Guidance in Radiology Report Summarization: An Empirical Evaluation and Error Analysis

Jan Trienes[†] **Paul Youssef**[‡] **Jörg Schlötterer**[†¶‡] **Christin Seifert**[†‡]

[†]University of Duisburg-Essen, University Hospital Essen
[¶]University of Mannheim [‡]University of Marburg

jan.trienes@uni-due.de
{paul.youssef, joerg.schloetterer, christin.seifert}@uni-marburg.de

Abstract

Automatically summarizing radiology reports into a concise impression can reduce the manual burden of clinicians and improve the consistency of reporting. Previous work aimed to enhance content selection and factuality through guided abstractive summarization. However, two key issues persist. First, current methods heavily rely on domain-specific resources to extract the guidance signal, limiting their transferability to domains and languages where those resources are unavailable. Second, while automatic metrics like ROUGE show progress, we lack a good understanding of the errors and failure modes in this task. To bridge these gaps, we first propose a domain-agnostic guidance signal in form of variable-length extractive summaries. Our empirical results on two English benchmarks demonstrate that this guidance signal improves upon unguided summarization while being competitive with domain-specific methods. Additionally, we run an expert evaluation of four systems according to a taxonomy of 11 fine-grained errors. We find that the most pressing differences between automatic summaries and those of radiologists relate to content selection including omissions (up to 52%) and additions (up to 57%). We hypothesize that latent reporting factors and corpus-level inconsistencies may limit models to reliably learn content selection from the available data, presenting promising directions for future work.

1 Introduction

The radiology report is an important tool for radiologists to communicate examination results with other clinicians. Typically, these reports contain three sections: the background section describing the exam and patient context, the findings section providing a detailed description of observations, and the impression section, which concisely summarizes the key findings (Kahn et al., 2009). In the clinical process, the impression is of high importance as it informs further treatments. However, writing the impression can be time-consuming and error-prone, which is why automatic text summarization systems can substantially improve the quality of clinical reporting (Gershanik et al., 2011).

Background: Technique: Chest, AP and lateral. Comparison: __ and __. History: Weakness and decreased blood sugar with leg swelling and tenderness.

Findings: The patient is status post coronary artery bypass graft surgery and apparently mitral valve replacement. The heart is mildly enlarged. The mediastinal and hilar contours appear unchanged. There is a slight interstitial abnormality, suggestive of a state of very mild congestion, but no new focal opacity. A left-sided pleural effusion has resolved although mild scarring or atelectasis persists. Bones are probably demineralized.

Impression: Findings suggesting mild pulmonary congestion. Resolution of small left-side pleural effusion.

BertAbs (unguided)
findings suggesting mild vascular congestion.

GSum Fixed (guidance = {●})
findings suggest mild vascular congestion.

GSum Variable (guidance = {●, ●})
findings suggest mild vascular congestion. resolution of left-sided pleural effusion.

Figure 1: Example radiology report. We guide abstractive summarization with extractive summaries. We propose to adapt the length of the guidance signal to each report rather than using a fixed setting across all reports which helps to accommodate varying target lengths.

From a summarization perspective, this task involves both an *extractive* component, where important findings are copied verbatim into the summary, and an *abstractive* component, forming those findings into a concise conclusion taking into account the full report (example in Figure 1). Although abstractive methods generate fluent and relevant summaries, they are prone to hallucinations and their output is difficult to control (Maynez et al., 2020; Kryściński et al., 2020; Huang et al., 2020). Therefore, current methods for radiology report summarization employ *guided text summarization* to control the summary content through carefully selected guidance signals such as salient ontology terms (Sotudeh et al., 2020), facts (Zhang et al., 2020b), and clinical entities (Hu et al., 2022).

While summary quality has improved steadily,

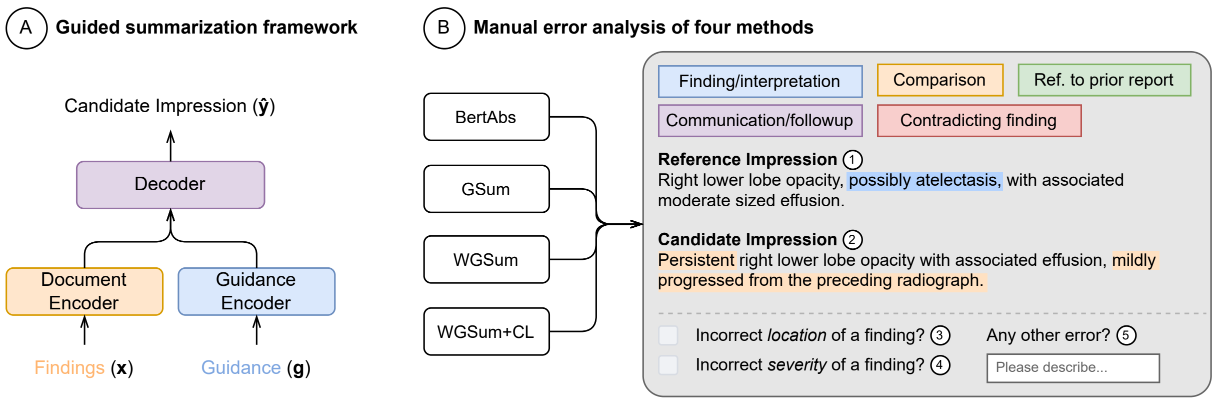

Figure 2: Paper overview. Left: We evaluate guided methods, where the decoder attends to both the input and a pre-selected guidance signal. Right: We task domain experts to identify errors in summaries of unguided and guided methods. Annotation is done on a span-level for omissions from the reference (①) and additions to the candidate (②), and on an instance-level when both texts report the same finding but with incorrect semantics (③,④). Annotators can flag any other error in free-form (⑤).

two key issues have received little attention. First, the success of current methods heavily relies on the availability and quality of the domain-specific guidance extractors (i.e., ontologies, fact extractors, clinical entity taggers). As these resources are expensive to develop, and as they are only available for a few languages and clinical domains, it is of great interest to investigate to what extent we can use domain-agnostic guidance signals to make guided radiology report summarization methods more easily adoptable. Second, while we do see improvements in automatic metrics and human assessments of coarse quality criteria such as fluency, correctness and completeness, we lack a good understanding and quantification of the errors and failure modes of current methods. We address the two issues as follows (overview in Figure 2).

Contribution 1: a variable-length extractive guidance signal (Sections 3 and 4). Motivated by the observation that summaries have a large extractive component, we investigate *extractive summaries* as guidance (Dou et al., 2021). We identify that the *length of the extractive summaries* is critical for the effectiveness of this guidance signal. Intuitively, generating longer summaries requires more guidance than shorter ones. Therefore, we outline two approaches to adapt the guidance length to each report: (i) a classifier that predicts a suitable length, and (ii) a threshold-based method. This variable-length guidance signal improves upon unguided summarization, is competitive with recent domain-guided methods, and cheap to adopt as it does not require any domain-specific resources.

Contribution 2: an error analysis (Section 5). We conduct an error analysis of unguided, guided and domain-guided methods to identify avenues for improvements of all methods. We ask domain experts to identify errors in the outputs of four systems and to characterize them along 11 fine-grained categories. This analysis uncovers three aspects: (1) despite guidance, a significant portion of candidates shows problems in content selection, (2) some content selection decisions are likely only explained by latent factors, (3) there are some dataset-level issues in MIMIC-CXR, including duplicate findings with different impressions, and impression segments without any grounding in the reports.

We make experiment code, full pre-processing pipeline for the datasets, and 1,200 expert assessments of model outputs publicly available.[1]

2 Related Work

Guided text summarization. To address shortcomings in content selection and factuality of neural abstractive summarization methods (Rush et al., 2015; Nallapati et al., 2016), guided methods aim to control the content of summaries through carefully selected guidance signals such as keywords (Li et al., 2018), sentences (Chen and Bansal, 2018), entities (Fan et al., 2018; Narayan et al., 2021), templates (Cao et al., 2018) and prompts (He et al., 2022).

In the radiology domain, Zhang et al. (2018) proposed to guide generation with the background section of reports using a dual-encoder model. With a

[1] github.com/jantrienes/inlg2023-radsum

similar architecture, subsequent work explored the use of salient ontology terms (MacAvaney et al., 2019; Sotudeh et al., 2020). Besides text-based guidance, Hu et al. (2021, 2022) propose a graph-guided decoder which attends both to the report text and to a word-graph of clinical entities. In contrast, we explore extractive summaries as guidance signal (Liu and Lapata, 2019; Dou et al., 2021). Crucially, this guidance signal can be extracted without any domain-specific resources such as ontologies and clinical entity taggers. To this end, our method is similar to the approach by Zhu et al. (2023), which guides summarization with reference summaries from the training set, selected at random or by an oracle. Similar to our approach, this guidance signal can vary in length.

Alternatives to guided summarization. Several parallel research lines aim to enhance radiology report summarization with a different methodological focus. First, several studies optimize factual consistency through reinforcement learning (Zhang et al., 2020b; Delbrouck et al., 2022) or post-hoc reranking (Xie et al., 2023). Second, Karn et al. (2022) devise an extract-then-abstract pipeline with multi-agent reinforcement learning. Last, recent work explores domain-adaptation techniques for pre-trained language models to better accommodate radiology reports (Cai et al., 2021; Van Veen et al., 2023). Our work is orthogonal to these efforts and future work could investigate how to combine them with guided summarization.

Evaluation of radiology report summarization. Evaluating text summarization systems is a long standing issue. Since automatic metrics have a limited correlation with human judgment (Fabbri et al., 2021), manual evaluation is still regarded as the gold standard. For the task of radiology report summarization, most manual evaluations focus on coarse criteria such as accuracy, completeness, conciseness, and readability (Zhang et al., 2018; Hu et al., 2022; Cai et al., 2021). Yet, these evaluations only provide limited insights into directions for improvement. To support the interpretation of automatic and manual evaluations, and to understand the pitfalls of current methods, we conduct an error analysis (van Miltenburg et al., 2021). In this line of work, Yu et al. (2022) evaluated the ability of automatic metrics to capture six fine-grained errors of radiograph-to-impression models. We extend this taxonomy in our error analysis.

3 Method

We formulate the task of summarizing radiology reports as follows. Given the findings section of a report, represented as a sequence of tokens $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, the goal is to generate an impression $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)$ that accurately summarizes the most significant findings. The guided summarization framework extends this problem setting with an additional input signal $\boldsymbol{g} = (g_1, g_2, \ldots, g_L)$ which aims to improve the quality of generated summaries by indicating salient information in $\boldsymbol{x}$.

3.1 Model and Extractive Guidance

As a concrete implementation of the guided summarization framework, we adopt GSum (Dou et al., 2021). This sequence-to-sequence model extends a transformer-based architecture for abstractive text summarization (Liu and Lapata, 2019) with an additional encoder for the guidance signal $\boldsymbol{g}$. To create a guidance-aware representation of the input, the decoder first attends to the encoded representation of $\boldsymbol{g}$, and afterwards to the whole input document $\boldsymbol{x}$ using cross-attention (Vaswani et al., 2017). The authors demonstrate that GSum is effective at controlling the content of summaries, leading to good results on several non-medical datasets.

Extractive guidance. While $\boldsymbol{g}$ can take any form, Dou et al. (2021) found the output of an extractive summarization to be highly effective. Intuitively, this guidance signal informs the model about which input sentences should be highlighted in a summary. An important implementation detail of GSum is the mechanism to obtain the extractive sentences. Dou et al. (2021) distinguish between the **oracle** setting and the **automatic** setting. In the oracle setting, the guidance sentences are greedily picked from $\boldsymbol{x}$ such that they maximize ROUGE with respect to $\boldsymbol{y}$ (Nallapati et al., 2017). In the automatic setting, this oracle is approximated by an extractive summarization method (BertExt, Liu and Lapata, 2019). The training labels for BertExt are derived using the same greedy matching, thus BertExt can be considered an approximation of the oracle guidance. Selecting the guidance signal from BertExt follows a top-k strategy: scoring all sentences for relevance and selecting the highest scoring sentences until a predefined length threshold is reached (Nallapati et al., 2017; Liu and Lapata, 2019). Following Dou et al. (2021), the oracle signal is used during training of GSum. During inference, we state explicitly whether we use the oracle or automatic variant.

3.2 Variable-length Extractive Guidance

We empirically find that extracting a fixed-length summary with the top-k approach has a negative impact on the effectiveness of GSum (Section 4.2). To address this problem, we propose two methods to select a variable-length extractive guidance signal from BertExt. Formally, for a given document $\boldsymbol{x}$ and its sequence of sentences $(s_1, \ldots, s_N)$, with s_i being the i-th sentence in $\boldsymbol{x}$, these methods have to select $L < N$ sentences as guidance $\boldsymbol{g}$.

Method 1: predicting oracle length. As described in Section 3.1, BertExt is trained to assign a label $y \in \{0, 1\}$ to each sentence s_i. The predicted probability $p(y = 1|s_i)$ indicates if s_i should be included in the summary. The ground-truth labels are derived from an extractive oracle $f_{\text{oracle}}(\boldsymbol{x}, \boldsymbol{y})$ which greedily selects a subset of sentences of length $[0, 3]$ that maximizes ROUGE against the gold summary $\boldsymbol{y}$ (Liu and Lapata, 2019). Instead of taking a fixed number of sentences with highest probability (top-k), we train a sequence-classification model to predict the length of the extractive oracle $f_{\text{approx}}(\boldsymbol{x}) = L \in [0, 3]$, and select the top-$L$ sentences as guidance signal.

Method 2: threshold calibration. Instead of considering the full ranked list of sentences, this method constrains selection with a threshold-based approach inspired by Jia et al. (2021). Recall that $p(y = 1|s_i)$ denotes the probability that BertExt assigns to the positive class. We then select the set of sentences that exceed a probability threshold T as guidance signal:

$$\boldsymbol{g} = \{s_i \in \boldsymbol{x} | p(y = 1|s_i) \geq T\}.$$

We optimize $T \in [0, 1]$ on a validation set to maximize ROUGE-1.

4 Technical Evaluation

RQ1. To what extent are extractive summaries an effective guidance signal for radiology report summarization?

RQ2. How does adapting the extractive guidance length to each report impact the overall quality of summaries?

4.1 Experimental Setup

Datasets. We use two public datasets of English chest x-ray reports: **MIMIC-CXR** (Johnson et al., 2019a) and **OpenI** (Demner-Fushman et al., 2015). Consistent with prior work (Zhang et al., 2018; Sotudeh et al., 2020; Hu et al., 2022), we retain reports with exactly one findings and one impression section, where both have an acceptable length (≥ 10 tokens in findings, ≥ 2 tokens in impression), and we discard the background section.[2] Following Hu et al. (2022), we use the official training, validation and test splits of MIMIC-CXR and a random split with a 70/10/20 ratio for OpenI. We use SPACY for tokenization and NLTK for sentence segmentation.[3] Table 1 reports the dataset statistics.

Aspect	MIMIC-CXR	OpenI
Reports	122,500 / 963 / 1,598	2,342 / 334 / 670
Avg. $\lvert\boldsymbol{x}\rvert_t$	56 ± 25.2	37 ± 16.4
Avg. $\lvert\boldsymbol{x}\rvert_s$	5.5 ± 1.9	4.6 ± 1.6
Avg. $\lvert\boldsymbol{y}\rvert_t$	15 ± 13.5	8 ± 8.1
Avg. $\lvert\boldsymbol{y}\rvert_s$	1.6 ± 0.9	1.4 ± 0.8
Novelty	73.4%	86.8%
Compression	73.8%	76.1%

Table 1: Statistics of the benchmark datasets, including the number of reports (train/valid/test), length and standard deviation in tokens/sentences ($\lvert\cdot\rvert_t$ and $\lvert\cdot\rvert_s$), novelty as average percentage of bigrams in impression $\boldsymbol{y}$, but not in findings $\boldsymbol{x}$, and average compression ($\frac{\lvert\boldsymbol{y}\rvert_t}{\lvert\boldsymbol{x}\rvert_t}$).

Baselines. We compare with three categories of baselines: (1) unguided methods, (2) vanilla GSum with fixed-length extractive guidance (Dou et al., 2021), and (3) domain-specific guided methods. Regarding unguided methods, we use **Oracle-Ext** (Nallapati et al., 2017) which greedily selects sentences from the findings that maximize ROUGE with respect to the impression. Furthermore, we use **BertExt** and **BertAbs** (Liu and Lapata, 2019) which are extractive/abstractive transformer-based models initialized with pre-trained BERT (Devlin et al., 2019). Regarding domain-specific methods, we compare with **WGSum** (Hu et al., 2021) which employs a graph-guided decoder to attend to a graph of clinical entities extracted with Stanza (Zhang et al., 2021), and with **WGSum+CL** (Hu et al., 2022) which refines this guidance signal through contrastive learning.

Automatic evaluation metrics. We evaluate the quality of generated impressions with ROUGE F_1 (Lin, 2004) to measure unigram and bigram overlap as a proxy for relevance (R-1, R-2) and the longest common subsequence as a proxy for fluency (R-L). In addition, we report BERTScore as

[2] To compare the relative utility of guidance signals, including the background section is not necessary. For completeness, we report results with background section in Section 5.3.

[3] `spacy.io` and `nltk.org`

a measure of soft-alignment (Zhang et al., 2020a). As factual correctness is critical, we also calculate a factuality F_1 (Zhang et al., 2020b; Hu et al., 2022, Fact.). This metric is based on a rule-based fact-extraction method, CheXpert (Irvin et al., 2019), which labels the status (present, absent, uncertain) of 14 radiological observations. By applying this procedure to both the reference and candidate summary, we can calculate a precision/recall of facts.

Implementation and hyperparameters. For all summarization models, we use the hyperparameters and code of the original papers. Below, we focus on deviations from those settings and report all hyperparameters in Appendix B.

For BertExt, BertAbs and GSum, we make three adaptations: (i) the summary length of BertExt is set to the average number of sentences selected by OracleExt, rounded to the nearest integer,[4] (ii) we reduce the training steps to 20,000 to account for the smaller datasets, and (iii) to address an exploding gradient problem, we reduce the initial learning rate by a factor of 10. For final testing, we take the checkpoint with lowest validation loss on MIMIC-CXR. On OpenI, we found the loss to be unstable, so opted to select models by validation R-1.

Regarding the guidance-length prediction models (Method 1 in Section 3.2), we experiment with two classifiers. First, a multinomial logistic regression classifier with unigram bag-of-words features (LR-APPROX). Second, as this model may be too simplistic to accurately predict the guidance length, we implement a transformer-based classifier (BERT-APPROX) on top of DistilBERT (Sanh et al., 2019).

4.2 Fixed-length Guidance (RQ1)

We first aim to understand if extractive summaries can be a useful guidance signal for radiology report summarization. To this end, we compare BertAbs (i.e., unguided) with GSum in its default configuration (Part 1 in Table 2).

We find that **GSum with fixed-length extractive guidance (Dou et al., 2021) does not generalize to the radiology domain.** Compared with BertAbs, effectiveness decreases by 4.5% and 3.2% in R-1 for MIMIC-CXR and OpenI, respectively. This is surprising as GSum is highly effective on multiple non-medical summarization benchmarks under the same experimental conditions (Dou et al., 2021). Our hypothesis is that highly varying summary lengths make the standard fixed-length guidance in GSum ineffective on this data.[5] We empirically verify this hypothesis in the following experiments.

Comparing oracle and automatic guidance. To get an upper-bound estimate for extractive guidance signals, we analyze GSum in an unrealistic oracle setting. Recall from Section 3.1 that during training of GSum, the guidance signal is extracted by OracleExt, whereas during inference guidance is extracted by BertExt with a summary length fixed to $k = 1$ across all reports. If we instead also use OracleExt as guidance extractor during inference, we see a substantial increase in all metrics (R-1 46.3→58.8 on MIMIC, and R-1 60.1→68.8 on OpenI, all metrics in Appendix Table 4). **This oracle experiment demonstrates (i) that GSum learned to rely on guidance, and (ii) that extractive summaries can be a highly effective guidance signal if selected in the right way.**

Given that GSum is effective when we use the oracle guidance (OracleExt), it is important to understand how this guidance signal differs from the automatically extracted guidance (BertExt). We find that a characterizing difference between the two guidance signals is the length of the resulting summaries. OracleExt produces summaries with 0/1/2/3 sentences for 2/52/32/14% of the MIMIC-CXR reports, and for 15/67/14/3% of the OpenI reports. This implies that a guidance signal with a length of $k = 1$ is too short for 46% of the MIMIC-CXR reports, whereas on OpenI it is too short for 17% and too long for 15%.

4.3 Variable-length Guidance (RQ2)

We next evaluate the utility of our proposed variable-length extractive guidance signal (Part 2 of Table 2). We make several observations.

First, we find that variable-length extractive guidance substantially improves the effectiveness of GSum. On MIMIC-CXR, our adaptation is also better than unguided summarization (BertAbs). In particular, we observe a large increase in factuality, which is critical in the clinical domain. While we see a similar improvement of GSum on OpenI, this guided summarization model does not improve over BertAbs. One potential reason is that OpenI is more abstractive than MIMIC-CXR, as indicated by the high degree of novelty (Table 1) and the relatively low scores of the extractive methods (BertExt, OracleExt in Table 2). This corroborates the findings by Dou et al. (2021), where GSum

[4] On MIMIC-CXR and OpenI $|\text{OracleExt}(\boldsymbol{x}, \boldsymbol{y})| \approx 1$.

[5] Appendix A.1 gives the length distribution of targets.

Method	MIMIC-CXR					OpenI				
	R-1	R-2	R-L	BS	Fact.	R-1	R-2	R-L	BS	Fact.
Part 1: Baselines and reproduction of GSum										
OracleExt	44.0	25.4	40.6	50.1	55.1	30.5	11.9	29.2	33.7	53.5
BertExt (Liu and Lapata, 2019)	32.7	18.1	30.0	41.9	44.5	23.6	7.4	22.6	32.2	**42.8**
BertAbs (Liu and Lapata, 2019)	48.4	34.1	46.6	58.8	47.3	62.0	52.7	61.7	69.2	39.3
GSum (Dou et al., 2021)	46.3	32.7	44.7	57.4	46.6	60.1	49.6	59.8	67.0	40.0
Part 2: GSum adapted with a variable-length guidance signal (ours)										
GSum w/ LR-Approx	48.9	34.2	47.0	59.1	48.2	62.0	51.2	61.6	67.9	41.7
GSum w/ BERT-Approx	49.4	34.5	47.4	59.5	50.6	62.5	51.6	62.2	68.4	39.6
GSum w/ Thresholding	**49.9**	34.3	**47.8**	**59.8**	49.0	62.2	50.8	61.8	68.6	40.4
Part 3: Comparison with domain-specific methods										
WGSum (Hu et al., 2021)	48.4	32.8	46.5	58.6	49.8	61.1	50.0	60.8	67.9	38.4
WGSum+CL (Hu et al., 2022)	49.5	**35.3**	47.8	59.5	**51.1**	**64.7**	**57.1**	**64.5**	**70.0**	37.2
WGSum (Hu et al., 2021)†	48.3	33.3	46.6	—	—	61.6	50.9	61.7	—	—
WGSum+CL (Hu et al., 2022)†	49.1	33.7	47.1	—	—	64.9	55.5	64.4	—	—

Table 2: Technical evaluation of unguided, guided and domain-guided methods on two datasets. Metrics are ROUGE-1/2/L, BERTScore (BS) and CheXpert factuality F1 (Fact). All results were obtained by re-implementing the models with the official code of respective papers, results directly cited are indicated with †.

BertExt length ($k = \cdot$)	MIMIC-CXR		OpenI	
	R-1 (Prec./Rec.)	$\lvert\hat{\boldsymbol{y}}\rvert$	R-1 (Prec./Rec.)	$\lvert\hat{\boldsymbol{y}}\rvert$
Fixed ($k = 1$)	32.7 (**38.5**/34.2)	1.0	**23.6** (**24.6**/26.9)	1.0
LR-APPROX	34.5 (35.7/40.0)	1.4	23.5 (23.9/27.2)	1.1
BERT-APPROX	35.2 (34.6/42.0)	1.5	23.5 (23.7/27.5)	1.1
Thresholding	**36.1** (34.1/**46.3**)	1.7	23.2 (22.9/**29.0**)	1.2
$k = \lvert$OracleExt$\rvert$	36.9 (35.3/44.2)	1.6	24.3 (23.2/29.2)	1.2

Table 3: Comparing strategies for extracting variable-length summaries with BertExt by measuring ROUGE against the gold summary. Average summary length $|\hat{\boldsymbol{y}}|$ given in sentences. All methods are tested as guidance signal for GSum in Table 2.

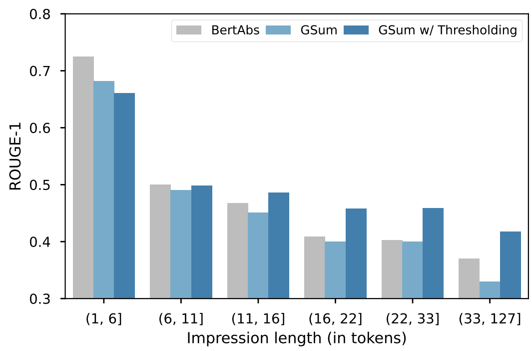

Figure 3: Evaluating summaries by target length on MIMIC-CXR (equal number of samples per bucket).

was less effective on more abstractive datasets. For future work, it would be interesting to study the interplay between the degree of abstraction, and the utility of extractive guidance signals.

Second, regarding the different strategies to obtain variable-length extractive summaries, we cannot conclude that one is superior over another. The classifier-based approaches (LR-Approx, BERT-Approx), and the thresholding-based approach (Thresholding) lead to similar results when the extracted guidance is used downstream in GSum. For each guidance extraction strategy, we calculate the ROUGE scores of the guidance signal with respect to the gold summaries. From Table 3, we see that all strategies have the desired effect of increasing content recall, with a smaller sacrifice in precision.

Third, to better understand how guidance influences the quality of summaries, we plot the R-1 scores across different target summary lengths (Figure 3). We find that variable-length guidance improves the quality of longer summaries, while for shorter targets, extractive guidance is not beneficial. By manual inspection, we find that short targets are standard phrasings of negative results (e.g., *"No evidence of acute findings"*), whereas longer targets have a higher extractive component by reiterating particular findings. In practice, it could be interesting to combine unguided and guided methods by letting the radiologist decide whether a long or short summary should be generated.

Comparison with domain-specific guided summarization (WGSum, WGSum+CL). Lastly, compared with the domain-specific guided methods (Part 3 of Table 2), we find on MIMIC-CXR that GSum with variable-length extractive guidance is just as effective as WGSum and WGSum+CL which use a graph of clinical entities. On OpenI, our approach improves over WGSum, but is slightly worse than WGSum+CL.

Summary of RQ1/RQ2. Overall our results show that extractive summaries are a promising guidance signal for clinical reports without requiring any domain-specific resources. We envision that this makes it easier to adopt guided summarization in other clinical domains and languages, for which domain-specific resources like ontologies and clinical NER models are not widely available.

5 Error Analysis

RQ3. What are the errors and failure modes of unguided and guided methods for radiology report summarization?

5.1 Evaluation Setup

Inspired by the Multidimensional Quality Metrics framework for evaluation of machine translation systems (Lommel et al., 2014), we conduct a span-based error annotation. We task annotators to highlight erroneous text spans and to classify them according to an error taxonomy. As a starting point, we use the taxonomy proposed by Yu et al. (2022). Based on two pilot runs, we extended this taxonomy from initially 6 to 11 fine-grained error categories (see Figure 4) and developed a definition and examples for each. Following Yu et al. (2022), we opt for a reference-based evaluation. We want to understand how the system generated summary differs from the clinician summary both in content and correctness of the presented facts. Therefore, our errors can be grouped into additions (spans in the candidate), omissions (spans in the reference), and binary choices for the correctness of presented facts. Further, we ask annotators to flag any additional errors they encounter as a free-form answer. We provide full annotation guidelines in Appendix D.

Materials. We randomly select 100 reports from the official test set of MIMIC-CXR which is stratified to cover both frequent and less frequent inputs/clinical observations (Johnson et al., 2019b). For each input, we generate four candidate summaries using BertAbs (representative of unguided systems), GSum w/ Thresholding (representative of systems with domain-agnostic guidance), and WGSum/WGSum+CL (representative of systems with domain-specific guidance). We present the reference summary and all candidates (in random order) at once to annotators to reduce effort and ensure consistent annotation of similar summaries. Each set of summaries is completed by three annotators resulting in 1,200 error annotations (100 reports × 4 candidates × 3 annotators). We form a "gold standard" from the triple annotation by majority voting (example aggregation in Appendix C).

Annotators. To account for the domain knowledge necessary for the annotation task, we hired 6 senior medical students in their fifth year of training. All annotators are fluent in English. We compensated annotators with 10.5€ per hour (standard rate for student assistants in Germany). The annotation took 23.1 hours (avg. 4.6 min/sample), plus additional time for pilot rounds and discussions.

5.2 Results (RQ3)

We report aggregated error counts and example annotations in Figure 4.[6]

Overall, we find that the prevalence of errors is comparable across the investigated methods, and that only 14–22% of generated summaries are error-free. The most common errors are omissions and additions of findings, which indicates that the models struggle to select relevant content (1a. 43–52%; 2a. 44–57%). Compared with unguided summarization, there is a slight trend that guided methods reduce the risk of omissions, while only WGSum+CL succeeds at doing this without sacrificing precision. Even though additions are common, they rarely contradict the reference (2e. 0–3%). Similarly, when both the reference and candidate present the same findings, errors related to their clinical correctness are rare (3. 5–8%; 4. 6–9%).

A surprising finding is the common omission and addition of clinicians' communications (1d. 16–20%; 2d. 3–8%). By manual inspection (examples in Figure 4), we find that these are specific actions that a clinician performed after the examination such as informing colleagues about the findings, or recommending additional analysis. Additions of this kind have likely no grounding in the underlying report. To successfully generate such statements, models would require additional context information or guidance from a user.

5.3 Discussion

Overall, our error analysis reveals that the key differences between model-generated impressions and radiologists' impressions relate to content selection (i.e., a tension between completeness/recall and

[6] To measure inter-annotator agreement (IAA), we calculate F_1 for span-annotations (Deleger et al., 2012) and Krippendorffs' Alpha for binary judgments (Krippendorff, 1970). Aggregated IAA: 1. Omissions: 0.61, 2. Additions: 0.60, 3. Incorrect Location: 0.25, and 4. Incorrect Severity: 0.41. IAA by error category for span-level annotations in Appendix C.

#	Error Category	M1 (%)	M2 (%)	M3 (%)	M4 (%)
0	No error	20 (20)	18 (18)	14 (14)	**22** (22)
Omissions from reference					
1a	Finding/interpretation	70 (52)	**58** (43)	62 (48)	64 (47)
1b	Comparison	23 (19)	**16** (15)	19 (16)	23 (19)
1c	Ref. to prior report	**1** (1)	3 (3)	2 (2)	2 (2)
1d	Communication/followup	20 (19)	**18** (16)	19 (17)	19 (17)
Total		114 (66)	**95** (58)	102 (63)	108 (61)
Additions to candidate					
2a	Finding/interpretation	**51** (44)	72 (57)	61 (50)	54 (46)
2b	Comparison	11 (8)	10 (9)	9 (9)	**7** (6)
2c	Ref. to prior report	**0** (0)	1 (1)	**0** (0)	**0** (0)
2d	Communication/followup	5 (5)	8 (6)	8 (8)	**4** (3)
2e	Contradicting finding	**0** (0)	1 (1)	3 (3)	1 (1)
Total		67 (49)	92 (63)	81 (58)	**66** (48)
Semantics of intersecting findings					
3	Incorrect location	**5** (5)	8 (8)	8 (8)	7 (7)
4	Incorrect severity	**6** (6)	7 (7)	7 (7)	9 (9)
5	Other error	31 (23)	**30** (23)	33 (29)	**30** (21)

Reference: Interval increase in vascular engorgement. No frank interstitial edema. No focal consolidations identified.
Candidate (M3): interval increase in pulmonary vascular congestion without evidence of interstitial edema. small right-sided pleural effusion.

Reference: Right lower lobe opacity, possibly atelectasis, with associated moderate sized effusion.
Candidate (M4): persistent right lower lobe opacity with associated effusion, mildly progressed from the preceding radiograph.

Reference: Multiloculated right pleural effusion unchanged since _. New linear and nodular opacities in the left upper lobe may represent carcinomatosis. Findings were relayed to Dr. _ by Dr. _ _ following review on _ at approximiately 11:00 via telephone.
Candidate (M1): stable appearance of multiple loculated right pleural effusion.

Reference: Unchanged size and position of right-sided hydropneumothorax over the last _-hour examination interval.
Candidate (M3): development of new right-sided hydropneumothorax in this patient with history of newly placed pigtail catheter. referring physician, _. _ was paged at 4:45 p.m.

Reference: Little change in the severe bronchiectasis and emphysema.
Candidate (M3): unchanged bibasilar bronchiectasis and bibasilar bronchiectasis.

Figure 4: Results of manual error analysis of 100 MIMIC-CXR reports. Left: number of times each error occurred per method (percent of reports in gray, least errors per row in bold). Right: example error annotations. Models: BertAbs (**M1**), GSum w/ Thresholding (**M2**), WGSum (**M3**), and WGSum+CL (**M4**) [best viewed in color].

relevance/precision). We offer two hypotheses to explain the models' difficulties in this area.

First, there may be latent factors that explain which findings are included in the impression. Among those factors could be patient demographics, the radiograph, prior exams and the clinical question. Typically, this information is available to radiologists through the electronic health records, and is partly documented in the background section of radiology reports. Early work explored using the background section as guidance (Zhang et al., 2018), but more recent work commonly excluded it in pre-processing (Sotudeh et al., 2020; Hu et al., 2021, 2022). We present evaluation results when including the background and observe an overall improvement in almost all metrics for abstractive methods (Appendix A.5). This improvement indicates that (i) additional context supports content selection, and (ii) it could be useful to explicitly model the background in guided summarization.

Second, we anecdotally observed a substantial degree of duplication in the MIMIC-CXR corpus, where reports with identical findings have different impressions (examples in Appendix A.6).[7] This may lead to corpus-level inconsistencies preventing models to reliably learn the selection of findings. We note that there can be numerous reasons for these duplication induced inconsistencies, including the presence of latent factors (see above) and remaining subjectivity/uncertainty in radiologists' assessments. We leave the investigation of this aspect of data quality and potential effects of training data deduplication for future work.

5.4 Limitations

We note two limitations of this error analysis.

First, the analysis is based on comparing candidate impressions with reference impressions. In the absence of the full clinical context, we argue that this is the most reliable benchmark for completeness and relevance of summaries. However, we recognize that we cannot draw any conclusions about the factuality of additions with respect to the full report. To give a first factuality estimate, we conducted a post-hoc analysis with RadNLI (Miura et al., 2021). Let x_i be a sentence in report $\boldsymbol{x} = (x_1, \ldots, x_n)$, and s be an addition span. If RadNLI predicts a contradiction for any (x_i, s) pair, we label this span as contradicting and neutral/entailed otherwise. We find that between 23.4% (BertAbs) and 29.3% (GSum w/ Thresholding) of additions are contradicting, indicating that factuality is another challenge for current models (details in Appendix A.7).

Second, the sample size was driven by time and resource constraints ($N = 100$). To estimate representativeness of this sample, we compare descriptive statistics of the sample with those of the whole test set (length, novelty, compression), and observe that these largely agree (see Appendix C). While we believe that this sample is sufficient to support

[7] 11.9% of the 122,500 MIMIC-CXR training reports have a findings section occurring more than once. Among those reports are only 1036 distinct impressions.

the qualitative conclusions about the failure modes of current methods, a larger study is warranted when the goal is to quantitatively compare the efficacy of different methods.

6 Conclusion

In this work, we revisited guided abstractive summarization of radiology reports. We demonstrated that extractive summaries can be an effective guidance signal for the task, if we allow the length of this guidance signal to vary across reports, and thereby make the gap between domain-agnostic and domain-specific guidance smaller. Furthermore, through a fine-grained error analysis of unguided and guided models we found that guidance successfully steers the content of summaries but that significant deficits in content selection persist.

We hope that this paper motivates future efforts on content selection mechanisms for radiology report summarization, their evaluation in other domains and languages, and on more comprehensive evaluation suites. We release our error annotations which can serve as a starting point for evaluating the efficacy of metrics in capturing these errors.

Ethical Considerations

Privacy sensitive datasets. Both the MIMIC-CXR dataset (Johnson et al., 2019a), and the OpenI dataset (Demner-Fushman et al., 2015) were fully de-identified by the dataset authors in compliance with applicable privacy laws (HIPAA). This includes the removal of any protected health information that may directly or indirectly identify a patient. Nevertheless, the data is still privacy sensitive, and special care was taken to only process it within secured computing infrastructure.

Intended use. We believe that the proposed methods can improve the workflow of clinicians both by reducing the documentation effort and encouraging higher-quality reporting, and thereby improving patient care. However, as our results and discussion show, state-of-the-art summarization methods may not have the desired level of quality that is needed in high-stakes domains such as the clinical context. Therefore, our work is not to be understood in the context of a system that can be deployed, but rather as a step toward a better understanding of the shortcomings of current text summarization methods and providing insight into how these can solved.

Supplementary Materials Availability Statement

- Detailed analysis, hyperparameters, and annotation guidelines are available in Appendices A to D.
- Source code to reproduce all experiments is available from `github.com/jantrienes/inlg2023-radsum/`
- The expert annotations of summarization errors are available from `github.com/jantrienes/inlg2023-radsum/` under the PhysioNet Credentialed Health Data License 1.5.0.
- The MIMIC-CXR (v2.0.0) dataset is available from `physionet.org/content/mimic-cxr/` under the PhysioNet Credentialed Health Data License 1.5.0.
- The OpenI dataset is available from `openi.nlm.nih.gov` (no license terms stated)

Acknowledgements

We thank three anonymous reviewers for their insightful comments, Dennis Aumiller for giving feedback on an earlier version of this work, and Jinpeng Hu for helping with the reproduction of WGSum+CL. We also thank our medical students for valuable discussions and their annotation effort.

References

Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, and Tianming Liu. 2021. ChestXRayBERT: A pretrained language model for chest radiology report summarization. *IEEE Transactions on Multimedia*, 25:845–855.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 152–161.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 675–686.

Jean-Benoit Delbrouck, Maya Varma, and Curtis P. Langlotz. 2022. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 2nd Machine Learning for Health symposium (ML4H), Extended Abstract Collection*.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal

Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, pages 144–153.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4830–4842.

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

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54.

Esteban F. Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, pages 465–469.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5879–5915.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 4980–4990.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4677–4688.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597.

Ruipeng Jia, Yanan Cao, Haichao Shi, Fang Fang, Pengfei Yin, and Shi Wang. 2021. Flexible non-autoregressive extractive summarization with threshold: How to extract a non-fixed number of summary sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13134–13142.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317).

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.

Charles E. Kahn, Curtis P. Langlotz, Elizabeth S. Burnside, John A. Carrino, David S. Channin, David M. Hovsepian, and Daniel L. Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schütze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1542–1553.

Klaus Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. 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.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding generation for abstractive text summarization based on key information guide network. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pages 55–60.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 2014(12):455–463.

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

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

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5288–5304.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 379–389.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2-NIPS)*.

Sajad Sotudeh, Nazli Goharian, and Ross W. Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1899–1905.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG)*, pages 140–153.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Qianqian Xie, Jiayu Zhou, Yifan Peng, and Fei Wang. 2023. FactReranker: Fact-guided reranker for faithful radiology report summarization. *CoRR*, abs/2303.08335.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 204–213.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2020b. 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 (ACL)*, pages 5108–5120.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2021. Biomedical and clinical english model packages for the stanza python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

Yunqi Zhu, Xuebing Yang, Yuanyuan Wu, and Wensheng Zhang. 2023. Leveraging summary guidance on medical report summarization. *CoRR*, abs/2302.04001.

A Detailed Analysis

To support replication, this section provides supplementary analysis on the results of the main part.

A.1 Target Impression Length Distribution and Evaluation by Target Length

We demonstrated in the main part that variable-length guidance helps to adapt to varying target lengths. To better interpret this result, we plot the length distribution of target summaries and the ROUGE-1 score by target-length interval in Figure 5. It can be observed that the length distribution has a long tail with a peak around 4-5 tokens. Impressions of this length are standard phrasings to indicate that no abnormalities were found (e.g., *"No evidence of acute findings"*).

A.2 Evaluating GSum in an Oracle Setting

As a supplement to the oracle experiment in Section 4.2, we provide all metrics for the three inference settings of GSum in Table 4: (i) automatic fixed-length guidance (i.e., extracted from BertExt with $k = 1$), (ii) automatic variable-length guidance but with an oracle length (i.e., BertExt with $k = |\text{OracleExt}(\boldsymbol{x}, \boldsymbol{y})|$), and (iii) oracle guidance (i.e., $\boldsymbol{g} = \text{OracleExt}(\boldsymbol{x}, \boldsymbol{y})$).

MIMIC-CXR	**R-1**	**R-2**	**R-L**	**BS**	**Fact.**
Guidance signal for GSum					
Fixed (Dou et al., 2021)	46.3	32.7	44.7	57.4	46.6
Oracle Length	51.7	36.3	49.6	61.2	52.4
Oracle Length + Content	58.5	42.0	56.2	66.0	60.0
OpenI					
Guidance signal for GSum					
Fixed (Dou et al., 2021)	60.1	49.6	59.8	67.0	40.0
Oracle Length	63.9	53.0	63.5	69.4	42.3
Oracle Length + Content	68.8	56.7	68.3	72.7	45.1

Table 4: Evaluating GSum in an oracle setting. *Fixed* is reproduced from Table 2.

	MIMIC-CXR					**OpenI**				
	R-1	R-2	R-L	B_P	B_R	R-1	R-2	R-L	B_P	B_R
BertExt with fixed-length summaries										
$k=1$	32.7	18.1	30.0	**45.2**	40.1	**23.6**	**7.4**	**22.6**	**33.6**	32.3
$k=2$	**34.1**	**18.6**	**31.3**	40.9	50.1	19.7	6.7	18.9	28.3	39.9
$k=3$	31.7	17.0	29.2	37.0	53.5	17.4	6.1	16.6	25.9	42.8
$k=4$	29.1	15.4	26.8	34.0	54.6	15.8	5.5	15.1	24.0	43.7
$k=5$	27.2	14.3	25.2	32.2	**54.9**	15.1	5.2	14.4	23.3	**44.1**
GSum with fixed-length guidance extracted from BertExt										
$k=1$	**46.3**	**32.7**	**44.7**	**64.6**	52.8	**60.1**	**49.6**	**59.8**	**67.0**	**68.5**
$k=2$	**46.3**	30.3	44.2	58.1	58.5	54.3	43.2	53.9	61.2	66.2
$k=3$	44.1	27.7	41.9	53.6	59.9	54.6	43.2	54.1	61.6	67.3
$k=4$	42.2	26.0	40.2	50.4	**60.2**	53.5	42.1	53.1	60.1	67.5
$k=5$	40.8	24.6	38.8	48.3	60.1	52.7	41.3	52.2	59.5	67.5

Table 5: Testing fixed-length summaries ($k \in [1, 5]$) for BertExt (first block) and as GSum guidance (second block). Metrics are ROUGE-1/2/L and BERTScore precision (B_P) and recall (B_R)

A.3 BertExt: Evaluating Fixed-length Settings

To evaluate if larger values for k in the fixed-summary length setting would improve the effectiveness of BertExt, we generate summaries for all settings of $k = \{1, ..., 5\}$. Analogously, we provide these summaries as guidance signal to GSum. Table 5 reports the results of this experiment. While we find that larger settings of k lead to an increase in recall, we see an equally strong drop in precision, both on BertExt and GSum which demonstrates the necessity of variable-length extractive guidance.

A.4 Evaluating Guidance Length Prediction

To predict the length of OracleExt in the variable-length guidance setting, we employ a logistic regression classifier and a BERT-based classifier (cf. Section 4.1). Detailed evaluation results for both classification models are given in Table 6.

A.5 Including the Background Section

To understand to what extent the background section carries important information for summarizing findings to impression, we prepend it to the findings section and retrain all models. It can be observed that this change improves most abstractive methods on both datasets (Figure 6). For extractive methods results stay largely on par or get worse, indicating that these models do not effectively integrate the background information.

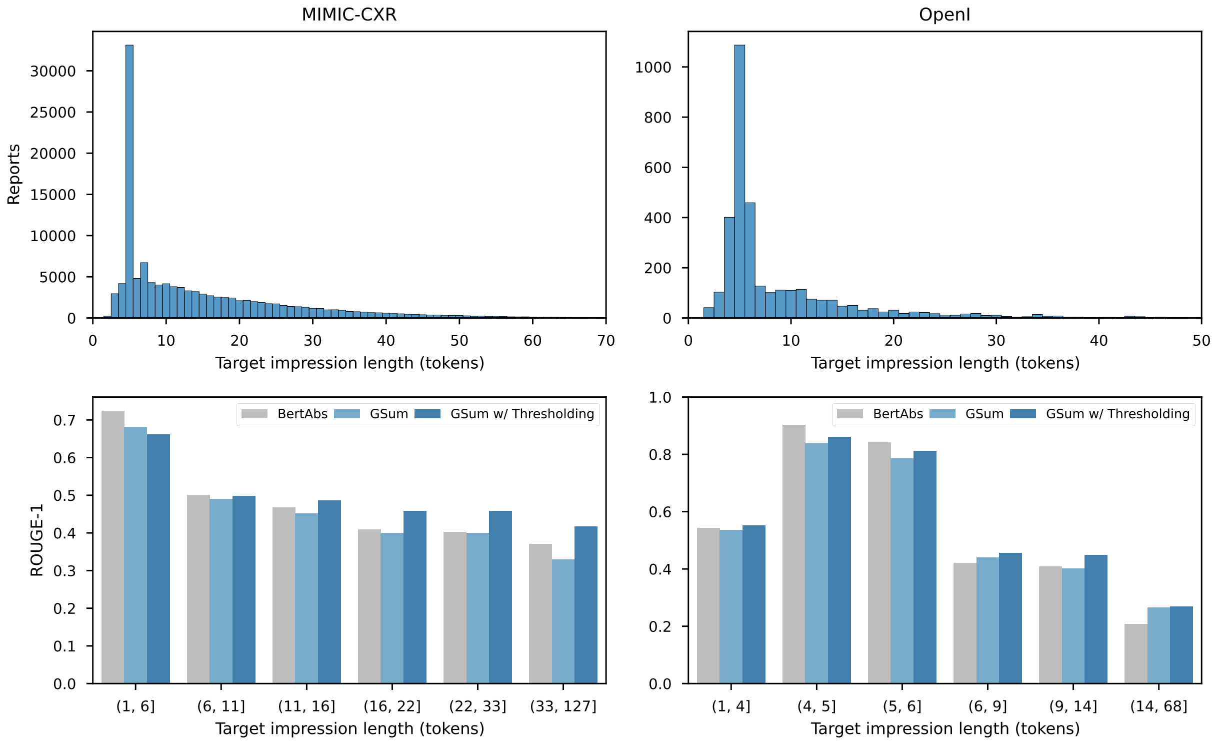

Figure 5: Top row: distribution of target impression lengths. Bottom row: ROUGE-1 by target length for BertAbs (unguided summarization), GSum (fixed-length guidance) and GSum w/ Thresholding (variable-length guidance).

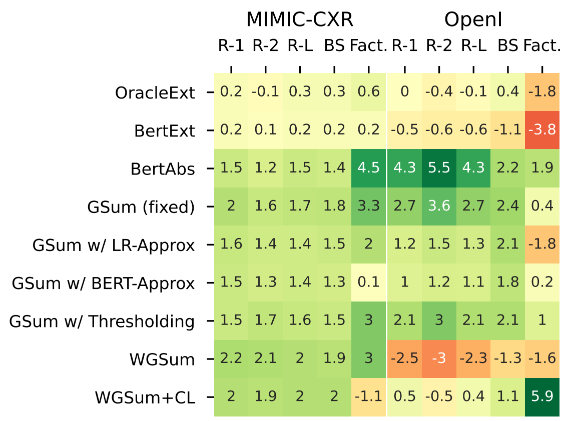

Figure 6: Training models to summarize both background and findings improves most abstractive methods. Scores as absolute delta to the same models without the background section (cf. Table 2).

A.6 Examples of Duplicated Findings and Impressions

We anecdotally observed a large degree of duplication within MIMIC-CXR which may cause corpus-level inconsistencies (see discussion in Section 5.3). This section further quantifies the degree of duplication and provides several examples. Throughout, we only consider instances of *exact* duplication. Of the 122,500 training reports in MIMIC-CXR, we find that 11.9% have a findings section occurring more than once. We present examples of duplicate findings with *different* impressions in Table 12. In addition, we calculate a *label entropy* over the probabilities that each impression occurs for a given finding. We posit that duplicate finding-impression pairs may negatively impact model training in two ways. First, for findings with a high label entropy, the training loss cannot not stabilize (i.e., it is not clear which impression the model should favor). Second, for findings with a low label entropy, the model may learn a kind of "majority vote," which in turn may render models not sensitive enough to generate useful summaries for slightly different findings. We leave further investigation of report duplication to future work.

A.7 Factuality of Additions

As discussed in Section 5.3, we use RadNLI (Miura et al., 2021) to get a first estimate for the factuality of additions marked by annotators in the error analysis. RadNLI obtained an accuracy of 77.8% on a test set of 480 manually labeled sentence pairs in MIMIC-CXR (Miura et al., 2021), which we consider sufficient for an initial exploration of the factuality of additions. Table 7 presents a breakdown of the RadNLI predictions for all addition

Target	LR-Approx	BERT-Approx	
	F-1 (Prec./Rec.)	F-1 (Prec./Rec.)	Support
$k = 0$	46.2 (80.0/32.4)	53.7 (60.0/48.6)	37
$k = 1$	71.1 (63.4/80.9)	71.7 (68.9/74.9)	824
$k = 2$	39.7 (43.1/36.9)	46.7 (45.1/48.4)	512
$k = 3$	30.9 (53.3/21.8)	43.2 (61.5/33.3)	225
Macro Avg.	47.0 (59.9/43.0)	53.9 (58.9/51.3)	1,598
On training set	52.3 (64.1/47.6)	62.5 (69.7/58.5)	122,500

(a) Dataset: MIMIC-CXR

Target	LR-Approx	BERT-Approx	
	F-1 (Prec./Rec.)	F-1 (Prec./Rec.)	Support
$k = 0$	77.7 (85.9/70.9)	84.0 (86.6/81.6)	103
$k = 1$	84.6 (77.2/93.6)	85.4 (79.8/92.0)	450
$k = 2$	19.8 (36.1/13.7)	28.4 (39.6/22.1)	95
$k = 3$	15.4 (50.0/9.1)	8.7 (100.0/4.5)	22
Macro Avg.	49.4 (62.3/46.8)	51.6 (76.5/50.1)	670
On training set	58.5 (83.3/54.3)	51.1 (53.0/51.0)	2,342

(b) Dataset: OpenI

Table 6: Precision, recall and F1 for length prediction of OracleExt. Scores are provided per class on the test set, and as macro-average for both the training and test set. Support indicates the number of samples in each class.

spans and models. It can be seen that the majority of additions is either neutral to the findings section, or entailed by it. Yet, between 23.4% and 29.3% of additions contradict at least one findings sentence, indicating that factuality of radiology report summarization methods can also further be improved.

A.8 Error Analysis: Responses to *Other* Category

We analyze the annotators' comments from the *other* error category, and categorize these errors into two-level hierarchy using a bottom-up approach. Our categorization alongside definitions, examples and counts is shown in Table 13.

B Replication Details for Modeling

We report hyperparameters of the summarization models in Table 9, and for models that predict the length of OracleExt (LR-APPROX/BERT-APPROX) in Table 10. All models were trained on NVIDIA RTX A6000 GPUs with 48GB of memory.

C Replication Details for Error Analysis

Sample statistics. For inclusion in the error analysis, samples were drawn uniformly at random from the official test set of MIMIC-CXR. We compare statistics of the sample with those of the full test set in Table 8.

Model	Entail	Neutral	Contradict
BertAbs	31.9%	44.7%	23.4%
GSum w/ Thresholding	34.5%	36.2%	29.3%
WGSum	32.0%	44.0%	24.0%
WGSum+CL	33.3%	41.2%	25.5%

Table 7: Factuality of additions in candidates (i.e., spans categorized as "2a Finding/interpretation"), as per RadNLI (Miura et al., 2021).

Aspect	Full Test Set	Sample
Reports	1,598	100
Avg. $\lvert \boldsymbol{x} \rvert_t$	70 ± 27.4	63 ± 20.4
Avg. $\lvert \boldsymbol{x} \rvert_s$	6.2 ± 1.9	5.7 ± 1.6
Avg. $\lvert \boldsymbol{y} \rvert_t$	19 ± 15.2	18 ± 12.4
Avg. $\lvert \boldsymbol{y} \rvert_s$	1.8 ± 1.0	1.7 ± 0.9
Novelty	69.8%	69.7%
CMP	71.9%	70.3%

Table 8: Statistics of the MIMIC-CXR test set and the sample used in the error analysis.

Aggregating span-based annotations. From the three annotations we form a "gold standard" as follows: for binary questions we take a majority vote. For span-based annotations, we first group (partially) overlapping spans, and then take a majority vote within each group. We provide an example for the majority voting of span-based annotations below. A1, A2, A3, denote annotators, and `[-eX-]` denotes an error of category X.

```
Tokens:  a   b  c   d   e   f   g   h
A1    : [-e1-]  [-----e2----]
A2    : [-e1-]  [-e1-] [-e2-]
A3    : [-e1-]                [--e1--]
---------------------------------------
Group :    1          2           3
---------------------------------------
Vote  : [-e1-]         [-e2-]
```

Inter-annotator agreement (IAA). We calculate F_1 for span-annotations (Deleger et al. (2012), categories 1 and 2), and Krippendorffs' Alpha (Krippendorff, 1970) for binary judgments (categories 3 and 4) and report the IAA by category in Table 11.

Parameter	BertExt	BertAbs	GSum	WGSum	WGSum+CL
Training Steps (MIMIC)	20,000	20,000	20,000	50,000	100,000
Training Steps (OpenI)	20,000	20,000	20,000	20,000	20,000
LR (Encoder)	2e−3	2e−4	2e−4	5e−2	2e−4
LR (Decoder)	n/a	2e−2	2e−2	5e−2	5e−2
Warmup (Encoder)	10,000	20,000	20,000	8000	10,000
Warmup (Decoder)	n/a	10,000	10,000	8000	7000
Dropout	0.1	0.2	0.2	0.1	0.2
Checkpoint freq. (MIMIC)	1000	2000	2000	2000	2000
Checkpoint freq. (OpenI)	1000	2000	2000	200	200
Decoding	n/a	Beam search	Beam search	Beam search	Beam search
Prediction length	n/a	$\geq$ 5 tokens	$\geq$ 5 tokens	$\geq$ 5 tokens	$\geq$ 5 tokens
Training GPUs	3	5	5	4	3
Inference GPUs	1	1	1	1	1
Base model	bert-base-uncased	bert-base-uncased	bert-base-uncased	None	dmis-lab/biobert-base-cased-v1.1
Parameters	120,512,513	180,222,522	205,433,914	82,260,794	221,600,069

Table 9: Hyperparameters of BertExt/BertAbs (Liu and Lapata, 2019), GSum (Dou et al., 2021), WGSum (Hu et al., 2021) and WGSum+CL (Hu et al., 2022). Training steps, warmup and learning rates were adapted as described in Section 4.1. Remaining parameters kept as in the original publications.

Parameter	Setting
LR-Approx	
Features	Bag-of-words, unigrams with minimum document-frequency of 5, tf-idf
Parameters	3718 (MIMIC-CXR), 592 (OpenI)
Regularization	L2 regularization with strength $C = 1$
Solver	SAGA
Max. Iterations	1000
BERT-Approx	
Checkpoint	`distilbert-base-cased`
Parameters	65,784,580
Optimizer	Adam
Learning rate	2e−5
Epochs	3
Dropout	0.2
Batch size	16
Checkpoint freq.	250
Hardware	6 GPUs

Table 10: Hyperparameters for guidance length prediction models.

#	Category	IAA	Count
	Omissions from reference		
1a	Finding/interpretation	0.64	774
1b	Comparison	0.34	236
1c	Ref. to prior report	0.23	43
1d	Communication/followup	0.83	216
Total		0.61	1269
	Additions to candidate		
2a	Finding/interpretation	0.66	718
2b	Comparison	0.44	155
2c	Ref. to prior report	0.08	17
2d	Communication/followup	0.65	72
2e	Contradicting finding	0.26	34
Total		0.60	996
3	Incorrect location	0.26	111
4	Incorrect severity	0.41	121

Table 11: Inter-annotator agreement (IAA) by category and total number of annotations before majority voting.

#	Finding	Dups.	%	$\|y^*\|$	H	Count	Top-5 Impressions
Most frequent duplicates							
1	PA and lateral views of the chest provided. There is no focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.	1141	0.93	26	0.12	1061	No acute intrathoracic process.
						45	No acute intrathoracic process
						3	No acute intrathoracic process. _, MD
						3	No acute intrathoracic process. Specifically, no pneumothorax.
						3	No evidence of pneumonia.
2	Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.	1033	0.84	34	0.11	974	No acute cardiopulmonary abnormality.
						24	No evidence of pneumonia.
						3	No radiographic evidence of pneumonia.
						2	No acute cardiopulmonary abnormality. No displaced fracture identified. If there is continued concern for a rib fracture, consider a dedicated rib series.
						1	Improving bibasilar atelectasis and decreasing bilateral effusions.
3	The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.	753	0.61	47	0.20	665	No acute cardiopulmonary process.
						15	No acute cardiopulmonary process. No focal consolidation to suggest pneumonia.
						8	No pneumonia.
						7	No evidence of pneumonia. No acute cardiopulmonary process.
						4	No acute cardiopulmonary process. No significant interval change.
Duplicates with highest impression entropy							
4	The heart is normal in size. The mediastinal and hilar contours appear within normal limits. There is no pleural effusion or pneumothorax. The lungs appear clear. Bony structures appear within normal limits.	25	0.02	2	0.99	14	No evidence of acute cardiopulmonary disease.
						11	No evidence of acute disease.
5	The lungs are clear. There is no pneumothorax. The heart and mediastinum are within normal limits. Regional bones and soft tissues are unremarkable.	25	0.02	2	0.94	16	Clear lungs with no evidence of pneumonia.
						9	Clear lungs.
6	The lungs are well expanded and clear. Hila and cardiomediastinal contours and pleural surfaces are normal.	23	0.02	15	0.92	6	Normal. No evidence of pneumonia.
						2	No evidence of pneumonia.
						2	Normal chest radiograph.
						2	No pneumonia.
						1	Normal. No evidence of mass.

Table 12: Examples of exact duplicates in the training set of MIMIC-CXR. In total, there are 14,596 reports with duplicated findings (11.9% of the training data). The table shows the number of reports with a given finding (**Dups.**), the relative frequency in the training set (**%**), the number of distinct impressions with this finding ($|\boldsymbol{y}^*|$), the entropy over the impression frequencies (**H**), and the top-5 impressions with their respective **Count**.

(Sub-)Category	Description	Example	Explanation	Count
1. Incorrect findings: the finding in the reference is replaced with a different and incorrect finding.				29
Finding	incorrectness affects the main finding.	no acute intrathoracic process.	The reference uses "cardiopulmonary process" instead of "intrathoracic process".	21
Past state	incorrectness affects a past state of the patient.	increased opacity in the right lung..	The reference mentions that the opacity is new and did not exist before.	7
Other	incorrectness affects other aspects.	bilateral pleural effusions,..., slightly improved...	The improvement is used to describe a second finding in the reference.	1
2. Imprecise findings: the description of the finding or some of its aspects is imprecise compared to the reference.				73
Finding	the description of the finding itself is imprecise compared to the reference.	...no displaced fractures are seen.	The reference uses "acute fractures" instead of "displaced fractures" (the reference is more general).	21
Location	the location of the finding is imprecise.	retrocardiac opacity compatible with pneumonia...	The references specifies the exact location: "Left lower lobe pneumonia".	21
Certainty	the summary is presented with a different degree of certainty.	bilateral middle lobe opacities could represent atelectasis or pneumonia.	The reference is certain about the finding.	9
Repetition	some findings are repeated.	unchanged bibasilar bronchiectasis and bibasilar bronchiectasis.	bibasilar bronchiectasis is mentioned twice.	6
Count	the count in the finding is imprise.	right pleural effusion.	The reference adds "Multiloculated", i.e., "Multiloculated right pleural effusion"	2
Size	the size of the finding is added/omitted/different.	multiple bilateral pulmonary nodules measuring up to 2. 5 cm.	The reference omits the size.	1
Other	other aspects about the finding are imprecise.	interval resolution of large right pleural effusion...	The reference includes other clinical information.	13
3. Minor/secondary: errors that do not affect the finding.				21
Limitation	some limitations of the examination are (not) mentioned.	no definite acute cardiopulmonary process.	The reference adds "based on this limited, portable examination".	15
Phone calls	The time of a telephone call is different.	...these findings were discussed with dr. _ by _ via telephone on _ at 4 : 45 pm.	The reference mentions a different time for the phone call.	4
Recommendation	errors related to recommendations.	short radiographic follow up is recommended within _ weeks to document resolution.	The reference omits "within _ weeks".	2

Table 13: Bottom-up categorization of errors from the *Other* category with descriptions, examples and counts.

D Annotation Guidelines

Introduction. We consider automatic impression generation for English radiology reports of chest imaging examinations. These reports conventionally have three sections (example in Figure 7).

1. **Background.** A description of the exam, patient information, and relevant prior exams.
2. **Findings.** A description or itemization of the radiologists' observations based on the radiographs.
3. **Impression.** A concise summary of the most important findings, including inferences and any recommendations.

FINAL REPORT

EXAMINATION: CHEST (PA AND LAT).

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

FINDINGS:
Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION: Large left pleural effusion

Figure 7: Example radiology report.

Study setup. We are looking to determine typical failures that automatic impression generation systems make. You will be shown a reference impression and four candidate impressions. The reference was written by a radiologist, and the candidates were generated by automatic systems. Your goal is to judge the accuracy of candidate impressions *based on a reference impression*. For each candidate, you will be asked to identify any errors that it may have.

Definition of "error." We define an error as a difference between the candidate and reference. An error can be one of the following:

1. Omissions
   (a) Omission of finding/interpretation
   (b) Omission of comparison describing a change from a previous examination
   (c) Omission of reference to prior report while making a comparison
   (d) Omission of next steps, recommendation, communications
2. Additions
   (a) Additional finding/interpretation
   (b) Mention a comparison that is not present in reference
   (c) Additional reference to prior report while making a comparison
   (d) Additional mention of next steps, recommendation, communications
   (e) Additional finding/interpretation which contradicts reference
3. Incorrect location/position of finding
4. Incorrect severity of finding
5. Other difference between candidate and reference (please describe...)

Table 14 shows an example for each error category.

Annotating errors as inline annotations. You are asked to annotate errors as *inline annotations*. For each error that you identify, first select the error category and then highlight the relevant text snippet with your mouse. This applies the category. If you have to remove an annotation, press on the highlight and use your backspace/delete key (Entf or ←Backspace). If one of the above categories occurs multiple times, please annotate all of them *separately* (see Figure 8). Some general guidelines:

- A candidate may have multiple errors, so please add all that apply.
- Some candidates will be the same, so please assign the same errors to all candidates.
- For additional findings that are plausible, pick 2a. Additional finding/interpretation. In the context of the full report, these additions may be correct. What this category aims to capture is that the system included information which the radiologist chose not to include. If a finding contradicts the reference, select 2e. Additional finding/interpretation which contradicts reference.
- Use 5. Other freely, especially if you find it difficult to assign any of the above categories. These remarks help us to better understand and characterize potential errors.
- You can ignore differences in word choice if they are synonymous. Example: *"may reflect developing consolidation"* is equal to *"could represent early consolidation."*

Finally, always use your best judgment when assessing the reports. If you are in doubt, you can add any questions/comments about the report or the error categories in the given box.

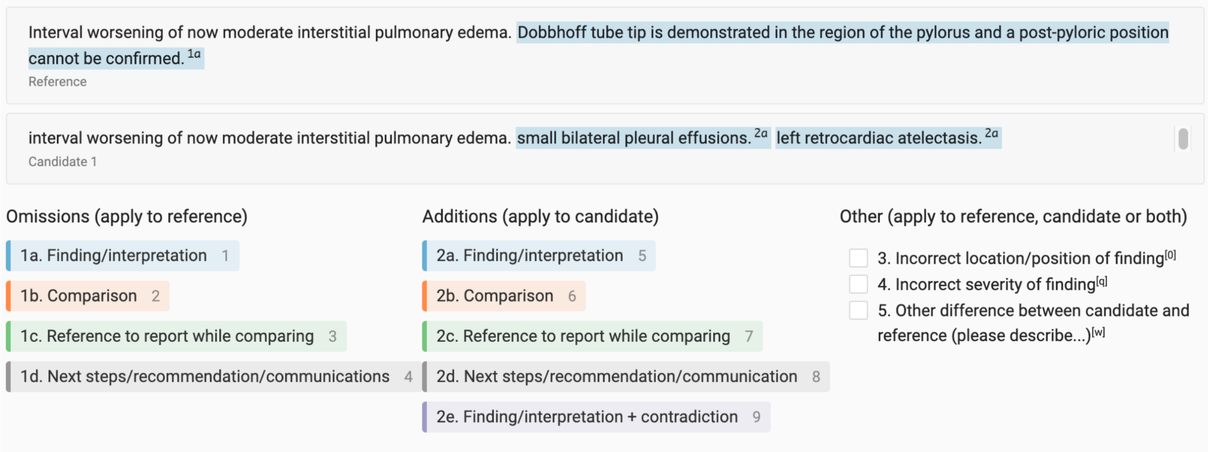

Figure 8: A candidate with two additional findings. Even though they are placed next to each other in the text, apply the category 2a. Additional finding/interpretation twice.

Corner Cases

How to annotate "3. Incorrect location/position of finding" and "4. Incorrect severity of finding"? Only apply if both reference and candidate mention a finding, *and* when there is a mismatch in severity/location. In the example below, both mention effusion, but the reference does not specify the size of effusion, whereas the candidate states that there are "small" effusions. Therefore, apply 4. Incorrect severity of finding.

> **Reference:** interval worsening of now moderate interstitial pulmonary edema. bilateral pleural effusions.
> **Candidate:** interval worsening of now moderate interstitial pulmonary edema. small bilateral pleural effusions.

Opacities vs. consolidation. Often, opacities are used in place of consolidation and vice versa. In those cases, apply 5. Other with a comment similar to "opacities not equal consolidation, but otherwise correct".

> **Reference:** Improved right lower medial lung peribronchial consolidation.
> **Candidate:** right lower medial lung peribronchial opacities have improved.

No acute abnormality vs. COPD. Does "no acute abnormality" contradict "COPD"? No, for the purposes of our evaluation, COPD is not an *acute* disease, so this is not contradicting. In the example below, following categories apply: (1) "COPD" is missing → 1a. Omission of finding/interpretation, (2) "opacity is resolved" → 1b. Omission of comparison describing a change from a previous examination, (3) "no acute cardiopulmonary abnormality" → 2a. Additional finding/interpretation.

> **Reference:** Left basilar opacity is resolved. COPD.
> **Candidate:** no acute cardiopulmonary abnormality.

Misleading grammar or sentence structure. In general, disregard grammatical errors. However, please pay attention to any *logical flaws* that arise because of grammar errors or a misleading sentence structure. In the example below, the *"and"* in the candidate implies that both "bronchiectasis" and "peribronchial consolidation" have improved, whereas the reference only states that the consolidation has improved. In those cases, apply 5. Other and add a comment similar to "logical error because of grammar."

> **Reference:** Bilateral lower lung bronchiectasis with improved peribronchial consolidation
> **Candidate:** bilateral lower lung bronchiectasis and peribronchial consolidation have improved since __.

Error	Reference	Candidate	Explanation
Omissions (apply to reference)			
1a. Omission of finding/interpretation	New left lower lobe infiltrate and effusion.	New left lower lobe infiltrate.	Effusion is missing.
1b. Omission of comparison describing a change from a previous examination	In comparison to _ exam, there is interval near-complete resolution of bilateral pleural effusion.	No evidence of acute cardiopulmonary process.	Resolution of effusion is not described, therefore the comparison is missing.
1c. Omission of reference to prior report while making a comparison	Increased pulmonary edema compared to _.	increased pulmonary edema.	While the candidate correctly states that the edema has increased, it lacks the reference to the prior report (or the date of it).
1d. Omission of next steps / recommendation / communications	No pneumothorax or pneumomediastinum. Recommend repeat PA and lateral imaging later today to verify these findings. Otherwise unremarkable chest radiograph. These findings were communicated to Dr. _ at 11:55 a.m. by telephone by Dr. _.	No pneumothorax or pneumomediastinum.	The candidate does not include the followup *(recommend repeat PA)* and the remark about a communication with another doctor *(These findings were communicated [...])*.
Additions (apply to candidate)			
2a. Additional finding / interpretation	Slight increased hazy opacities at the right lung base which may reflect developing consolidation.	slightly increased hazy opacities at the right lung base which may represent atelectasis or developing consolidation.	Atelectasis is not mentioned in the reference. This finding is not contradicting the reference. It may be correct in the context of the full report. *Same as 1a, but in the other direction.*
2b. Mention a comparison that is not present in reference	Mild to moderate pulmonary edema, increased from _.	Mild to moderate pulmonary edema, increased from _. Stable cardiomegaly.	"Stable" suggests that the state of a finding was compared to a previous examination. This comparison is not made in the reference. *Same as 1b, but in the other direction.*
2c. Additional reference to prior report while making a comparison			*Same as 1c, but in the other direction.*
2d. Additional mention of next steps / recommendation / communications			*Same as 1d, but in the other direction.*
2e. Additional finding / interpretation which contradicts reference	Unchanged size and position of right-sided hydropneumothorax.	Development of new right-sided hydropneumothorax	Unchanged vs. development of new
Incorrect location, Incorrect Severity, Other			
3. Incorrect location/position of finding	New left lower lobe infiltrate	New right lower lobe infiltrate	Left vs. right
4. Incorrect severity of finding	In comparison prior exam, there is near-complete resolution of bilateral pleural effusion	In comparison to _ exam, there is resolution of bilateral pleural effusion	Near complete vs. resolved
5. Other	Slight increased hazy opacities at the right lung base which may reflect developing consolidation	Slight increased hazy opacity at the right lung base which may reflect developing consolidation	Difference in multiplicity
5. Other	left picc terminates within the upper svc.	left picc terminates within the proximal svc.	Ambiguous location
5. Other	No acute abnormalities identified to explain patient's cough and asthma flare.	no acute abnormalities identified to explain patient's cough.	Asthma flare is a symptom, which was not mentioned in the candidate.

Table 14: Examples for all error categories.