A Meta-Evaluation of Faithfulness Metrics for Long-Form Hospital-Course Summarization

Griffin Adams GRIFFIN.ADAMS@COLUMBIA.EDU
Computer Science
Columbia University
New York, New York, US

Jason Zucker JZ2700@CUMC.COLUMBIA.EDU
Columbia University Irving Medical Center
Columbia University
New York, NY, US

Noémie Elhadad NOEMIE.ELHADAD@COLUMBIA.EDU
Computer Science and Biomedical Informatics
Columbia University
New York, NY, US

Abstract

Long-form clinical summarization of hospital admissions has real-world significance because of its potential to help both clinicians and patients. The factual consistency of summaries–their faithfulness–is critical to their safe usage in clinical settings. To better understand the limitations of state-of-the-art natural language processing (NLP) systems, as well as the suitability of existing evaluation metrics, we benchmark faithfulness metrics against fine-grained human annotations for model-generated summaries of a patient's Brief Hospital Course. We create a corpus of patient hospital admissions and summaries for a cohort of HIV patients, each with complex medical histories. Annotators are presented with summaries and source notes, and asked to categorize manually highlighted summary elements (clinical entities like conditions and medications as well as actions like "following up") into one of three categories: "Incorrect," "Missing," and "Not in Notes." We meta-evaluate a broad set of faithfulness metrics–proposed for the general NLP domain–by measuring the correlation of metric scores to clinician ratings. Across metrics, we explore the importance of domain adaptation (e.g. the impact of in-domain pre-training and metric fine-tuning), the use of source-summary alignments, and the effects of distilling a single metric from an ensemble. We find that off-the-shelf metrics with no exposure to clinical text correlate well to clinician ratings yet overly rely on copy-and-pasted text. As a practical guide, we observe that most metrics correlate best to clinicians when provided with one summary sentence at a time and a minimal set of supporting sentences from the notes before discharge.

1. Introduction

A significant factor for clinician burnout is the Electronic Health Record (EHR), the information overload it produces, and the documentation burden it requires (Shanafelt et al., 2016; Moy et al., 2021). A study of US physicians revealed that doctors spent 27% of working hours with patients and nearly 50% of their time on EHR and desk work, in addition to 1-2 hours at night, spent mostly on documentation (Sinsky et al., 2016). Clinician burnout can have damaging consequences not only for clinicians (National Academies of

Sciences, 2019), due to, among other factors, increased rates of depression (Maslach and Leiter, 2016) and interrupted work-life balance (Kroth et al., 2019)), but also patients, due to an increased risk of errors (Salvagioni et al., 2017; Panagioti et al., 2018).

In the inpatient setting, the Discharge Summary (Kind and Smith, 2008) is a particularly tedious and time-consuming note to write (Chan et al., 2014). Yet, it is a critical piece of documentation. Written at the end of a patient's hospital admission, the Discharge Summary ensures continuity of care (Kripalani et al., 2007; O'Leary et al., 2009). Its timely availability has been shown to have a direct impact on patient quality of care, including the rate of hospital readmission (Van Walraven et al., 2002). A key mandatory section of the Discharge Summary is the "Brief Hospital Course," which, in a paragraph of variable-length, recounts in a narrative form the events occurred during the patient stay, and why they happened. Composing the hospital-course summary is a cognitively difficult task for clinicians. They must review a high number of clinical notes and reports entered during the patient stay and synthesize them into a long paragraph. It is even more challenging when an admission is complex—the case for patients with comorbidities or chronic conditions.

Automated summarization can support clinicians in this difficult task. An automatically generated hospital course summary can act as a first draft for a clinician and ensure that the critical elements of the patient stay are not missed in the potentially overwhelmingly large amount of notes produced during the patient stay. Generating a high-quality hospital course narrative is difficult and ensuring its faithfulness is paramount. It requires synthesizing and fusing information from diverse note types, while remaining consistent: adhering to temporal constraints, providing sufficient context to avoid misleading patient characterizations, and even resolving source note errors.

Long-form summarization is an active topic of research in the general NLP domain (Guo et al., 2021; Phang et al., 2022), yet most faithfulness metrics have been developed on shorter datasets Kryscinski et al. (2020a); Durmus et al. (2020); Wang et al. (2020); Deng et al. (2021b); Yuan et al. (2021); Laban et al. (2022); Ribeiro et al. (2022). In the clinical domain, there are additional open questions, including the performance of modern summarization models and whether existing evaluation metrics are truly reflective of clinical quality. In this paper, we examine the performance of an established long-form abstractive summarization model on the task of hospital course summarization, as well as the quality of existing faithfulness metrics when compared to clinicians' judgments. To this end, we fine-tune a long-range transformer (Longformer Encoder-Decoder (LED) (Beltagy et al., 2020) on a large dataset of Hospital Course summaries, pertaining to all in-patient hospital admissions at a large healthcare institution. On a held-out set of admissions of complex patients (patients with HIV) (Levy-Fix et al., 2020), we rely on experts (clinicians) to collect annotations of LED summaries based on the notes written before discharge.

We then meta-evaluate a large set of existing summarization evaluation metrics (including BARTScore (Yuan et al., 2021), BERTScore (Zhang et al., 2019), Entailment-based CTC (Deng et al., 2021a) and SummaC (Laban et al., 2022)) by measuring their correlation to human annotations. Since these metrics were mostly developed on single document general-domain corpora, we identify three key dimensions pertinent to adaptation to the task of long-form, multi-document clinical summarization: **domain adapation** (pre-training and metric fine-tuning), **length of inputs**, and **length of outputs**. For length-based dimensions, we explore the impact of source-summary alignments and summary granularity

(sentence-level versus summary-level). We find that metrics tend to correlate best with human annotations when provided summary sentences one at a time, and when only the most relevant content (high precision source-summary alignments) is provided. Metrics which are trained on clinical text do not perform as well as metrics trained on general corpora. This can be explained by the fact that general domain metrics over-rely on the level of copy-and-paste, which provides a good, but brittle, proxy for faithfulness. In-domain adaptation of metrics will likely be critical to clinical summaries generated by Large Language Models (LLM), such as GPT-4 (Bubeck et al., 2023; Nori et al., 2023). On news article summarization, LLMs have been shown to rely less on copy-and-paste (extractiveness) and exhibit more lexical diversity (Goyal et al., 2022). Rather than adapt metrics to clinical text by training on references, we find it advantageous to learn directly from system summaries. We use an ensemble of our baseline metrics to produce a pseudo faithfulness score on system summaries and distill a metric from these noisy ground-truth labels. Our distilled metric has a higher correlation than all baseline metrics to expert annotation labels.

Our primary contributions are: **(1)** We collect fine-grained faithfulness annotations for the the task of hospital-course summarization, which contains substantially longer inputs than previous clinical annotation efforts; **(2)** We benchmark existing faithfulness metrics against these annotations, as well as explore practical considerations of adapting general domain metrics to long-form clinical narratives; **(3)** We analyze the confounding role of copy-and-paste (extractiveness) and show how a simple lexical statistic can be complementary to more complex metrics, including one distilled from an ensemble of other metrics.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work highlights the following insights relevant to other machine learning in health endeavors, specifically those in clinical natural language processing (clinical NLP).

- Evaluation of faithfulness in AI-generated clinical texts needs to account for the documentation practice of copy and paste. Clinicians often rely on copy-and-paste when authoring clnical notes. On one hand, the redundancy of clinical text due to copy and paste impacts models trained on clinical texts. On the other, modern NLP metrics tend to conflate copy-and-paste (extractiveness) with faithfulness. Taken together, a model which over-relies on copy-and-paste may appear faithful without having requisite understanding of clinical concepts. If undetected by humans and metrics, this could lead to too much trust in brittle systems.

- Most clinical text generation papers evaluate new models and baselines with metrics created for general NLP tasks. We demonstrate that these metrics correlate poorly to clinical experts' assessments of quality. Clinical NLP practictioners should focus on the co-development of clinical systems and metrics suited to these systems, and prioritize human evaluation.

- For many real-world clinical NLP tasks, beyond the one discussed in this paper, inputs are long. Note bloat is one factor, but so is the fact that patient records are longitudinal and dense with complex information. Our analysis demonstrates that only a small fraction of the inputs are necessary to evaluate the faithfulness of text generated based on long inputs. This is a highly convenient result, as it suggests

that clinical evaluation metrics can be trained and used on more granular alignments, rather than computationally expensive, larger-scale inputs.

- In the general NLP domain, the development of metrics has focused on one-size fits all solutions: a single metric which can identify all possible errors. Yet, when analyzing clinical text, error are highly diverse–some involve obvious mistakes like an incorrect dosages, while others are more subtle, omitting a key detail about a patient or misrepresenting the chronology of a disorder. By showing that an ensemble of metrics can capture diverse error types more effectively than a single metric, we hope to motivate the development of specialized metrics.

2. Related Work

Faithfulness Metrics. Metrics to assess faithfulness can be roughly distilled into the following categories: QA-based (Wang et al., 2020; Fabbri et al., 2022; Durmus et al., 2020), entailment based metrics from natural language inference (NLI) (Falke et al., 2019) or synthetic data (Kryscinski et al., 2020b; Deng et al., 2021b; Utama et al., 2022), fact-based, reference-free overlap (Goodrich et al., 2019), and those which directly learn from human judgments (Ribeiro et al., 2022) (similar to BLEURT (Sellam et al., 2020) approach for machine translation). Most of these metrics have been developed on single-document news summarization datasets, such as CNN / DailyMail (Hermann et al., 2015; See et al., 2017) and Xsum (Narayan et al., 2018). Faithfulness metrics proposed for clinical summary evaluation have typically come from the overlap category and focus on concept alignment between summaries and the source input (Zhang et al., 2020; Tang et al., 2022).

Evaluation of Clinical Note Summarization. Moen et al. (2014) evaluate extractively generated Discharge Summaries based on content criteria guidelines and benchmark ROUGE against these coverage-focused annotations. Recent work on human evaluation of clinical summarization has focused on self-contained, single-document tasks: including radiology report summarization (MacAvaney et al., 2019; Zhang et al., 2020) and echocardiogram conclusions (Tang et al., 2022). For these shorter tasks, summary-level assessments are collected, in the form of pairwise ranking (Tang et al., 2022) or point-wise assessments (MacAvaney et al., 2019) on a Likert Scale. Moramarco et al. (2021) examine brief descriptions of SOAP notes for mock patient encounters and compare fact-based overlap between reference and system-generated summaries. Most closely related to our work, Moramarco et al. (2022) perform a human evaluation on a more self-contained, conditional clinical note generation task: generating a SOAP note from consultation transcripts. They rely on a dataset of mock patient-doctor conversations and corresponding SOAP notes. Annotators were asked to post-edit notes to correct errors, as well as manually highlight spans with incorrect or omitted information. Automatic metrics were then benchmarked against post-editing time, as well as the number of incorrect and omitted spans. Our work differs as we define a typology of errors with more categories, consider more diverse faithfulness metrics, and, given much longer clinical narratives, explore the impact of using source-summary alignments and different summary granularities.

Split	Admissions	Source		Reference	
		Notes	Tokens	Sentences	Tokens
Training Data for Summarization Model	82k	41	18.4k	11.6	207
Training Data for Evaluation Metrics	2.7k	40	19.1k	12.5	243
Held-Out Human Annotation Data	29	24	11.7k	12.1	211

Table 1: Data Statistics for training the summarization LED model, the subset used for training evaluation metrics, as well as the subset used for Human Annotation. The statistics for `Source` and `Reference` lengths represent the averages per each admission.

3. Data

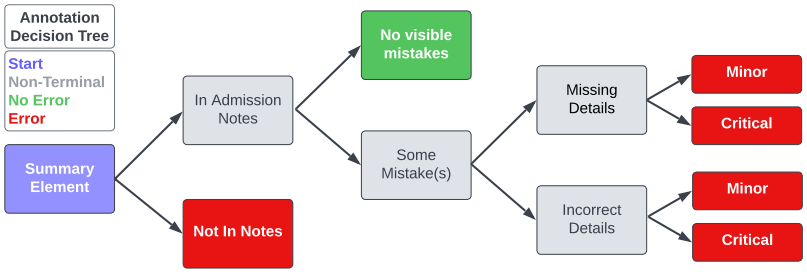
The data is comprised of clinical notes from the Electronic Health Record (EHR) for all in-patient admissions at a large healthcare institution (Columbia University Irving Medical Center in New York City) from 2010-2014. The inputs are all notes between the patient's Admission to the Hospital before Discharge (excluding the Discharge Summary).[1]. The gold-standard references are the Brief Hospital Course section, which is extracted with regexes from the Discharge Summary.

Training Data. We show training data statistics in the first row of Table 1. We delineate between the full training set, which is used to train the summarization models and the subset of the training set which is used for fine-tuning evaluation metrics in-domain. The subset filters for HIV patients which mirrors the filtering done to produce the human evaluation cohort (discussed directly below).

Human Evaluation Cohort. The training set comprises both HIV and non-HIV patients while the human annotation test set is solely HIV. We choose HIV patients as they typically have multiple co-morbidities and, concomitantly, complex hospital courses (Gallant et al., 2017). To select the set of admissions and corresponding summaries to be annotated, we consider all admissions of HIV patients. We filter out admissions that are in the bottom and top deciles for number of source notes and summary reference length. Based on annotator availability, we sample 29 summaries for annotation (245 sentences), which, in total, cover a total of 703 source notes from the EHR.

Generating Summaries for Annotation. We fine-tune a Transformer Encoder-Decoder with sparse attention (Longformer Encoder-Decoder, LED) (Beltagy et al., 2020). The LED handles inputs up to 16,384 tokens. To fit all inputs (the average input length from Table 1 is $18.4k$), we train a simple bi-LSTM model to rank each sections and, during inference, retain the top 100 sections. Filtering and fine-tuning details and hyper-parameters are provided in Appendix B.

1. We select all note types as the input for summarization due to the fact that we found decreases in performance when filtering out certain note types. Content selection is implicitly performed by the model during fine-tuning.



Summary sentence with *incorrect* summary element bolded	Error Type	Explanation Provided by Annotator
The patient was discharged with a plan to **follow up** with PMD	**Not In Notes**	No follow-up mentioned in notes
Patient had a **foley catheter** placed with immediate 500 CC return	**Incorrect Minor**	Foley catheter wasn't linked to 500 CC return. That was the straight catheter (mentioned elsewhere).
The patient being admitted to establish HIV care.		Patient has HIV but not reason for admission.
Rest - as per assessment	**Incorrect *Critical***	Randomly copied from notes. The patient is gravely ill.
Patient was started on ... Azithromycin for **MAC ppx**.		Incorrect indication. Should be a suspected infection.
Given **history**, it appears that this has been longstanding.	**Missing *Minor***	Unclear which of patient's problems *history* refers to
Patient **ambulated independently**		Patient required a walker for assistance
Patient is on day 2 of 9 for **HCAP coverage**	**Missing *Critical***	Unclear what timeline is from discharge summary
Patient was admitted to the ID service on 1/1/2001		Reason for admission (spitting up blood) not mentioned

Figure 1: Annotation Decision Tree with examples for each error type. Examples have been modified to remove any protected health information (PHI).

4. Collecting Annotations

At a high-level, the annotation task consisted of assigning an error category (or No Error) to each Summary Element (defined below) in a system output, based solely on clinical knowledge and all patient's clinical notes from the hospital admission.

Summary Elements. As in other faithfulness work (Goyal and Durrett, 2021), we decided to collect fine-grained annotations and experimented with different granularities while piloting the study. We found that entities (used in Cao et al. (2022)) were too granular, noisy, and incomplete on clinical notes. Syntactic parses were unreliable on our text as well. On the other hand, sentence-level annotation (Wang et al., 2020; Durmus et al., 2020; Pagnoni et al., 2021) was insufficiently fine-grained given the length and information density of many sentences. As such, the authors of the paper manually extracted `Summary Elements` (SE), which consist of standard medical concepts and actions, as well as compound concepts. Standard medical concepts included Disorders, Medications, Procedures, and Treatments, while actions capture phrases like "discharged to home" and "plans to follow up". We merge compound entities into a single **SE**: "alkanization of urine".

Error Categories. For each SE, annotators were asked to identify and categorize errors. As represented as a decision tree in Figure 1, annotators were first asked to confirm whether or not the summary element is "hallucinated": `Not in Notes`. If the SE can be found in

the notes, they either deem it correct: `No visible mistakes` or denote an inconsistency in its usage. For these intrinsic-focused errors, we delineate between `Incorrect Details` and `Missing Details`. A SE has `Incorrect Details` if it can be found in the source notes yet contains information that does not reflect what is written in the notes. This category encapsulates numerical errors (dosages, dates), mis-representations of symptoms ("afebrile" is incorrect if patient had a fever), fusion errors (an incorrect indication for a drug), among others. An SE has a `Missing Details` error if the summary omits important information about the SE, which could lead to misleading conclusions being formed about the patient's true hospital course.

Severity of Errors. For `Incorrect` and `Missing`, as in Moramarco et al. (2022), we ask annotators to distinguish between Minor and Critical errors. We provide annotators with examples of both kinds of errors and define `Critical` as a mistake which could negatively impact the patient's present and future treatment. Minor is an exclusionary category defined as "Not Critical".

Annotators. We recruited 6 clinical practitioners, with IRB-approved access to the patient data, to annotate the summaries in `Eval - HIV`. Each annotator was compensated at a rate of $30 / hour. 4/6 of the annotators self-identify as female, with the other two as male. 4/6 self-identify as "White", and 1 each as "Black or African" and "Other". 2 annotators are attending physicians, 3 are in medical residency, and 1 is a fellow. They have a combined 25 years of medical practice. Each expert annotated summaries for a minimum of one hour at the same computer and location with the guidance of the authors of the paper, who were available for technical guidance and clarification questions. Collectively, the task was carried out over $\sim$ 10 hours across 4 days.

Description of Interface. We develop a custom annotation interface within Prodigy. The interface presented each annotator with one summary at a time. For viewing ease, summaries were split such that one sentence was shown per line. Summary Elements (SE) were highlighted. For each SE, annotators would select the appropriate error category (or No Error) and then double click or highlight the SE span. On a separate browser page, we displayed the source notes for the patient visit, which were hosted locally on a custom, light-weight app. The left-hand side of the display showed section headers and free text for each note. Notes were sorted by date and annotators could search for a note by its title on a drop-down menu. Section headers were indexed and searchable to allow for efficient navigation of long notes. On the right hand side of the webpage, we enabled free-text search across notes. Each note was pre-indexed such that all mentions of matching search terms across notes could be quickly surfaced. We extracted all concepts with CLAMP NLP, highlighted them in the interface, and allowed for annotators to trigger a concept-based search query by double-clicking on the concept span in the note. Our code will be made public for a camera-ready version.

5. Error Analysis

Distribution of Errors. Table 2 shows the number of SE per summary and per sentence, as well as the breakdown of SE into each error category. 18% of SEs are marked as having *Any mistake*, of which the predominant category is `Incorrect` (11% versus 3% and 4%

	Per Summary	Per Sent	% of All SE
All Summary Elements (SE)	27.10	3.21	-
Incorrect SE	2.86	0.34	11%
Missing SE	0.93	0.11	3%
Not In Notes SE	1.03	0.12	4%
Any Mistake SE	4.83	0.57	18%

Table 2: Statistics on annotated Summary Elements (SE), broken down by error category.

for `Missing` and `Not in Notes`). In Table 2, Minor and Critical are lumped together and contribute equally.

Qualitative Analysis. As shown in Figure 1, incorrect errors often result from improper fusion of concepts: ("foley catheter" with "500 CC return", "Azithromycin" with "MAC ppx", and "admitted" with "HIV care"). Incorrect errors can also be perfectly extractive. "Rest - as per assessment" is copied verbatim from a previous note, yet is incorrect because, at the time of discharge, the patient is actually gravely ill, which contradicts the recommendation. `Missing Errors` are also quite extractive (see analysis in §H) and tend to occur from the reverse problem: insufficient fusion. The model fails to fuse related concepts when they appear in different contexts. Specifically, the model fails to make the following links: use of a "walker" is relevant to his "ambulat[ion]", that the "HCAP coverage" duration should be related to the note timestamp, and that "admitted to ID service" should be linked to the reason for admission—"spitting up blood".

Severity of Errors. The majority of `Incorrect` errors were marked as **Critical** (57%), whereas a minority for `Missing` (37%). As implicated by Figure 1, the difference between Critical and Minor errors is very subtle. Typically, the justifications for each, as provided by the annotators, were highly specific to the patient in question. This is interesting as it represents a non-standard definition of consistency, one which is grounded on a more holistic view of the patient journey.

In Appendix F, we demonstrate that errors get worse as the summaries grow longer. This points to decoder-side degeneration for models tasked with producing very long clinical narratives.

6. Evaluation Metrics

6.1. Task-Specific Concerns

Broadly speaking, we identify three high-level challenges for evaluating long-form clinical summaries, which are distinct from those faced when evaluating single-document new summaries: **(1) Domain Adaptation**, **(1) Long Outputs**, **(3) Long Inputs**.

Domain Adaptation. The first challenge relates to adapting metrics, typically trained and used on general domain data, to clinical text. We cannot adapt all metrics, especially metrics (Sellam et al., 2020; Ribeiro et al., 2022) which directly learn from news summary annotation benchmarks (Wang et al., 2020; Pagnoni et al., 2021; Fabbri et al., 2021; Laban et al., 2022). Domain-specific pre-training can improve performance of downstream models

on many tasks (Gururangan et al., 2020), including clinical (Alsentzer et al., 2019a), yet the impact of in-domain exposure is less studied when meta-evaluating faithfulness metrics. As such, we implement three versions of each metric with increasing levels of domain adaptation: `Off-The-Shelf` (fully out-of-domain), `Tuned In-Domain` (pre-trained out-of-domain, tuned-in-domain), and `Double In-Domain` (pre-trained and tuned in-domain). For in-domain pre-training, we rely on existing models pre-trained on clinical or biomedical corpora, specific to each dataset. For in-domain metric tuning, we use the `Train - HIV` data from Table 1. Specific Training details are provided in §6.2.

Output Lengths. Given previous work (Adams et al., 2021) detailing the lack of inter-sentence discourse markers in clinical narratives, we evaluate each sentence independently. Performing meta-evaluation of metrics at the sentence-level also increases the level of support (29 vs 245).

Input Lengths. Our inputs contain $\sim$ 30,000 tokens. Conditioning evaluation on the entire source is computationally expensive and often undesirable (e.g., entailment models are trained on short premises). Existing faithfulness metrics tend to struggle with long inputs (Honovich et al., 2022), likely due to a high noise-to-signal ratio. Lebanoff et al. (2019a) demonstrate with human annotations that only a handful of sentences from the source text are relevant to a given summary sentence.

Yet, computing source-summary alignments (Ernst et al., 2021) is particularly challenging for clinical text because 1) massive redundancy from copy-and-paste (Hirschtick, 2006); 2) lexical variation in discussing semantically identical concepts (abbreviations, acronyms, etc.) (Adams et al., 2020); 3) the need for complete context when assessing missing or misleading information. To explain 3), if a summary includes an outdated lab measurement, simply returning that single lab value as the alignment would provide a false sense of clinical correctness. The full chronology is needed.

Given this complexity, we separately evaluate the impact of alignment granularity (2-3 sentences to the whole input) on metric tuning and inference. Each method aligns a summary sentence to a subset of sentences from the source. Duplicate source sentences are removed. Table 10 shows the average number of aligned sentences from the source notes by each alignment method.

Alignments - Granular. `ROUGE-TopK` takes the $k = 5$ highest ROUGE-aligned sentences (average of R1, R2, RL F-1), while `ROUGE-Gain` follows Lebanoff et al. (2019b) and maximizes the relative ROUGE gain of adding each additional sentence to the current set of aligned sentences. To account for lexical variation and noise, we also build alignments with BERTScore (BS) from in-domain weights (see description of BERTScore model used in §6.2). `BS-TopK` selects the k source sentences with the highest F-1 BS vis-a-vis the summary sentence. `BS-Gain` follows the approach in (Adams et al., 2022) in which a coverage weight is assigned to each token in the summary sentence and updated based on the maximal alignment so far.

Alignments - Entity-Chain. Given a summary sentence, we define an alignment method based on `Entity-Chains` (Barzilay and Elhadad, 1997; Narayan et al., 2021) as the set of sentences in the source with at least one medical concept (a CUI from the Unified Medical Language System (UMLS) aligned to any of the CUIs in the summary sentence. Appendix

C details how entities are extracted and linked to the UMLS, as well as how synonymous concepts are identified and merged.

Alignments - Section-Level. To avoid fragmented alignments pulled from different notes, we also consider the Top-1 most aligned section as its own alignment. In particular, we select the section with the highest average ROUGE-{1, 2, L} overlap vis-a-vis each sentence in the summary.

Alignments - Full Input. The conventional approach is to pass the whole source as input. Most of our inputs surpass both short and long transformer token limits. As needed for each metric, then, for `Full Input` alignments for each summary sentence, we select the source sentences with the highest ROUGE-{1, 2} overlap vis-a-vis summary sentence until a target token limit is reached.

6.2. Metrics

We describe each metric at a high-level and then detail domain adaptation.

BERTScore. High-Level. BERTScore (Zhang et al., 2019) computes a greedy soft-alignment, based on BERT hidden state similarities, between a reference and a hypothesis text. As in Pagnoni et al. (2021), we compute a *reference-free* BERTScore: in our case, the hypothesis is a summary sentence and the reference its aligned source sentences. **Domain-Adaptation.** For `Off-The-Shelf`, we use `RoBERTA-Large` (Liu et al., 2019). There is no task-specific training for BERTScore so we report a single `In-Domain` variant. Specifically, we use a RoBERTA-Large model trained on biomedical (PubMed and PubMed Central) and clinical (MIMIC-III) text (Lewis et al., 2020)[2].

BARTScore. High-Level. BARTScore (Yuan et al., 2021) computes the length-normalized log likelihood of a summary conditioned on the input. We measure BARTScore for each sentence based on its aligned source inputs. **Domain Adaptation.** For `Off-The-Shelf`, we use a BART-Large model fine-tuned on CNN/DailyMail news summaries[3]. For `Tuned In-Domain` and `Double In-Domain`, we fine-tune BART-based models on `Train - HIV` corpus. The targets are single summary sentences and the inputs are their aligned source sentences. We fine-tune a separate model each alignment method from §6.1 to analyze the impact of alignment granularity on *training* metrics. For `Double In-Domain`, we initialize fine-tuning on `Train - HIV` with the BART-based ReDRESS model from Adams et al. (2022) [4]. For `Tuned In-Domain`, we initialize fine-tuning from `BART-Base` (to match ReDRESS). Using the Trainer from the Transformers library (Wolf et al., 2020), we fine-tune each model in batches of 16 for $10,000$ steps with a learning rate of $3e-5$ (200 warmup steps followed by linear decay). We use a label smoothing factor of 0.1.

2. The model weights (RoBERTa-large-PM-M3-Voc-large) can be freely downloaded and used with HuggingFace.
3. `facebook/bart-large-cnn` from HuggingFace.
4. ReDRESS is pre-trained on a novel entity-based de-noising objective on unlabeled clinical text (MIMIC-III discharge summaries). The model weights are accessible on HuggingFace as "griffin/redress-clinical-hallucination-generator".

CTC. High-Level. Compression, Transduction, Creation (CTC) (Deng et al., 2021a) defines a unified series of weakly supervised methods to evaluate system outputs on several NLG tasks. For summary faithfulness, the `CTC Score` represents the average number of tokens predicted as "fake" given the source. To train the CTC model, spans from reference summaries are masked-and-filled with a separate language model: the generator. **Domain Adaptation.** For `Off-The-Shelf`, we use `D-cnndm`, a RoBERTA-Large model fine-tuned for CTC consistency **d**iscrimination on the CNN/Dailymail dataset. For domain adapation, we corrupt summary sentences from `Train - HIV` and learn to discriminate based on source alignments. To generate fake tokens (the generator), we first train a mask-infiller (BART-base) on all discharge summaries in MIMIC-III. We use the same span mask procedure from CTC (based on a dependency parse) to align the training objective with its usage. We discuss generator training details and example outputs in Appendix D. For `Double In-Domain`, we initialize the CTC Discriminator from the same biomedical RoBERTA model used for the `In-Domain` BERTScore (Lewis et al., 2020). For `Tuned In-Domain`, we initialize tuning from `RoBERTA-Large` (to match the initialization for `Off-The-Shelf`).

Entailment. High-Level. Faithful summaries should be entailed by the source text. **Domain Adaptation**. For `Off-The-Shelf`, we use a state-of-the-art entailment consistency model: SummaC (Laban et al., 2022). SummaC computes a faithfulness score for a summary sentence by computing separate entailment scores for all source sentence-summary sentence pairs and then aggregating. For `In-Domain`, we use the SciFIVE Model[5] with SOTA performance on the MedNLI dataset (Romanov and Shivade, 2018)–clinician-annotated entailment corpus whose premises come from MIMIC-III. SciFive is provided the summary sentence and its aligned source text as input, and generates a label: {`contradiction, neutral, entailment`}. To be able to compute correlations to human annotations, we convert each class label to an integer in the set $\{-1, 0, 1\}$.

7. Meta-Evaluation of Existing Metrics

Separately for each sentence of each summary in the human annotation set (245), we compute a human error rate `HErr`: defined as the fraction of summary elements (SE) in the sentence marked as either `Not In Notes`, `Incorrect`, or `Missing`[6]. We report the instance-level Pearson (Cohen et al., 2009) correlation coefficient between `HErr` and metric scores (two 245 length vectors). In Appendix H, we breakdown metric correlations separately by error category: `Incorrect`, `Missing`, and `Not in Notes`. At a high-level, metrics are unable to capture `Missing` errors as they require the deepest understanding of patient histories. `Not in Notes` are easiest as they require a more surface-level comparison of words and concepts between summary and source notes.

5. `razent/SciFive-large-Pubmed_PMC-MedNLI` on HuggingFace.
6. Unless explicitly stated, we do not distinguish between error type or severity (Minor, Critical) for the meta-evaluation.

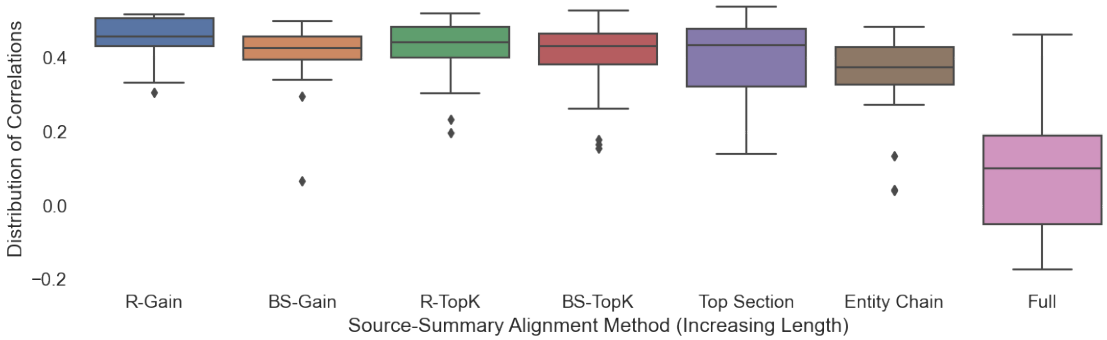


Figure 2: The effect of alignment granularity on the distribution of instance-level Pearson correlations to human judgments across a wide range of metric variants (42). Correlations are more stable across metrics (higher average, higher minimum, and less overall variation) when the inputs (source-summary alignments) are shorter in length.

7.1. Finding the Optimal Source Granularity

Research Question. How much of the source text (averaging $< 20k$ tokens across > 40 notes) is necessary to provide as input for a metric to achieve a high correlation with clinicians?

Experimental Setup. To answer this question, we vary the number of source sentences provided to *every* metric and variant from §6.2 and analyze its impact on performacne (`HErr`).

Findings. Figure 2 reveals that metrics have higher correlations to human judgment when the inputs to the metric are shorter (with ROUGE-Gain being the shortest and having highest average Pearson Correlation of .46). The standard deviation of average instance-level correlations grows monotonically as alignments grow longer. Also, using the entire source is the most volatile (minimum of $-.17$). These findings strongly suggest that scoring summaries based on the full source input (often hundreds of notes) is not only computationally unnecessary, but detrimental.

7.2. Optimal Alignments for Metric Tuning

Research Question. §7.1 reveals that shorter source alignments are preferable when *using* metrics. Is the story the same when *tuning* metrics?

Experimental Setup. To answer this question, we breakdown metric performance (correlation to `HErr`) by the alignment method used for metric *tuning* and, separately, for *usage*. We consider 4 metrics (`Tuned In-Domain` and `Double In-Domain` variants for BARTScore and CTC). Each training instance is a summary sentence from `Train - HIV` and its aligned source context.

		Usage Alignment						
		R-Gain	BS-Gain	R-TopK	BS-TopK	Top Section	Entity Chain	Tune Avg
Tune Alignment	R-Gain	**.467**	.449	.458	.449	.397	.344	.427
	BS-Gain	**.458**	.387	.427	.382	.396	.351	.400
	R-TopK	**.449**	.440	.442	.446	.408	.387	.428
	BS-TopK	**.460**	.411	.435	.407	.416	.387	.419
	Top Section	**.469**	.440	.463	.446	.427	.379	.437
	Entity Chain	.452	.450	**.469**	.438	.407	.379	.432
	Usage Avg	.459	.429	.449	.428	.408	.371	

Table 3: Each row represents the Source-Summary alignments computed for metric *tuning*, whereas the columns denote the alignment method for inference (*usage*). Each cell represents the instance-level metric correlation to the Human Error Rate, averaged across four metric variants (BARTScore and CTC, `Tuned In-Domain` and `Double Domain`). The row-wise max is **bolded** and column-wise is underlined.

Findings. Each cell in Table 3[7] represents an average of instance-level correlations to `HErr` across 4 metric variants (2 for BARTScore, 2 for CTC). Looking at the row-wise maximum values (**cells**), we notice that 5/6 involve using the shortest alignment (`R-Gain`) for metric *usage*. This aligns with our analysis above in §7.1. Yet, the optimal alignment method for metric tuning is much less clear. If anything, we notice that 4/6 of the column-wise maximum values (cells) come from models tuned models from one of the two longest alignment methods (`Top Section` and `Entity Chain`). Additionally, on average, the diagonal values (shaded in gray) do not outperform the non-shaded regions. Taken together, at a high-level, this analysis suggests that additional context may be helpful when learning metrics, yet, when using a metric, providing higher precision contexts are preferable.

7.3. Effect of Summary Granularity

Research Question. For our meta-analysis, we measure faithfulness at the summary sentence level. As such, we have been scoring summaries sentence-by-sentence (`Sentence-Level`). Yet, for some metrics with localized predictions, we can process the entire summary and then extract sentence-level scores (`Summary-Level`). Which method leads to higher metric correlations?

BARTScore Experimental Setup. `Sentence-Level` is the default approach for all metrics, as detailed in §6.2. `Summary-Level` BARTScore involves processing the full summary conditioned on aligned source sentences. For this setting, we simply treat the summary as a "single sentence" and align it to the source sentences. Yet, these source alignments often exceed the BART context window (1,024 tokens). To handle longer inputs, we replace BART with an LED model (which scales up BART to much longer sequences–up to 16k–with sparsified self-attention). We fine-tune for 10,000 steps on `HIV - Train`) using the same LED hyper-parameters from Appendix B.

7. `Full` is not shown because it was not implemented for CTC due to token context restrictions for RoBERTA of 512.

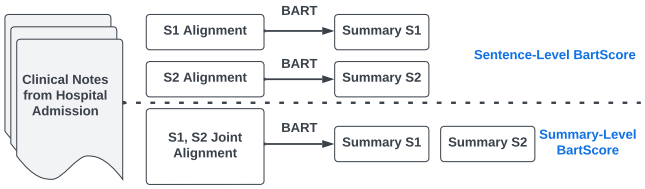


Figure 3: Sentence-Level BARTScore (BART-based) versus Summary-Level (LongFormer Encoder-Decoder (LED)). The LED scales BART to long inputs (> 1024 tokens). While Summary-Level generates a full summary, BARTScores are computed separately for each sentence by extracting logits from sentence boundaries.

Summary Granularity	Source Alignment	Pearson Correlation
Summary Level	ROUGE-Gain	.438
	ROUGE-TopK	.424
Sentence Level	ROUGE-Gain	.516
	ROUGE-TopK	.481

Table 4: BARTScore correlation to human faithfulness labels by summary granularity (processing the full summary at once as opposed to sentence-by-sentence).

BARTScore Findings. Table 4 reveals that `Sentence-Level` BARTScore (with separate alignments computed per sentence) is preferable to processing `Summary-Level` (.516 / .481 versus .438/.424). This relates to the previous finding in §7.1. In both cases, tighter alignment between the inputs and outputs passed to a metric is preferable.

i

7.4. Curious Case of In-Domain Training on Clinical Text

Research Question. There is a wealth of evidence to demonstrate the beneficial impact of in-domain pre-training on clinical (Alsentzer et al., 2019b; Lehman et al., 2023) and biomedical (Gu et al., 2021) downstream tasks. Yet, to our knowledge, no previous work examines the benefits of in-domain pre-training on clinical evaluation metrics. Is domain adaptation: at the pre-training level, and at the task-specific fine-tuning level, necessary for developing clinical faithfulness metrics?

Experimental Setup. We breakdown instance-level metric correlations by the level of domain adaptation: `Off-The-Shelf`, `Tuned In-Domain`, and `Double In-Domain`. We consider `BARTScore`, `CTC`, and `Entailment`. Please see 6.2 for the specific model weights used.

Findings. Table 5 shows a curious trend: that increasing levels of metric domain adaptation is associated with lower correlation to faithfulness annotations at the metric-level and across metrics (average declines .501 → .478 → .468). Below, we link this outcome to summary extractiveness.

Domain Adaptation	Metric	Pearson Correlation
Off The Shelf	BARTScore	.539
	CTC	.507
	Entailment	.453
	Average	**.501**
Tuned In-Domain	BARTScore	.522
	CTC	.462
	Entailment	.450
	Average	**.478**
Double In-Domain	BARTScore	.516
	CTC	.439
	Entailment	.450
	Average	**.468**

Table 5: The impact of domain adaptation of metrics on correlation to human assessments.

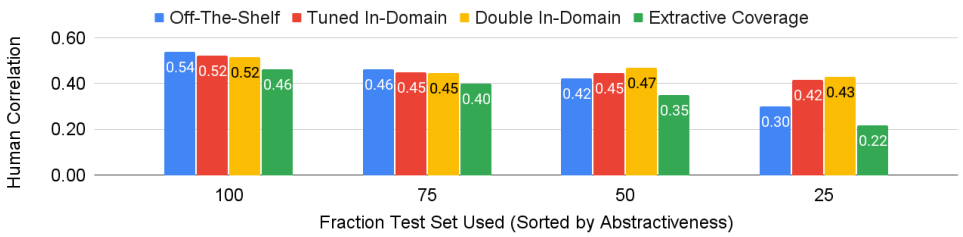


Figure 4: Impact of extractiveness on correlation to clinician annotations. BARTScore variants with different levels of in-domain training are shown, along with Extractiveness (Coverage). Coverage shows the steepest decline in correlation to human labels as average coverage declines, followed by the BARTScore variant most correlated to it (`Off-The-Shelf`). Metrics with exposure to clinical text best on the more abstractive subsets.

Spurious Correlates Hypothesis. Durmus et al. (2022) find that reference-free metrics over rely on spurious correlates: variables which are highly correlated to human annotations on a biased test set, yet less correlated on a more realistic, diverse data distribution. Identifying such correlates is important because it suggests a metric brittleness which may not be captured by simple correlation analysis. As in their work, we focus on summary extractiveness (Grusky et al., 2018) as the potentially spurious correlate. In Appendix Figure 8, we reveal a clear pattern between metric correlation to extractiveness and correlation to the human error rate. In particular, across Coverage (top) and Density (bottom), high correlations to extractiveness are positively related to the correlation with the human error rate. Additionally, we see that in-domain training de-correlates metrics to extractiveness (`Tuned-In-Domain` and `Double In-Domain`. To examine why this might be the case, we examine the extractiveness of reference versus system summaries and a clear bias emerges.

Table 6 shows that references are substantially more extractive in terms of both coverage (percentage of unigrams copied from the source) and density (average squared length of copy-and-pasted fragments) (Grusky et al., 2018). In other words, clinicians write more

Summary	Coverage	Density
Reference	0.88	12.04
Model-Generated	0.95	39.12

Table 6: Model-Generated summaries are *substantially* more extractive (Coverage, Density) than the references on which they are trained. This creates a train-test mismatch for metrics, which are fine-tuned on abstractive summaries and meta-evaluated on extractive ones.

abstractive summaries than the Longformer. To more closely approximate more abstractive, clinician-authored summaries, we examine changes in correlations to human judgments as we filter for more abstractive subsets of the test set. We sort system summary sentences in the test set by coverage and filter for smaller and smaller subsets (making the average coverage lower). Figure 4 reveals that in-domain BARTScore metrics start to outperform when summaries are more abstractive (.30 → .42 → .43 for the smallest bucket, i.e., the top 25% most abstractive sentences in the eval set).

Domain Adaptation	Metric	Pearson Correlation
	Coverage (Cov)	.457
Off The Shelf	BARTScore + Cov	.542
	CTC + Cov	.522
	Entailment + Cov	.524
	Average	**.529**
Tuned In-Domain	BARTScore + Cov	.547
	CTC + Cov	.523
	Entailment + Cov	.535
	Average	**.535**
Double In-Domain	BARTScore	.547
	CTC + Cov	.514
	Entailment + Cov	.535
	Average	**.532**

Table 7: The impact of domain adaptation on metric correlation to human assessments when combining with an easy-to-compute extractiveness statistic (coverage).

Domain-Adapted Metrics are Complementary to Coverage. Recent work demonstrates the efficacy of ensembling de-correlated metrics (Kasai et al., 2022; Colombo et al., 2022). As such, we explore forming a combined metric g given a raw metric score f and a coverage *cov* score: $g = \frac{1}{2} * \left(\frac{f-\mu_f}{\sigma_f^2} + \frac{cov-\mu_{cov}}{\sigma_{cov}^2} \right)$. μ and σ represent mean and standard deviations for f and *cov* across all summary sentences. Table 7 reveals that when combining metrics with coverage, In-Domain adaptation slightly helps. `Off-The-Shelf` averages across three metrics (+ Cov) are .529 versus .535 and .532 for `Tuned In-Domain` and `Double In-Domain`, respectively. Yet, the improvement in correlation is still relatively minor.

8. Distilling a New Metric from Existing Metrics

Adapting to System Outputs with Knowledge Distillation. Even when combined with coverage, domain adaptation does not help much. This may be due to a train-test mismatch. Model summaries use far more copy-and-paste than reference summaries. The above metrics are all trained solely on references yet meta-evaluated on system summaries. To bridge this gap, we can learn a metric directly from system summaries. To do so, we need ground-truth faithfulness labels. The human annotation set is not large enough to train a model, so instead, we rely on the `Train - HIV` subset. We first generate summaries with the LED model and then produce pseudo-faithfulness targets for each sentence. To produce pseudo targets, as shown in Figure 7, we identify a subset of In-Domain metrics with desired attributes: high-correlation to human labels and relatively low correlation to coverage. We then score each summary sentence with each metric in the ensemble, normalize the scores on a per-metric basis, and then average them to produce pseudo-target f for each training instance. We then train a student model, which receives as input a concatenation of a model-generated summary sentence and its aligned source context, and outputs a scalar: f' using the `[CLS]` hidden state. The student is trained with a standard MSE loss: $|f' - f|^2$ and is initialized from clinical/biomedical RoBERTA (Lewis et al., 2020). We train in batches of 8 for $10,000$ steps with a learning rate of $1e-5$ (200 warmup steps, followed by linear decay). For usage, we can *optionally* combine the distilled score with the coverage score.

Via distillation of metrics which are relatively de-correlated with coverage, the goal is two-fold: to learn a single model that achieves a higher correlation on its own to other single-metric variants, and is complementary to coverage when combined.

Metric	Pearson Correlation
Best Single Metric	.539
Best Single Metric + Cov	.547
Distilled Metric	.564
Distilled + Cov	**.573**

Table 8: Distilling a metric from the subset of metrics which are relatively less correlated to extractiveness (coverage) yields higher correlation with human labels than any other single metric. Additionally, combining the distilled metric with (+ Cov) obtains yields superior correlations to all single metric + coverage variants.

Table 8 reveals that the Distilled metric outperforms the best baseline metric variant (.564 vs .539) and, because it is distilled from metrics which are relatively de-correlated with coverage, can be combined at inference with coverage to achieve an even higher correlation (.573). We ran a one-sided Williams Test (Graham and Baldwin, 2014) to estimate the significance of increase in correlation to human labels from `Best Single Metric + Cov` to `Distilled + Cov`. The p-value was .081. As such, we cannot state that the impact of distillation is statistically significant at $p < 0.05$. But, the sample size is small (245).

Multi-Metric Ensembles. Previously, we reported promising performance of our proposed Distilled metric–both on its own and combined with an extractiveness statistic. Yet,

Metric	Pearson Correlation	
	Single	Avg In Ensemble
Coverage (Cov)	.457	.544
BARTScore	.539	.550
CTC	.507	.546
Entailment	.453	.539
BERTScore	.482	.535
Reviser	.324	.528
FactScore	.444	.536
Distilled	**.564**	.556
Best Ensemble	N/A	**.583**

Table 9: Comparing the correlation to human annotations of single metrics, as well as the average correlation of ensembles of metrics that include a given metric. Lastly, we include the correlation of the best performing metric ensemble (Coverage, BARTScore, Distilled).

ideally, we would also want a metric that improves correlation when ensembled with other metrics. To this end, we enumerate all possible ensembles from a set which includes the coverage statistic and 7 metrics: our distilled model and our 6 implemented metrics (BARTScore, BERTScore, CTC, Entailment, FactScore, ReDRESS)[8]. This provides us with $\sum_{n=1}^{N=8} \binom{N}{n} = 255$ unique ensembles, of which each metric takes part in 128. Table 9 shows correlation of metrics to `HErr` for metrics on their own (`Single`), as well as the average correlation to `HeRR` for metric ensembles which include a given metric (`In Ensemble`). Firstly, the metric rankings induced by `Single` and `In Ensemble` are mostly in agreement. Distilled outperforms all baselines on its own (.564) as well as its average correlation when used in an ensemble (.556). The last row of Table 9 shows the correlation of the ensemble with the highest correlation to `HErr`: Coverage, BARTScore, and Distilled. To test significance of the `In Ensemble` results, we bootstrap 95% confidence intervals (CI) for each metric's average `In Ensemble` correlation (1000 samples with replacement from vectors of size 128) and find that the average correlation when `Distilled` is a part of an ensemble is significantly higher ($p < 0.05$) than the average correlation of any of the other 6 metrics (when part of an ensemble). These results demonstrate that `Distilled` is useful on its own and is complementary to other metrics. More broadly speaking, the relative out-performance of ensembling (`In Ensemble` over `Single`) supports the notion that, when developing a metric, it is more useful to focus on complementarity to existing metrics, rather than solo performance (Colombo et al., 2022).

9. Discussion

To guide future research, we distill the findings from our paper into a few salient recommendations.

8. We report the best performing variant across in-domain pre-training / tuning and source-summary alignment methods.

- When evaluating long-form clinical summaries, prioritize finding the minimal set of supporting context for each summary sentence, given that precise alignments improve performance.
- Always report a metric's correlation to extractiveness–the level of copy-and-paste. If it is really high, it might still correlate well with human judgments, but be very brittle and perform very poorly when evaluated on summaries with more abstraction and novel re-phrasing.
- Given the shortcomings of automatic metrics, coupled with the intense cost of collecting human annotations, future work should focus on the role automatic metrics can play in making clinician annotation more efficient. Calibrating metric confidence to accuracy can play a role.
- The intended use case of a model-generated summary should play a role in how to evaluate it. If a model generated summary serves as a draft that is edited by a clinician, then efforts on faithfulness detection should focus on those which are subtle and likely to be missed.

Limitations A primary limitation is the relatively small size of the annotation test set. We consider 29 summaries across a single system. This points to two things: the time-consuming nature of evaluating long-form clinical summaries, as well as the difficulties on collecting annotations for clinical text. Clinical text cannot be precisely judged by non-experts. On the other hand, clinicians are experiencing more burnout than ever and large-scale recruitment remains very difficult, especially in academic settings. Also, due to HIPAA requirements, access to PHI is restricted. We could only recruit from an internal pool of clinicians with pre-existing data access. Future work should be devoted to leveraging automatic metrics to enhance annotation efficiency without introducing bias.

10. Conclusion

We collect fine-grained faithfulness annotations of Hospital Course summaries from clinicians and benchmark metrics against them. For each metric, we consider dimensions relevant to long-form clinical summarization: domain adaptation, input lengths, and output lengths. We find that metrics over-rely on the level of copy-and-paste in summaries. Moreover, metrics struggle with errors which require deep clinical knowledge (such as missingness and errors in the source notes). Learning from explicit human feedback will likely be necessary for deployment in real-world, human-in-the-loop clinical settings in order to more tightly align metric behavior with clinical reasoning (Wei et al., 2022) and to safeguard against errors which could negatively affect patient outcomes.

Acknowledgments

Many thanks to our clinical annotators, as well as the reviewers for their thoughtful comments. We also want to call out Alex Fabbri for his helpful advice on meta-evaluation. This research was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH), as well as the National Institute of Allergy and Infectious Diseases

(NIAID), under Award Number T15LM007079. The content is solely the responsibility of the authors and does not reflect the official views of the National Institutes of Health.

References

Griffin Adams, Mert Ketenci, Shreyas Bhave, Adler Perotte, and Noémie Elhadad. Zero-shot clinical acronym expansion via latent meaning cells. In *Machine Learning for Health*, pages 12–40. PMLR, 2020.

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.382. URL https://aclanthology.org/2021.naacl-main.382.

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.296.

Griffin Adams, Bichlien Nguyen, Jake Smith, Yingce Xia, Shufang Xie, Anna Ostropolets, Budhaditya Deb, Yuan-Jyue Chen, Tristan Naumann, and Noémie Elhadad. What are the desired characteristics of calibration sets? identifying correlates on long form scientific summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10520–10542, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.587.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019b.

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. *Proceedings of Intelligent Scalable Text Summarization Workshop*, 1997.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.236. URL https://aclanthology.org/2022.acl-long.236.

Samuel Chan, Andrew P Maurice, Clifford W Pollard, Stephen J Ayre, Darren L Walters, and Helen E Ward. Improving the efficiency of discharge summary completion by linking to preexisiting patient information databases. *BMJ Open Quality*, 3(1):u200548–w2006, 2014.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. The glass ceiling of automatic evaluation in natural language generation. *arXiv preprint arXiv:2208.14585*, 2022.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.599. URL https://aclanthology.org/2021.emnlp-main.599.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.599. URL https://aclanthology.org/2021.emnlp-main.599.

Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.454. URL https://aclanthology.org/2020.acl-main.454.

Esin Durmus, Faisal Ladhak, and Tatsunori 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)*, pages 1443–1454, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.102. URL https://aclanthology.org/2022.acl-long.102.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.25. URL https://aclanthology.org/2021.conll-1.25.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.57. URL https://aclanthology.org/2021.naacl-main.57.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.187. URL https://aclanthology.org/2022.naacl-main.187.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1213. URL https://aclanthology.org/P19-1213.

Joel Gallant, Priscilla Y Hsue, Sanatan Shreay, and Nicole Meyer. Comorbidities among us patients with prevalent hiv infection—a trend analysis. *The Journal of infectious diseases*, 216(12):1525–1533, 2017.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175, 2019.

Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.114. URL https://aclanthology.org/2021.naacl-main.114.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

Yvette Graham and Timothy Baldwin. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1020. URL https://aclanthology.org/D14-1020.

Max Grusky, Mor Naaman, and Yoav Artzi. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL https://aclanthology.org/N18-1065.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*, 2021.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

Robert E Hirschtick. Copy-and-paste. *Jama*, 295(20):2335–2336, 2006.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.259. URL https://aclanthology.org/2022.naacl-main.259.

Amy JH Kind and Maureen A Smith. Documentation of mandated discharge summary components in transitions from acute to subacute care. *Advances in patient safety: new directions and alternative approaches (Vol. 2: culture and redesign)*, 2008.

Sunil Kripalani, Frank LeFevre, Christopher O Phillips, Mark V Williams, Preetha Basaviah, and David W Baker. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *Jama*, 297(8):831–841, 2007.

Philip J Kroth, Nancy Morioka-Douglas, Sharry Veres, Stewart Babbott, Sara Poplau, Fares Qeadan, Carolyn Parshall, Kathryne Corrigan, and Mark Linzer. Association of electronic health record design and use factors with clinician stress and burnout. *JAMA network open*, 2(8):e199609–e199609, 2019.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Revisiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-5413. URL https://aclanthology.org/D19-5413.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics, pages 2175–2189, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1209. URL https://aclanthology.org/P19-1209.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models? *arXiv preprint arXiv:2302.08091*, 2023.

Gal Levy-Fix, Jason Zucker, Konstantin Stojanovic, and Noémie Elhadad. Towards patient record summarization through joint phenotype learning in hiv patients. *arXiv preprint arXiv:2003.11474*, 2020.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.17. URL https://aclanthology.org/2020.clinicalnlp-1.17.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.334. URL https://aclanthology.org/2021.naacl-main.334.

Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1500. URL https://aclanthology.org/P19-1500.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016, 2019.

Christina Maslach and Michael P Leiter. Understanding the burnout experience: recent research and its implications for psychiatry. *World psychiatry*, 15(2):103–111, 2016.

Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski, and Sanna Salanterä. On evaluation of automatically generated clinical discharge summaries. In *PAHI*, pages 101–114, 2014.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, and Ehud Reiter. Towards objectively evaluating the quality of generated medical summaries. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 56–61, Online, April

2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.6.

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

Amanda J Moy, Jessica M Schwartz, RuiJun Chen, Shirin Sadri, Eugene Lucas, Kenrick D Cato, and Sarah Collins Rossetti. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *Journal of the American Medical Informatics Association*, 28(5):998–1008, 2021.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492, 2021. doi: 10.1162/tacl_a_00438. URL https://aclanthology.org/2021.tacl-1.88.

& Medicine & others National Academies of Sciences, Engineering. *Taking action against clinician burnout: a systems approach to professional well-being*. National Academies Press, 2019.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Kevin J O'Leary, David M Liebovitz, Joseph Feinglass, David T Liss, Daniel B Evans, Nita Kulkarni, Matthew P Landler, and David W Baker. Creating a better discharge summary: improvement in quality and timeliness using an electronic discharge summary. *Journal of Hospital Medicine: An Official Publication of the Society of Hospital Medicine*, 4(4): 219–225, 2009.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.383. URL https://aclanthology.org/2021.naacl-main.383.

Maria Panagioti, Keith Geraghty, Judith Johnson, Anli Zhou, Efharis Panagopoulou, Carolyn Chew-Graham, David Peters, Alexander Hodkinson, Ruth Riley, and Aneez Esmail.

Association between physician burnout and patient safety, professionalism, and patient satisfaction: a systematic review and meta-analysis. *JAMA internal medicine*, 178(10): 1317–1331, 2018.

Jason Phang, Yao Zhao, and Peter J Liu. Investigating efficiently extending transformers for long input summarization. *arXiv preprint arXiv:2208.04347*, 2022.

Prodigy. Prodigy: an annotation tool for ai, machine learning, 2020. URL https://prodi.gy/.

Leonardo Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.236. URL https://aclanthology.org/2022.naacl-main.236.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187. URL https://aclanthology.org/D18-1187.

Denise Albieri Jodas Salvagioni, Francine Nesello Melanda, Arthur Eumann Mesas, Alberto Durán González, Flávia Lopes Gabani, and Selma Maffei de Andrade. Physical, psychological and occupational consequences of job burnout: A systematic review of prospective studies. *PloS one*, 12(10):e0185781, 2017.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://aclanthology.org/P17-1099.

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

Tait D Shanafelt, Lotte N Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P West. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. In *Mayo Clinic Proceedings*, volume 91, pages 836–848. Elsevier, 2016.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3): 331–336, 2018.

Liyan Tang, Shravan Kooragayalu, Yanshan Wang, Ying Ding, Greg Durrett, Justin F. Rousseau, and Yifan Peng. EchoGen: Generating conclusions from echocardiogram notes. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 359–368, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.35. URL https://aclanthology.org/2022.bionlp-1.35.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.199. URL https://aclanthology.org/2022.naacl-main.199.

Carl Van Walraven, Ratika Seth, Peter C Austin, and Andreas Laupacis. Effect of discharge summary availability during post-discharge visits on hospital readmission. *Journal of general internal medicine*, 17:186–192, 2002.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.6. URL https://aclanthology.org/2022.findings-naacl.6.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison,

Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.458. URL https://aclanthology.org/2020.acl-main.458.

Appendix A. Alignment Statistics

Alignment Method	Number of Source Sents
ROUGE-Gain	1.1
BS-Gain	1.8
ROUGE-TopK	5.0
BERT-TopK	5.0
Top Section	13.2
Entity Chain	15.3
Full	921.2

Table 10: The average # of source sentences aligned to each summary sentence by method. $K = 5$.

Table 10 shows the average number of source sentences aligned to each summary sentence by the methods described in §6.1.

Appendix B. LED Training Details

Coarse Filtering. The average length of the inputs ($\sim$ 30,000 tokens) exceeds the maximum sequence length even for transformer models with sparse attention mechanisms designed for long input sequences (Dai et al., 2019; Zaheer et al., 2020; Guo et al., 2021).

Similarly to Liu and Lapata (2019), we learn a simple bi-LSTM model which learns the relevance of each section, to predict the average ROUGE-1 and ROUGE-2 recall of each section vis-a-vis the reference. In particular, we pass a bi-LSTM over the tokens in each section and compute a soft cross-entropy loss between the gold-standard ROUGE-2 recall and the predicted logit (sigmoid(score)). Then, we score each section and filter for the top-K sections. The top 100 sections are provided by an oracle during training and by the model for evaluation.

Fine-Tuning. We fine-tune the Longformer Encoder-Decoder (LED) for 10 epochs with a batch size of 1 and gradient accumulation steps of 16. We set the maximum learning rate to $3e-5$ (tuned in range the range of $1e-6$ to $1e-3$) with a warmup of 200 steps with linear decay. The maximum input size was set to 16,384 and outputs were produced with minimum length of 64, maximum length of 1,024, trigam-blocking, and a beam size of 4 with length penalty 4.0. Training took 8 days on 1 NVIDIA RTX 3090 GPU (24GB).

Appendix C. Entity Extraction

We extract and link entities to the Unified Medical Language System (UMLS (Bodenreider, 2004)) with CLAMP (Soysal et al., 2018) and embed each entity mention with SapBERT (Liu et al., 2021) and first merge all entity mentions which share the same CUI from the UMLS. Exact match of two entities by CUI is far too strict given the size of the UMLS vocabulary as well as extraction noise from abbreviations, acronyms, etc. (Adams et al., 2020). Then, we treat two distinct CUIs as synonyms based on a random forest classifier. The authors of this paper manually labeled 1,000 pairs of entities sampled from 10 different admissions, from a held-out set. The labels were `Unrelated`, `Related`, `Synonyms`. *Ceftriaxone* is `Related` to *antibiotics* since it is in the class of antibiotic, while it is a synonym of *Rocephin*, its brand name. We split the 1,000 manually labeled examples into an 80-20 train-test split and compute features for all pairs of unique CUIs. They include similarity scores (cosine similarity) between CUIs, where CUI embeddings are provided by a pre-trained section-level CUI2Vec model on our corpus, as well as maximum pairwise alignments between mentions from different CUI sets: cosine similarity between SapBERT mention embeddings and lexical similarity (IDF overlap and string levenshtein distance), and finally, binary indicators for TUI and semantic group status from the UMLS.

Appendix D. CTC Generator Details

We use the same masking procedure used to train the CTC model to align the pre-training with the use case and use a BART-Base model to train for 500,000 steps with a batch size of 50 and maximum learning rate of $2.2e-4$, linearly decaying after 200 warmup steps. We show an example of the improvement in Mask-Infilling in Figure 5.

Appendix E. Other In-Domain Metrics

ReDRESS. ReDRESS (Adams et al., 2022) uses a novel hybrid approach that incorporates entity-swapping into a de-noising framework to generate synthetic corruptions on clinical text. Contrastive learning is used to teach another model to reverse the synthetic

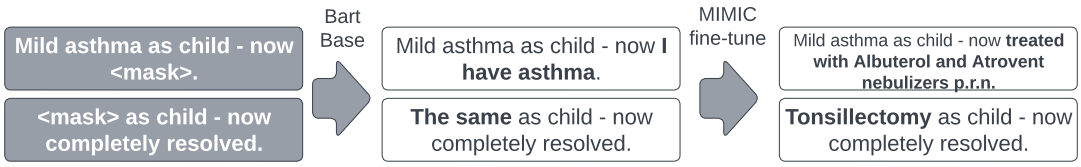


Figure 5: The improvement in Mask-And-Fill completions after fine-tuning in-domain (MIMIC-III Discharge summaries) for just 500,000 steps. Syntactic spans are masked according to the procedure in Deng et al. (2021a).

hallucinations. We adapt it as a faithfulness metric by revising model outputs conditioned on aligned source context and then measuring the revision intensity, e.g., how much was each summary edited to become faithful. We return the BERTScore F-1 between revised and un-revised summaries as the `ReDRESS-Score`: a higher score suggests fewer edits are necessary to re-write the summaries such that they are faithful.

FactScore. As in Adams et al. (2023), FactScore is based on the state of the art model (MultiVERS (Wadden et al., 2022)) trained on the SciFact dataset (Wadden et al., 2020). SciFact is an expert-annotated dataset of 1,409 sentence-level scientific claims. Each summary sentence is scored conditioned on its aligned source sentences (which are varied according to the methods described in §6.1). The `FactScore` is the probability that the MultiVERS assigns to the `SUPPORTED` label.

Appendix F. Impact of Position in Summary on Summary Faithfulness

Similarly to degeneration in unconditional generation tasks (Holtzman et al., 2019). we can measure whether or not quality (as measured by faithfulness) declines at different summary positions. Figure 6 plots the percentage of SE marked with any error by the sentence position in the summary. A clear trend emerges of an increasing error rate as summaries grow longer. This may point to a task-agnostic factor: scaling limitations from full self-attention within the decoder, or task-specific factors: a shift in topics. Figure 6 shows the overall number of SEs decreasing by sentence position. From qualitative analysis, we, in fact, observe a topic shift: from dense history of present illness history recounting (diagnosis-heavy) to concise descriptions of procedures and, finally, any post-discharge instructions.

Appendix G. The Issue of Spurious Correlates

Figure 8 demonstrates that metrics trained on clinical text are, interestingly, less reliant on extractiveness–a proxy for the level of copy-and-pasted text, than non-clinical variants.

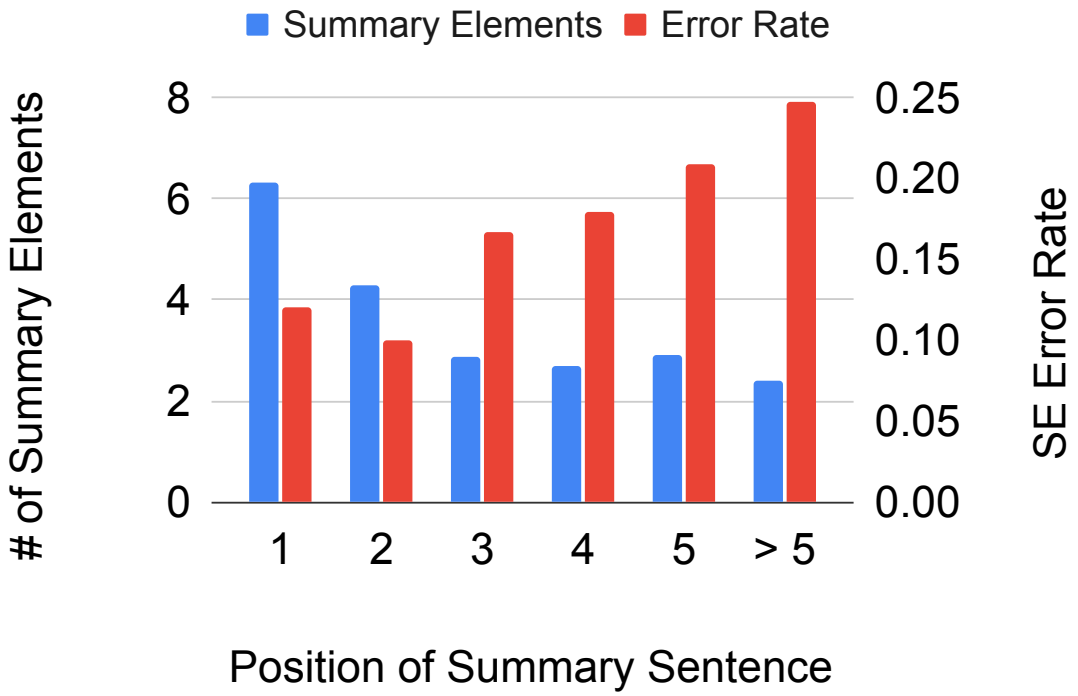


Figure 6: Increasing error rate as summary length increases. There are more SEs at the beginning of summaries, which tend to involve longer sentences and many cover lists of diagnoses for the patients (HPI).

Appendix H. Correlation by Metric Type

Previously, we meta-evaluated metrics against the percentage of summary elements (SE) with *any* error. In this section, we breakdown metric correlations separately by error category: `Incorrect`, `Missing`, and `Not in Notes`. We analyze metrics at the sentence-level against the percentage of Summary Elements in the sentence marked with a certain error. To provide more granular insights, we breakdown error type correlations by Domain Adaptation, Source-Summary Alignment methods, and metric classes (BARTScore vs CTC, etc). Figure 9 shows that `Missing` is the hardest for metrics (the instance-level correlations of metrics to fraction Missing across metric variants), which makes sense given its negligible correlation with Coverage (.021). Not in Notes are the simplest as they tend to be most associated with lexical overlap: .391 Pearson correlation between coverage and fraction of SE's in a sentence identified as `Not in Notes`. `Incorrect` errors can be subtle and are less correlated to coverage than `Missing`: .249. The over-reliance of these metrics on copy-and-paste obfuscates their actual ability to reason over clinical narratives.

Metric-Wise. Figure 10 breaks down correlations to human judgments by metric and errory category. The primary take-away is that metric performance (here, correlation) does

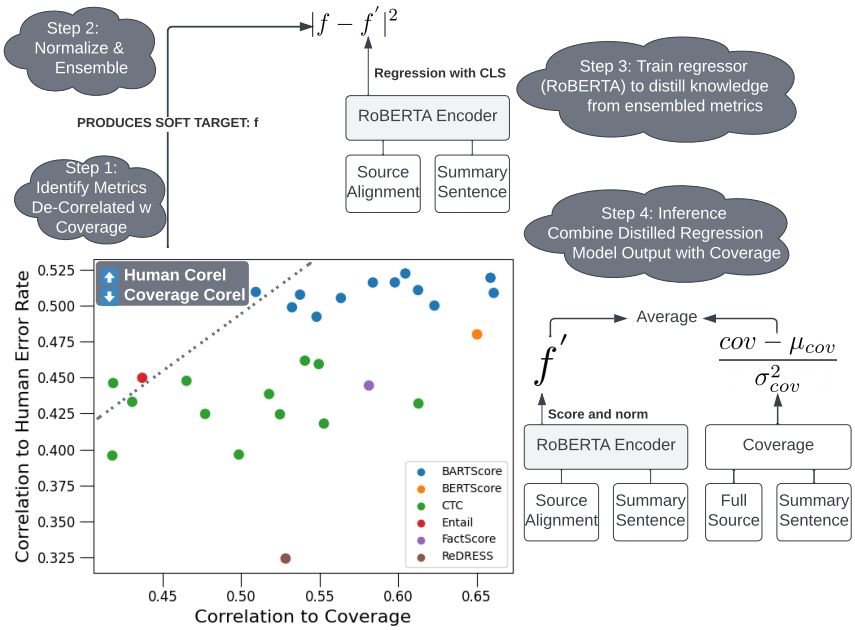


Figure 7: **Step 1**: Identify Optimal Metrics for Knowledge Distillation: High Correlation to Human Labels and Low Correlation to Extractive Coverage. **Step 2**: Normalize and ensemble (average) to produce produce soft targets f on the `Train - HIV` cohort. **Step 3**: Train a model (RoBERTA) as a regressor (f') against the ensembled soft targets f. **Step 4**: Create a combined metric: `Distilled + Coverage`, which combines the score from the RoBERTA model–distilled from metrics relatively less correlated with coverage–with a normalized coverage score.

not exhibit monotonicity across error categories. Excluding Distilled, BARTScore is best at identifying `Any Error`, while Entailment outperforms on `Incorrect Errors`, and CTC performs best on `Not in Notes`. As discussed before, all metrics perform poorly on identifying missing content. CTC learns to identify extrinsic hallucinations so its strong performance on `Not in Notes` makes sense. Entailment metrics are trained on NLI datasets, which target the kinds of logic and inconsistency errors found in `Incorrect`. All metrics struggle with `Missing`. Taken together, these findings reveal that there is no one-size fits all solution to evaluation and we believe that metrics should be designed to fit the particular needs of a system and dataset (Pagnoni et al., 2021). Reporting a single score for meta-evaluation obscures important differences across categories, as well as ignores the potential complementarity of different metrics. Given the potential of ensembling, targeted metrics–which out-perform on one category–may be more valuable to real-world use cases than "jack of all trades, master of none"-type metrics.

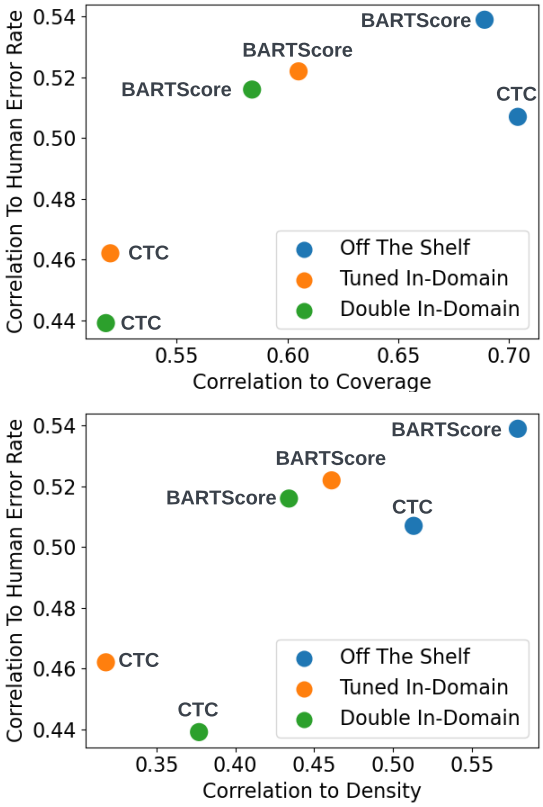


Figure 8: Relationship between Correlation To Extractiveness and Correlation to Human Performance. Each dot represents the best performing (highest correlation) score across each source-summary alignment (see §7.1).

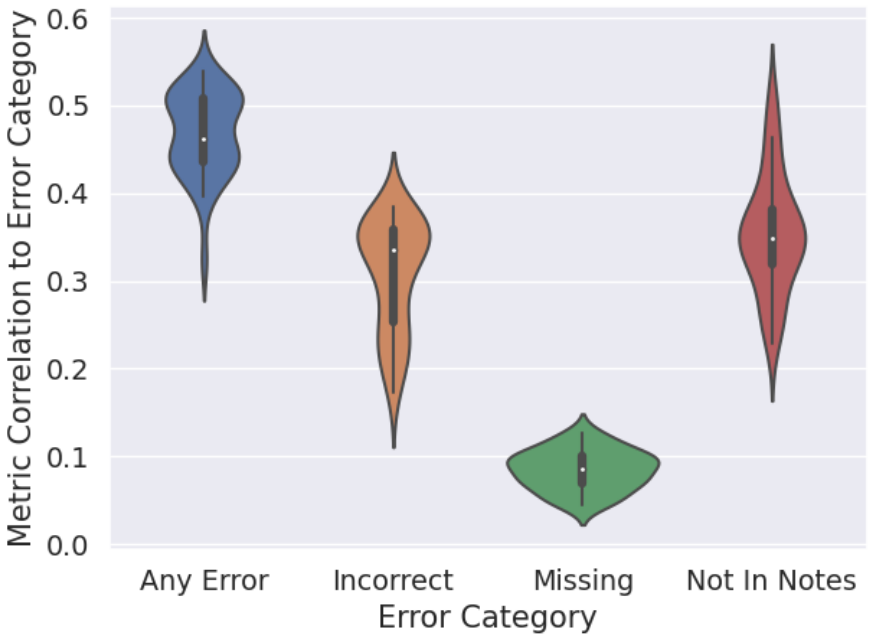


Figure 9: Distribution of Metric Correlations to Human annotations by Category (includes Minor and Critical).

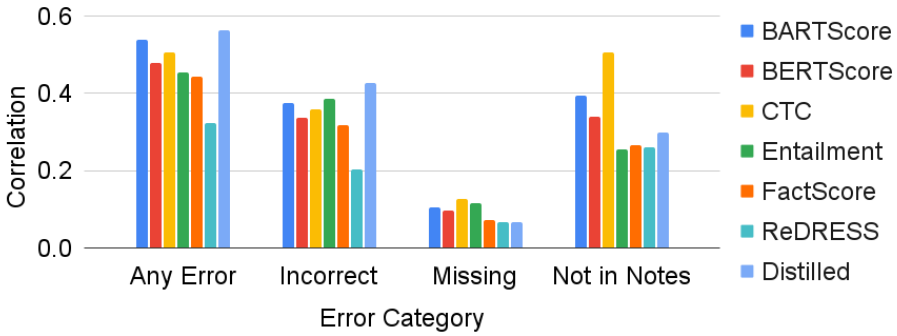


Figure 10: Metric Correlations to Human Judgments by Error Category for each class of metrics from §6.2.