Automatic vs. Human Evaluation: Benchmarking 26 Automatic Metrics for Clinical Discharge Summaries

Yu-Wen Chen[1], Anne A.H. de Hond[1], Laura Veerhoek[2], Ruben Peters[2], Karel G.M. Moons[1], Artuur M. Leeuwenberg[1]
[1]Julius Center for Health Sciences and Primary Care,
University Medical Center Utrecht, Utrecht University
[2]Team AI, Operations department, University Medical Center Utrecht
{y.chen-4, a.a.h.dehond, l.p.veerhoek, r.peters-7, k.g.m.moons, a.m.leeuwenberg-15}@umcutrecht.nl

Abstract

Automatic evaluation metrics are widely used to assess the quality of clinical summaries. However, most existing automatic metrics were not designed to capture clinically critical errors and were primarily developed and validated on English data, raising concerns about their applicability in non-English clinical settings. In this work, we examine how well commonly used automatic metrics reflect clinically critical criteria in Dutch. We benchmark 26 evaluation metrics on real-world LLM-generated clinical discharge summaries and correlate their scores with clinically critical errors, including hallucinations, omissions, and trivial facts. Our results show that standard automatic metrics do not capture hallucination errors in this setting. Specific lightweight source-based metrics may provide useful signals for identifying omission-related errors and trivial facts, highlighting their potential for prioritizing human review efforts in real-world settings.

1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated significant potential to automate clinical summarization tasks (Ganzinger et al., 2025; Oliveira et al., 2025; Asgari et al., 2025) while achieving quality comparable to clinician-authored summaries (Williams et al., 2025) and reducing physician burnout (Grolleau et al., 2026). However, the deployment of such systems depends on reliable evaluation, as deficiencies in the summary quality may adversely affect patient care and care transition (Patel and Lam, 2023; de Hond et al., 2024).

Human evaluation remains the gold standard for assessing LLM-generated clinical summaries, with particular emphasis on hallucination, omission, revision, faithfulness/confidence, bias/harm, groundedness, and fluency (Croxford et al., 2025). However, assessing these dimensions requires domain experts, making the process expensive, time-consuming, and difficult to scale. Automatic evaluation metrics have therefore become the evaluation approach of choice in clinical summarization studies. Most clinical summarization studies report metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2001), METEOR (Banerjee and Lavie, 2005), or BERTScore (Zhang et al., 2019), even though these metrics were initially developed to assess lexical or semantic similarity rather than clinically meaningful quality.

This raises the question: To what extent do commonly used automatic metrics capture the quality dimensions that matter in clinical summarization? Previous studies in both clinical and non-clinical domains have shown that surface-based automatic metrics exhibit poor to moderate correlation with human judgments and do not capture the critical factuality and faithfulness issues (i.e., statements in the generated summaries are not supported by the source document or are factually incorrect) (Sai et al., 2020; Fabbri et al., 2021; Gehrmann et al., 2023; Ben Abacha et al., 2023; Adams et al., 2023; Hegselmann et al., 2024). These limitations are particularly concerning in the clinical domain, where LLMs may generate unsupported claims about medical conditions and incorrect medical advice (Hegselmann et al., 2024), or omit clinically important medical facts and diagnostic decisions (Croxford et al., 2025), both of which can adversely affect patient safety. Nevertheless, because of their efficiency and scalability, automatic metrics remain predominant and omnipresent in model development and benchmarking. Understanding whether they can serve as reliable indicators of human evaluation, or at least help prioritize summaries requiring expert review, is therefore an important methodological challenge.

This challenge is specifically pronounced for non-English text, for which applicable metrics remain scarce, as most automatic metrics are

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 544–558
October 17-21, 2026 ©2026 Association for Computational Linguistics

language-specific and trained almost primarily on English data. Additionally, the existing evidence on the correlation between automatic and human evaluation in the clinical domain is almost entirely limited to English (Croxford et al., 2024; Moramarco et al., 2022). As multilingual LLMs are increasingly adopted for clinical applications in languages such as Dutch, German, Korean, and Kinyarwanda (Ganzinger et al., 2025; Mehri et al., 2026; Song et al., 2025; Rutunda et al., 2026), it is unclear whether conclusions drawn from English evaluation studies generalize across languages given the linguistic variation and differences in clinical documentation practices.

In this study, we aim to understand whether widely used automatic metrics can provide a meaningful indication of clinical quality and whether their performance is comparable across English and Dutch. We deliberately focus on well-established automatic metrics and do not include LLM-based metrics (e.g., question answering- or question generation-based approaches, or LLM-as-a-judge) in this work, as they represent a separate evaluation paradigm in which generative models are used to directly approximate human assessment rather than computing fixed, surface-level statistics. In addition, from a practical perspective, we aim to investigate the performance of automatic metrics that are easier to implement, monitor, and widely adopted than LLM-based metrics in real-world clinical settings. Our contributions are as follows:

- A comprehensive benchmark of 26 automatic evaluation metrics on a real-world Dutch clinical summarization dataset with expert annotations of hallucinations, omissions, and trivial facts.
- A systematic analysis of the correlation between automatic metrics and human judgment on clinically critical criteria, offering insights on their potential role as a preliminary quality screening tool prior to human evaluation.
- A comparison of the performance of automatic summarization metrics across an English and Dutch application

2 Related work

Automatic evaluation for clinical summarization. This automatic evaluation task has largely relied on metrics originally developed for general natural language generation (NLG) systems, including ROUGE, BLEU, METEOR, and BERTScore (Schmidtova et al., 2024). These metrics remain evaluation standards in the MultiClinSumm Shared Task (Rodríguez-Ortega et al., 2025; Nentidis et al., 2026) and have been widely adopted in recent studies on clinical summarization (Ganzinger et al., 2025; Tang et al., 2023; Aali et al., 2025).

However, only a limited number of studies (Van Veen et al., 2024; Williams et al., 2025; Fraile Navarro et al., 2025) reported how well these automatic metrics align with human judgments on clinically relevant dimensions such as correctness (i.e., containing less false information) and completeness (i.e., completely capturing important information).

Domain-specific evaluation methods have leveraged LLM-as-a-judge to assess medically critical errors. MED-OMIT (Schumacher et al., 2025) identifies information omissions with importance stratification, while DocLens (Xie et al., 2024) verifies entailment relationships between the source document, generated summary, and reference summary. Despite these advances, the development and validation of LLM-based evaluations have focused almost exclusively on English clinical summarization.

Correlation between automatic metrics and human evaluation. Although widely in use, general automatic summarization metrics are particularly limited in detecting clinically critical errors (Hegselmann et al., 2024; Croxford et al., 2024). Human evaluation, therefore, remains the gold standard evaluation method because it can directly capture clinically relevant criteria that automatic metrics often struggle to capture. Existing evaluation protocols typically rely on expert ratings of criteria such as accuracy, completeness, clarity, and conciseness (Bedi et al., 2025; Mehri et al., 2026) or on fine-grained annotation of clinically relevant errors, including hallucinations and omissions (Asgari et al., 2025; Ben Abacha et al., 2023).

Some prior evaluation work has explicitly examined metric-human correlations for English clinical note summarization and diagnosis generation along those criteria, showing that correlation strength varies across task types (Moramarco et al., 2022; Croxford et al., 2024; Ben Abacha et al., 2023). However, little is known about this metric-human correlation for non-English clinical summarization, particularly on clinically critical criteria such as hallucinations, omissions, and trivial facts. To our

knowledge, this work is the first to investigate correlations between automatic metrics and human judgments on these criteria in non-English clinical summarization.

3 Data

Two datasets are used for our correlation study. Dataset and annotation statistics are provided in Table 1 and Table 2.

3.1 Dutch Clinical Discharge Summary

The first dataset (de Hond et al., 2025) comprises patient records and corresponding discharge summaries from the Intensive Care Unit (ICU), Neonatal ICU (NICU), and Cardiology (CAR) department at the University Medical Center Utrecht, Netherlands, collected in April and May 2025. The data consisted of patient records, with matching physician- and GPT-generated discharge letters[1]. GPT-4-turbo was used to generate the discharge letters, accessed through a secure Azure OpenAI deployment for patient data protection. Three error categories were defined: (1) hallucinations, referring to facts generated in the discharge letter that were not present in the patient record; (2) omissions, referring to missing but relevant facts from the patient record; and (3) trivial facts, referring to non-essential information included in the discharge letter. Annotations were performed by two senior medical students and each error was subsequently verified and stratified according to its clinical severity by at least one physician. Errors considered likely to impact clinical decision-making are marked as "severe", whereas errors considered to have little to no impact on clinical decision-making are marked as "not severe".

Usability was rated by physicians on a 5-point Likert scale based on the statement: “This draft discharge letter can be used by a physician in clinical practice with minimal adjustments.” The total error count was defined as the number of unique and correct annotations across both annotators. Sixteen overlapping examples were used to assess the Spearman correlation in the number of annotated errors between the two annotators, which may act as an approximate upper bound for the automatic metrics (Table 2).

[1]Out of 318 examples, 145 were removed due to incomplete generated discharge letters, followed by a further 48 excluded due to incomplete annotations, resulting in 97 complete examples.

Dutch Clinical Discharge Summary (N=97)

Document	Mean (SD)	Median	IQR
Patient record	4581 (5018.2)	2974	1763-5435
Generation	320.3 (100.4)	304	240-384
Human ref	243.9 (116.3)	226	169-298

Annotations: Hallucination, Omission, Trivial Facts; Usability (Likert 1–5)

Hallucinations-Generated-DI (N=100)

Document	Mean (SD)	Median	IQR
Patient record	394 (153)	361.5	284-469
Generation	149 (46)	157	114-180
Human ref	153 (38)	134	124-180

Annotations: Hallucination

Table 1: Mean, median, and interquartile range (IQR) of the number of tokens for each document type in both datasets.

	Halluc.		Omission		Triv.		MIMIC
	All	Sev.	All	Sev.	All	Sev.	Halluc.
ρ	0.43	0.48	0.6	0.7	0.8	0.85	–
Mean	0.94	0.73	4.01	2.95	2.41	1.72	1.14
Median	1	0	2	1	2	1	0
SD	1	1	5	4	2	2	2
Max	11	11	22	22	9	9	7
IQR	0–1	0–1	0–6	0–4	1–4	0–3	0–1
%0	48	57	26	37	20	34	50

Table 2: Spearman’s correlation (ρ) between annotated error counts of the two annotators and summary statistics for error count (“All” vs. severe “Sev.” errors). "%0" denotes the percentage of examples with zero annotated errors.

3.2 Hallucinations-Generated-DI

The second dataset (Hegselmann et al., 2024) consists of patient records (i.e., preprocessed Brief Hospital Course (BHC) from MIMIC-IV-Note v2.2), doctor-written summaries (i.e., preprocessed Discharge Instructions from MIMIC-IV-Note v2.2), and LLM-generated summaries[2]. Hallucinations were labeled by two trained medical students and further categorized as unsupported, contradicted, or incorrect facts. The total number of hallucinations per summary across these categories was used in the analyses.

[2]Five LLMs were used to generate discharge summaries: Llama2-70b, Llama2-70b-orig, Llama2-70b-cleaned, gpt4-zero-shot, gpt4-orig, gpt4-cleaned. The cleaned suffix indicates the model was trained or prompted with hallucination-free examples. Twenty preprocessed BHC cases were used as input to five models, yielding a total of 100 generated summaries.

4 Method

4.1 Automatic Metrics

The selection of metrics was based on previous literature reviews and commonly used NLG evaluation metrics (Sai et al., 2020; Schmidtova et al., 2024; de Hond et al., 2026). We include metrics that (1) are applicable to non-English data[3], (2) do not rely on LLM-based methods, and (3) have publicly available implementations. A total of 26 automatic metrics are identified and evaluated against counts of human-annotated errors. We categorize metrics according to the criteria they measure and the data required for the metric calculation (Table 3). Detailed descriptions of all evaluation metrics are provided in Appendix B.

4.1.1 Reference-based

This set of metrics requires a reference standard, often a human-written clinical summary, for comparison. Metrics are grouped into three main categories based on the aspect of quality they capture. **Lexical similarity** metrics measure surface-level overlap between generated and reference summaries using n-gram and character-level matching, as well as token-level distributional comparisons. **Semantic similarity** metrics capture meaning equivalence beyond exact lexical matches, which is particularly important in abstractive summarization where the same content may be expressed using different wording or structure. **Simplification** metrics evaluate the goodness of words that are added, deleted, and kept by the system (i.e., LLMs) and are well-suited to abstractive summarization involving compression and abstraction.

4.1.2 Source-based

Source-based metrics use both the source text (i.e., patient record) and generated summaries as input. These metrics may be more practical in real-world settings as they do not depend on human-written references, which are often costly to obtain in clinical settings. **Extractiveness** measures the extent to which the summary reuses text spans (i.e., extractive fragments) from the source. **Degree of abstraction** captures the proportion of novel tokens not present in the source. **Uniformity** evaluates whether information in the summary is evenly drawn from across the source document, addressing layout bias where important content is concentrated in specific sections (e.g., early sentences).

4.1.3 Generation only

This set of metrics quantifies the intrinsic linguistic quality of the generated text without additional references. **Redundancy** measures repetition in generated text, computed as the average ROUGE-L across all pairs of distinct sentences. **Semantic coherence** captures the logical flow and sequential consistency between sentences in the generation. **Diversity** measures lexical variation in terms of word and n-gram usage. **Fluency** is quantified using perplexity under a fixed reference language model, EuroLLM (Martins et al., 2025). Another fluency metric is readability, which reflects the ease or difficulty of text comprehension.

4.2 Statistical analysis of automatic metrics against human-annotated errors

Correlations are computed between each automatic metric and the total count of annotated hallucinations, omissions, trivial facts, and usability scores. Because error counts are discrete with many tied observations and relationships are not assumed to be linear, Kendall's tau correlation is used as the primary measure of correlation. Metrics are ranked based on the absolute coefficient $|\tau_b|$, while the sign of τ_b is used for the direction of associations. Spearman's correlations are also computed.

To further examine whether metrics with stronger correlations to error annotations also distinguish between summaries with and without errors, Mann–Whitney U tests are used. For each error category, summaries are divided into two groups: error-free and error-present summaries containing at least one error. The three metrics with the highest absolute correlation coefficients $|\tau_b|$ for each error category (i.e., the top three metrics in each column of Table 4) are selected, and their metric scores are compared between the two groups. To account for multiple comparisons, p-values are adjusted using the Benjamini-Hochberg procedure to control the false discovery rate (FDR).

5 Result

Table 4 reports Kendall's tau correlations between automatic metrics and human evaluations. Spearman correlation coefficients are reported in Appendix A and show consistent trends. Metrics with the highest absolute correlation coefficients remain identical across error categories in both analyses.

[3]Metrics trained primarily in general-domain English text (e.g., BLEURT (Sellam et al., 2020), BARTScore (Yuan et al., 2021)) are not included

Required Reference	Evaluation Criteria	Method	Metrics
Reference-based	Lexical similarity	Word-level overlap	BLEU (Papineni et al., 2001), METEOR (Banerjee and Lavie, 2005), ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, ROUGE-SU (Lin, 2004), NIST (Doddington, 2002)
		Character-level overlap	chrF (Popović, 2015), chrF++ (Popović, 2017)
		Distribution comparison	Jensen–Shannon divergence (Lin, 1991)
	Semantic similarity	Embedding-based similarity	BERTScore (Zhang et al., 2019), WEEM4TS (Hailu et al., 2020), MoverScore (Zhao et al., 2019), Word Mover's Distance (WMD) (Kusner et al., 2015), Sentence Mover Similarity (SMS) (Clark et al., 2019), Sentence-BERT (Reimers and Gurevych, 2019)
	Simplification	Edit-based comparison	SARI(Xu et al., 2016)
Source-based	Extractiveness	N-gram/factoid ratio	Coverage, Compression, Density (Grusky et al., 2018), Abstractivity (Bommasani and Cardie, 2020)
	Degree of abstraction	Source–generation ratio	Abstraction n-grams (Tang et al., 2023)
	Uniformity	Entropy-based	Uniformity (Koh et al., 2023)
Generation-only	Redundancy	Self-overlap	ROUGE-L on generation (Bommasani and Cardie, 2020)
	Semantic coherence	Language model probability	BERT-based sentence-level probability (Bommasani and Cardie, 2020)
	Diversity	Lexical diversity	Distinct-1, Distinct-2 (Li et al., 2016), Distinct-3, MTLD (McCarthy and Jarvis, 2010), Self-BLEU (Zhu et al., 2018)
	Fluency	Language model	Perplexity (Vinyals and Le, 2015)
		Linguistic features	Flesch Reading Ease Index (Flesch, 1948)

Table 3: Automatic evaluation metrics employed in this study, categorized by required reference type, evaluation criteria, and method.

For hallucinations in the Dutch Clinical Discharge Summary dataset, all metrics exhibit low correlations; none of the absolute coefficients is greater than 0.2. Despite low correlations, two generation-only metrics, Self-BLEU ($|\tau_b| = 0.179$, $n = 97$, $p < .05$) and readability ($|\tau_b| = 0.167$, $n = 97$, $p < .05$), show the largest absolute coefficients with all and severe hallucination errors, respectively. Generation length ($|\tau_b| = 0.169$, $n = 97$, $p < .05$), included as a reference feature rather than an evaluation metric, shows the second-largest absolute coefficient with all hallucination errors.

For omissions, source-based metrics show moderate correlations with omission errors ($0.198 \leq |\tau_b| \leq 0.529$, $n = 97$, $p < .01$), where compression ($|\tau_b| = 0.529$, $n = 97$, $p < .001$) exhibit the strongest correlation with all omissions errors, followed by coverage ($|\tau_b| = 0.486$, $n = 97$, $p < .001$). Three reference-based metrics, ROUGE-W, METEOR, and BLEU unigram, have weaker correlations with omission errors ($0.262 \leq |\tau_b| \leq 0.389$, $n = 97$, $p < .001$).

For trivial facts, reference-based BLEU unigram ($|\tau_b| = 0.393$, $n = 97$, $p < .001$) and source-based compression ($|\tau_b| = 0.348$, $n = 97$, $p < .001$) have the largest absolute coefficients for both severity categories, respectively. Generation length also shows a relatively strong association, particularly for all trivial facts ($|\tau_b| = 0.31$, $n = 97$, $p < .001$). For usability, none of the absolute coefficients is higher than 0.3. The largest correlation is observed for source-based metric uniformity ($|\tau_b| = 0.243$, $n = 97$, $p < .01$).

For the MIMIC dataset, only hallucinations without severity stratification are available for analysis. None of the absolute coefficients is higher than 0.3, with the largest absolute correlation observed for reference-based metric Jensen-Shannon divergence ($|\tau_b| = 0.273$, $n = 100$, $p < .001$), followed by two generation-only metrics, MTLD ($|\tau_b| = 0.268$, $n = 100$, $p < .001$) and readability ($|\tau_b| = 0.256$, $n = 100$, $p < .001$).

The analysis of score differences between error-free and error-present outputs using Mann-Whitney

U tests is presented in Appendix C, considering only the top three metrics with the highest absolute coefficients per column in Table 4. Hallucination-free outputs exhibit significantly higher Self-BLEU scores than outputs containing at least one hallucination (n = 97, FDR-adjusted $p < .05$, considering all severity). Omission-free outputs exhibit significantly lower compression and coverage scores than omission-containing outputs (n = 97, FDR-adjusted $p < .001$, considering all severity). Outputs free of trivial facts have significantly higher compression and BLEU unigram scores than outputs containing trivial facts (n = 97, FDR-adjusted $p < .001$, considering all severity). For the MIMIC dataset, hallucination-free outputs have significantly higher Jensen-Shannon divergence and MTLD scores and lower readability scores than outputs with at least one hallucination (n = 100, FDR-adjusted $p < .01$).

6 Discussion

Across datasets and languages, we observe weak correlations between automatic evaluation metrics and hallucination errors. Readability consistently showed relatively larger correlation coefficients with hallucination errors in both Dutch and English datasets, although these correlations remain weak. These findings suggest that hallucination-related quality issues are not adequately captured by standard automatic metrics, consistent with previous work demonstrating that similarity-based metrics are poorly aligned with factuality and faithfulness in abstractive summarization (Maynez et al., 2020). This discrepancy is expected, as hallucination and textual similarity measure fundamentally different aspects of summary quality. A high similarity score does not necessarily indicate a factually faithful summary. Our findings further question the suitability of traditional automatic evaluation metrics for assessing clinical summarization systems when factual correctness is a primary concern.

In contrast, source-based metrics show greater promise for identifying omission-related and trivial facts quality issues. Compression and coverage exhibit moderate correlations with omission errors in the Dutch Clinical Discharge Summary dataset. Higher compression and coverage values are associated with increased omission-related errors, indicating that summaries that compress longer patient records more aggressively are more likely to omit clinically relevant information. In the opposite direction, lightly compressed summaries are more likely to contain trivial, irrelevant information. Generation length, included as a sanity check for length-related confounding, shows a stronger association with trivial facts than with omissions. This indicates that shorter summaries alone do not necessarily imply more omissions; rather, the ratio between the patient records and the generated summaries provides additional information about omissions. In clinical practice, this obvious, yet useful result may prompt a physician to double-check the contents of a clinical summary for omission of important information or inclusion of irrelevant information.

Weak correlations between automatic metrics and usability ratings are observed in the Dutch dataset. This result is perhaps unsurprising, as usability is inherently subjective and depends on human judgment regarding whether a summary is clinically useful. Such judgments likely reflect aspects beyond surface-level linguistic characteristics, making them difficult to capture using text-based features alone.

The highest correlated metrics, such as compression, coverage, BLEU unigram, and readability, have discriminative ability for separating error-free and error-present summaries, highlighting their practical utility for identifying error-present examples for human review. However, the direction of these correlations may flip depending on the nature of the dataset inherent to its purpose. For example, readability is higher for hallucination-free summaries in the Dutch Clinical Discharge Summary dataset but lower in the English MIMIC dataset. This discrepancy could be due to the fact that, in the MIMIC dataset, summaries were intended for laypeople rather than other physicians. The LLMs were prompted to "help the patient understand their medical records" and to "avoid using medical jargon," (Hegselmann et al., 2024) requiring them to translate complex medical information into simplified language. While this approach may improve readability, the simplification process could also increase the risk of hallucinations by encouraging the model to paraphrase or infer information beyond the patient records.

Identifying automatic metrics that generalize across datasets and languages remains challenging. The lack of consistent correlations between automatic metrics and hallucination errors across datasets may be due to the differences in dataset composition in terms of annotation protocols, defi-

		Dutch Clinical Discharge Summary (N=97)							MIMIC (N=100)
		Hallucinations		**Omissions**		**Trivial facts**		**Usability**	**Hallucinations**
Error severity		All	Sev.	All	Sev.	All	Sev.		
Reference-based									
Lexical similarity	BLEU	-0.055	-0.062	0.009	0.035	-0.100	0.066	0.059	0.058
	BLEU (unigram)	-0.046	-0.034	0.330	0.262	**-0.393**	-0.309	0.175	0.177
	BLEU (bigram)	-0.110	-0.104	0.183	0.158	-0.268	-0.146	0.194	0.175
	BLEU (trigram)	-0.063	-0.066	0.117	0.135	-0.207	-0.066	0.110	0.191
	BLEU (4gram)	-0.081	-0.045	0.080	0.110	-0.236	-0.036	0.048	0.014
	chrF+	0.034	-0.064	-0.124	-0.139	0.136	0.174	0.058	-0.100
	chrF++	0.038	-0.060	-0.115	-0.129	0.135	0.168	0.053	-0.113
	JS div.	0.065	0.076	-0.271	-0.181	0.282	0.230	-0.212	**-0.273**
	METEOR	-0.016	-0.027	-0.331	-0.267	0.173	0.289	-0.018	-0.080
	NIST	-0.012	-0.071	0.101	0.060	-0.114	-0.081	0.061	0.037
	ROUGE-s4	-0.095	-0.052	-0.069	-0.058	-0.156	-0.024	0.084	0.012
	ROUGE-su4	-0.086	-0.057	-0.038	-0.043	-0.182	-0.062	0.124	0.019
	ROUGE-W-1.2	-0.072	-0.083	-0.389	-0.367	0.074	0.128	-0.008	-0.033
	ROUGE1	-0.071	-0.069	0.064	0.024	-0.229	-0.149	0.206	0.014
	ROUGE2	-0.063	-0.042	-0.055	-0.035	-0.109	0.018	0.051	-0.003
	ROUGEL	-0.066	-0.102	-0.138	-0.203	-0.154	-0.096	0.130	0.024
Semantic similarity	BERTScore (F1)	-0.032	-0.034	0.219	0.164	-0.149	-0.146	0.214	0.043
	BERTScore (Precision)	-0.020	-0.013	0.255	0.190	-0.187	-0.174	0.206	0.158
	BERTScore(Recall)	-0.033	-0.041	0.160	0.114	-0.129	-0.124	0.210	-0.079
	WEEM4TS	-0.036	-0.061	0.163	0.111	-0.079	-0.097	0.102	-0.021
	MoverScore	-0.092	-0.151	0.060	-0.034	-0.148	-0.148	0.087	0.024
	WMD	0.111	0.102	-0.112	-0.130	0.032	0.030	-0.166	0.091
	Sentence-BERT	0.047	-0.056	0.184	0.124	-0.087	-0.114	0.210	0.218
	SMS	-0.079	0.025	0.194	0.199	-0.195	-0.082	0.062	-0.133
Simplification	SARI	-0.021	-0.029	-0.008	-0.015	-0.087	-0.015	0.149	-0.030
Source-based									
Extractiveness	Compression	0.090	0.122	**0.529**	**0.446**	-0.322	**-0.348**	0.147	0.165
	Coverage	0.055	0.048	0.486	0.434	-0.246	-0.195	0.137	-0.043
	Density	0.021	0.003	0.323	0.362	-0.138	-0.053	0.075	-0.076
	Abstractivity	0.121	0.074	-0.262	-0.267	0.293	0.156	-0.032	-0.040
Degree of abstraction	Abstraction (unigram)	-0.028	-0.006	-0.525	-0.434	0.269	0.241	-0.149	-0.001
	Abstraction (bigram)	-0.039	-0.028	-0.370	-0.412	0.122	0.069	-0.082	0.126
Uniformity	Uniformity	0.049	0.026	0.268	0.198	-0.203	-0.207	**0.243**	-0.023
Generation only									
Redundancy	ROUGE1 (red.)	0.023	-0.024	0.104	-0.009	-0.025	-0.137	0.041	0.102
	ROUGE2 (red.)	0.140	0.098	0.134	0.067	-0.115	-0.147	0.041	0.165
	ROUGEL (red.)	0.040	-0.008	0.127	0.009	-0.055	-0.171	0.038	0.169
Semantic coherence	Coherence	0.108	0.071	0.190	0.183	-0.046	-0.109	0.051	-0.153
Diversity	Distinct-1	0.032	0.017	-0.057	0.005	0.175	0.145	0.053	-0.098
	Distinct-2	0.041	0.027	-0.053	0.013	0.196	0.159	0.047	-0.181
	Distinct-3	0.052	0.036	-0.054	0.011	0.203	0.163	0.047	-0.153
	MTLD	-0.055	0.011	-0.073	0.006	-0.062	0.045	-0.096	-0.268
	Self-BLEU	**-0.179**	-0.136	-0.041	-0.015	0.003	0.078	0.050	0.073
Fluency	Perplexity	-0.048	-0.031	0.053	0.033	-0.021	-0.061	0.016	-0.251
	Readability	-0.125	**-0.167**	-0.255	-0.274	0.094	0.075	-0.076	0.256
Others									
	Generation length	0.169	0.106	-0.045	-0.020	0.310	0.197	0.014	-0.150

Table 4: Kendall's tau correlations between human evaluation criteria and automatic metrics across datasets. Bold values indicate the highest absolute correlation coefficient within each column.

nitions of hallucination, or dataset characteristics. The relatively small sample size may lead to less precise estimates of the coefficients and should be considered when interpreting their magnitude. Furthermore, the lower inter-annotator agreement for hallucination errors may also affect the reliability of the correlations. This suggests that the results may not be stable and should be interpreted with caution.

From a practical perspective, our findings highlight the continued value of simple, computationally inexpensive metrics such as compression, coverage, and readability. Although none of these metrics alone provides a reliable assessment of summary quality, they do not require costly human-written reference summaries, making them attractive for real-world assessment scenarios and continuous monitoring of LLM-based applications. Rather than serving as standalone quality indicators, these lightweight metrics could be used to flag potentially problematic summaries prioritized for human review, thereby supporting more efficient quality assurance in real-world settings.

7 Conclusion

We benchmark 26 evaluation metrics on real-world LLM-generated clinical discharge summaries in both Dutch and English datasets, assessing their correlations with clinically critical errors, including hallucinations, omissions, and trivial facts. Across datasets and languages, we find no consistent strong correlations between standard automatic metrics and hallucinations. In contrast, we show that simple source-based metrics exhibit moderate correlations with omissions and trivial facts despite their lightweight design. Overall, our findings challenge the adequacy of current automatic evaluation practices for hallucination detection in clinical summarization and highlight the potential of simple source-based metrics to prioritize human review for clinical summaries that may contain errors.

Limitations: LLM-based metrics were deliberately not included in the analysis, as we focused on metrics that are readily developed and can be easily monitored in real-world clinical settings in this work. In future work, we plan to evaluate discharge summaries across multiple quality dimensions using both open and proprietary LLMs as judges on Dutch datasets. Our analysis is restricted to largely language-agnostic automatic metrics. Existing factuality evaluation approaches are predominantly English-centric and rely on QA- or NLI-based models that do not generalize reliably to Dutch clinical text. This highlights the need for factuality evaluation methods for non-English clinical summarization. Although many metric implementations are publicly available, discrepancies in preprocessing, tokenization, and input formatting may reduce reproducibility and comparability across studies. Standardized and transparently reported evaluation protocols would improve the comparability of automatic evaluations.

Supplementary Materials Availability Statement: Source code is available on Github[4]. The Dutch Clinical Discharge Summary dataset cannot be made available because it contains private patient data. Hallucinations-Generated-DI is available through PhysioNet[5].

[4] https://github.com/yuwenchen31/automatic_evaluation_clinical_summary

[5] https://physionet.org/content/ann-pt-summ/1.0.1/

Acknowledgements: KGMM receives research funding from the Dutch Research Council (2026/KIC/02228102; KICH3.LTP.24.003). YWC, AdH and AML receive research funding from the Netherlands Organisation for Health Research and Development (11510012410002).

References

Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S Tehrani, Jangwon Kim, and Akshay S Chaudhari. 2025. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association*, 32(3):470–479.

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A Meta-Evaluation of Faithfulness Metrics for Long-Form Hospital-Course Summarization. *Proceedings of Machine Learning Research*, 219:2–30.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, and 62 others. 2025. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. *arXiv preprint*. ArXiv:2505.23802 [cs].

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An Investigation of Evaluation Methods in Automatic Medical Note Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic Evaluation of Summarization Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Emma Croxford, Yanjun Gao, Brian Patterson, Daniel To, Samuel Tesch, Dmitriy Dligach, Anoop Mayampurath, Matthew M. Churpek, and Majid Afshar. 2024. Development of a Human Evaluation Framework and Correlation with Automated Metrics for Natural Language Generation of Medical Diagnoses. Pages: 2024.03.20.24304620.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2025. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2(1):6.

Anne de Hond, Ruurd Kuiper, Isa Spiero, Yu-Wen Chen, Demy Idema, Johanna Damen, Maarten van Smeden, Karel Moons, and Artuur Leeuwenberg. 2026. Validation of large language models in healthcare: A scoping review. Manuscript under review.

Anne de Hond, Tuur Leeuwenberg, Richard Bartels, Marieke van Buchem, Ilse Kant, Karel GM Moons, and Maarten van Smeden. 2024. From text to treatment: the crucial role of validation for generative large language models in health care. *The Lancet Digital Health*, 6(7):e441–e443.

Anne de Hond, Laura Veerhoek, Ruben Peters, Daniel Claes Vijlbrief, David Sprenkeler, Dino Ahmetagic, Sanul Wilmer, Pieter van der Hoeven, Marjon Smit, Marc Platenkamp, Karel GM Moons, Maarten van Smeden, Artuur M. Leeuwenberg, and Ilse M. J. Kant. 2025. Hallucinations, Omissions, and Usability: A Comparison between GPT-Generated and Physician Discharge Letters for Clinical Use.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* -, page 138, San Diego, California. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

R. Flesch. 1948. A new readability yardstick. *The Journal of Applied Psychology*, 32(3):221–233.

David Fraile Navarro, Enrico Coiera, Thomas W. Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15(1):1195.

Matthias Ganzinger, Nicola Kunz, Pascal Fuchs, Cornelia K. Lyu, Martin Loos, Martin Dugas, and Thomas M. Pausch. 2025. Automated generation of discharge summaries: leveraging large language models with clinical data. *Scientific Reports*, 15(1):16466.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166.

François Grolleau, April S. Liang, Timothy Keyes, Stephen P. Ma, Thomas Lew, Tridu R. Huynh, Natasha Steele, Philip Chung, Paige Qin, Gowri Chandra, Stephanie F. Wang, Evan Mullen, Lauren Carpenter, Mita Hoppenfeld, Matthew Morrin, Baffour A. Kyerematen, Nerissa Ambers, Nikesh Kotecha, Emily Alsentzer, and 4 others. 2026. Physician-Reported Safety Outcomes of AI-Generated Hospital Course Summaries. *JAMA network open*, 9(5):e2616556.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Tulu Tilahun Hailu, Junqing Yu, and Tessfu Geteye Fantaye. 2020. A Framework for Word Embedding Based Automatic Text Summarization and Evaluation. *Information*, 11(2).

Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2024. A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models. *arXiv preprint*. ArXiv:2402.15422 [cs].

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2023. An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics. *ACM Computing Surveys*, 55(8):1–35. ArXiv:2207.00939 [cs].

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966. PMLR.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, 255:53–62.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Tarannom Mehri, Tun Nadalini, Anne H. Hoekman, Tom P. Van Der Laan, Katerina Kagialari, Robert K. Wagner, Job N. Doornberg, Rosanne C. Schoonbeek, Charlotte M.H.H.T. Bootsma-Robroeks, M. Aalderink, R. Van Den Berg, M.T.P. Besouw, A.V. Biere, F.A.J.A. Bodewes, A.L. Boerboom, M.A.J. Borgdorff, M.H. De Borst, M. Bouhuys, B.R. Brandsema, and 48 others. 2026. Assessing the quality of AI-generated and physician-written discharge summaries: evaluation of an EHR-integrated tool in a Dutch academic hospital. *eBioMedicine*, 127:106247.

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)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodriguez-Ortega, Eduard Rodriguez-Lopez, Natalia Loukachevitch, Igor Rozhkov, Elena Tutubalina, Grigorios Tsoumakas, George Giannakoulas, Dimitris Dimitriadis, Alexandra Bekiaridou, Athanasios Samaras, Vasiliki Patsiou, Giorgio Maria Di Nunzio, Nicola Ferro, Stefano Marchesin, Marco Martinelli, and 2 others. 2026. BioASQ at CLEF2026: The Fourteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge. In *Advances in Information Retrieval*, pages 315–324, Cham. Springer Nature Switzerland.

Juliana Damasio Oliveira, Henrique D. P. Santos, Ana Helena D. P. S. Ulbrich, Julia Colleoni Couto, Marcelo Arocha, Joaquim Santos, Manuela Martins Costa, Daniela Faccio, Fabio O. Tabalipa, and Rodrigo F. Nogueira. 2025. Development and evaluation of a clinical note summarization system using large language models. *Communications Medicine*, 5(1):376.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Sajan B Patel and Kyle Lam. 2023. ChatGPT: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Miguel Rodríguez-Ortega, Eduard Rodríguez-López, Salvador Lima-López, Carlos Escolano, Audrey Mash, Maite Melero, Lorenzo Pratesi, Laura Vigil-Gimenez, Leticia Fernandez, Eulàlia Farré-Maduell, and Martin Krallinger. 2025. Overview of MultiClinSum Task at BioASQ 2025: Evaluation of Clinical Case Summarization Strategies for Multiple Languages: data, evaluation, resources and results.

Samuel Rutunda, Gwydion Williams, Kleber Kabanda, Francis Nkurunziza, Solange Uwiduhaye, Eulade Rugegamanzi, Cyprien Nshimiyimana, Vaishnavi Menon, Mira Emmanuel-Fabula, Alastair K. Denniston, Xiaoxuan Liu, Emery Hezagira, and Bilal A. Mateen. 2026. Large language models for frontline healthcare support in low-resource settings. *Nature Health*, 1(2):191–197.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A Survey of Evaluation Metrics Used for NLG Systems. *arXiv preprint*. ArXiv:2008.12009 [cs].

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Elliot Schumacher, Daniel Rosenthal, Dhruv Naik, Varun Nair, Luladay Price, Geoffrey Tso, and Anitha Kannan. 2025. MED-OMIT: Extrinsically-Focused Evaluation Metric for Omissions in Medical Summarization. In *Proceedings of the 4th Machine Learning for Health Symposium*, pages 897–922. PMLR.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ji Woo Song, Junseong Park, Ji Hoon Kim, and Seng Chan You. 2025. Large Language Model Assistant for Emergency Department Discharge Documentation. *JAMA Network Open*, 8(10):e2538427.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv preprint*. ArXiv:1506.05869 [cs].

Christopher Y. K. Williams, Charumathi Raghu Subramanian, Syed Salman Ali, Michael Apolinario, Elisabeth Askin, Peter Barish, Monica Cheng, W. James Deardorff, Nisha Donthi, Smitha Ganeshan, Owen Huang, Molly A. Kantor, Andrew R. Lai, Ashley Manchanda, Kendra A. Moore, Anoop N. Muniyappa, Geethu Nair, Prashant P. Patel, Lekshmi Santhosh, and 5 others. 2025. Physician- and Large Language Model–Generated Hospital Discharge Summaries. *JAMA Internal Medicine*, 185(7):818–825.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. DocLens: Multi-aspect Fine-grained Evaluation for Medical Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, Bangkok, Thailand. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. *arXiv preprint*. ArXiv:2106.11520 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *ArXiv*.

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.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. *arXiv preprint*. ArXiv:1802.01886 [cs].

A Appendix

		Dutch Clinical Discharge Summary (N=97)							MIMIC (N=100)
		Hallucinations		**Omissions**		**Trivial facts**		**Usability**	**Hallucinations**
Error Severity		All	Sev.	All	Sev.	All	Sev.		
Reference-based									
Lexical similarity	BLEU	-0.071	-0.08	0.015	0.051	-0.133	0.079	0.079	0.076
	BLEU (unigram)	-0.06	-0.044	0.456	0.355	**-0.521**	-0.411	0.239	0.236
	BLEU (bigram)	-0.149	-0.137	0.255	0.216	-0.358	-0.193	0.252	0.235
	BLEU (trigram)	-0.084	-0.089	0.177	0.193	-0.277	-0.095	0.145	0.250
	BLEU (4 gram)	-0.109	-0.057	0.122	0.162	-0.318	-0.047	0.068	0.02
	chrf+	0.043	-0.083	-0.175	-0.19	0.177	0.228	0.067	-0.129
	chrf++	0.049	-0.078	-0.16	-0.175	0.176	0.222	0.06	-0.147
	JS div.	0.083	0.1	-0.369	-0.235	0.393	0.316	-0.278	**-0.36**
	METEOR	-0.017	-0.034	-0.443	-0.358	0.241	0.385	-0.024	-0.109
	NIST	-0.017	-0.091	0.134	0.079	-0.142	-0.093	0.076	0.061
	ROUGE-s4	-0.121	-0.068	-0.098	-0.08	-0.201	-0.026	0.104	0.014
	ROUGE-su4	-0.113	-0.074	-0.047	-0.059	-0.252	-0.083	0.161	0.023
	ROUGE-W-1.2	-0.092	-0.102	-0.519	-0.481	0.098	0.167	-0.009	-0.04
	ROUGE1	-0.089	-0.087	0.088	0.028	-0.320	-0.204	0.267	0.022
	ROUGE2	-0.083	-0.055	-0.089	-0.052	-0.157	0.022	0.071	-0.004
	ROUGEL	-0.088	-0.135	-0.194	-0.276	-0.215	-0.13	0.171	0.029
Semantic similarity	BERTScore (f1)	-0.037	-0.041	0.292	0.217	-0.206	-0.193	0.277	0.057
	BERTScore (precision)	-0.025	-0.018	0.343	0.256	-0.250	-0.231	0.271	0.210
	BERTScore (recall)	-0.038	-0.052	0.220	0.154	-0.173	-0.162	0.276	-0.102
	WEEM4TS	-0.047	-0.081	0.226	0.149	-0.103	-0.124	0.137	-0.027
	MoverScore	-0.118	-0.195	0.09	-0.051	-0.187	-0.184	0.114	0.033
	WMD	0.143	0.131	-0.149	-0.177	0.05	0.042	-0.227	0.117
	Sentence-BERT	0.06	-0.071	0.249	0.166	-0.121	-0.152	0.279	0.267
	SMS	-0.107	0.03	0.275	0.278	-0.261	-0.103	0.081	-0.185
Simplification	SARI	-0.028	-0.038	-0.009	-0.024	-0.123	-0.022	0.193	-0.034
Source-based									
Extractiveness	Compression	0.118	0.156	**0.696**	**0.579**	-0.448	**-0.459**	0.185	0.225
	Coverage	0.071	0.063	0.645	0.573	-0.348	-0.266	0.177	-0.06
	Density	0.031	0.003	0.444	0.484	-0.191	-0.07	0.105	-0.096
	Abstractivity	0.162	0.099	-0.370	-0.368	0.408	0.215	-0.048	-0.061
Degree of abstraction	Abstraction (unigram)	-0.035	-0.006	-0.693	-0.575	0.376	0.330	-0.196	0
	Abstraction (bigram)	-0.053	-0.036	-0.508	-0.550	0.171	0.092	-0.114	0.164
Uniformity	Uniformity	0.071	0.037	0.367	0.272	-0.286	-0.284	**0.313**	-0.029
Generation only									
Redundancy	ROUGE1 (red.)	0.033	-0.031	0.13	-0.026	-0.031	-0.174	0.062	0.136
	ROUGE2 (red.)	0.181	0.127	0.198	0.093	-0.172	-0.201	0.05	0.214
	ROUGEL (red.)	0.054	-0.008	0.161	0.005	-0.072	-0.220	0.047	0.231
Semantic coherence	Coherence	0.145	0.093	0.268	0.247	-0.067	-0.144	0.069	-0.167
Diversity	Distinct-1	0.051	0.026	-0.084	0.013	0.245	0.197	0.075	-0.129
	Distinct-2	0.059	0.034	-0.074	0.021	0.274	0.214	0.067	-0.232
	Distinct-3	0.072	0.044	-0.071	0.023	0.283	0.221	0.068	-0.196
	MTLD	-0.067	0.016	-0.091	0.018	-0.094	0.061	-0.116	-0.352
	Self-BLEU	**-0.240**	-0.175	-0.056	-0.013	0.014	0.11	0.066	0.102
Fluency	Perplexity	-0.068	-0.041	0.072	0.044	-0.019	-0.08	0.02	-0.321
	Readability	-0.158	**-0.208**	-0.360	-0.373	0.132	0.108	-0.096	0.335
Others									
	Generation length	0.218	0.134	-0.063	-0.015	0.404	0.266	0.027	-0.198

Table 5: Spearman correlations between human evaluation criteria and automatic metrics across datasets. Bold values indicate the highest absolute correlation coefficient within each evaluation column.

B Appendix

Detailed descriptions of automatic metrics used in this study.

B.1 Lexical similarity

- BLEU / NIST: N-gram precision between generation and reference. NIST is a variant of BLEU that weights rare n-grams more heavily.
- METEOR: Alignment between generation and reference using exact word matches, stemming, and synonymy
- ROUGE-N: Recall of overlapping n-grams
- ROUGE-L / ROUGE-W (weighted LCS): Longest common subsequence similarity (weighted for ROUGE-W)
- ROUGE-S (skipbigram) / ROUGE-SU (skipbigram plus unigram): Skip-bigram (i.e., pairs of words in the same sentence order, with arbitrary words in between) overlap
- chrF / chrF++: Character n-gram overlap (up to 6 characters) between generation and reference. chrF++ includes unigram and bigram overlaps.
- Jensen-Shannon divergence: The statistical similarity between unigram probability distributions.

B.2 Semantic similarity

- BERTScore / Sentence BERT: Semantic similarity with contextual token/sentence embeddings
- WEEM4TS: Combination of unigram recall and bigram precision between generation and reference with exact wording and semantic embeddings
- Word Mover's Distance (WMD) / Sentence Mover Similarity (SMS) / MoverScore: Semantic distance or transport cost between word or sentence embeddings of the generation and reference

B.3 Simplification

- SARI: The goodness of words that are added, deleted, and kept by the system (i.e., LLMs)

B.4 Extractiveness

- Coverage: The extent to which a summary is derivative of a text. It is calculated as the percentage of words in the summaries that are part of an extractive fragment[6].
- Compression: The ratio between the length of patient records and the summary
- Density: How much of a summary is made up of copied spans from the source, giving more weight to longer extractive fragments
- Abstractivity: Derived from density and reflects how novel and non-extractive a generated summary is

B.5 Degree of abstraction

- Abstraction n-grams: The percentage of novel n-grams (i.e., words that are not in the patient records) in the generated summaries

B.6 Uniformity

- Uniformity: Normalized entropy of decile positions of salient unigrams in the patient record, where salient unigrams are the top 20 keywords extracted[7] from the generated summary, excluding stopwords

B.7 Redundancy

- ROUGE-L on generation: ROUGE-L F1-score across all pairs of distinct sentences in the generated text

B.8 Semantic coherence

- BERT-based sentence-level probability: The probability of each successive sentence conditioned on the preceding one using the Bidirectional Encoder Representations from Transformers (BERT) language model

B.9 Diversity

- Distinct-1, Distinct-2, Distinct-3: The number of distinct unigrams, bigrams, and trigrams while accounting for long sentences in generated summaries
- Measure of textual lexical diversity (MTLD): The average number of words it takes to reach a target type-token ratio (TTR) stabilization (set at 0.72)

[6] Extractive fragments were first identified as the set of shared sequences of tokens in patient records and generated summaries.

[7] NLTK-RAKE is used to extract keywords.

- selfBLEU: The degree of repetition in generated text by computing how similar a sentence is to other generated sentences.

B.10 Fluency

- Perplexity: A language model's uncertainty in predicting the next token in a sequence. Lower perplexity indicates better next-token prediction and higher confidence under the model's learned distribution.

- Flesch Reading Ease Index:

  $$206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW})$$

  ASL = average sentence length (words/sentences).

  ASW = average syllables per word (syllables/words).

C Appendix

Error type	Metric	U	No error		Error		FDR-adjusted p
			size	Median score	size	Median score	
All hallucinations	Abstractivity	1042	47	1 ↓	50	1	0.339
	Readability	1435	47	38.07	50	32.89	0.073
	Self-BLEU	1471	47	1.34	50	0.78	0.042*
Severe hallucinations	Readability	1497	55	38.07	42	32.22	0.019*
	Self-BLEU	1395	55	1.22	42	0.84	0.091
	MoverScore	1374	55	0.45	42	0.44	0.118
All omissions	Abstractivity	1244	25	1	72	1	0.008**
	Compression	264	25	4.95 ↓	72	12.90	2.87e-06***
	Coverage	328	25	0.66 ↓	72	0.75	1.46e-05***
Severe omissions	Abstractivity	1419	36	1	61	1	0.023*
	Compression	533	36	6.01 ↓	61	12.61	7.48e-05***
	Coverage	492	36	0.69 ↓	61	0.77	2.61e-05***
All trivial facts	Abstractivity	398	19	0.99 ↓	78	1	0.003**
	Compression	1235	19	23.20	78	7.53	2.61e-05***
	BLEU unigram	1270	19	43.35	78	25.24	1.40e-05***
Severe trivial facts	Compression	1559	33	17.54	64	7.30	2.93e-04***
	BLEU unigram	1596	33	39.02	64	24.64	1.03e-04***
	JS divergence	574	33	0.25 ↓	64	0.28	4.92e-04***
Hallucinations (MIMIC)	JS divergence	1720	50	0.26	50	0.23	0.002**
	MTLD	1720	50	74.07	50	58.38	0.002**
	Readability	875	50	70.65 ↓	50	77.74	0.010**

Table 6: Mann-Whitney U test results comparing metric scores between error-free and error-present outputs. For each error type, we report the three metrics with the highest absolute coefficients (|τbl) from Table 4. The symbol ↓ in the No Error column indicates that the metric scores are lower in the error-free group than in the error-present group. In the MIMIC hallucination analysis, Jensen–Shannon divergence and MTLD yielded identical Mann–Whitney U statistics and p-values because both metrics produced the same rank-sum separation between the error-free and error-present groups, with identical group sizes (n = 50 per group). Significance are denoted by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).