An Investigation of Evaluation Metrics for Automated Medical Note Generation

Asma Ben Abacha
Microsoft Health AI
abenabacha@microsoft.com

Wen-wai Yim
Microsoft Health AI
yimwenwai@microsoft.com

George Michalopoulos
Microsoft Health AI
georgemi@microsoft.com

Thomas Lin
Microsoft Health AI
tlin@microsoft.com

Abstract

Recent studies on automatic note generation have shown that doctors can save significant amounts of time when using automatic clinical note generation (Knoll et al., 2022). Summarization models have been used for this task to generate clinical notes as summaries of doctor-patient conversations (Krishna et al., 2021; Cai et al., 2022). However, assessing which model would best serve clinicians in their daily practice is still a challenging task due to the large set of possible correct summaries, and the potential limitations of automatic evaluation metrics. In this paper, we study evaluation methods and metrics for the automatic generation of clinical notes from medical conversations. In particular, we propose new task-specific metrics and we compare them to SOTA evaluation metrics in text summarization and generation, including: (i) knowledge-graph embedding-based metrics, (ii) customized model-based metrics, (iii) domain-adapted/fine-tuned metrics, and (iv) ensemble metrics. To study the correlation between the automatic metrics and manual judgments, we evaluate automatic notes/summaries by comparing the system and reference facts and computing the factual correctness, and the hallucination and omission rates for critical medical facts. This study relied on seven datasets manually annotated by domain experts. Our experiments show that automatic evaluation metrics can have substantially different behaviors on different types of clinical notes datasets. However, the results highlight one stable subset of metrics as the most correlated with human judgments with a relevant aggregation of different evaluation criteria.

1 Introduction

In recent years, the volume of data created in healthcare has grown considerably as a result of record keeping policies (Kudyba, 2010). The documentation requirements for electronic health records significantly contribute to physician burnout and work-life imbalance (Arndt et al., 2017). Automatic generation of clinical notes can help healthcare providers by significantly reducing the time they spend on documentation, and allowing them to spend more time with patients (Payne et al., 2018). It can also improve the clinical notes' accuracy by reducing errors and inconsistencies in documentation, leading to patient records with higher quality.

A reliable evaluation methodology is necessary to build and improve clinical note generation systems, but faces the two traditional limitations of evaluating Natural Language Generation (NLG) systems. On one hand, human-expert evaluation, considered to be the most reliable way to evaluate NLG systems, can be both time-consuming and expensive. On the other hand, evaluating the performance of natural language generation (NLG) systems automatically can be challenging due to the complexity of human language.

Several metrics have been proposed to evaluate the performance of NLG systems, including lexical N-gram based metrics and embedding-based metrics that measure the similarity between a system's generated text and one or more reference texts using pre-trained language models.

While several research efforts studied and compared automatic evaluation metrics on many open-domain and domain-specific datasets such as the CNN/DailyMail and TAC datasets (Lin, 2004a; Owczarzak et al., 2012; Peyrard, 2019; Fabbri et al., 2021a; Deutsch et al., 2022), very few research works addressed the adequacy of evaluation metrics to the task of clinical note generation, where e.g., omitting critical medical facts in the generated text is a more significant failure point. To the best of our knowledge, only one research paper addressed this task based on one synthetic dataset of 57 mock consultation transcripts and summary notes (Moramarco et al., 2022).

In this paper, we study evaluation methods and metrics for the automatic generation of clinical

Findings of the Association for Computational Linguistics: ACL 2023, pages 2575–2588
July 9-14, 2023

notes from medical conversations, including their correlations with human assessments of factual omissions and hallucinations. We also propose new task-specific metrics and we compare them to SOTA evaluation metrics across several clinical text summarization datasets.
Our contributions are as follows:

- We study the relevance and impact of a wide-range of existing automatic evaluation metrics in clinical note generation.
- We propose and study four types of evaluation metrics for the task of automatic note generation: knowledge-graph embedding-based metrics, customized model-based metrics, domain-adapted/fine-tuned metrics, and ensemble metrics[1].
- We compare these metrics with SOTA metrics by performing a wide evaluation with 21 metrics according to different criteria such as factual correctness, hallucination, and omission rates.
- To perform a fact-based evaluation of the generated notes, we annotate seven datasets of automatically generated clinical notes using key phrase- and fact-based annotation guidelines that we use to compute reference manual scores for the correlation study[2].

2 Related Work

Different evaluation metrics are commonly used to evaluate text summarization and generation including ROUGE-N (Lin, 2004b), BERTScore (Zhang* et al., 2020), MoverScore (Zhao et al., 2019), BARTScore (Yuan et al., 2021), and BLEURT (Sellam et al., 2020). Other metrics have been also proposed for evaluating factual consistency and faithfulness (Durmus et al., 2020; Maynez et al., 2020; Wang et al., 2020; Pagnoni et al., 2021; Zhang et al., 2022).

To study their effectiveness, several efforts focused on comparing automatic metrics such as ROUGE and BLEU based on their correlation with human judgments (Graham, 2015), and showed that automatic evaluation of generated summaries still has several limitations and biases (Hardy et al., 2019; Fabbri et al., 2021b). Furthermore, in (Bhandari et al., 2020), the authors showcase that the effectiveness of an evaluation metric depends on the task (e.g. summarization) and on the application scenario (e.g. system-level/ summary level).

Despite observations of frequent disagreements in manual evaluation campaigns (Howcroft et al., 2020), expert-based evaluation remains an effective method to assess the performance of automatic metrics, especially in specialized domains. However, it relies on the availability of domain experts to rate the summaries and relevant datasets. Recently, (Moramarco et al., 2022) studied the task of medical note generation on a small set of 57 transcript-note pairs, manually annotated by clinicians. Their experiments showed that character-based Levenshtein distance, BERTScore, and METEOR performed best for evaluating automatic note generation in that dataset.

3 Evaluation Methodology

To assess the relevance and suitability of automatic evaluation metrics for the task of clinical note generation, we create expert-based annotations for critical aspects such as factual consistency, hallucinations, and omissions. We then assess each metric in light of its correlation with manual scores generated from the expert annotations.

3.1 Fact-based Annotation

We define a fact as information that cannot be written in more than one sentence (e.g., *"Family history is significant for coronary artery disease."*). Medical facts include problems, allergies, medical history, treatments, medications, tests, laboratory/radiology results, and diagnoses. We also include the patient age, gender, and race, and expand the critical facts to the patient and his family.

Annotators extracted individual facts from both the reference and system summaries in the form of subject-predicate-object expressions and following the above fact definition.

Comparing between a reference and hypothesis summary, and referencing the source text/conversation if required, the annotators were additionally tasked to identify overlapping and non-overlapping facts to one of several categories which were later automatically counted. These included:

- Critical Omissions: the number of medical facts that were omitted,

[1]We publish the source code and fine-tuned checkpoint at: https://github.com/abachaa/EvaluationMetrics-ACL23

[2]We also release the manual annotations: https://github.com/abachaa/EvaluationMetrics-ACL23

Dataset	#Summary Pairs	#Words /Summary	#Words /Reference	Annotations
MTS-DIALOG	400	15	36	Facts
MEDIQA-RRS	182	18	28	Facts
CONSULT-FACTS	54	203	214	Facts
$\text{CONSULT}_{\text{HPI}}$	3,397	333	336	Key Phrases
$\text{CONSULT}_{\text{ASSESSMENT}}$	3,141	149	177	Key Phrases
$\text{CONSULT}_{\text{EXAM}}$	2,144	163	137	Key Phrases
$\text{CONSULT}_{\text{RESULTS}}$	540	38	15	Key Phrases

Table 1: Annotated datasets of clinical summaries and reference notes.

- Hallucinations: the number of hallucinated facts. Hallucinations are factual errors that do not exist in the source text and cannot be supported by the source facts (e.g., added dates, names, or treatments).

- Correct Facts: the number of correct facts according to the input conversation and the reference summary, and

- Incorrect Facts: the number of incorrect facts outside of hallucinations. Incorrect facts include values and attributes that are incorrectly copied from the source (e.g., date with a wrong year, wrong age, or dose).

Three trained annotators with medical background participated in the annotation process. Inter-annotator agreement for these computations are shown in Table 7 and Table 8 in Appendix A.

3.2 Key Phrase-based Annotation

The key phrase- and fact-based annotations use different ways of representing information in clinical notes. While the fact-based annotation compares semantic triples (e.g., "Back pain stopped 8 days ago" vs. "Low back pain started 8 years ago"), the key-phrase annotations involved labeling incorrect words and phrases; for instance: "back pain" (instead of "lower back pain"), "stopped" (instead of "started"), or "8 days" (instead of "8 years"). This method is more conducive in a production environment where errors can be attributable to specific parts of the report; the same labeling method is often also used for feedback to the author of the note in our different human quality review settings.

In our annotation setups, the key phrase-based annotation operated on text span highlights, while the fact-based annotation required more steps as the annotators were required to write the system and reference facts based on the system and reference summaries before comparing their counts.

Using highlights, critical hallucinations and incorrect information can be identified; meanwhile omissions were marked by identifying a required insertion of information in a corresponding location of the note. However, unlike the previous annotation, repeats of the same incorrect facts may be counted more than once if they appear multiple times. The labels produced here were from CONSULT-FULL dataset (cf. Section 3.4), with a reported average agreement of critical hallucinations, omissions, and inaccuracies was at 0.80 F1 score, relaxed overlap between 12 annotator pairs.

3.3 Reference Scores

From the fact-based annotations, we compute the following reference scores:

$$FactualPrecision = \frac{\#CorrectFacts}{\#SystemFacts}$$

$$Factual\ Recall = \frac{\#Correct\ Facts}{\#Reference\ Facts}$$

$$Hallucination\ Rate = \frac{\#Hallucinated\ Facts}{\#SystemFacts}$$

$$Omission\ Rate = \frac{\#Omitted\ Facts}{\#Reference\ Facts}$$

- *System Facts = Correct + Incorrect + Hallucinated*

From the key phrase-based annotations, we compute the normalized hallucination and omission counts:

$$HallucinationCount = \frac{\#Hallucinated\ key\ phrases}{\#System\ Summary\ Words}$$

$$OmissionCount = \frac{\#Omitted\ key\ phrases}{\#Reference\ Summary\ Words}$$

3.4 Datasets

Publicly available datasets on medical note generation and clinical text summarization are rare compared to open-domain data. For this study, we use three main collections:

- The MTS-DIALOG collection of 1.7k pairs of doctor-patient dialogues and associated clinical notes (Ben Abacha et al., 2023). System summaries are generated using the BART model (Lewis et al., 2020).
- The MEDIQA-RRS dataset includes 182 pairs of clinical notes and system summaries randomly selected from the MEDIQA-RRS collection (Ben Abacha et al., 2021).
- An in-house collection of medical notes (called CONSULT-FULL) from multiple specialties with system summaries generated using a pointer-generator transformer model from doctor-patient conversations (Enarvi et al., 2020).

We followed the fact-based annotation guidelines to annotate the MTS-DIALOG and MEDIQA-RRS datasets, and a random subset from the CONSULT-FULL collection, called CONSULT-FACTS.

To study the relevance of the automatic metrics to the individual sections of clinical notes, we also split the CONSULT-FULL collection into four subsets: $\text{CONSULT}_{\text{HPI}}$, $\text{CONSULT}_{\text{ASSESSMENT}}$, $\text{CONSULT}_{\text{EXAM}}$, and $\text{CONSULT}_{\text{RESULTS}}$, which include summaries associated with the HPI, Assessment, Exam, and Results sections, and we annotated them manually at a phrase level.

Table 1 provides statistics about the datasets.

4 Task-specific Evaluation Metrics

We study four different types of evaluation metrics for the task of automatic clinical note generation, that take into account the specificities of the medical domain by: (i) using embeddings built from medical Knowledge graphs (e.g, UMLS), (ii) adapting model-based metrics (e.g., BERTScore) by increasing the weights of medical terms, (iii) fine-tuning a model-based metric on a large collection of clinical notes, and (iv) building linear ensembles based on normalization and averaging of different metrics.

4.1 Knowledge-Graph Embedding-based Metrics

Our first approach, called MIST, relies on knowledge embeddings generated by a Knowledge-Graph Embedding (KGE)-based model. Knowledge graphs provide additional semantic information that can support language understanding, especially in the medical domain where both terminologies and facts might not be common enough to be captured by contextual embeddings.

To build medical KGE, we use a generative adversarial networks model (Cai and Wang, 2018) trained on concepts and relations from the Unified Medical Language System (UMLS) (Lindberg et al., 1993; Bodenreider, 2004).

The MIST metric relies on the embeddings of the medical concepts recognized in the texts to compute the similarity between the reference clinical notes and the automatically generated summaries.

To link the clinical notes to the UMLS concepts, we extract medical concepts by combining the scispaCy (Neumann et al., 2019) and MedCAT (Kraljevic et al., 2021) entity linking models.

We compute the final recall-oriented MIST value using the graph-based embeddings (G_c) of each concept c recognized in the reference and system summaries and the cosine similarity, as follows, for a set of reference concepts R and a set of system concepts S:

$$MIST(S,R) = \frac{1}{|R|}\sum_{c\in S} max_{r\in R}\ cos(G_c, G_r) \quad (1)$$

4.2 Finetuning-based Metric

Our second approach relies on fine-tuning model-based metrics on relevant large medical collections of family medicine and orthopaedic notes. In particular, we started with the BLEURT-512 model (Sellam et al., 2020) and fine-tuned it using a quality score, derived from an assigned *error score*[3] from an internal quality review grading. The derived *quality score* was calculated by the following equation:

$$quality = 1 - \frac{error\_score}{max\_sentlen(summary, reference)} \quad (2)$$

A total of 6,367 family medicine and orthopaedic encounters were used for fine-tuning. To maximize

[3]This error score is calculated by a weighted sum of critical and non-critical errors, as well as spelling/grammar/style errors annotated by domain expert labelers. The weight scheme is given in Appendix B, Table 9.

diverse pairings as well as to satisfy BLEURT's maximum sequence length constraint, we fine-tuned at the level of each note's HPI, EXAM, RESULTS, and ASSESSMENT sections (with empty sections removed), resulting in 17,852 pairs. We fine-tuned over one epoch at default parameters. We call the resulting metric based on this model: ClinicalBLEURT.

4.3 Customized Model-based Metrics

4.3.1 Medical Weighted Evaluation Metrics

Our third approach relies on designing new customized model-based metrics that assign a higher weight to the term with a medical meaning. These medical weighted metrics will allow us to examine whether words with a medical meaning can be more indicative for sentence similarity than common words for the task of automated medical note generation. Specifically, we update the scoring policy of two popular evaluation metrics, by providing a higher weight to the words in the summaries that have a medical meaning:
(i) BARTScore (Yuan et al., 2021) which uses a seq-seq model to calculate the log probability of one text y given another text x, and
(ii) BERTScore (Zhang* et al., 2020) which computes a similarity score for each token in the candidate summary with each token in the reference.
For both metrics, firstly we identify all the words, in the candidate and in the reference summary, which have a clinical meaning defined in UMLS using the MedCAT toolkit (Kraljevic et al., 2021). We then modify the scoring policy of both evaluation metrics to a weighted scoring policy where the weight for all the medical words is higher to provide a stronger incentive to the evaluation model to take in consideration these words during the evaluation of a candidate summary. Specifically, the BARTScore metric is updated to:

$$MedBARTScore = \sum_{i=1}^{m} w_t \log p(y_t|y < t, x) \quad (3)$$

where x is the source sequence and y $(y_1, ..., y_m)$ are the tokens of the target sequence of length m.

We also update the BERTScore for a pair of a reference summary x and candidate summary $\hat{x}$ to:

$$MedBERTScoreP = \frac{1}{|\hat{x}|} \sum_{\hat{x}_i \in \hat{x}} w_x \max_{x_j} x_i^\top \hat{x_j}$$

where, for both metrics, $w = 1$ for all the non-medical words and $w_t = 1 + \alpha$ for all the words with a medical meaning, where α is an additional weight value for these words. After experimenting with different values in the $[0.1, 1.5]$ range, we found that the best α value was 1.0 for the weight policy.

4.3.2 Sliding Window Policy

The main disadvantage of the previously mentioned model-based metrics over the traditional evaluation metrics (e.g., ROUGE) is that they can only encode texts that have length less than the encode-limit of the pre-trained models that are based on. For example, the encode-limit for a BERT-based metric is 512 tokens. However, real-world summaries and clinical notes may contain more than 512 tokens. For instance, our analysis in the CONSULT-FULL dataset shows that 31% of the summaries have more than 512 tokens. We, therefore, create a variation of the BERTScore metric where we use a sliding window approach with the offset size of 100 tokens to encode overlength summaries.

Our sliding window policy is to first split the initial sentence into segments of at most 512 tokens with an overlap size of 100 tokens. Afterward, we calculate the embeddings of these segments independently and concatenate the results to get the original document representation.

This metric will be referred to as MedBERTScore-SP in the Results section.

4.4 Ensemble Metrics

To take advantage of the different perspectives brought by knowledge graph-based metrics, contextual embedding-based metrics, and lexical metrics, we tested different ensembles of normalized metric values. We selected the top-2 performing ensemble metrics for further experiments; $MIST_{Comb1}$ and $MIST_{Comb2}$:

$$Z_m(x) = \frac{x - \mu_m}{\sigma_m} \quad (4)$$

$$MIST_{Comb1}(x) = \frac{1}{3} \sum_{m \in C_1} Z_m(m(x)) \quad (5)$$

$$MIST_{Comb2}(x) = \frac{1}{3} \sum_{m \in C_2} Z_m(m(x)) \quad (6)$$

with $Z_m(x)$ the normalized Z_{score} of a metric m, μ_m the mean value of m over the summaries set, σ_m the standard deviation of m, C_1 = {MIST, ROUGE-1-R, BERTScore} and C_2 = {MIST, ROUGE-1-R, BLEURT}

5 Evaluation Setup

We used the deberta-xlarge-mnli model (He et al., 2021) as the base model for BERTScore and the BLEURT-20 checkpoint for the BLEURT metric, that correlate better with human judgment than the default variants based on recent experiments. For BARTScore metric, we used the BART model that was trained on the ParaBank2 dataset (Hu et al., 2019) which was provided by the authors.

From the designed and tested 50+ metrics and variants (e.g. our new metrics and variants, open-domain metrics, ensemble metrics), we selected the top 21 metrics to study and analyze their performance on the different datasets. The selection was based on the performance of these metrics and their Pearson correlation scores with human judgments on the MTS-DIALOG and the CONSULT-FULL datasets. Our first tests also included open-domain fact-based metrics such as FactCC (Kryściński et al., 2019) (trained on the CNN/DailyMail dataset) and QA metrics such as QUALS (Nan et al., 2021) (developed using XSUM and CNN/DailyMail) but they did not perform well due to the differences between open-domain and clinical questions/answers.

The experiments were performed on one 80GB NVidia A100 GPU.

6 Performance of Evaluation Metrics

We compute the Pearson correlation scores between the automatic metrics and the reference scores. When both manual factual scores (F), hallucination (H), and omission rates (O) are available, we compute an aggregate score:

$$AggregateScore = \frac{1}{4}(2F - H - O) \quad (7)$$

The intuition behind this score is that both omissions (O) and hallucinations (H) are critical criteria but they need to be taken into account in the context of factual correctness (F).

The results on the MTS-DIALOG dataset are presented in Table 2, where the ensemble metric MIST-Comb1 achieved the best correlation with manual scores on Factual F1, Factual Recall, and Omission Rate, with respective correlation values of 0.61, 0.64, and -0.71. The new MedBARTScore metric achieved the best correlation with human assessment for both Factual Precision and Hallucination Rate with 0.46 and -0.46 correlation values.

Table 3 presents the Pearson correlations between the automatic metrics and reference scores on the CONSULT-FACTS dataset. Compared with the results on the MTS dataset, ROUGE-N variants achieved high correlation scores in all categories. In particular, ROUGE-1-R and ROUGE-L-R have the best scores for Factual F1 and Factual Recall. ROUGE-1-F and ROUGE-L-F have the best scores for Factual Precision. ROUGE-1-P and ROUGE-L-P have the best correlations with the Hallucination Rate. BERTScore-R and the ensemble metric MIST-Comb2 achieved the highest correlations with manual scores for the Omission Rate.

On the larger CONSULT-FULL dataset, ROUGE-N results followed a similar pattern on the $\text{CONSULT}_{\text{HPI}}$, $\text{CONSULT}_{\text{ASSESSMENT}}$, $\text{CONSULT}_{\text{EXAM}}$, and $\text{CONSULT}_{\text{RESULTS}}$ subsets, as presented in Table 5, with ROUGE-1-P, ROUGE-2-P, and ROUGE-L-P having the highest correlations with the Omission Rate in the $\text{CONSULT}_{\text{ASSESSMENT}}$ dataset, and the Hallucination Rate in the $\text{CONSULT}_{\text{RESULTS}}$ dataset. This could be explained in part by the fact that the reference notes in the CONSULT-FULL dataset have been created from initial drafts produced by summarization models which increases the likelihood of word overlap.

The fine-tuned ClinicalBLEURT metric achieves the highest correlation scores for the Hallucination Rate in the $\text{CONSULT}_{\text{HPI}}$, $\text{CONSULT}_{\text{EXAM}}$, and $\text{CONSULT}_{\text{RESULTS}}$ datasets. The new medical metrics MedBERTScore-P and MedBERTScore-PS have the highest correlations for Hallucination and Omission Rates on the $\text{CONSULT}_{\text{ASSESSMENT}}$ and $\text{CONSULT}_{\text{EXAM}}$ datasets, respectively.

Table 4 presents the Pearson correlations between the automatic metrics and reference scores on the MEDIQA-RRS dataset, where ROUGE-1-P has the highest correlation with Factual Precision and Hallucination Rate with 0.40 and -0.39. The new MIST metric has the highest correlation scores with Factual Recall and Factual F1 with 0.73 and 0.66, respectively.

Table 6 presents the average scores of the 21 metrics across all datasets. On specific evaluation criteria, the new MedBARTScore metric performed the best on average on correlating with low Hallucinate Rate, with a correlation score of -0.38, and Factual Precision with an average correlation score of 0.45. Both MIST-Comb2 and BERTScore-R have the highest Aggregate Score

Reference / Automatic	↑ Factual P	↑ Factual R	↑ Factual F1	↓ Hallucination	↓ Omission	↑ Aggregate Score
SOTA Metrics						
ROUGE-1-P	0.14	-0.09	-0.04	-0.16	0.06	0.00
ROUGE-1-R	0.10	0.57	0.53	0.02	-0.60	0.41
ROUGE-1-F	0.13	0.39	0.40	-0.08	-0.44	0.33
ROUGE-2-P	0.12	0.05	0.07	-0.12	-0.12	0.10
ROUGE-2-R	0.12	0.34	0.34	-0.09	-0.39	0.29
ROUGE-2-F	0.12	0.28	0.29	-0.10	-0.33	0.25
ROUGE-L-P	0.13	-0.08	-0.05	-0.15	0.07	0.00
ROUGE-L-R	0.10	0.56	0.51	0.02	-0.58	0.40
ROUGE-L-F	0.13	0.38	0.38	-0.08	-0.41	0.31
BERTScore-P	0.10	0.11	0.15	-0.18	-0.23	0.18
BERTScore-R	0.07	<u>0.62</u>	<u>0.59</u>	0.02	**-0.71**	**0.47**
BERTScore-F	0.09	0.44	0.45	-0.08	-0.56	0.38
BLEURT	0.11	0.48	0.47	-0.08	-0.59	0.40
BARTScore	<u>0.37</u>	0.09	0.19	<u>-0.34</u>	-0.26	0.25
New Metrics						
MedBERTScore-P	0.28	-0.16	-0.02	-0.27	-0.32	0.14
MedBERTScore-SP	0.28	-0.16	-0.02	-0.27	-0.32	0.14
MedBARTScore	**0.46**	0.13	0.24	**-0.46**	-0.27	0.30
ClinicalBLEURT	0.19	0.22	0.19	-0.06	-0.20	0.16
MIST	0.02	0.46	0.45	0.08	-0.51	0.33
MIST-Comb1	0.08	**0.64**	**0.61**	0.05	**-0.71**	**0.47**
MIST-Comb2	0.09	0.60	0.58	0.01	-0.68	0.46

Table 2: **MTS-DIALOG**: Pearson's correlation coefficients between the automatic and manual scores. Best results are highlighted in bold and second best are underlined.

Reference / Automatic	↑ Factual P	↑ Factual R	↑ Factual F1	↓ Hallucination	↓ Omission	↑ Aggregate Score
SOTA Metrics						
ROUGE-1-P	0.63	0.32	0.50	**-0.73**	-0.46	0.55
ROUGE-1-R	0.59	**0.80**	**0.79**	-0.39	-0.84	0.70
ROUGE-1-F	**0.70**	0.70	0.78	-0.55	-0.79	**0.73**
ROUGE-2-P	0.56	0.33	0.45	-0.60	-0.43	0.48
ROUGE-2-R	0.55	0.73	0.71	-0.39	-0.78	0.65
ROUGE-2-F	0.62	0.62	0.68	-0.49	-0.70	0.64
ROUGE-L-P	0.63	0.33	0.51	**-0.73**	-0.47	0.56
ROUGE-L-R	0.60	**0.80**	**0.79**	-0.40	-0.84	0.71
ROUGE-L-F	**0.70**	0.70	0.78	-0.56	-0.79	**0.73**
BERTScore-P	0.62	0.47	0.58	-0.56	-0.60	0.58
BERTScore-R	0.60	**0.80**	0.78	-0.37	**-0.85**	0.70
BERTScore-F	0.66	0.69	0.74	-0.49	-0.79	0.69
BLEURT	0.61	0.67	0.71	-0.49	-0.76	0.67
BARTScore	0.61	0.34	0.51	-0.66	-0.41	0.52
New Metrics						
MedBERTScore-P	0.63	0.47	0.59	-0.57	-0.60	0.59
MedBERTScore-SP	0.63	0.47	0.59	-0.57	-0.61	0.59
MedBARTScore	0.61	0.35	0.51	-0.67	-0.42	0.53
ClinicalBLEURT	0.04	0.15	0.08	0.09	-0.15	0.05
MIST	0.08	0.44	0.31	0.08	-0.44	0.25
MIST-Comb1	0.48	0.78	0.72	-0.26	-0.81	0.63
MIST-Comb2	0.53	**0.80**	0.75	-0.33	**-0.85**	0.67

Table 3: **CONSULT-FACTS**: Pearson's correlation coefficients between the automatic and manual scores.

Automatic \ Reference	↑ Factual P	↑ Factual R	↑ Factual F1	↓ Hallucination	↓ Omission	↑ Aggregate Score
SOTA Metrics						
ROUGE-1-P	**0.40**	-0.10	0.00	**-0.39**	-0.30	0.17
ROUGE-1-R	0.22	0.55	0.57	-0.22	**-0.74**	0.53
ROUGE-1-F	0.31	0.39	0.47	-0.31	-0.69	0.49
ROUGE-2-P	0.34	0.04	0.10	-0.32	-0.36	0.22
ROUGE-2-R	0.20	0.46	0.47	-0.18	-0.66	0.45
ROUGE-2-F	0.25	0.37	0.41	-0.23	-0.63	0.42
ROUGE-L-P	0.37	-0.11	-0.02	-0.36	-0.29	0.15
ROUGE-L-R	0.20	0.54	0.55	-0.21	-0.73	0.51
ROUGE-L-F	0.29	0.38	0.44	-0.29	-0.69	0.47
BERTScore-P	0.31	-0.07	0.03	-0.30	-0.33	0.17
BERTScore-R	0.17	0.56	0.58	-0.21	-0.73	0.53
BERTScore-F	0.29	0.32	0.38	-0.30	-0.64	0.43
BLEURT	0.33	0.46	0.51	-0.29	-0.69	0.50
BARTScore	0.38	0.15	0.23	-0.37	-0.39	0.31
New Metrics						
MedBERTScore-P	0.32	-0.04	0.05	-0.31	-0.35	0.19
MedBERTScore-SP	0.32	-0.04	0.05	-0.31	-0.35	0.19
MedBARTScore	0.29	0.03	0.13	-0.28	-0.30	0.21
ClinicalBLEURT	0.27	0.11	0.10	-0.26	-0.09	0.14
MIST	0.11	**0.73**	**0.66**	-0.10	-0.52	0.49
MIST-Comb1	0.18	0.67	**0.66**	-0.19	-0.72	**0.56**
MIST-Comb2	0.24	0.64	0.65	-0.23	-0.72	**0.56**

Table 4: **MEDIQA-RRS**: Pearson's correlation coefficients between the automatic and manual scores. Best results are highlighted in bold and second best are underlined.

	HPI Section		Assessment Section		Exam Section		Results Section	
	Hallucination	Omission	Hallucination	Omission	Hallucination	Omission	Hallucination	Omission
SOTA Metrics								
ROUGE-1-P	-0.23	-0.21	-0.45	**-0.30**	-0.19	-0.17	-0.18	-0.23
ROUGE-1-R	-0.20	-0.18	-0.33	-0.21	-0.19	-0.15	-0.09	-0.19
ROUGE-1-F	-0.24	-0.22	-0.37	-0.25	-0.21	-0.18	-0.11	-0.20
ROUGE-2-P	-0.25	-0.21	-0.46	**-0.30**	-0.24	-0.17	-0.18	**-0.24**
ROUGE-2-R	-0.22	-0.19	-0.37	-0.24	-0.21	-0.18	-0.12	-0.23
ROUGE-2-F	-0.25	-0.21	-0.41	-0.27	-0.23	-0.18	-0.13	-0.22
ROUGE-L-P	-0.23	-0.21	-0.45	**-0.30**	-0.20	-0.17	-0.18	-0.23
ROUGE-L-R	-0.20	-0.18	-0.33	-0.21	-0.19	-0.15	-0.09	-0.20
ROUGE-L-F	-0.24	-0.22	-0.38	-0.25	-0.21	-0.18	-0.11	-0.20
BERTScore-P	-0.23	-0.21	-0.46	-0.27	-0.22	**-0.20**	-0.12	-0.23
BERTScore-R	-0.22	-0.18	-0.31	-0.19	-0.22	-0.16	-0.05	-0.16
BERTScore-F	-0.24	-0.20	-0.39	-0.23	-0.22	-0.19	-0.08	-0.20
BLEURT	-0.20	-0.20	-0.37	-0.23	-0.18	-0.13	-0.10	-0.21
BARTScore	-0.26	-0.21	-0.42	-0.29	-0.27	-0.19	-0.16	-0.21
New Metrics								
MedBERT-P	-0.23	-0.21	**-0.47**	-0.27	-0.21	**-0.20**	-0.10	-0.23
MedBERT-SP	-0.23	-0.22	**-0.47**	-0.28	-0.22	**-0.20**	-0.10	-0.23
MedBART	-0.26	**-0.23**	-0.46	-0.29	-0.25	-0.19	-0.16	-0.23
ClinicalBLEURT	**-0.30**	-0.19	-0.29	-0.25	**-0.31**	-0.18	**-0.25**	-0.19
MIST	-0.07	-0.05	-0.12	-0.16	-0.09	-0.09	0.02	-0.02
MIST-Comb1	-0.18	-0.15	-0.27	-0.20	-0.19	-0.15	-0.04	-0.13
MIST-Comb2	-0.18	-0.17	-0.30	-0.22	-0.18	-0.15	-0.06	-0.15

Table 5: **CONSULT-FULL**: Pearson's correlation coefficients between the automatic and manual scores on the $\text{CONSULT}_{\text{HPI}}$, $\text{CONSULT}_{\text{ASSESSMENT}}$, $\text{CONSULT}_{\text{EXAM}}$, and $\text{CONSULT}_{\text{RESULTS}}$ datasets. Unlike Tables 2-4 which present the fact-based results, here, Hallucination and Omission are measured at the key-phrase level.

SOTA Metrics	↑ Factual P	↑ Factual R	↑ Factual F1	↓ Hallucination	↓ Omission	↑ Aggregate Score
ROUGE-1-P	0.39	0.04	0.15	-0.35	-0.23	0.22
ROUGE-1-R	0.30	0.64	0.63	-0.20	-0.46	0.48
ROUGE-1-F	0.38	0.49	0.55	-0.27	-0.43	0.45
ROUGE-2-P	0.34	0.14	0.21	-0.31	-0.27	0.25
ROUGE-2-R	0.29	0.51	0.51	-0.23	-0.41	0.41
ROUGE-2-F	0.33	0.42	0.46	-0.26	-0.39	0.39
ROUGE-L-P	0.38	0.04	0.15	-0.34	-0.23	0.22
ROUGE-L-R	0.30	0.63	0.62	-0.20	-0.45	0.47
ROUGE-L-F	0.37	0.49	0.53	-0.27	-0.42	0.44
BERTScore-P	0.34	0.17	0.25	-0.30	-0.31	0.28
BERTScore-R	0.28	0.66	0.65	-0.19	**-0.47**	**0.49**
BERTScore-F	0.35	0.48	0.52	-0.26	-0.44	0.44
BLEURT	0.35	0.54	0.56	-0.25	-0.44	0.45
BARTScore	**0.45**	0.19	0.31	-0.37	-0.29	0.32
New Metrics						
MedBERTScore-P	0.41	0.09	0.20	-0.32	-0.33	0.26
MedBERTScore-SP	0.41	0.09	0.20	-0.32	-0.33	0.27
MedBARTScore	**0.45**	0.17	0.29	**-0.38**	-0.28	0.31
ClinicalBLEURT	0.17	0.16	0.13	-0.08	-0.15	0.12
MIST	0.07	0.55	0.47	-0.02	-0.28	0.31
MIST-Comb1	0.25	**0.70**	**0.66**	-0.15	-0.45	0.48
MIST-Comb2	0.29	0.68	**0.66**	-0.18	-0.46	**0.49**

Table 6: Average scores of the 21 automatic metrics across all datasets. Best results are highlighted in bold and second best are underlined.

of 0.49 followed by MIST-Comb1 and ROUGE-1-R. The same set of metrics has similar positive results on the MTS-DIALOG, MEDIQA-RRS, and CONSULT-FACTS datasets. Using the dataset-specific Aggregate Score, we observe that MIST-Comb1, MIST-Comb2, BERTScore-R, and ROUGE-1-R perform well on Factual correctness while maintaining a stable/good performance on being indicative of lower hallucination and omission rates. These datasets are substantially different from each other: long clinical notes for CONSULT-FACTS (with 214 words/note), concise impression sections from radiology reports for MEDIQA-RRS(with 18 words/summary), and different types of sections from different specialities for MTS-DIALOG (15 words/summary), which suggests that this set of metrics can be relied upon for the evaluation of clinical note generation.

7 Conclusion

While finding a relevant and generic evaluation metric for NLG systems remains a challenging task, our study shows that the solution to the problem is likely to be domain- and task-specific. In particular, metrics that did well on capturing factual accuracy did not necessarily capture critical aspects in clinical note generation such as hallucinations and key medical fact omissions. Our experiments also show that language-model based metrics and metric ensembles can outperform SOTA N-gram based measures such as ROUGE when reference summaries are not biased. The extensive measurements and new metrics evaluated in this paper are valuable for guiding decisions on which metrics will be most effective for researchers to use going forward in their Automated Medical Note Generation scenarios.

Limitations

While our research and empirical results support specific evaluation metrics for the task of clinical note generation according to a given evaluation criteria, more results, including testing on additional datasets are needed to further validate these findings. Our manual annotations followed clear and structured guidelines, but could still contain some level of annotator bias and have an average Pearson inter-annotator-agreement of 0.67 (Tables 7 and 8).

Ethics Statement

No protected health information will be released with the created annotations. Annotators were paid a fair hourly wage consistent with the practice of the state of hire.

Acknowledgements

We thank the anonymous reviewers and area chair for their valuable feedback. We also thank our annotators for their help with the manual evaluation.

References

Brian G. Arndt, John W. Beasley, Michelle D. Watkinson, Jonathan L. Temte, Wen-Jan Tuan, Christine A. Sinsky, and Valerie J. Gilchrist. 2017. Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis P. Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021*, pages 74–85. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Liwei Cai and William Yang Wang. 2018. KBGAN: adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1470–1480. Association for Computational Linguistics.

Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. 2022. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6234–6247, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 6038–6052. Association for Computational Linguistics.

Esin Durmus, He He, and Mona T. Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5055–5070. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021b. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3381–3392. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 169–182. Association for Computational Linguistics.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale,

diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Stephan Kudyba. 2010. *Healthcare Informatics: Improving Efficiency and Productivity*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, NTCIR-4, National Center of Sciences, Tokyo, Japan, June 2-4, 2004*. National Institute of Informatics (NII).

Chin-Yew Lin. 2004b. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Donald A. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32:281–291.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos-Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5739–5754. Association for Computational Linguistics.

Feng Nan, Cícero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen R. McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6881–6894. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. *CoRR*, abs/2104.13346.

Thomas H. Payne, W. David Alonso, J. Andrew Markiel, Kevin Lybarger, and Andrew A. White. 2018. Using voice to create hospital progress notes: Description of a mobile application and supporting system integrated with a commercial electronic health record. *Journal of Biomedical Informatics*, 77:91–96.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5008–5020. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Shiyue Zhang, David Wan, and Mohit Bansal. 2022. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *CoRR*, abs/2209.03549.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Inter-Annotator Agreements (IAA)

annotations	kappa	f1	f1(tol=1)	f1(tol=2)	pearson
crit-ommissions	0.29	0.48	0.65	0.75	0.75
hallucinations	0.46	0.73	0.87	0.92	0.97
correct-facts	0.12	0.13	0.30	0.40	0.79
incorrect-facts	0.58	0.73	0.90	1.00	0.89

Table 7: Averaged pairwise IAA for the annotation of 20 transcript-section pairs from the CONSULT-FACTS dataset.

annotations	kappa	f1	f1(tol=1)	f1(tol=2)	pearson
crit-ommissions	0.26	0.34	0.66	0.85	0.81
hallucinations	0.36	0.76	0.96	0.98	0.34
correct-facts	0.07	0.16	0.60	0.82	0.76
incorrect-facts	0.06	0.64	0.79	0.90	0.07

Table 8: Averaged pairwise IAA for the annotation of 34 summary-note pairs from the MEDIQA-RRS dataset.

B Finetuning-based Metric: Weight scheme

error_type	original weight	normalized weight
critical	3	1
non-critical	1	$\frac{1}{3}$
spelling/grammar	$\frac{1}{4}$	$\frac{1}{12}$

Table 9: Error score weights used in production for evaluating produced notes during a QA review. The normalized versions of the weights are used in our calculations so that the number of errors will not exceed over 1 per sentence unless there is more than 1 critical error.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.4

C ☑ Did you run computational experiments?

6

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.