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

Griffin Adams  
*Computer Science*
*Columbia University*
*New York, New York, US*

Jason Zucker  
*Columbia University Irving Medical Center*
*Columbia University*
*New York, NY, US*

Noémi Elhadad  
*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
Adams Zucker Elhadad

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 (Krippalani 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 adaption (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 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 clinical 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-relied 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 practitioners 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 mis-representing 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).

**Human Faithfulness Evaluation.** Assessing faithfulness is a challenging task to automate with metrics (Bhandari et al., 2020), which underscores the importance of collecting high-quality human evaluation annotations (Lux et al., 2020; Wang et al., 2020; Kryscinski et al., 2020b; Maynez et al., 2020; Huang et al., 2020; Fabbri et al., 2021b; Pagnoni et al., 2021; Goyal and Durrett, 2021; Cao and Wang, 2021a; Cao et al., 2022). Additionally, given the relatively small size of each separate evaluation, it can be useful for training and/or meta-evaluation to aggregate them into larger benchmark datasets (Fabbri et al., 2021b; Laban et al., 2022). Based on low inter-annotator agreements for summary-level faithfulness annotations (Lebanoff et al., 2019a; Kryscinski et al., 2020a), recent work has focused more on fine-grained annotations at the entity (Cao and Wang, 2021b; Cao et al., 2022), sentence (Pagnoni et al., 2021), and span level (Maynez et al., 2020). These studies tend to show higher annotator agreement and allow for an understanding of the typology of error distributions across datasets. Sophisticated error taxonomies are generally formulated by examining system outputs (e.g., card-sorting exercises (Lux et al., 2020)) and tend to demarcate error types on two fronts: where the error is located (broken down by syntactic roles) and where it likely comes from (intrinsic or extrinsic). Zhang et al. (2022) challenges the widely held notion that extractive summaries are perfectly faithful by analyzing conflicts in the discourse across sentences.

**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
Faithfulness Metrics for Clinical Summarization

### 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.

<table>
<thead>
<tr>
<th>Split</th>
<th>Admissions</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data for Summarization Model</td>
<td>82k</td>
<td>41</td>
<td>11.6</td>
</tr>
<tr>
<td>Training Data for Evaluation Metrics</td>
<td>2.7k</td>
<td>40</td>
<td>12.5</td>
</tr>
<tr>
<td>Held-Out Human Annotation Data</td>
<td>29</td>
<td>24</td>
<td>12.1</td>
</tr>
</tbody>
</table>

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.

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). 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).

---

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.
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.
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 ~ 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

<table>
<thead>
<tr>
<th></th>
<th>Per Summary</th>
<th>Per Sent</th>
<th>% of All SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Summary Elements (SE)</td>
<td>27.10</td>
<td>3.21</td>
<td>-</td>
</tr>
<tr>
<td>Incorrect SE</td>
<td>2.86</td>
<td>0.34</td>
<td>11%</td>
</tr>
<tr>
<td>Missing SE</td>
<td>0.93</td>
<td>0.11</td>
<td>3%</td>
</tr>
<tr>
<td>Not In Notes SE</td>
<td>1.03</td>
<td>0.12</td>
<td>4%</td>
</tr>
<tr>
<td>Any Mistake SE</td>
<td>4.83</td>
<td>0.57</td>
<td>18%</td>
</tr>
</tbody>
</table>

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

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% 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 §II) 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 “ambulation”, that the “HCP 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”.

8
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., 2021a; 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 ∼ 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 \(\S 6.1\) 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.

**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 discrimination on the CNN/Dailymail dataset. For domain adaption, 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

---

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”.
5. razent/SciFive-large-Pubmed_PMC-MedNLI on HuggingFace.
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.

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. 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.

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 performance ($HErr$).

6. Unless explicitly stated, we do not distinguish between error type or severity (Minor, Critical) for the meta-evaluation.
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.

### 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.

<table>
<thead>
<tr>
<th>Usage Alignment</th>
<th>R-Gain</th>
<th>BS-Gain</th>
<th>R-TopK</th>
<th>BS-TopK</th>
<th>Top Section</th>
<th>Entity Chain</th>
<th>Tune Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Gain</td>
<td>.467</td>
<td>.449</td>
<td>.458</td>
<td>.449</td>
<td>.397</td>
<td>.344</td>
<td>.427</td>
</tr>
<tr>
<td>BS-Gain</td>
<td>.458</td>
<td>.387</td>
<td>.427</td>
<td>.382</td>
<td>.396</td>
<td>.351</td>
<td>.400</td>
</tr>
<tr>
<td>Top Section</td>
<td>.449</td>
<td>.440</td>
<td>.442</td>
<td>.446</td>
<td>.408</td>
<td>.387</td>
<td>.428</td>
</tr>
<tr>
<td>BS-TopK</td>
<td>.460</td>
<td>.411</td>
<td>.435</td>
<td>.407</td>
<td>.416</td>
<td>.387</td>
<td>.419</td>
</tr>
<tr>
<td>Entity Chain</td>
<td>.469</td>
<td>.440</td>
<td>.463</td>
<td>.446</td>
<td>.427</td>
<td>.379</td>
<td>.437</td>
</tr>
<tr>
<td>Usage Avg</td>
<td>.452</td>
<td>.450</td>
<td>.469</td>
<td>.438</td>
<td>.407</td>
<td>.379</td>
<td>.432</td>
</tr>
</tbody>
</table>

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.

**Findings.** Each cell in Table 3 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,

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

![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.](image)

**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.

<table>
<thead>
<tr>
<th>Summary Granularity</th>
<th>Source Alignment</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary Level</td>
<td>ROUGE-Gain</td>
<td>.438</td>
</tr>
<tr>
<td>Level</td>
<td>ROUGE-TopK</td>
<td>.424</td>
</tr>
<tr>
<td>Sentence Level</td>
<td>ROUGE-Gain</td>
<td>.516</td>
</tr>
<tr>
<td>Level</td>
<td>ROUGE-TopK</td>
<td>.481</td>
</tr>
</tbody>
</table>

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.
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.

<table>
<thead>
<tr>
<th>Domain Adaptation</th>
<th>Metric</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off The Shelf</td>
<td>BARTScore</td>
<td>.539</td>
</tr>
<tr>
<td></td>
<td>CTC</td>
<td>.507</td>
</tr>
<tr>
<td></td>
<td>Entailment</td>
<td>.453</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.501</td>
</tr>
<tr>
<td>Tuned In-Domain</td>
<td>BARTScore</td>
<td>.522</td>
</tr>
<tr>
<td></td>
<td>CTC</td>
<td>.462</td>
</tr>
<tr>
<td></td>
<td>Entailment</td>
<td>.450</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.478</td>
</tr>
<tr>
<td>Double In-Domain</td>
<td>BARTScore</td>
<td>.516</td>
</tr>
<tr>
<td></td>
<td>CTC</td>
<td>.439</td>
</tr>
<tr>
<td></td>
<td>Entailment</td>
<td>.450</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.468</td>
</tr>
</tbody>
</table>

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

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.

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.
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.

(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>
<thead>
<tr>
<th>Summary</th>
<th>Coverage</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>0.88</td>
<td>12.04</td>
</tr>
<tr>
<td>Model-Generated</td>
<td>0.95</td>
<td>39.12</td>
</tr>
</tbody>
</table>

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.

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 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-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: \( g = \frac{1}{2} \left( f - \frac{\mu_f}{\sigma_f} + \frac{cov - \mu_{cov}}{\sigma_{cov}} \right) \). \( \mu \) and \( \sigma \) 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.
Table 7: The impact of domain adaptation on metric correlation to human assessments when combining with an easy-to-compute extractiveness statistic (coverage).

### 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.

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
### 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>
<thead>
<tr>
<th>Metric</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Single Metric</td>
<td>.539</td>
</tr>
<tr>
<td>Best Single Metric + Cov</td>
<td>.547</td>
</tr>
<tr>
<td>Distilled Metric</td>
<td>.564</td>
</tr>
<tr>
<td>Distilled + Cov</td>
<td>.573</td>
</tr>
</tbody>
</table>

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).

### 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).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage (Cov)</td>
<td>.457</td>
</tr>
<tr>
<td>BARTScore</td>
<td>.539</td>
</tr>
<tr>
<td>CTC</td>
<td>.507</td>
</tr>
<tr>
<td>Entailment</td>
<td>.453</td>
</tr>
<tr>
<td>BERTScore</td>
<td>.482</td>
</tr>
<tr>
<td>Reviser</td>
<td>.324</td>
</tr>
<tr>
<td>FactScore</td>
<td>.444</td>
</tr>
<tr>
<td>Distilled</td>
<td>.564</td>
</tr>
<tr>
<td>Best Ensemble</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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).

**Multi-Metric Ensembles.** Previously, we reported promising performance of our proposed Distilled metric–both on its own and combined with an extractiveness statistic. Yet, ideally, we would also want a metric that improves correlation when combined 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). 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 each ensemble.

---

8. We report the best performing variant across in-domain pre-training / tuning and source-summary alignment methods.
an ensemble (.556). The last row of Table 9 shows the correlation of the ensemble with the highest correlation to \texttt{HErr}: Coverage, BARTScore, and Distilled. To test significance of the \texttt{In Ensemble} results, we bootstrap 95% confidence intervals (CI) for each metric’s average \texttt{In Ensemble} correlation (1000 samples with replacement from vectors of size 128) and find that the average correlation when \texttt{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 \texttt{Distilled} is useful on its own and is complementary to other metrics. More broadly speaking, the relative out-performance of ensembling (\texttt{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.

- 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


Faithfulness Metrics for Clinical Summarization


Adams Zucker Elhadad


Adams Zucker Elhadad


Faithfulness Metrics for Clinical Summarization


Appendix A. Alignment Statistics

<table>
<thead>
<tr>
<th>Alignment Method</th>
<th>Number of Source Sents</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-Gain</td>
<td>1.1</td>
</tr>
<tr>
<td>BS-Gain</td>
<td>1.8</td>
</tr>
<tr>
<td>ROUGE-TopK</td>
<td>5.0</td>
</tr>
<tr>
<td>BERT-TopK</td>
<td>5.0</td>
</tr>
<tr>
<td>Top Section</td>
<td>13.2</td>
</tr>
<tr>
<td>Entity Chain</td>
<td>15.3</td>
</tr>
<tr>
<td>Full</td>
<td>921.2</td>
</tr>
</tbody>
</table>

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 ($\text{sigmoid(score)}$). 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 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 ac-
cording to the methods described in §6.1). The FactScore is the probability that the MultiVERS assigns to the SUPPORTED label.

![Graph showing the relationship between the position of the summary sentence and the error rate.](image)

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 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.

**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 error category. The primary take-away is that metric performance (here, correlation) does 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
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.

out-perform on one category—may be more valuable to real-world use cases than “jack of all trades, master of none”-type metrics.