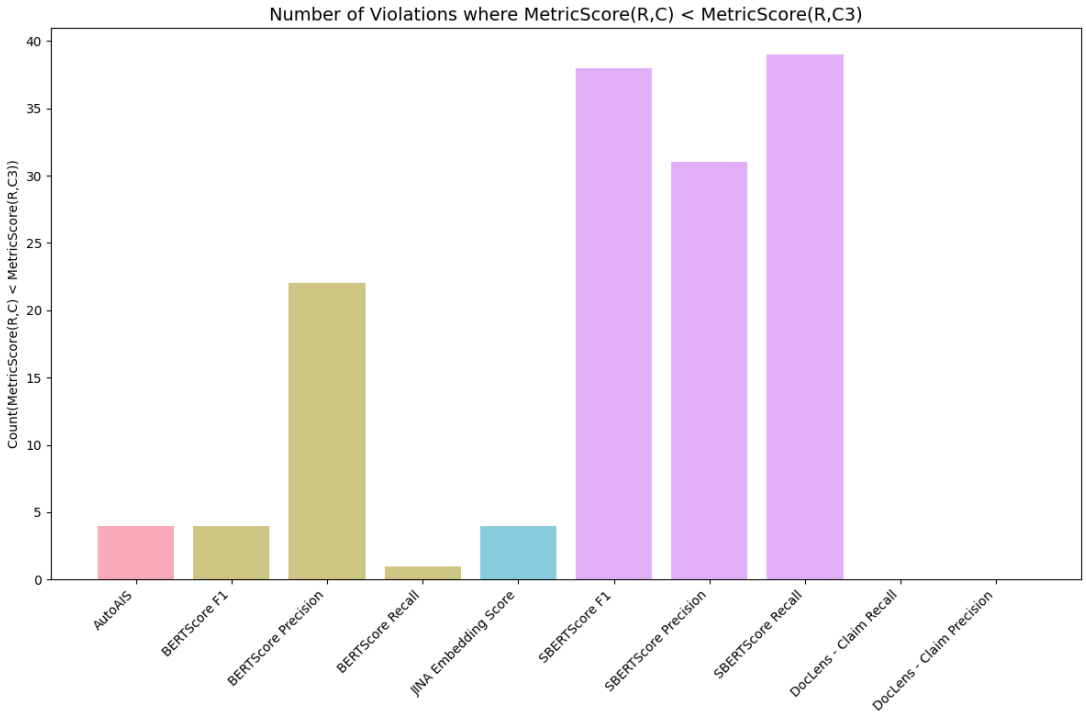

Figure 13: Each bar counts the instance where the Metric scored the degraded candidate C3 higher than candidate C, out of a total of 67 examples. DocLens has the lowest number of such violations.

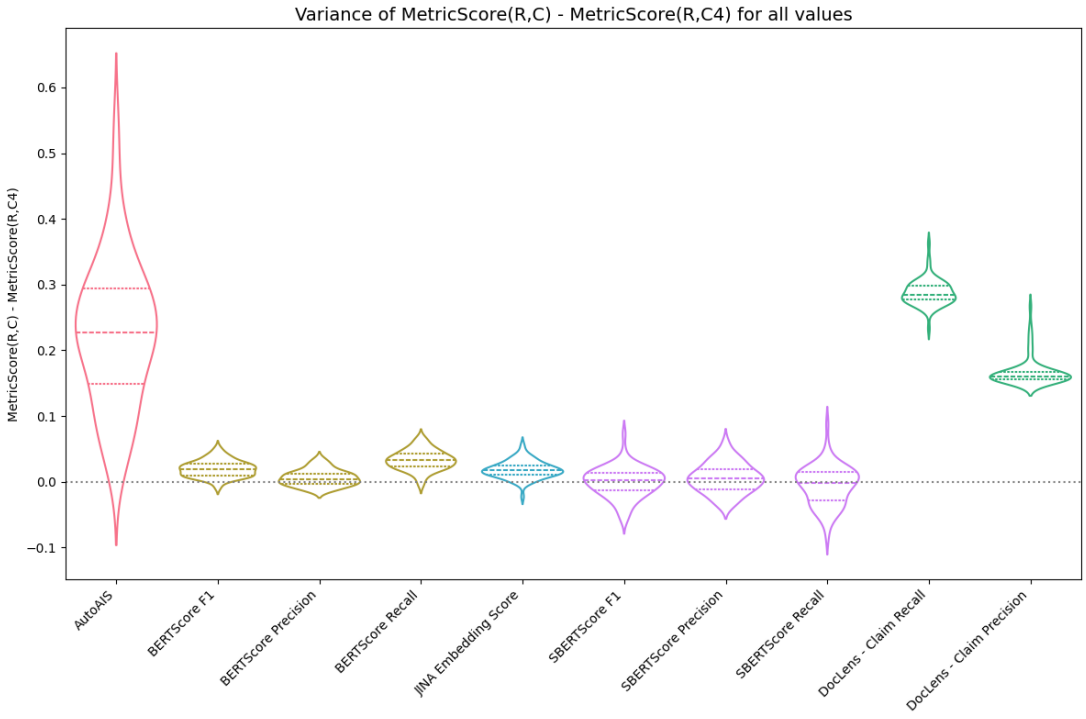

Figure 14: Each violin plots the value corresponding to $E_i(R, C) - E_i(R, C_4)$ along with the mean, and quartiles of the distribution denoted by dotted lines.

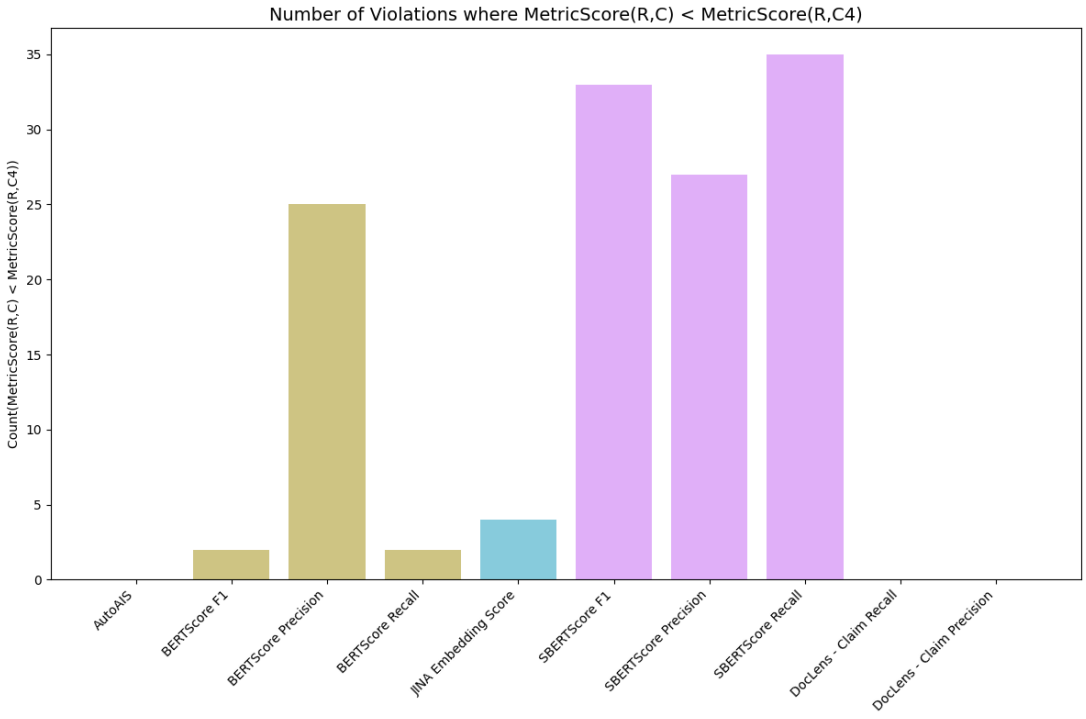

Figure 15: Each bar counts the instance where the Metric scored the degraded candidate C4 higher than candidate C, out of a total of 67 examples. DocLens has the lowest number of such violations.

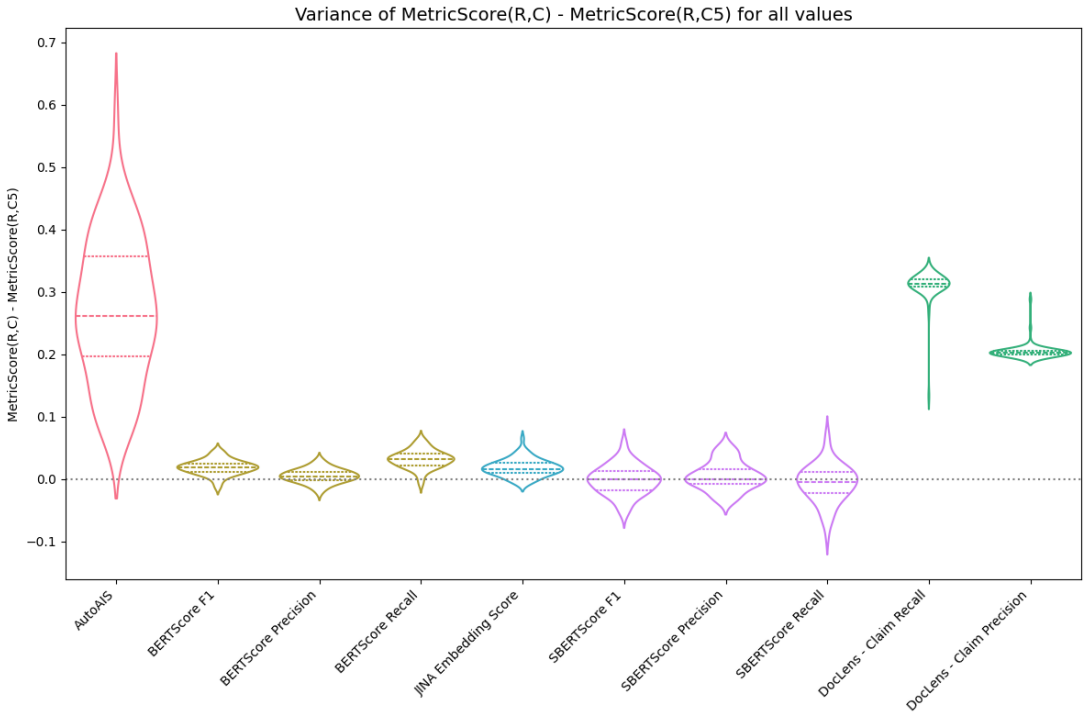

Figure 16: Each violin plots the value corresponding to $E_i(R, C) - E_i(R, C_5)$ along with the mean, and quartiles of the distribution denoted by dotted lines.

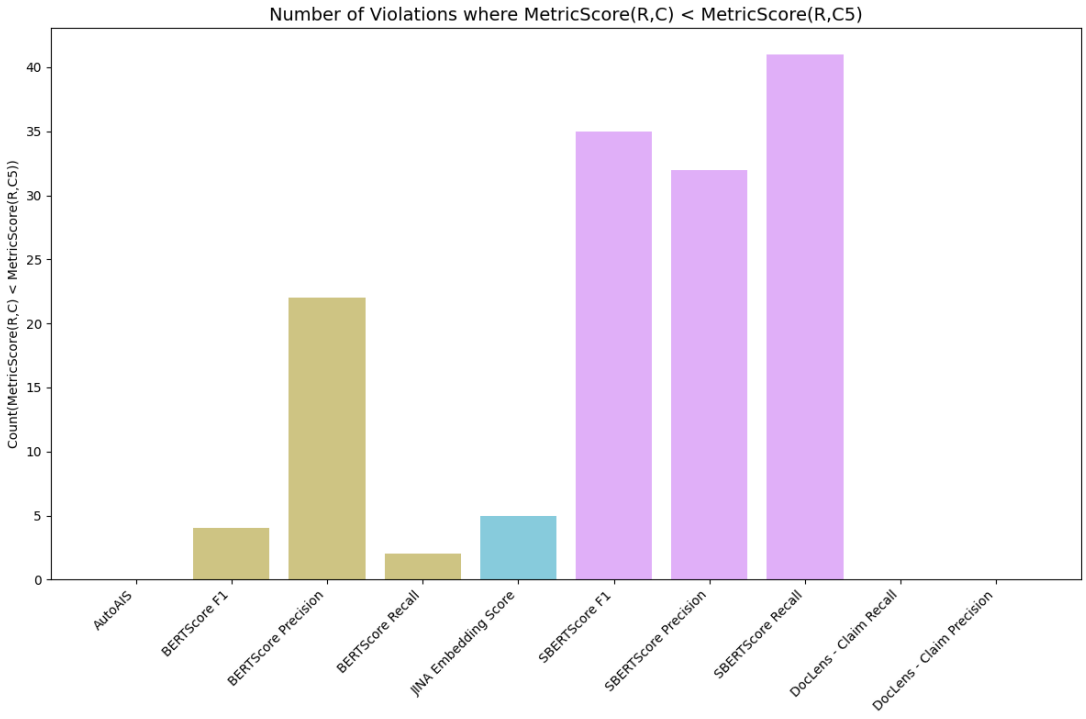

Figure 17: Each bar counts the instance where the Metric scored the degraded candidate C5 higher than candidate C, out of a total of 67 examples. DocLens has the lowest number of such violations.

Label	Content
Reference (R)	CC: Right upper arm pain. HPI: Ms. Russell is a 39-year-old female who presents today for an evaluation of severe right upper arm pain. She states she was playing volleyball yesterday and fell and landed on her arm. She has been taking ibuprofen every 6 hours but it does not help. She rates her pain 9/10. She denies any numbness or tingling in her arm. She has a history of gallstones and takes Pepcid. She has a past surgical history of a lumbar fusion 6 years ago. She denies any allergies. EXAM Examination of the right upper extremity shows swelling and erythema of the right shoulder. Tenderness over the right shoulder. Normal pulses. RESULTS X-rays of the right humerus, 2 views, obtained on today's visit show a proximal humerus fracture. IMPRESSION Right proximal humerus fracture. PLAN At this point, I discussed the diagnosis and treatment options with the patient. I have recommended we place her into a long arm cast and sling. She will follow up in 2 weeks for repeat imaging. At that point, we may need to discuss surigical options. I will prescribe Lortab, 5 mg, #20 (twenty) to take every 6 hours as needed for pain.All questions were answered
Candidate (C)	Chief Complaint: Pain in the right upper arm. History of Present Illness: Ms. Russell, a 39-year-old woman, reports severe pain in her right upper arm following a fall while playing volleyball yesterday. Despite taking ibuprofen every six hours, her pain remains intense, rated at 9 out of 10. She denies experiencing any numbness or tingling in the arm. Her medical history includes gallstones, and she is on Pepcid. She underwent lumbar fusion surgery six years ago and has no known allergies. Examination: The examination revealed swelling and redness in the right shoulder, along with tenderness in the same area. Her pulses are normal.Results:X-rays taken of the right humerus show a fracture in the proximal part of the bone.Impression:Fracture of the right proximal humerus. Plan: The diagnosis and treatment options were discussed with Ms. Russell. It was advised that she be placed in a long arm cast and sling. A follow-up appointment is scheduled in two weeks for additional imaging, and potential surgical options may be considered at that time. She has been prescribed Lortab, 5 mg, to take every 6 hours as needed for pain relief.All her questions were addressed.
Candidate (C_3)	Ms. Russell, a 39-year-old female, presents today for an evaluation of severe right upper arm pain following a fall while playing volleyball yesterday. Despite taking ibuprofen every 6 hours, her pain remains severe, rated at 9/10, and ibuprofen has not been effective. Ms. Russell denies any numbness or tingling in her arm. She has a medical history of gallstones and takes Pepcid, and she has a past surgical history of a lumbar fusion 6 years ago. Additionally, she is allergic to penicillin .Upon examination of the right upper extremity, there is no swelling or erythema in the right shoulder, but there is tenderness over the shoulder. Ms. Russell has normal pulses. X-rays of the right humerus obtained today reveal a proximal humerus fracture. Following a discussion regarding the diagnosis and treatment options, Ms. Russell is recommended to have a long arm cast and sling and will follow up in 2 weeks for repeat imaging. There may be a need to discuss surgical options at that time. For pain management, Lortab, 5 mg, has been prescribed, with instructions to take it every 6 hours as needed. All of Ms. Russell's questions were answered during the visit.

Table 1: The text highlighted in red shows the errors introduced by the degradation process in Candidate C_3