Interspeech 2024
1-5 September 2024, Kos, Greece

Reference-Free Estimation of the Quality of Clinical Notes Generated from Doctor-Patient Conversations

Mojtaba Kadkhodaie Elyaderani[1], *John Glover*[2], *Thomas Schaaf*[1]

[1]Solventum Health Information Systems, USA
[2]Solventum Health Information Systems, Ireland
{mkadkhodaieelyaderani, jglover, tschaaf}@solventum.com

Abstract

This paper describes a simple yet robust approach to performing reference-free estimation of the quality of automatically-generated clinical notes derived from doctor-patient conversations. In the absence of human-written reference notes, this approach works by generating a diverse collection of "pseudo-reference notes" and comparing the generated note against those pseudo-references. This method has been applied to estimate the quality of clinical note sections generated by three different note generation models, using a collection of evaluation metrics that are based on natural language inference and clinical concept extraction. Our experiments show the proposed approach is robust to the choice of note generation models, and consistently produces higher correlations with reference-based counterparts when compared against a strong baseline method.
Index Terms: quality estimation, conversational AI, natural language generation, doctor-patient conversation

1. Introduction

Producing detailed clinical documentation based on doctor-patient conversations is a necessary yet burdensome task for physicians [1], and is often cited as one of the leading causes of physician burn-out [2, 3]. One way to reduce the documentation workload on physicians is to hire medical scribes, who assist in writing clinical notes [4]. However, this option is costly and difficult to scale [5], putting it beyond the reach of many practitioners. This has lead to the emergence of the "ambient clinical documentation" framework, where the conversation between doctor and patient is recorded and transcribed (with the patient's permission) and passed to a clinically-trained Language Model (LM) which generates the corresponding note [6, 7, 8, 9, 10].

Despite the recent improvements in their performance, modern LMs still make many errors that are unacceptable in a medical setting, and can generate clinical notes that are of poor quality. For example, they may miss critical information, contain hallucinated content, or include important information in wrong note sections [11]. Delivering poor-quality notes to physicians can be an extra burden, potentially resulting in a more time-consuming note creation process than simply starting from scratch. It also runs the risk of introducing errors that may be detrimental to patient care. Therefore, it is vital to be able to evaluate the quality of generated notes *before* showing them to physicians, so there is the option to take appropriate corrective action.

In this work we study how we can perform reference-free estimation of the quality of generated clinical notes. As we find that the most reliable automatic summarization metrics are typically reference-based [12], our goal is to replicate their performance in the reference-free setting. A common way to approach this problem is to score a fixed set of examples with a reference-based metric, then use the values as targets to train a reference-free method [13]. However, training-based approaches are computationally costly to create [14] and may not be robust to data distribution shifts [15].

In this study we instead investigate if we can generate a diverse set of "pseudo-reference notes" and use them as substitutes for human-written references, allowing us to use whichever reference-based metrics are appropriate at inference. Similar to [16], we use stochastic generations of a model to evaluate the factuality of an output generation. However, we expand the scope of our approach to beyond factuality and use it to estimate the quality of generated notes from multiple complementary and non-overlapping aspects. Furthermore, we demonstrate that the stochastic generations do not have to be obtained from the same model that generates the output note. Our experimental results reveal that our proposed approach outperforms an alternative training-based solution, which is developed based on the work of [17], and is robust to data distribution shifts.

2. Method

In an ambient clinical documentation setting, the doctor-patient conversation (DoPaCo) is typically first transcribed by an automatic speech recognition (ASR) system, and the resulting transcript is then provided as the input to a LM which creates a clinical note. The generated note is then presented to the doctor for their final review and submission to an electronic health record (EHR) system.

Assume that o denotes the output note generated by the language model for a conversation between a doctor and a patient. Furthermore, let r denote the final note for this encounter that is submitted to the EHR by the doctor. It is a common practice to structure a clinical note in multiple sections, including History of Present Illness (HPI), Physical Examination (PE), and Assessment and Plan (AP) [18]. Let us use o_s and r_s to denote the section s of o and r, respectively, where $s \in \{\text{HPI}, \text{PE}, \text{AP}, \cdots\}$.

Since the output note o is prone to typical errors made by language models, it is important to estimate its quality before presenting it to the doctor. Assume that $\{m_1, m_2, \cdots, m_k\}$ denotes a set of evaluation metrics that can jointly measure the quality of different sections of an output note. Such metrics should assess complementary and non-overlapping aspects of note quality. Most existing evaluation metrics rely on having a reference note. So, we assume our metrics $\{m_1, m_2, \cdots, m_k\}$ require having the reference note r to evaluate the output note.

10.21437/Interspeech.2024-1907

2.1. Proposed quality estimation approach

In the ambient clinical documentation setting, the reference note r is not available at the time of presenting the output note o to a doctor. Therefore, our proposed quality estimation approach works by first generating a diverse set of pseudo-reference notes $\{p^1, p^2, \cdots, p^n\}$. Then, treating each pseudo-reference note p^j, for $1 \leq j \leq n$, as a proxy for the missing reference note r, we approximate the reference-based value of every evaluation metric by the average of its values when individual pseudo-references replace the reference. For a metric $m \in \{m_1, m_2, \cdots, m_k\}$ and a note section s, we do

$$m\left(o_s, r_s\right) \approx \frac{1}{n} \sum_{j=1}^{n} m\left(o_s, p_s^j\right).$$

In the above equation, the quality of a section of the output note is estimated independent of other sections. This is to make sure the content of note sections are not misplaced in wrong sections. We name this approach as QUESST, which stands for QUality Estimation via Semantic STability.

2.2. Baseline quality estimation approach

To show the efficacy of QUESST, we adopt the BRIO training method in [17] to develop a baseline model for quality estimation. Assume that for a conversation and a note section s we have the reference note r_s and a set of pseudo-reference notes $\{p_s^1, p_s^2, \cdots, p_s^n\}$. Given an evaluation metric of interest $m \in \{m_1, m_2, \cdots, m_k\}$, one can order the pseudo-references in terms of their quality scores returned by this metric. Let $\pi(\cdot)$ be the permutation function over the set $\{1, 2, \cdots, n\}$ that indicates the ordering implied by m, i.e., assume

$$m\left(p_s^{\pi(1)}, r_s\right) \geq m\left(p_s^{\pi(2)}, r_s\right) \geq \cdots \geq m\left(p_s^{\pi(n)}, r_s\right).$$

BRIO proposes a training paradigm that calibrates the output probability scores of a sequence-to-sequence (Seq2Seq) model so that it preserves this ordering *without* requiring the reference note section r_s. To explain this, assume that the transcript of a conversation, denoted by t, and one of the pseudo-references p_s^j is passed to a Seq2Seq model and let $f(t, p_s^j)$ be the length-normalized estimated log-probability score returned by this model. As argued in [17], BRIO training calibrates the probability scores $\left\{f(t, p_s^1), f(t, p_s^2), \cdots, f(t, p_s^n)\right\}$, so that the ordering of these scores matches with the one implied by the metric m. Therefore, to incur a zero ranking loss over the transcript t and the pseudo-reference set $\{p_s^1, p_s^2, \cdots, p_s^n\}$, the model scores need to satisfy the following condition[1]

$$f\left(t, p_s^{\pi(1)}\right) \geq f\left(t, p_s^{\pi(2)}\right) \geq \cdots \geq f\left(t, p_s^{\pi(n)}\right).$$

Since the probability scores of a BRIO calibrated model follow the ordering implied by the metric of interest and a reference note is not required to compute those scores, one can use BRIO for reference-free quality estimation. Training quality estimation models for natural language generation via learning to rank is also studied in [19]. In addition, learning to rank is also utilized to train evaluation models for machine translation [20].

[1] Assuming the ranking loss margins in Eq. 8 of [17] are set to zero.

2.3. Metrics

Two different sets of metrics are used throughout this work. One set uses the Unified Medical Language System (UMLS) repository [21] to assess the clinical concept matching between an output and a reference (or a pseudo-reference) note. We extract the UMLS concepts from an output note and compare them with those obtained from a (pseudo-)reference using QuickUMLS [22], computing UMLS precision, UMLS recall, and UMLS f-measure scores [8].

The second set of metrics used are NLI factuality, NLI coverage, and NLI h-mean. These metrics are based on the Natural Language Inference (NLI) [23] model from [24], with further fine-tuning on manually-annotated de-identified private clinical data. To produce the NLI factuality score we follow the approach in [25]. We first decompose the (pseudo-)reference and output notes into sentences, score all sentence pairs using the NLI model, and take the maximum entailment score for each sentence to be the factuality score. The score for the full note is the average of the individual sentence scores. We perform a similar operation to produce the NLI coverage score, but with the positions of the (pseudo-)references and outputs switched when passed in to the NLI model. The NLI h-mean is defined as the harmonic mean of NLI factuality and NLI coverage.

3. Results

We start this Section by describing how we have trained a note generation model, named Pegasus-X-Gen, that will be utilized to create pseudo-reference notes required for doing quality estimation using QUESST. Then we explain how, for a choice of an evaluation metric, the Pegasus-X-Gen model is further calibrated via the BRIO training framework to obtain a metric-specific quality estimation model, named Pegasus-X-BRIO, that will serve as our baseline for quality estimation. We use a proprietary dataset throughout this Section, since there are no suitable public datasets that have the required properties, i.e., long-form medical conversations and clinical notes, and describe in sufficient detail how the results can be reproduced.

3.1. Pegasus-X-Gen model

We use 23828 pairs of DoPaCos and clinical notes to fine-tune the Pegasus-X-Large model [26] for clinical note generation. All of these DoPaCos are Orthopedics visits, carried out in English in the United States, and are transcribed using Amazon Transcribe Medical[2] service[3]. We employ an internal note parsing algorithm to extract different sections of clinical notes. Specifically, the content of the HPI, PE, and AP sections are extracted for model fine-tuning. We create a unique token corresponding to each of the extracted sections, which is added to the model's vocabulary, and prepended to the start of each note section in the training data. The average and standard deviation of the length of the transcripts and extracted note sections across the dataset are (after being tokenized by Pegasus-X tokenizer) 1761 ± 955 for the transcript, 151 ± 69 for the HPI section, 171 ± 137 for the PE section, and 175 ± 97 for the AP section.

We train the Pegasus-X-Gen model with maximum input and output lengths of 5120 and 512 tokens, respectively. We use a batch size of 8, an initial learning rate of $1e^{-3}$, warmup over 100 steps, use a polynomial learning rate schedule over a maximum of 30 epochs, and perform early-stopping using vali-

[2] https://aws.amazon.com/transcribe/medical/

[3] Word Error Rate for DoPaCos generally ranges from 15% - 40%.

dation set loss. To generate a given section at inference time we use the ASR transcript as input, then prompt the decoder with the appropriate section token. In all the experiments, except for those in Section 3.5, the number of pseudo-references is set to be 16 and they are generated by the Pegasus-X-Gen model using diverse beam search [27], with a diversity penalty of 0.4.

3.2. Pegasus-X-BRIO model

Using BRIO training method outlined in Section 2.2, we further calibrate our Pegasus-X-Gen model to obtain different metric-specific variants of our Pegasus-X-BRIO model. To prepare each variant of this model, we first adopt one of the six metrics introduced in Section 2.3 as the metric of interest. Second, for every DoPaCo in an unseen dataset of 1259 Orthopedics visits and a note section from the set {HPI, PE, AP}, we utilize the Pegasus-X-Gen model to generate 16 pseudo-references as detailed in Section 3.1. Third, we score the sixteen generated pseudo-references using the adopted metric of interest. The sixteen scored pseudo-references for every encounter are then sorted according to their metric scores. Finally, to improve the memory and computational efficiency of BRIO training and inspired by the approach of [28], only the top and bottom pseudo-references are selected from the sorted list for BRIO calibration.

All of the six variants of Pegasus-X-BRIO are trained for 10 epochs. Following the setup of BRIO [17], we use the Adam optimizer with an inverse-square-root learning rate scheduler, but set the maximum learning rate to $1e^{-3}$ and use 200 warm-up steps. Since the resulting models will be used for quality estimation, we only keep the contrastive component of the BRIO loss function and ignore the generation (MLE) component. We leverage an adaptive margin in the contrastive loss, where for any training pair of pseudo-references, the ranking margin is set to the difference between the corresponding metric scores. For checkpoint selection, we measure the Spearman correlation between the rankings imposed by the checkpoint model and the adopted metric over a separate validation dataset and select the endpoint with the highest correlation.

3.3. Dataset for evaluation

A separate dataset of 90 Orthopedics examples is used to evaluate the performance of QUESST and Pegasus-X-BRIO models for quality estimation. Three different note generation models, Pegasus-X-Gen, AWS HealthScribe[4], and GPT-4[5], are leveraged to create output notes for the evaluation examples. Patients' protected health information are removed from this dataset to prevent the leakage of sensitive data.

Unlike the case of pseudo-references, where diverse beam search was employed to generate them with Pegasus-X-Gen, the output notes from this model are created by the standard beam search, with a beam size of 3. In both cases of output and pseudo-reference notes, the maximum encoder and decoder input lengths are set to 5120 and 512, respectively. Also, a length penalty of 0.8 and a no-repeat-n-gram-size of 20 are used.

To generate output notes with GPT-4 we follow the prompting approach of [29]. For every evaluation example, we first identify the most similar encounter in the dataset, where the cosine similarity of the embedded transcripts is used as the measure of closeness. Then, the DoPaCo transcript of the evaluation example, along with the reference note of the most similar example, are used to prompt GPT-4 to generate an output note.

[4]https://aws.amazon.com/healthscribe/

[5]https://openai.com/research/gpt-4

3.4. Reference-based versus reference-free correlations

Figure 1 shows scatterplots for reference-based NLI h-mean scores (vertical axis) versus NLI h-mean estimations (horizontal axis) of the QUESST approach. Each row corresponds to a particular section of clinical notes and every circle corresponds to a clinical encounter from the evaluation dataset. The left-hand side plots are for the output notes generated using Pegasus-X-Gen, whereas the right-hand side plots are for the GPT-4 notes.

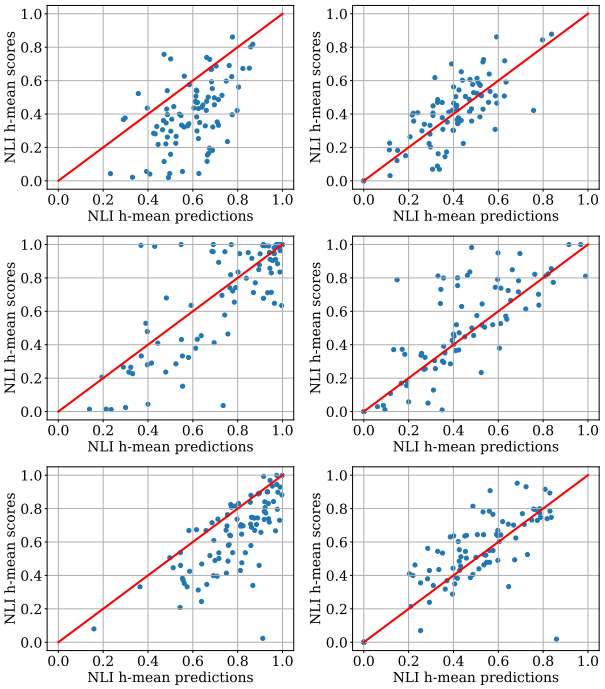

Figure 1: *Correlation plots for different sections of notes generated by Pegasus-X-Gen (left column) and GPT-4 (right column) models. From top to bottom, the pairs of scatterplots correspond to HPI, PE, and AP sections, respectively. All pseudo-references are generated using Pegasus-X-Gen.*

Interestingly, we observe a high correlation between the reference-based scores and their corresponding QUESST predictions across note sections, regardless of the note generation model. However, when the notes are generated by Pegasus-X-Gen, QUESST predictions are relatively biased towards higher values than their reference-based counterparts, which is why in the left-hand side plots there are more points lying below the diagonal line. This could be because in this case the same generation model is used to generate the output and pseudo-reference notes. The bias of quality estimation approaches towards their own generations is extensively studied in [30].

Table 1 shows Spearman correlations between the (reference-based) metric scores and corresponding (reference-free) predictions for the output notes generated by Pegasus-X-Gen, AWS HealthScribe, and GPT-4 models as outlined in Section 3.3. For every metric, the correlations for QUESST and the corresponding Pegasus-X-BRIO model, calibrated for that metric, are reported. With only one exception[6], QUESST consistently outperforms Pegasus-X-BRIO across different metrics,

[6]The exception happens for NLI coverage scores of HPI sections generated by Pegasus-X-Gen model, where QUESST and Pegasus-X-BRIO have achieved 0.48 and 0.50 correlations, respectively.

Table 1: *Spearman correlation for different note generation systems (Pegasus-X-Gen, HealthScribe, and GPT-4). For every metric, the top row corresponds to the QUESST approach of Section 2.1, and the bottom row is for the BRIO approach, where the Pegasus-X-BRIO model calibrated for that metric is used, as discussed in Section 2.2. The number before parentheses for each entry is computed over the entire de-identified evaluation dataset. The 95 percent confidence intervals, for 1000 bootstrap samples drawn from the evaluation dataset, are reported in parentheses. HealthScribe does not generate PE sections at the time of preparing this paper. The last row reports the mean and standard deviation of the length of generated note sections tokenized using Pegasus-X tokenizer.*

Metric	*Approach*	**HPI**			**PE**		**AP**		
		Pegasus-X-Gen	HealthScribe	GPT-4	Pegasus-X-Gen	GPT-4	Pegasus-X-Gen	HealthScribe	GPT-4
NLI factuality	QUESST	.44 (.24, .62)	.70 (.55, .81)	.64 (.45, .77)	.53 (.33, .68)	.74 (.59, .85)	.54 (.35, .68)	.84 (.74, .90)	.73 (.57, .83)
	BRIO	.13 (-.10, .35)	.42 (.21, .60)	.46 (.28, .62)	.29 (.08, .48)	.52 (.35, .67)	.21 (-.01, .42)	.53 (.35, .67)	.37 (.16, .54)
NLI coverage	QUESST	.48 (.29, .64)	.62 (.47, .73)	.66 (.52, .78)	.71 (.56, .82)	.81 (.70, .89)	.77 (.66, .84)	.71 (.57, .81)	.76 (.62, .86)
	BRIO	.50 (.32, .64)	.50 (.34, .63)	.43 (.22, .61)	.38 (.19, .55)	.55 (.36, .70)	.49 (.30, .64)	.42 (.22, .57)	.53 (.35, .68)
NLI h-mean	QUESST	.50 (.32, .66)	.65 (.49, .75)	.72 (.59, .81)	.69 (.52, .82)	.79 (.65, .88)	.73 (.59, .82)	.77 (.64, .86)	.75 (.59, .86)
	BRIO	.47 (.27, .63)	.58 (.44, .67)	.51 (.34, .65)	.49 (.28, .65)	.61 (.44, .74)	.49 (.30, .65)	.55 (.39, .68)	.53 (.31, .69)
UMLS precision	QUESST	.58 (.41, .70)	.62 (.46, .75)	.57 (.39, .71)	.57 (.41, .72)	.73 (.57, .84)	.48 (.29, .65)	.77 (.65, .85)	.68 (.51, .81)
	BRIO	.22 (.02, .40)	.43 (.23, .59)	.32 (.10, .50)	.29 (.07, .49)	.25 (.00, .46)	.19 (-.02, .39)	.41 (.21, .59)	.29 (.07, .51)
UMLS recall	QUESST	.63 (.47, .74)	.54 (.35, .70)	.57 (.39, .70)	.67 (.53, .80)	.76 (.64, .85)	.73 (.56, .84)	.68 (.56, .77)	.73 (.58, .84)
	BRIO	.28 (.10, .45)	.20 (-.01, .36)	.33 (.11, .51)	.32 (.11, .49)	.55 (.36, .71)	.47 (.26, .65)	.30 (.09, .48)	.48 (.27, .65)
UMLS f-measure	QUESST	.64 (.47, .76)	.66 (.51, .77)	.61 (.45, .73)	.67 (.53, .79)	.78 (.66, .86)	.72 (.56, .82)	.76 (.63, .84)	.72 (.56, .84)
	BRIO	.49 (.30, .64)	.38 (.17, .55)	.32 (.09, .52)	.55 (.40, .69)	.46 (.27, .63)	.50 (.30, .66)	.43 (.24, .59)	.40 (.18, .59)
Average across metrics	QUESST	.55 (.37, .69)	.63 (.47, .75)	.63 (.47, .75)	.64 (.48, .77)	.77 (.64, .86)	.66 (.50, .77)	.76 (.63, .84)	.73 (.57, .84)
	BRIO	.35 (.15, .52)	.42 (.23, .57)	.39 (.19, .57)	.39 (.19, .56)	.49 (.30, .65)	.39 (.19, .57)	.44 (.25, .60)	.43 (.22, .61)
Mean and std. of length	N.A.	122 ± 50	199 ± 36	100 ± 35	143 ± 114	84 ± 69	131 ± 70	125 ± 42	113 ± 51

note sections, and generation models. Furthermore, its performance is quite robust to the choice of the generation model.

Looking at the two rows that report average correlations across metrics implies that both QUESST and Pegasus-X-BRIO exhibit weaker performance (lower correlations) for the notes generated by Pegasus-X-Gen. This could be attributed to the self-bias of quality estimation methods as described earlier.

In addition, comparing the confidence intervals of those two rows for QUESST and Pegasus-X-BRIO suggests that the intervals are more overlapping when the output notes are generated by Pegasus-X-Gen[7], which means that Pegasus-X-BRIO competes better with QUESST in this case. This can be attributed to the fact that Pegasus-X-BRIO models are calibrated with Pegasus-X-Gen notes. Therefore, Pegasus-X-BRIO models are sensitive to data distribution changes, and to improve their performance for new generation models we need to further calibrate them.

3.5. Pseudo-reference generation parameters

Next, we explore the impact of the number of pseudo-references and the diversity penalty on the performance of QUESST. The left-hand side plot in Figure 2 shows the Spearman correlations for UMLS metrics as the number of pseudo-references, generated by Pegasus-X-Gen model using diverse beam search, varies from 2 to 40. The output notes evaluated here are the HPI sections generated by the same model for our evaluation DoPaCos. The diversity penalty is fixed at 0.4. As the plot shows, increasing the number of pseudo-references initially improves the average correlation scores for the three metrics. However, after around 30 pseudo-references, the performance begins to drop. This could be because having to generate a large number of diverse pseudo-references forces the model to hallucinate.

This observation is confirmed by the right-hand side plot, where the number of pseudo-references is fixed at 16 and the diversity penalty is increased from 0.05 to 1.4. As the plot shows, imposing a large diversity penalty degrades the correlations.

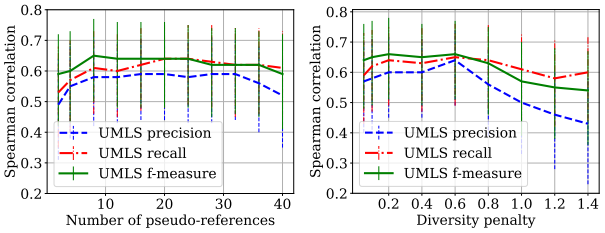

Figure 2: *Impact of the number of pseudo-references (left plot) and the diversity penalty (right plot) on the Spearman correlations for the HPI sections generated by Pegasus-X-Gen.*

4. Conclusion

We introduced QUESST, a novel, reference-free approach for estimating the quality of clinical notes generated automatically from doctor-patient conversations. By leveraging diverse "pseudo-reference" notes, this method overcomes the limitations of requiring human-written references and demonstrates robustness across different note generation models. The proposed approach achieves significantly higher correlations with reference-based quality estimates compared to the BRIO baseline, paving the way for efficient and reliable assessment of AI-powered healthcare documentation systems. This enables further research into integrating such automatic quality estimation methods into real-world clinical workflows and applications.

In spite of the good correlations obtained by QUESST, the current study is limited to one model for pseudo-reference generation. How much the performance of QUESST depends on the choice of the pseudo-reference generation model still remains as an open question. This study also focused on diverse beam search for pseudo-reference generation. Extending to other generation methods should be pursued in a future work.

[7]For instance, for the HPI section the confidence interval overlap goes from $(0.52 - 0.37)/(0.69 - 0.15) \approx 0.28$ for Pegasus-X-Gen to around 0.19 and 0.18 for HealthScribe and GPT-4, respectively.

5. References

[1] A. D. Misra-Hebert, L. Amah, A. Rabovsky, S. Morrison, M. Cantave, C. A. Sinsky, and M. B. Rothberg, "Medical scribes: how do their notes stack up?" *Journal of Family Practice*, vol. 65, no. 3, pp. 155–160, 2016.

[2] A. A. Wright, I. T. Katz *et al.*, "Beyond burnout—redesigning care to restore meaning and sanity for physicians," *N Engl J Med*, vol. 378, no. 4, pp. 309–311, 2018.

[3] G. Kumar and A. Mezoff, "Physician burnout at a children's hospital: incidence, interventions, and impact," *Pediatric Quality & Safety*, vol. 5, no. 5, 2020.

[4] R. Gidwani, C. Nguyen, A. Kofoed, C. Carragee, T. Rydel, I. Nelligan, A. Sattler, M. Mahoney, and S. Lin, "Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial," *The Annals of Family Medicine*, vol. 15, no. 5, pp. 427–433, 2017.

[5] K. J. Walker, W. Dunlop, D. Liew, M. P. Staples, M. Johnson, M. Ben-Meir, H. G. Rodda, I. Turner, and D. Phillips, "An economic evaluation of the costs of training a medical scribe to work in emergency medicine," *Emergency Medicine Journal*, vol. 33, no. 12, pp. 865–869, 2016.

[6] S. Enarvi, M. Amoia, M. Del-Agua Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, L. Rubini, M. Ruiz, G. Singh, F. Stemmer, W. Sun, P. Vozila, T. Lin, and R. Ramamurthy, "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*. Association for Computational Linguistics, Jul. 2020, pp. 22–30.

[7] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, "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)*. Association for Computational Linguistics, Aug. 2021, pp. 4958–4972.

[8] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, "Leveraging pretrained models for automatic summarization of doctor-patient conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3693–3712.

[9] C. Grambow, L. Zhang, and T. Schaaf, "In-domain pre-training improves clinical note generation from doctor-patient conversations," in *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, 2022, pp. 9–22.

[10] T. Haberle, C. Cleveland, G. L. Snow, C. Barber, N. Stookey, C. Thornock, L. Younger, B. Mullahkhel, and D. Ize-Ludlow, "The impact of nuance dax ambient listening ai documentation: a cohort study," *Journal of the American Medical Informatics Association*, p. ocae022, 2024.

[11] F. Moramarco, A. P. Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Savkov, and A. Belz, "Human evaluation and correlation with automatic metrics in consultation note generation," in *ACL 2022: 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022, pp. 5739–5754.

[12] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *Journal of Artificial Intelligence Research*, vol. 77, pp. 103–166, 2023.

[13] D. Jiang, Y. Li, G. Zhang, W. Huang, B. Y. Lin, and W. Chen, "Tigerscore: Towards building explainable metric for all text generation tasks," *arXiv preprint arXiv:2310.00752*, 2023.

[14] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[15] E. Durmus, F. Ladhak, and T. B. Hashimoto, "Spurious correlations in reference-free evaluation of text generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1443–1454.

[16] P. Manakul, A. Liusie, and M. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9004–9017.

[17] Y. Liu, P. Liu, D. Radev, and G. Neubig, "Brio: Bringing order to abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2890–2903.

[18] L. Buttman and K. Kingsley, *Ultimate Medical Scribe Handbook: General Edition*. CreateSpace Independent Publishing Platform, 2013.

[19] O. Dušek, K. Sevegnani, I. Konstas, and V. Rieser, "Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking)," in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 369–376.

[20] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2685–2702.

[21] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[22] L. Soldaini and N. Goharian, "Quickumls: a fast, unsupervised approach for medical concept extraction," in *Proc. of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016, pp. 1–4.

[23] I. Dagan, O. Glickman, and B. Magnini, "The pascal recognising textual entailment challenge," in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 177–190.

[24] S. Zhang and M. Bansal, "Finding a balanced degree of automation for summary evaluation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6617–6632.

[25] J. Glover, F. Fancellu, V. Jagannathan, M. R. Gormley, and T. Schaaf, "Revisiting text decomposition methods for NLI-based factuality scoring of summaries," in *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 97–105.

[26] J. Phang, Y. Zhao, and P. J. Liu, "Investigating efficiently extending transformers for long input summarization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3946–3961.

[27] A. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search for improved description of complex scenes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[28] M. Ravaut, S. Joty, and N. Chen, "Summareranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4504–4524.

[29] J. Giorgi, A. Toma, R. Xie, S. Chen, K. An, G. Zheng, and B. Wang, "Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 323–334.

[30] D. Deutsch, R. Dror, and D. Roth, "On the limitations of reference-free evaluations of generated text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 10 960–10 977.