Learning to Revise References for Faithful Summarization

Griffin Adams[1]* **Han-Chin Shing**[2] **Qing Sun**[2]
Christopher Winestock[2] **Kathleen McKeown**[1,2] **Noémie Elhadad**[1]
{griffin.adams, noemie.elhadad}@columbia.edu
{hanchins, qinsun, winestock, mckeownk}@amazon.com
[1]Columbia University, New York, NY [2]Amazon AWS AI, Seattle, WA

Abstract

In real-world scenarios with naturally occurring datasets, reference summaries are noisy and may contain information that cannot be inferred from the source text. On large news corpora, removing low quality samples has been shown to reduce model hallucinations. Yet, for smaller, and/or noisier corpora, filtering is detrimental to performance. To improve reference quality while retaining all data, we propose a new approach: to selectively rewrite *un*supported reference sentences to better reflect source data. We automatically generate a synthetic dataset of positive and negative revisions by corrupting supported sentences and learn to revise reference sentences with contrastive learning. The intensity of revisions is treated as a controllable attribute so that, at inference, diverse candidates can be over-generated-then-rescored to balance faithfulness and abstraction. To test our methods, we extract noisy references from publicly available MIMIC-III discharge summaries for the task of hospital-course summarization, and vary the data on which models are trained. According to metrics and human evaluation, models trained on revised clinical references are much more faithful, informative, and fluent than models trained on original or filtered data.

1 Introduction

The tendency of abstractive systems to produce unfaithful summaries is well-studied (Maynez et al., 2020), yet less attention is paid to the role of the data on which the models are trained. This is problematic for two reasons: **(1)** many corpora are naturally occurring–not created for training models–and, as such, are noisy (Kryscinski et al., 2019) and without "inherent quality guarantees" (Bommasani and Cardie, 2020); **(2)** noisy data is detrimental to training faithful models (Dušek et al., 2019).

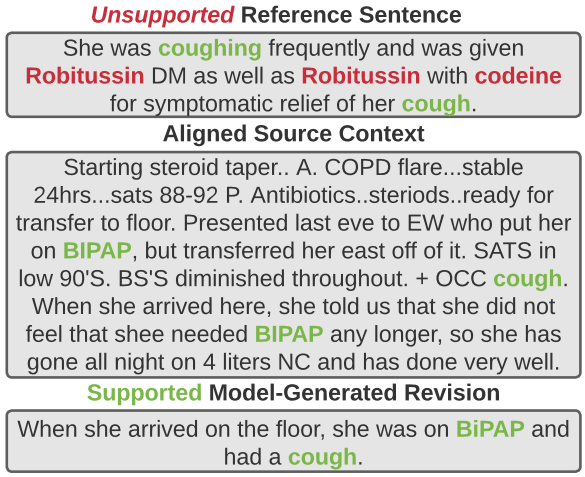

Figure 1: Example of a revised reference sentence. **Robitussin** and **codeine** are edited out of the sentence, while **cough** is correctly kept and a new supported entity **BiPAP** is added. The model is trained on synthetic data to reconstruct well-supported sentences based on context and diverse model-generated hallucinations.

A common approach to deal with training noise is *filtering*: to identify and ignore low quality text at the reference (Kang and Hashimoto, 2020; Matsumaru et al., 2020; Nan et al., 2021a; Narayan et al., 2021) or span level (Goyal and Durrett, 2021). Yet, these methods largely work because they are applied to clean, large-scale corpora. For instance, after removing references with "entity hallucinations", Nan et al. (2021a) still have 855k (92% of the original) training examples for Newsroom, 286k (99%) for CNN/DM, 135k (66%) for Xsum.

We consider a noisier, lower resource setting (clinical summarization) and propose a new approach: to revise–not remove–noisy reference content. First, we align each reference sentence to 1-5 sentences in the source text and classify it as *supported* (to be left alone) or *unsupported* (to be revised). Our objective is to revise all unsupported reference sentences in such a way that retains faithful content, removes unfaithful content, and, as

* This project was completed during an NLP research internship with Amazon Comprehend Medical.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4009–4027
December 7-11, 2022

needed to preserve length, adds relevant context. An example output is shown in Figure 1. In a coherent sentence, the model removes unsupported entities (**Robitussin**, **codeine**) and introduces a twice mentioned concept from the context (**BiPAP**).

To learn this revision task, we need examples of supported and unsupported reference sentences for the same context. Without directly observing it, we generate synthetic data. At a high-level, we take each supported sentence, corrupt it to form a diverse set of unsupported alternatives, and use this mix of real and synthetic data to create examples of (un)faithful revisions for contrastive learning.

As a test case, we consider a task of real-world significance–summarizing a hospital admission–and extract a corpus from a noisy source–notes from the Electronic Health Record (EHR). We experiment with the publicly available MIMIC-III dataset (Johnson et al., 2016). As in Adams et al. (2021), we treat the Brief Hospital Course (BHC) section of the discharge summary as a reference summary and all notes prior to discharge as source. Data coverage is a huge issue as only 60% of reference summary entities can be found in the source.

The contributions of this work are: **(1)** Proposing a new method to address variable reference quality: reference revision, which, as a data preprocessing step, is model agnostic and complementary to other faithfulness approaches; **(2)** Showing that training on revised references can improve faithfulness while also improving informativeness and fluency; **(3)** Providing code[1], pre-processed datasets, and models for alignment, corruption, revision, post-hoc editing, and generation of clinical summaries[2]; **(4)** Beyond its primary use case of data pre-processing, demonstrating that reference revision can have standalone value: as a post-hoc editor and a pre-training objective for faithfulness.

Related Work

Faithfulness. Efforts to address faithfulness have focused on smarter models (Huang et al., 2020), more suitable metrics (Durmus et al., 2020), content-plan editing (Narayan et al., 2021), and post-hoc interventions: ranking (Falke et al., 2019) and editing (Cao et al., 2020; Dong et al., 2020; Zhu et al., 2021). Synthetic errors (Kryscinski et al., 2020) are useful for optimizing (Cao and Wang, 2021) and evaluating (Goyal and Durrett, 2020) faithfulness, yet are best supplemented with fine-grained annotations (Goyal and Durrett, 2021).

The impact of training data noise on faithfulness is less studied. The most common proposal is to identify low quality samples–with entailment (Matsumaru et al., 2020; Goyal and Durrett, 2021) or entity overlap (Nan et al., 2021a; Narayan et al., 2021)–and drop them. These filtering methods tend to improve faithfulness yet can degrade informativeness. Kang and Hashimoto (2020) address the data hunger issue by first training on all samples before implementing Loss Truncation–ignoring high log loss datapoints. This is effective on a relatively clean Gigaword corpus yet untested on noisier corpora in which hallucination behavior from the unfiltered data may be difficult to unlearn. Filtering can be "insufficient to train a competitive model" when high-quality data is limited (Filippova, 2020). Our proposed method takes advantage of all available data while seeking to redress the underlying issue.

Clinical Summarization. Faithfulness is less studied for clinical summarization because most proposed methods are extractive (Pivovarov and Elhadad, 2015; Moen et al., 2016; Alsentzer and Kim, 2018). Abstractive approaches tend to focus on finer temporal granularities, e.g., synthesizing a single radiology report (MacAvaney et al., 2019; Sotudeh Gharebagh et al., 2020; Zhang et al., 2020b) and doctor-patient conversations (Krishna et al., 2020; Joshi et al., 2020; Zhang et al., 2021a).

Most similar, Shing et al. (2021) survey extract-then-abstract approaches to section-specific discharge summarization on MIMIC-III. They measure factuality with entity overlap and remove poorly supported references. When analyzing a proprietary EHR-derived, hospital-course summarization corpus, Adams et al. (2021) perform oracle extractive analysis and confirm, as we do, that EHR-derived summary references are highly noisy.

2 Data

The publicly available MIMIC-III dataset contains de-identified clinical records from patients admitted to Beth Israel Deaconess Medical Center (Johnson et al., 2016). Hospital-admission summarization is a challenging task: in at most a paragraph, a summary must discuss what happened to the patient during the admission, why it happened, and what needs to happen next. To create a dataset for

[1] https://github.com/amazon-research/summary-reference-revision

[2] The datasets are accessible via PhysioNet (Goldberger et al., 2000) and models via HuggingFace (Wolf et al., 2020). Please refer to our GitHub README for further details.

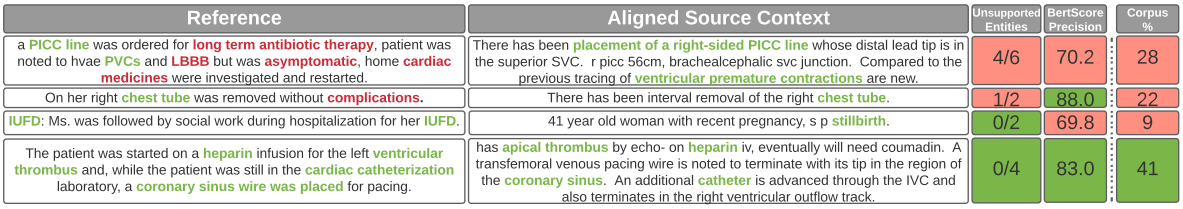

Figure 2: From top to bottom: examples of source-reference alignments which fail to meet both criteria for supportedness (only supported entities and a high BERTScore precision) (28% of corpus reference sentences), just entity overlap (22%), just low BERTScore (9%), and the remaining which meet both (41%). Only the last reference sentence is classified as *supported* and, for this paper, treated as a gold standard when training revisions.

	Statistic	Value
Global	Notes	1.38M
	Unique Patients	47,553
Per Admission (Avg. #s)	Notes	29
	Note types	3.3
	Source sentences	703
	Source tokens	7,553
	Reference sentences	23.5
	Reference tokens	370
Extractive Analysis	Coverage	46.0
	Density	1.2
	Compression Ratio	20

Table 1: Hospital-Admission Summarization Dataset.

this task, we treat the Brief Hospital Course section of the discharge summary as a reference and all notes authored during the patient's stay as the source. Table 1 shows that source documents are long (on average, 7.5K tokens and 703 sentences) and, while references are also long (370 tokens on average and 23.5 sentences), there is a high degree of word-level compression (∼20x). Coverage and density metrics reveal high levels of abstraction[3].

Amazon Comprehend Medical is used to extract entities of the following semantic types: diagnoses (using the ICD-10 classification), medications (RxNorm), procedures, treatments, and tests. To determine whether or not an entity is supported, we compute a similarity score for all mention pairs based on a combination of ontological (RxNorm/ICD-10 codes) and lexical overlap (embedding and exact match) (see Appendix A).

3 Building Source-Reference Alignments

We link each reference sentence to a subset of source sentences to identify the minimal context necessary for revision and determine which sentences need to be revised. We select no more than five sentences because Lebanoff et al. (2019b) find that reference sentences tend to reflect content from very few source sentences. We follow their approach to greedily select sentences with high ROUGE (Lin, 2004) overlap and minimize redundancy by removing covered tokens after each step. Yet, given high levels of abstraction (abbreviations (Adams et al., 2020), misspellings), we increase semantic coverage by replacing ROUGE with BERTScore precision (Zhang et al., 2020a) and adding additional sentences, as needed, to cover all supported entities. Please refer to Appendix B for intuition, notation, and an example.

Classifying References. We treat reference sentences with 0 unsupported entities and a BERTScore precision with respect with its aligned source evidence of ≥ 0.75 as supported. The remaining are unsupported. 417,318 (41%) reference sentences qualify as supported and the remaining 595,300 (59%) unsupported: 47% fail both thresholds (280,839), 38% have hallucination(s) with a high BERTScore (225,423), and 15% have no hallucinations but a low BERTScore (88,189). Figure 2 reveals why both BERTScore and entity overlap are needed to identify full coverage. The first sentence has unsupported entities and poor embedding-based scores. The second is semantically similar yet missing a critical concept (complications) which cannot be inferred from the context. The third has full entity coverage (IUFD is a term for stillbirth) yet BERTScore is low because there is no mention of social work. Only the final sentence is covered by both metrics and treated as supported.

4 Learning to Revise Unsupported Text

The goal is to re-write these *un*supported reference sentences such that they are supported, i.e., covered by the source notes. To learn this revision task without a gold standard ground-truth, we take each supported reference sentence, inject noise to create

[3]Please refer to Grusky et al. (2018) for details on extractive analysis, including formulas for density and coverage.

unsupported, yet realistic, alternatives (§4.1), and then use this mix of real and synthetic data to create a supervised set of positive and negative revisions to support a contrastive learning objective (§4.2)[4].

4.1 Generating *Synthetic* Hallucinations

(D)esiderata. Based on Figure 2, unsupported sentences look normal (unfaithful only in the context of available data) **(D1)**; contain many hallucinated entities **(D2)**; exhibit a wide range of semantic divergence from the aligned source context **(D3)**; and in spite of clear differences, are topically similar to aligned context **(D4)**. To illustrate **D3**, we note that the second sentence could be revised by simply removing the bigram "without complications", yet the first and third sentences require more substantial re-writing to remove unfaithful content while still remaining informative and coherent.

High-Level. The simplest way to construct, and control for, hallucinations is to perform entity swaps (Kryscinski et al., 2020; Zhao et al., 2020; Zhang et al., 2021b; Chen et al., 2021). Yet, this can produce disfluent text which is easily detectable (Goyal and Durrett, 2021). In contrast, generating from a LLM produces fluent (Zhou et al., 2021), more diverse text (Cao and Wang, 2021), yet without as much control over hallucinations. Given our desiderata, we combine entity swaps **(D2)** into a generative framework **(D1, D3)** and incorporate a set of topical entities to avoid excessive semantic drift **(D4)**. Our proposed method is called **ReDRESS**: **re**ference **d**ist**r**actor **e**ntity **s**et **s**wapping.

Training Objective. The **ReDRESS** backbone is a BART encoder-decoder model (Lewis et al., 2020a). BART is trained as a denoising autoencoder to reconstruct a corrupted sentence: $p(s|f(s))$, where f is an arbitrary noise function(s). Across natural language tasks, the BART authors find span deletion to be the most effective noise. **ReDRESS** also uses span deletion but adds an extra noise function: entity swaps. Specifically, for each sentence s, we extract a set of topically related entities $\boldsymbol{e_s}$ and then exchange entities between s and $\boldsymbol{e_s}$. Let us denote span deletion as f and the swap transformation as $g(s, \boldsymbol{e_s}, k) \rightarrow \boldsymbol{e_s}^{+k}_{-k}, s^{+k}_{-k}$, where k represents the number of entities exchanged. To vary the level of corruption (**D3**), we sample a different k for each example such that, on average, half of the entity mentions in s are swapped out for entities in $\boldsymbol{e_s}$. The final pre-training objective is $p(s|k, \boldsymbol{e_s}^{+k}_{-k}, f(s)^{+k}_{-k})$. k is represented by a special token and added to the input to make entity swaps a controllable aspect of generation. Each component (k, $\boldsymbol{e_s}^{+k}_{-k}$, $f(s)^{+k}_{-k}$) is separated by a special `<sep>` token and passed to the BART encoder. During training, the decoder learns to reconstruct s by re-writing $f(s)^{+k}_{-k}$ such that it reverses the k entity swaps performed between $\boldsymbol{e_s}$ and s and fills in a deleted text span[5].

Inference. To use **ReDRESS** to generate a plausible, corrupted version of a sentence s, we apply f to s and sample k to apply g to both s and its distractor set $\boldsymbol{e_s}$. Two key **m**odifications are introduced to discourage the model from reconstructing s: **(m1)** entities removed from s are not added to $\boldsymbol{e_s}$, and **(m2)** k swaps are implemented, yet the model is provided $k+1$ as the swap code to trick the model into performing an additional swap than is required for reconstruction. Using the notation above, **ReDRESS** generates $(k+1, \boldsymbol{e_s}_{-k}, f(s)^{+k}_{-k}) \rightarrow \hat{s}$ using standard beam search decoding. Without access to the original entities from s in $\boldsymbol{e_s}$ **(m1)**, the model looks for plausible, *hallucinated* alternatives. In Appendix C, we show that **(m1, m2)** increase the diversity of synthetic hallucinations $\hat{s}$ to mirror the variable level of coverage observed in the data.

Implementation Details. We train **ReDRESS** from `bart-base` on a large set of unlabeled sentences extracted from MIMIC-III discharge summaries. To find the distractor set $\boldsymbol{e_s}$ specific to each s, we first retrieve the sentences most similar to s (using BioSentVec (Chen et al., 2019) embeddings to create a Faiss index (Johnson et al., 2017)). The first 25 unique concepts extracted from the set of nearest neighbors form the distractor set $\boldsymbol{e_s}$ for s.

4.2 Learning a Revision Model

We apply **ReDRESS** to the set of supported reference summary sentences to generate positive and negative revision examples for contrastive learning.

Notation. Let $\boldsymbol{r}$ represent a reference sentence with aligned source sentences $\boldsymbol{S}$. $\hat{\boldsymbol{r}}_{\boldsymbol{n}}$ is a corrupted version of $\boldsymbol{r}$ generated by **ReDRESS** with random seed $\boldsymbol{n}$. Given N diverse outputs, each generated

[4] We do not rely on *un*supported sentences during training because they are set aside for inference. To use them, we would need to synthetically construct supported alternatives, which is not possible without first knowing how to revise.

[5] $\boldsymbol{e_s}$ is provided as input yet is not part of the target output. In other words, the output of the model is a natural sentence.

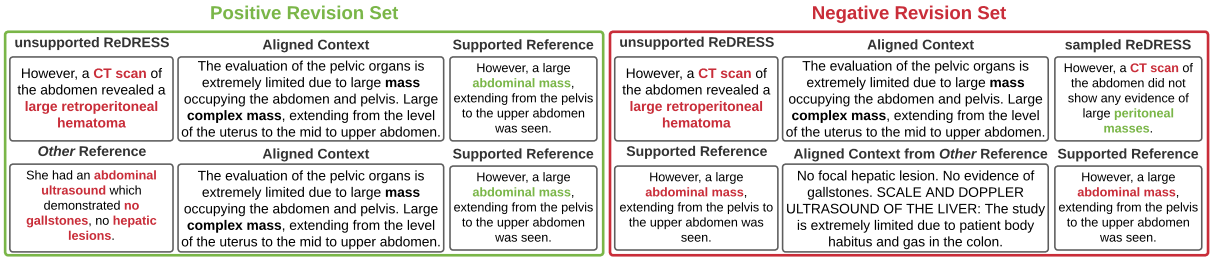

Figure 3: Synthetic positive and negative sets for revision training. The encoder input is the concatenation of the input (first box) and source context (second box), while the (un)faithful revision target is the third. Entities from inputs and targets are colored as unsupported ***relative*** to the provided context. *Sampled ReDRESS* is a randomly sampled synthetic hallucination, while *Unsupported ReDRESS* is the sample most unsupported by *Aligned Context*. Figure 9 (Appendix D) visualizes how this data is obtained from ReDRESS and within-example misalignments.

from its own sampled set of corruptions, $\hat{\boldsymbol{r}}_{\boldsymbol{u}}$ is the most **u**nsupported (lowest BERTScore precision).

Training. The input to the **reviser** model is the concatenation of a noisy input and aligned source context, while the output is a (un)faithful revision. We rely on **ReDRESS** to generate noisy inputs *and* unfaithful outputs. **ReDRESS** hallucinations require moderate levels of revision to reconstruct the original supported input. This makes sense for most of the *observed* unsupported sentences. Yet, sometimes, a reference sentence is almost entirely unsupported. In such cases, the model should effectively learn to ignore it and just summarize the aligned context. To simulate this more drastic scenario, we also retrieve a **random** reference sentence ($\boldsymbol{r}^*$), and its aligned source ($\boldsymbol{S}^*$), from the same example. Our ablation study in Table 4 shows that both sources of hallucinated content (**ReDRESS**-generated and random mis-alignments) are complementary, necessary to achieve the best downstream summary results.

Using the notation above, as tuples of format (input, context, target), the positive set is: $(\hat{\boldsymbol{r}}_{\boldsymbol{u}}, \boldsymbol{S}, \boldsymbol{r})$ and $(\boldsymbol{r}^*, \boldsymbol{S}, \boldsymbol{r})$. The negative set is: $(\hat{\boldsymbol{r}}_{\boldsymbol{u}}, \boldsymbol{S}, \hat{\boldsymbol{r}}_{\boldsymbol{c}\in\boldsymbol{N}})$ and $(\boldsymbol{r}, \boldsymbol{S}^*, \boldsymbol{r})$, where $\hat{\boldsymbol{r}}_{\boldsymbol{n}\in\boldsymbol{N}}$ is a randomly selected corruption. In other words, for the positive set, we learn to generate a supported reference $\boldsymbol{r}$ from its aligned context ($\boldsymbol{S}$) and either a **ReDRESS** output ($\hat{\boldsymbol{r}}_{\boldsymbol{u}}$) or another reference sentence ($\boldsymbol{r}^*$) as the synthetic, unsupported input. For the negatives, we discourage the model from generating (1) a synthetic hallucination ($\hat{\boldsymbol{r}}_{\boldsymbol{n}\in\boldsymbol{N}}$), and (2) itself if unsupported. Figure 3 shows an example from the training data.

As in ConSeq (Nan et al., 2021b), we optimize the likelihood of positives ($\boldsymbol{Z}^+$) and unlikelihood (Welleck et al., 2020) of negatives ($\boldsymbol{Z}^-$):

$$\mathcal{L}_{contrast} = \mathbb{E}_{\boldsymbol{Z}^+} log(p_\theta(\boldsymbol{r_{out}}|\boldsymbol{r_{in}}, \boldsymbol{S})) - \mathbb{E}_{\boldsymbol{Z}^-} log(1 - p_\theta(\boldsymbol{r_{out}}|\boldsymbol{r_{in}}, \boldsymbol{S})) \quad (1)$$

$\boldsymbol{r_{in}}$ stands for the noisy reference input and $\boldsymbol{r_{out}}$ the revision target (positive or negative). We concatenate r_{in} and S with a special <SEP> token as the input, in addition to two key revision codes ($input_{frac}$ and $source_{frac}$) which we discuss next.

Controlling Revision Intensity. Some sentences require minimal edits to be fully supported while others require major edits. This variance is difficult to learn without an explicit control for it. Qualitatively, we noticed a tendency of the revision model to over-edit mostly supported sentences and under-edit highly unsupported content. Given this, we introduce the notion of revision intensity, parameterized by the fraction of words in the revision copied from the input ($input_{frac} = \frac{|\boldsymbol{r_{out}} \cap \boldsymbol{r_{in}}|}{|\boldsymbol{r_{out}}|}$), and the fraction copied from the aligned context ($source_{frac} = \frac{|\boldsymbol{r_{out}} \cap \boldsymbol{S}|}{|\boldsymbol{r_{out}}|}$). Intense revisions tend to require a larger lexical shift from input to source: a low $input_{frac}$ and a high $source_{frac}$. During training, we bin the fractions into deciles and include them as style codes prefixed to the encoder. Our ablation study in Table 4 shows that controlling the intensity of revisions to support diverse candidate generation, followed by re-scoring, has a massive impact on downstream summaries.

Inference. We apply the trained reviser model to all unsupported reference sentences in the summarization training dataset. In particular, we concatenate each unsupported sentence as r_{in} to its aligned context S for beam search decoding. For this set of sentences, the desired revision intensity codes are unknown because no ground-truth revision exists

(r_{out}). As a proxy, we fix $input_{frac} = \frac{|r_{in} \cap S|}{|r_{in}|}$, which instructs the model to remove words from the input proportional to its lexical overlap with S. Then, we vary $source_{frac}$ and over-generate 10 revision candidates with different levels of copy-paste from S and re-rank each candidate to select a final revision. In this way, the codes are useful both as a control mechanism and as a prompt for diverse generation. We experiment with two different scoring functions for re-ranking, which are discussed below as part of the experimental setup.

Implementation Details. The **reviser** is trained from `bart-base` on the subset of references sentences classified as supported (417k, from §3), and then used to over-generate revisions for the 595k *un*supported sentences. The top scoring revision replaces the original sentence in the training data.

5 Experimental Setup

We design experiments around our central hypothesis: *for a setting (long form hospital-course summarization), in which high-quality reference data is limited, reference revision is the best data-centric intervention to improve model faithfulness.* As such, we restrict the set of comparison methods to model-agnostic methods which explicitly address data quality. Based on our thorough literature review, we consider two classes of baselines: those which **filter** low quality data (Kang and Hashimoto, 2020; Narayan et al., 2021; Matsumaru et al., 2020; Nan et al., 2021a; Goyal and Durrett, 2021), and those which **control** for it (Filippova, 2020).

Reference Revision Strategies. We experiment with two different functions to re-score over-generated candidate revisions: **Less Abstractive** selects the one with the highest BERTScore precision, while **More Abstractive** adds a penalty, based on the extractive fragment density (Grusky et al., 2018), to encourage more abstraction. We also consider a baseline revision approach: **Fully Extractive**, which replaces each unsupported reference sentence with the source sentence with the highest BERTScore overlap. Even though our dataset is highly abstractive, this is necessary to justify the complexity of abstractive revision.

Baselines. **(1) Filtered.** We experiment with three heuristics for low quality: references where no Admission Note is available in the source documents (**No Admission**) (Shing et al., 2021), references where a significant portion of the content is unsupported by the source notes (< 0.75 token coverage or entity hallucination rate[6] of $> 10\%$) (**Unsupported**), and **Halluc. Ents**, which masks the training loss over spans of hallucinated reference entities. **Halluc. Ents** is inspired by Goyal and Durrett (2021) who use a factuality model to ignore negatively entailed dependency arcs[7]. Given the poor performance of other filtering strategies, we did not implement entailment-based filtering (Matsumaru et al., 2020). We also implement **Loss Truncation** (Kang and Hashimoto, 2020), which involves training for a fixed number of steps on the full training data before skipping high log loss examples. We grid-searched for the optimal number of warmup steps (2k) and the fraction of examples to drop (0.6) during truncation. **(2) Control Hallucination.** We implement the method in Filippova (2020): group training data into quality buckets based on token coverage, and control for hallucinations with encoder-prefixed style codes.

Training Details. We fine-tune downstream summarization models from BART (Lewis et al., 2020b) and the encoder-decoder Longformer (Beltagy et al., 2020). We train all models for 10,000 steps or until convergence on the validation set. Some methods use revised or filtered training data yet all use the same 1,195 validation examples and evaluation test set (1,190). Please refer to Appendix E for hyperparameters and more training details.

Metrics. To measure source faithfulness, we compute the entity **hallucination rate** (HR) using the soft matching heuristic described in §A, **BERTScore** (BS) using in-domain weights from Clinical BERT (Alsentzer et al., 2019), and the fraction of summary sentences predicted as **entailed** by SciFive (Phan et al., 2021) fine-tuned on MedNLI (Romanov and Shivade, 2018)[8]. To capture entity coverage, we record **faithful-adjusted recall** (FaR): the fraction of non-hallucinated reference entities included in the output (Shing et al., 2021).

6 Results

Please refer to Appendix for basic statistics on the revised training datasets (§F), BART summariza-

[6] Nan et al. (2021a); Narayan et al. (2021) rely on entities.

[7] They use the dependency-arc entailment (DAE) model (Goyal and Durrett, 2020) to identify inconsistent spans. Without such a model for clinical text, we use unsupported entities.

[8] MedNLI is a clinician-annotated entailment dataset whose premise sentences come from MIMIC-III. SciFive is a biomedical T5 model that achieves SOTA performance on MedNLI.

	Reference Version	Quality Strategy	Hallucination Rate (HR) ↓	BERTScore P / R / F1 (BS) ↑			Entail. ↑	Faithful-Adjusted Recall (FaR) ↑
LONGFORMER	**Original**	N/A	**36.8**	**82.3**	69.5	75.2	**48.4**	**48.2**
		Control Halluc.	36.5	83.3	70.2	76.0	51.5	49.0
	Filtered (Baselines)	No Admission	20.1	87.8	70.4	78.0	61.6	41.2
		Unsupported	**18.4**	87.6	70.8	78.1	**61.6**	**46.9**
		Loss Truncation	36.3	83.1	69.9	75.8	**51.7**	47.4
		Halluc. Ents	33.8	83.5	69.8	75.9	55.0	47.7
	Revised (Ours)	Fully Extractive	**5.4**	94.5	73.2	82.3	78.6	**52.4**
		Less Abstractive	**3.8**	**94.6**	73.1	82.3	**83.7**	**54.0**
		More Abstractive	**5.6**	92.1	73.0	81.3	76.3	**57.1**

Table 2: Summarization quality metrics across reference quality mitigation strategies (original, filtered, control, revised). The Longformer Encoder-Decoder (LED) model is used for fine-tuning. **Numbers** discussed below.

tion results (§H) and an example of over-generated revisions (by varying the $source_{frac}$ code from §4.2), which enables us to optimize the abstractive-faithful tradeoff during revision re-ranking (§I).

Impact of Revisions on Summarization. Table 2 confirms that filtering improves faithfulness (**Filtered - Unsupported** lowers the HR from 36.8 to 18.4 and improves entailment from 48.4 to 61.6), yet degrades coverage (48.2 vs 46.9 FaR). Masking hallucinated entities from log loss training (**Filtered - Halluc. Ents**) only slightly improves faithfulness, which underscores the difficulty in assigning token-level credit to pervasively noisy references. **Loss Truncation** leads to worse performance except on entailment (48.4 vs 51.7), which can likely be attributed to: 1) learning from fewer examples (from truncation); 2) hallucination patterns learned from the full data warmup are not unlearned during truncation; and 3) log loss might not capture faithfulness as well as more direct measures, i.e., entity overlap (used by **Halluc. Ents**).

In comparison, all revision approaches yield dramatic improvements in faithfulness (e.g., for **Less Abstractive**, 3.8/94.6.3/83.7 vs 36.8/82.3/48.4). These precision-oriented gains do not come at the expense of coverage (as defined by FaR), which actually jumps from 48.2 to 54.0. Abstractive revision (**Less** and **More**) outperforms **Fully Extractive** on coverage (e.g., 54.0/57.1 vs 52.4). Surprisingly, despite Fully Extractive revision being perfectly faithful, Less Abstractive revision leads to more faithful models (e.g., 3.8/83.7 vs 5.4/78.6 for HR and entailment, respectively), which suggests the need to re-write, not replace, unsupported content. Out of the revised results, More Abstractive has the best coverage (56.3/57.1) while being competitive on faithfulness (5.6 vs 5.4/3.8 HR).

Reference Version	Quality Strategy	Con.	Rel.	Fl.	Coh.
Original	N/A	1.5	2.2	3.9	**3.5**
Filtered	Unsupported	3.2	2.8	3.3	3.3
Revised	More Abstractive	**3.5**	**3.0**	**4.0**	2.7

Table 3: Average rating (1-5) assigned by a domain expert (clinician) according to Consistency (Con.), Relevance (Rel.), Fluency (Fl.), and Coherence (Coh.).

Human Evaluation. We rely on the protocol from Fabbri et al. (2021) to procure expert judgments on *Consistency*, *Relevance*, *Fluency*, and *Coherence*. An experienced clinician was shown the source notes for 10 examples and 3 randomly shuffled Longformer outputs side-by-side (10 x 3), and asked to rank each summary on a 1-5 Likert Scale. Please refer to Appendix (§G) for more detail on the protocol. We include the most faithful baseline according to automatic metrics: **Filtered - Unsupported**, as well as **Original** and **Revised**.

Table 3 shows that the model trained on Abstractively revised references produces the most Consistent, Relevant, *and* Fluent summaries. Training on Filtered references improves consistency and relevance but degrades fluency given the relative paucity of training data (a similar finding for data-to-text generation is noted by Filippova (2020)). Assessed coherence is lower for models trained on Abstractively Revised data, yet we note that for 3/10 summaries, it held the highest coherence rating. Additionally, there was large variance in coherence ratings within each system (1.34 standard deviation), as each had at least one summary rated as 1 and at least one as 5. Consistency had the lowest (1.04). Adams et al. (2021) argue that discourse is hard to evaluate because "clinical summaries naturally exhibit frequent, abrupt topic shifts".

Qualitative Analysis. Figure 4 shows an example from the human annotation test set for which we requested additional qualitative feedback. "Origi-

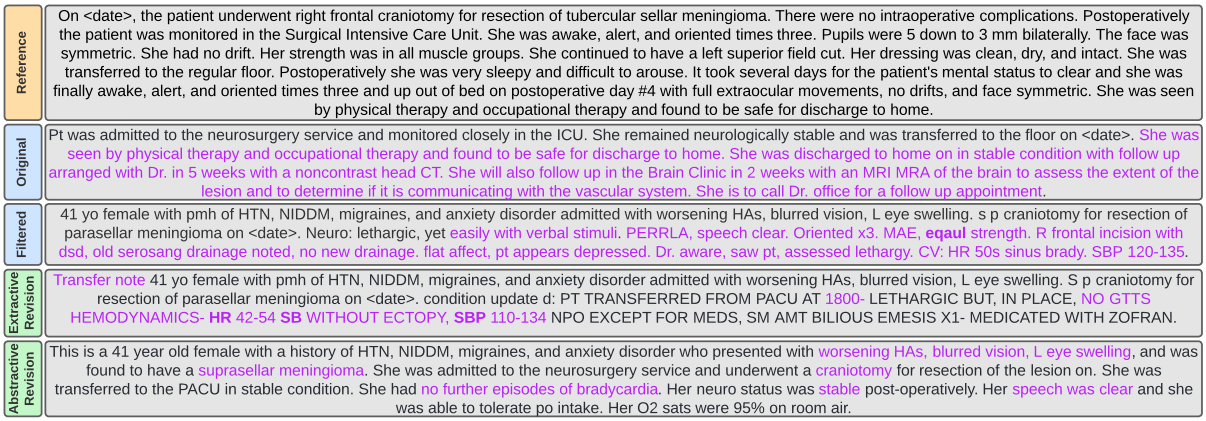

Figure 4: An example Reference, along with outputs from models trained on un-revised references (Original and Filtered), as well as Extractive and Abstractive revisions. Times and ages have been modified to ensure anonymity. Source notes are not shown for space reasons. **Purple Snippets** represent text mentioned in our analysis below.

nal" was inconsistent because it fabricates the last 4/6 sentences, operating almost as an unconditional LM. "Filtered" was less relevant than "Abstractive" because the second half of Filtered (starting with "PERRLA") contains non-essential status information. "Filtered" fluency was discounted for a misspelling ("eqaul" rather than "equal") and a missing word ("easily *roused* with verbal stimuli").

The annotator noted high fluency for Abstractive Revision (5) because the summary mimics the narrative style of the original references, rather than a sequence of choppy lists as in Filtered and Extractive summaries. Consistency is high (5) because the facts presented can be directly inferred. Coherence and consistency scored 5 because it follows a conventional arc from presenting illness (headache, vision, eye swelling), diagnosis (meningioma) and treatment (craniotomy), and ends with the patient's status (stable - no bradycardia, clear speech).

We also include "Extractive" to further emphasize the need to train on abstractive revisions. The "Extractive" summary does not match the target style of the Brief Hospital Course and includes many undesirable artefacts from source notes: note type metadata ("Transfer note"), long lists of non-essential data points (HR, SB, SBP), and unclear section boundaries ("1900- Lethargic...").

To better understand slightly lower coherence results for "Abstractive Revision", we include an extra example in the Appendix (§I) for which coherence was rated as a 3. We hypothesize that it stems from a sentence-level revision strategy which does not consider the position in the summary. As such, heavily revised references can deviate slightly from a conventional structure for the hospital course.

Reviser Training Objective (Each Ablation is Separate)	**BERTScore P / R / F1 (BS)** ↑	**Entail.** ↑
Full (last row of Table 2)	**92.1 / 73.0 / 81.3**	**76.3**
w/o ReDRESS hallucinations	90.8 / 72.4 / 80.4	72.1
w/o Random *other* alignments	90.9 / 72.8 / 80.7	73.8
w/o All Negatives (no contrast)	90.9 / 72.8 / 80.7	72.4
w/o Revision Codes	87.6 / 71.7 / 78.7	62.5

Table 4: Separately removing key components of the reviser training objective (from Equation 1) hurts the *downstream* performance of Longformer summaries.

Ablation Analysis. Which parts of the revision pipeline are necessary to improve *downstream* summary performance? Separately, we remove ReDRESS hallucinations (top sequences from Figure 3), randomly sampled *other* alignments (bottom sequences), and all negative examples (right side). We also train a model without control codes for revision intensity (*w/o* Revision Codes) and, in turn, at inference time, generate a single revision rather than vary the codes to over-generate-then-rescore. Table 4 reveals that both sources of synthetic revision training data contribute to downstream performance (BS F1/Entailment of 80.4/72.1 and 80.7/73.8, individually, vs 81.3/76.3 combined). This aligns with Cao and Wang (2021) who also demonstrate "the importance of covering diverse types of errors in negative samples". Eliminating unlikelihood training of synthetic negatives reduces summary BS F1/Entailment from 81.3/76.3 to 80.7/72.4. Removing the ability to over-generate-then-rescore diverse revision candidates by varying style codes (*w/o* Revision Codes) has the largest impact on downstream performance (78.7/62.5). A model trained without codes tends to insufficiently edit highly unsupported sentences.

Pretrain Weights	Hallucination Rate (HR) ↓	BERTScore P / R / F1 (BS) ↑			Entail. ↑	Faithful-Adjusted Recall (FaR) ↑
`bart-base`	25.6	85.6	70.7	77.3	**56.0**	44.4
ReDRESS	**22.3**	86.0	70.7	77.5	**55.2**	**44.7**
***w/o* Entity Swap**	**26.9**	85.7	70.3	77.1	54.4	**42.1**
Reviser	23.0	86.0	71.0	77.7	56.7	47.9

Table 5: Assessing the usefulness of ReDRESS & reviser models for pre-training. We separately fine-tune models from `bart-base`, ReDRESS, and reviser checkpoints on *Filtered - No Admission* data. *w/o* Entity Swap restricts noise to span-deletion, similar to the optimal configuration in the original BART paper. **Numbers** discussed below.

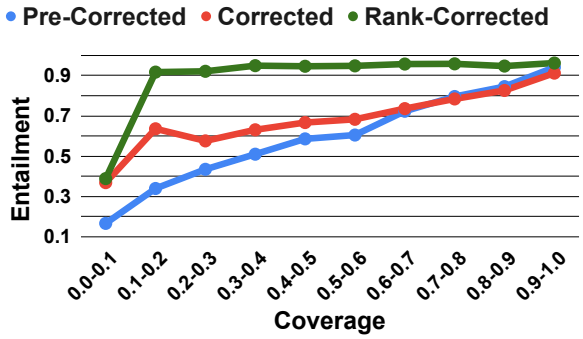

Figure 5: Faithfulness before and after correcting summaries with the Reviser, controlling for extractiveness.

Reviser as a Post-Hoc Editor. As a natural extension, we experiment with using the revision model to correct system outputs. In particular, we take summaries from the model trained on **Original** references, and then separately feed each predicted sentence and aligned context to the reviser for generation. It would be easy to increase faithfulness by copying the context verbatim. Yet, we are more interested to see if the reviser can increases *effective* faithfulness (Ladhak et al., 2021), i.e., controlling for extractiveness. For each *predicted* sentence, we over-generate revision candidates with different targeted levels of extraction (by varying the $source_{frac}$ code defined in Section 4. Then, we bin sentences by extractiveness (coverage) and record faithfulness (fraction of sentences predicted as entailed) within each bin. We include separate plots for **Corrected**, which includes each over-generated candidate, and **Rank Corrected**, which selects the top candidate by entailment prediction (breaking ties according to the most abstractive). Figure 5 demonstrates that reviser-corrected summaries are much more effectively faithful, although, naturally, the gap shrinks as all outputs become highly extractive and mostly entailed. The ability to re-rank and select from diverse candidates makes a huge difference, as evidenced by the large gap between **Corrected** and **Rank Corrected**.

ReDRESS/Revision as Pre-Training Objectives. Beyond data cleaning, do our proposed methods to generate hallucinations (ReDRESS), and then edit them out (reviser), hold intrinsic value for the task as pre-training objectives? One can argue that both models are trained on faithfulness objectives: based on context, ReDRESS learns to add/remove/integrate entities, and the reviser to edit out synthetic hallucinations. Table 5 shows that fine-tuning from ReDRESS and reviser–both trained from checkpoints of `bart-base`–improves all evaluation metrics vis-a-vis `bart-base` (except slight entailment decrease from ReDRESS (56.0 to 55.2)). *w/o* Entity Swapping is a denoising baseline in which corruptions are limited to span deletion and word reordering. On its own, then, incorporating entity swaps into pre-training (the full version of the ReDRESS framework), causes HR to drop (26.9 to 22.3) and NSP/FaR to rise (81.1/42.1 to 83.3/44.7).

7 Limitations

Our sentence-level revision strategy does not consider the impact of changing one sentence on the reference as a whole, which can impact coherence. This can be addressed in future work by exploring summary-level revision (especially for corpora with short summaries) or incorporating coherence into revision re-scoring and/or contrast set creation. Given the complexity and expertise required for the task, the length of both source notes and summaries, and a mandatory license for data access, only 30 human ratings were procured from a single annotator. The assessment still took four full days.

8 Conclusion

We propose a new approach to mitigating the downstream impact of noisy references on summarization models. We learn a model to improve the coverage of existing references from a large clinical corpus of EHR notes, and re-train models on revised references. Results show that reference revision is a more effective intervention for improving faithfulness than quality filtering for this task.

9 Ethical Considerations

Deidentification. Our summarization corpus is extracted from a publicly avilable database of real-world, de-identified clinical records: MIMIC-III v1.4 (Johnson et al., 2016). Even though it is HIPAA-compliant, we make sure no Protected Health Information (PHI) is shared with the public.

Intended Use & Failure Modes. The goal of this paper is to make progress toward automatic summarization of a patient's hospital admission. Deploy such a system in a real-world clinical setting has its own set of ethical and procedural concerns. Robustness, Fairness, and Trust is vital to any NLP system, especially one deployed in a hospital setting where lives are at risk. As with many NLP datasets, our MIMIC-III dataset likely contains biases, which may be perpetuated by its use. It is important to analyze the underlying population to identify demographic, social, and economic discrepancies vis-a-vis the broader population of interest. Model-generated errors could be harmful to patient safety and even negatively affect outcomes. There are lessons to be learned from existing clinical decision support tools (Pivovarov et al., 2016; Chen et al., 2020). Furthermore, there is a moral hazard from deploying clinical systems, in which clinicians start to over-rely on a system at the expense of their own judgment (Goddard et al., 2012). EHRs are also living systems and deploying a summarization system within it necessitates evolving with the EHR and the underlying population.

Acknowledgments

We thank the reviewers for the insightful, actionable insights, as well as Michael Elhadad and Alex Fabbri for their feedback on earlier drafts.

References

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

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

Emily Alsentzer and Anne Kim. 2018. Extractive summarization of ehr discharge notes. *ArXiv preprint*, abs/1810.12085.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

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

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical machine learning in health care. *arXiv e-prints*, pages arXiv–2009.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

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

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

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

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

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *ArXiv preprint*, abs/1702.08734.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Daniel Kang and Tatsunori Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations. *ArXiv preprint*, abs/2005.01795.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, 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, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2021. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. *ArXiv preprint*, abs/2108.13684.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019a. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019b. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. 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 2019, Paris, France, July 21-25, 2019*, pages 1013–1016. ACM.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

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

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial intelligence in medicine*, 67:25–37.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization. *ArXiv preprint*, abs/2104.07606.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *ArXiv preprint*, abs/2106.03598.

Rimma Pivovarov, Yael Judith Coppleson, Sharon Lipsky Gorman, David K Vawdrey, and Noémie Elhadad. 2016. Can patient record summarization support quality metric abstraction? In *AMIA Annual Symposium Proceedings*, volume 2016, page 1020. American Medical Informatics Association.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *ArXiv preprint*, abs/2104.13498.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross 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*, pages 1899–1905, Online. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021a. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.

Sen Zhang, Jianwei Niu, and Chuyuan Wei. 2021b. Fine-grained factual consistency assessment for abstractive summarization models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 107–116, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Bioword-vec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6(1).

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis 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*, pages 5108–5120, Online. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A Merging Entities

For each pair of entity mentions in the source text and reference, we compute a code overlap score. Let entity e_x have codes $\boldsymbol{c_x}$ and tokens $\boldsymbol{t_x}$ and e_y have codes $\boldsymbol{c_y}$ and tokens $\boldsymbol{t_y}$, the pairwise code overlap score is:

$$CodeOverlap(e_x, e_y) = \frac{|\boldsymbol{c_x} \cap \boldsymbol{c_y}|}{|\boldsymbol{c_x}| + |\boldsymbol{c_y}|}$$

Then, we compute embedding cosine similarity between mention tokens with BioWordVec (Zhang

et al., 2019), filtering out stopwords and punctuation. Let

$$EmbedOverlap(e_x, e_y) = cosine(E(t_x), E(t_y))$$

Finally, we compute the TF-IDF overlap ($TF\_IDF(t_x, t_y)$) to compensate for occasional noise in embedding space. We define the aggregate score as the average of embed, code, and TF-IDF overlaps. For entity types with no available codes (treatments, procedures, and tests), we only examine mention overlap. Based on validation on a manually labeled held-out set, we classify e_x and e_y as synonyms iff any of the following thresholds are met: $CodeOverlap(e_x, e_y) \geq 0.4$, $EmbedOverlap(e_x, e_y) \geq 0.75$, or $AggOverlap(e_x, e_y) \geq 0.4$.

A.1 Analyzing Unsupported References

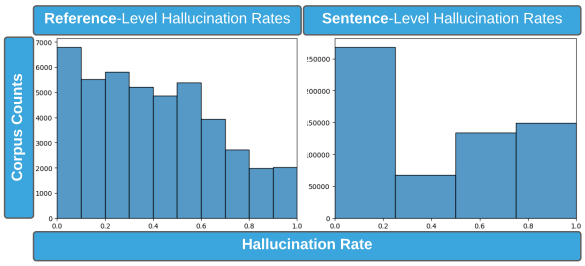

Figure 6: The distribution of reference-level (left) and reference *sentence*-level (right) hallucination rates–the fraction of entity mentions not present in source text.

Identifying Correlates We collect a wide-range of visit-level statistics and compute the Pearson correlation coefficient with respect to the hallucination rate, defined as the fraction of entities in the reference text not present in the available source documentation. Unsurprisingly, lexical overlap (unigram coverage) is highly correlated (88%) with the hallucination rate. Examples with well-supported references contain more source notes and distinct note types. The MIMIC-III notes do not cover time a patient spent outside the ICU. Interestingly enough, however, the number of days spent outside the ICU has zero direct correlation with the hallucination rate.

Distribution of Hallucinations. We examine the example and sentence-level distribution of hallucinations before devising a revision strategy. Figure 6 reveals a degree of uniformity in the example-level hallucination rate and, to a lesser extent, sentence-level[9]. The figure indicates that the faithfulness issue is not concentrated to just a few very low coverage examples, or sentences. As such, example-level quality filtering is noisy since there is no clear coverage boundary and relatively few references contain zero entity hallucinations ($< 2k$). These two basic plots inform two key design choices: **(1.)** to address quality at the sentence-level rather than the reference; **(2.)** to enforce diversity of faithfulness in synthetic hallucinations.

B Alignment Algorithm

Figure 7 provides an example alignment with improvement filtering and an extra extraction step to ensure full entity coverage.

B.1 Notation

Let $\langle (\boldsymbol{S_1}, \boldsymbol{R_1}), ..., (\boldsymbol{S_N}, \boldsymbol{R_N}) \rangle$ represent the corpus, consisting of source-reference pairs for N unique patient ICU admissions. Let $\boldsymbol{S_n} = \langle \boldsymbol{s_1^n}, ..., \boldsymbol{s_{|S_n|}^n} \rangle$ represent the sentences extracted from the source input for the n^{th} example and, likewise, $\boldsymbol{R_n} = \langle \boldsymbol{r_1^n}, ..., \boldsymbol{r_{|R_n|}^n} \rangle$ the reference sentences. Similarly, $\boldsymbol{s_i^n} = \langle x_1, ..., x_{|s_i^n|} \rangle$ is the tokenized sequence for the i^{th} source sentence from the n^{th} example, and $\boldsymbol{r_j^n} = \langle \hat{x}_1, ..., \hat{x}_{|r_j^n|} \rangle$ the tokenization of the j^{th} reference sentence.

Given very long inputs, we link each reference sentence $\boldsymbol{r_j^n}$ ($n \in N$, $j \in \boldsymbol{R_n}$), to a small subset (≤ 5) of source sentences corresponding to the same example. Due to the abstractiveness of the data (acronym usage, shorthand, etc.), as well as redundancy from copy-and-paste (Hirschtick, 2006), we align sentences using a new approach which combines BERTScore and subword-level coverage, rather than the conventional approach of lexical overlap with ROUGE (Lebanoff et al., 2019a; Liu and Lapata, 2019). Given a candidate alignment pair: a reference sentence $\boldsymbol{r_j^n}$ with K tokens and a source sentence $\boldsymbol{s_i^n} \in \boldsymbol{S_n}$ with L tokens, for each reference token $\hat{x}_k$, we find its closest match in $\boldsymbol{s_i^n}$:

$$align(\hat{x}_k, \boldsymbol{s_i^n}) = \max_{1 \leq \ell \leq L} cos(h(\hat{x}_k), h(x_\ell))$$

where $h(x)$ represents the contextualized BERT

[9] Sentence-level hallucination rates are shown only for multi-entity sentences to get a sense of the non-binary distribution. 14% of reference sentences have no entities, and 30 % have one. Single entity sentences have a hallucination rate of 41%.

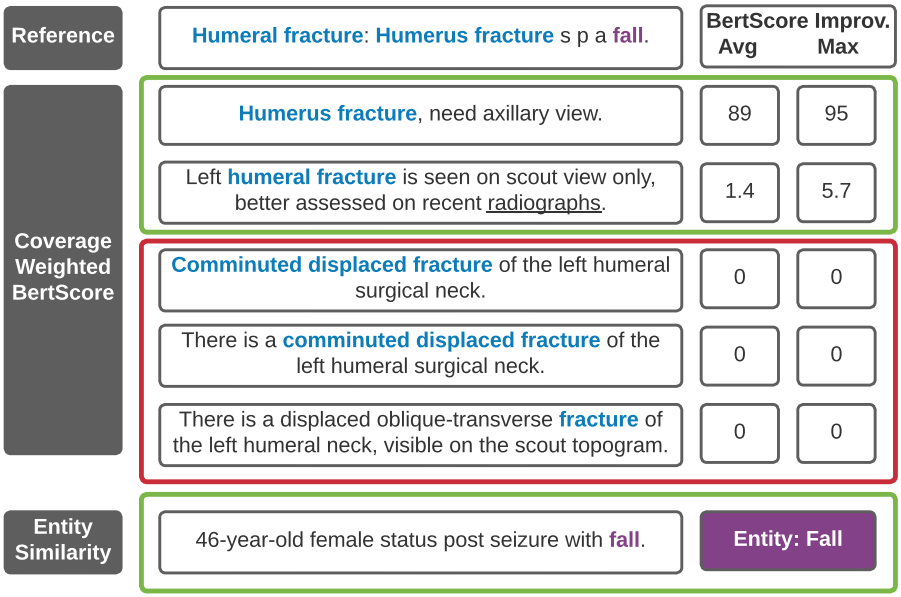

Figure 7: Source-Reference Alignment. 5 source sentences are greedily extracted with a coverage-weighted BERTScore heuristic. Non-influential sentences are discarded (low BERTScore improvement). Finally, in the case of missing clinical concepts (**fall**), we find the source sentence whose contextualized representation of the entity span (**fall**) is closest to the reference usage (the last word of the top reference row).

embedding[10]. Based on these greedy alignments, we extract sentences for T steps. At $t = 0$, we initialize an importance vector $\boldsymbol{w}$ of length K, to all ones. Then, at each timestep, we select $\boldsymbol{s^*} \in \boldsymbol{S_n}$ which maximizes the importance-weighted BERTScore:

$$\boldsymbol{s^*} = \underset{\boldsymbol{s^*} \in \boldsymbol{S_n}}{\text{argmax}}\left(\frac{\sum_{k=1}^{K} w_{tk} align(\hat{x}_k, \boldsymbol{s^*})}{\sum_{k=1}^{K} w_{tk}}\right)$$

After each timestep, we keep track of the best alignment score for each reference token via the importance vector. In particular, we update each token's importance by the inverse of its best coverage score: $w_{t+1,1} = min(w_{t1}, 1 - align(\hat{x}_1, \boldsymbol{s^*}))$ (formula shown for first element). Similarly to Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), the importance vector deprioritizes well-covered tokens for future extractions. We also remove $\boldsymbol{s^*}$ from $\boldsymbol{S_n}$ to ensure it is only extracted once. After retrieving the top K using this procedure, we only use sentences for the final alignment set for which the average coverage improvement ≥ 0.01 or the max ≥ 0.05, where improvement is defined as the reference token-level increase in coverage of the latest extraction over the previous max from prior extractions: $max(0, \boldsymbol{w^t} - \langle align(\hat{x}_1, s^*), ..., align(\hat{x}_K, s^*)\rangle)$.

[10]The mean-pool of the last four layers of ClinicalBERT.

Infrequently, the medical concepts extracted from the aligned sentences do not cover all the concepts in the reference sentence. In this case, for each missing concept, we filter for the subset of source sentences containing that concept, and add the sentence with the highest pairwise similarity of contextualized concept embeddings–the mean of hidden states corresponding to the entity span.

C ReDRESS Details

Pre-Training Data. We pre-train on a large unlabeled sentence corpus: all sentences extracted from MIMIC-III discharge summaries, excluding notes related to patients in the summary test set. To minimize EHR-related noise, we filter out sentences without any clinical concepts and those found in non-narrative sections related to structured data, demographics, and/or administration (billing codes, dates, times, signatures, lab values).

Intrinsic Evaluation of ReDRESS. **ReDRESS** combines entity swapping and span-infilling–so we compare it approaches that do one or the other. For pure entity swaps: **(1.) Swap Random** randomly removes entities and replaces them with a random one of the same type from the training data. Given the long tail of rare entities, we sample the replacement entity by its empirical frequency in the corpus. **(2). Swap Related** follows an identi-

Model	Hallucination Rate (HR) ↑	BERTScore F1 (BS) ↓	Coherence (NSP) ↑	Diversity ↑
Swap Random (Baseline)	45	91.5	70.0	16
Swap Related (Baseline)	32	93.2	73.9	14
Span Infill + Entity Swap (ReDRESS)	38	87.0	72.6	23
***w/o* Entity Swap**	21	90.1	67.7	21
***w/o* Add-1 Inference Trick**	28	88.8	72.4	22
***w/o* Entity Hiding Inference Trick**	27	89.7	73.4	19

Table 6: Intrinsic evaluation of ReDRESS model (Span Infill + Entity Swap) on *supported* reference sentences. For an upper bound, the original sentences have coherence of 75.8. We define **diversity** as one minus the pairwise unigram coverage score (Grusky et al., 2018) between two synthetic hallucination samples for the same input. Greater error diversity has been shown to be useful for synthetic data generation in other summarization work (Cao and Wang, 2021). Each ablation (*w/o*) is performed independently of the others.

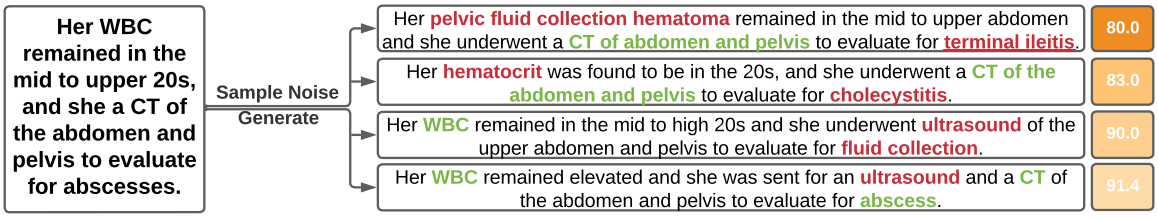

Figure 8: **ReDRESS** model outputs. Green represents non-hallucinated entities, red not present in input sentence, and red also not present in any of the source notes. The orange box shows BERTScore F1 vis-a-vis original. Due to the topical nature of the distractor set, all hallucinations except *terminal ileitis* exist elsewhere in the source notes.

cal procedure with the exception that replacement entities are sampled from the related distractor set. **Span Fill** is a version of ReDRESS model without entity swaps (and no pre-pended distractor set). Corruption is limited to span removal and word order shuffling. Each baseline/ablation follows the same approach: over-generate five candidate hallucinations by re-sampling noise levels.

Based on Table 6, the unified ReDRESS model–**Span Infill + Entity Swap**–achieves diverse hallucinations while maintaining topical consistency. Interestingly, the prepended distractor set greatly improves the coherence of the outputs (**Span Fill** vs. **Span Fill + Ent Swap**) because the distractor set is topically consistent. Entity swaps alone produce incoherent sentences. The corpus-level hallucination rate is 40% which we are nearly able to achieve (38%) with ReDRESS while maintaining topicality. The first ablation (*w/o* Entity Swap) removes entity swapping entirely from training and inference. This dramatically reduces the rate of hallucinations (38 to 21) and reduces coherence dramatically. Coherence is reduced because the topical entities from the distractor set help to keep the generation on track. More specifically, the prepended set allows us to not have to include source context, as in Cao and Wang (2021), to avoid excessive semantic drift from mask-and-fill generation. We also can see that the *Add-1* inference trick is working as expected. Removing it (*w/o Add-1* Inference Trick) leads to 10% lower hallucination rate (38 vs 28) without compromising coherence. In other words, we instruct the model to implement an additional entity swap and it appears to be doing so. Figure 8 demonstrates the diversity from the standpoint of metrics (BERTScore) and meaning.

D Revision Model Details

The end-to-end *reviser* training pipeline is visually shown in Figure 9.

E Summary Training Details

For training, we use two abstractive models: BART (Lewis et al., 2020b) and the encoder-decoder Longformer (LED) (Beltagy et al., 2020) - a scaled up version of Bart to handle longer sequences via windowed local self-attention and a constant global attention. We fine-tune pretrained checkpoints (`facebook/bart-base` and `allenai/led-base-16384` from the HuggingFace transformers library (Wolf et al., 2020)) for a maximum of 10,000 training steps, a maximum learning rate of $2e-5$ with a linear warmup of 200 steps, followed by linear decay to 0, and a batch size of 16. For Longformer, we

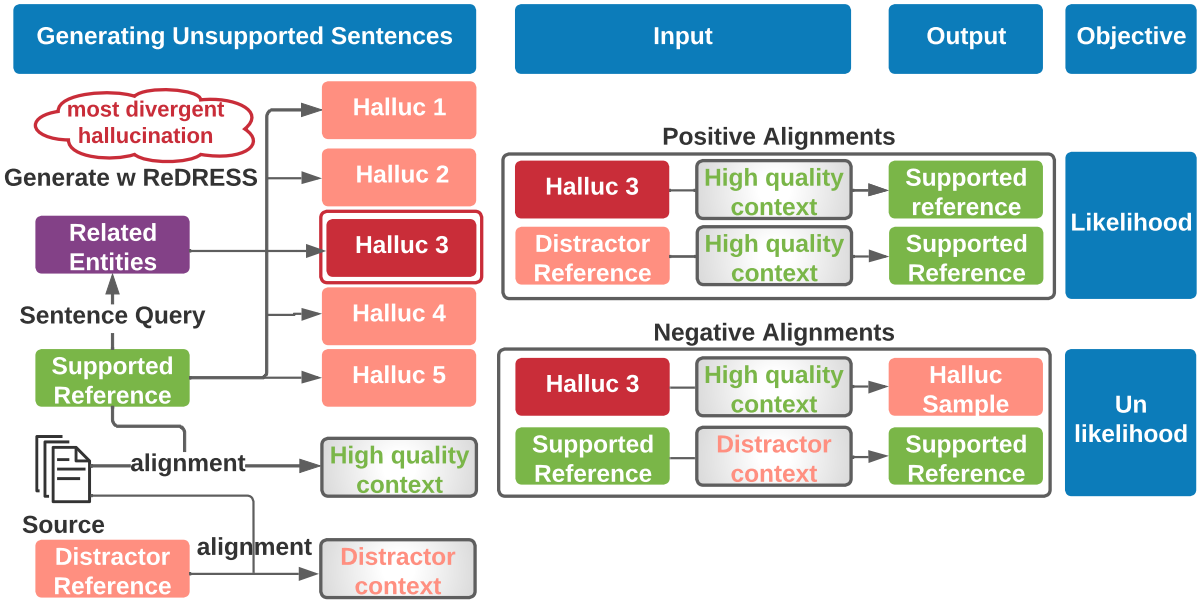

Figure 9: A high-level depiction of the **reviser** training pipeline, which includes generating data from ReDRESS hallucinations and sampling within-example mis-alignments (a distractor reference and distractor context).

use a maximum input length of 10,000. The maximum encoder length is 16,384 yet we could only fit 10,000 tokens onto a single 16GB V100 GPU. Training took approximately 1 day to complete 10 epochs on a single V100. For Bart, we use the maximum input length of 1,024. To handle longer input sequences, we rely on oracle filtering by taking the sentences with the largest average of ROUGE-1, ROUGE-2, and ROUGE-L F1 vis-a-vis the original reference. For models trained on revised data, oracle extraction is computed based on revised references during training yet the original during testing. This puts it a slight disadvantage to the other models which do not have this train-test mismatch. Yet, this mismatch only strengthens the empirical hypothesis considering the performance gains we see. During generation, we set beam search to 4, use trigram blocking, and set the maximum output length to 1,024.

F Statistics on Revised Training Datasets

Table 7 describes statistics from original, filtered, and revised versions of the training data. As expected, the Filtered datasets (5.7k/6.0k) are smaller than the Original and Revised sets (45k), indicating the high number of noisy references. The Filtered references are more faithful than Original, as reflected by entity-level source precision (e.g., 91.5/95.7 vs 60.4). In comparison, the abstractively revised references are more concise than Original and Filtered at the token (e.g., 300/272 vs 370 and 420/381) and entity level (e.g., 29/26 vs 50 and 64/55), yet contain a larger fraction of the entities present in the source (36.8/35.0 versus 26.8 and 32.7/28.9). While the fully extractive references have perfect average precision (by construction), the abstractive revisions are close, 97.6/96.8, and contain almost twice as many relevant entities (36.8/35.0 vs 19.6).

G Human Evaluation Setup

As described in Fabbri et al. (2021), we solicit summary feedback from an in-house clinical expert on 4 critical dimensions: Consistency, Relevance, Fluency, and Coherence. The annotator was provided the following guidance on each metric:

- **Consistency**: The rating measures whether the facts in the summary are consistent with the facts in the original article. Consider whether the summary does reproduce all facts accurately and does not make up untrue information.
- **Relevance**: The rating measures how well the summary captures the key points of the article. Consider whether all and only the important aspects are contained in the summary.
- **Fluency**: The rating measures the quality of individual sentences: are they well-written

Reference Version	Quality Strategy	# Training Examples	Avg. # Tokens	Avg. # Entities	Entity-Level Source Overlap Avg. Precision	Avg. Recall
Original	Original	45k	370	50	60.4	26.8
Filtered	No Admission	5.7k	420	64	91.5	32.7
	Unsupported	6.0k	381	55	95.7	28.9
Revised (Ours)	Fully Extractive	45k	295	28	100	19.6
	Less Abstractive	45k	300	29	97.6	36.8
	More Abstractive	45k	272	26	96.8	35.0

Table 7: Training datasets obtained from the revision strategies according to size (number of references), average number of tokens and entities in the references, and entity-level overlap with sources (average precision and recall).

	Reference Version	Quality Strategy	Hallucination Rate (HR) ↓	BERTScore P / R / F1 (BS) ↑			Entail. ↑	Faithful-Adjusted Recall (FaR) ↑
BART	**Original**	N/A	**38.9**	81.3	69.2	74.7	**43.6**	**47.7**
		Control Halluc.	40.3	81.7	69.2	74.8	43.2	46.6
	Filtered (Baselines)	No Admission	25.6	85.6	70.7	77.3	56.0	44.4
		Unsupported	**22.9**	86.5	71.1	77.9	**59.6**	**47.2**
		Loss Truncation	40.9	81.3	69.1	74.6	**51.6**	45.1
		Halluc. Ents	37.6	82.3	69.1	75.0	48.6	46.4
	Revised (Ours)	Fully Extractive	**9.1**	92.3	72.8	81.2	72.1	52.1
		Less Abstractive	7.3	91.8	72.9	81.1	72.5	**56.3**
		More Abstractive	**7.4**	90.7	72.2	80.3	69.2	56.3

Table 8: Summarization quality metrics across reference quality mitigation strategies (original, filtered, control, revised) used for training BART summarization models.

and grammatically correct? Consider the quality of individual sentences.

- **Coherence**: The rating measures the quality of all sentences collectively: do they fit together and sound naturally? Consider the quality of the summary as a whole.

The annotator was asked to assess summaries with independent rankings for each of the 4 metrics along a 1-5 Likert Scale. Given the complexity of the task and resource constraints, we sampled a set of 10 summaries from the test set. We first discard length outliers: the 10% of examples with the fewest source tokens (insufficient input) as well as the 10% of examples with most source tokens (too difficult to evaluate), before sampling at random. It took the expert 4 days to review 10 patient charts.

H BART Summarization Results

For robustness, we evaluate our methods by fine-tuning from BART models, as well as Longformer Encoder-Decoder (LED) models. Table 8 reveals similar findings for BART models as Table 2 revealed for Longformer. These findings demonstrate a degree of invariance to summary compression ratios, since BART only accepts 1,024 tokens, yet the maximum target output was 1,024 tokens for both.

I Qualitative Analysis

We show an Abstractive example which received low assessed coherence (1) (other scores were all 3) to better understand its relative under-performance:

> A small right sub pectoral subcutaneous emphysema is mildly increased from the recent outside hospital exam. He was found to have a small right pneumothorax and multiple rib fractures on the right. A chest x-ray was obtained which showed a small anterior mediastinal hematoma likely tracking from the adjacent rib fractures. A CT of the chest was obtained to assess for progression of PTX contusions. There was no evidence of lung contusion or laceration on chest CT. He had no further episodes of bradycardia while in the Trauma ICU. A repeat chest CT was obtained on <date>.

The Brief Hospital Course should roughly follow the same arc: presenting problem, investigation, diagnosis, treatment, (any complications and treatment), and, possibly, any follow-up plan. Deviation from this standard was penalized with lower coherence scores. The above summary does not follow this arc and the last sentence does not provide new information regarding the chest CT.

Reviser Outputs at Different Intensities. In Figure 10, we show diverse outputs from the reviser model, with each output conditioned on a different extractiveness code ($source_{frac}$ from §4.2). We see a relatively smooth interpolation from abstractive to extractive. At low $source_{frac}$ codes, revisions include hallucinations regarding the third

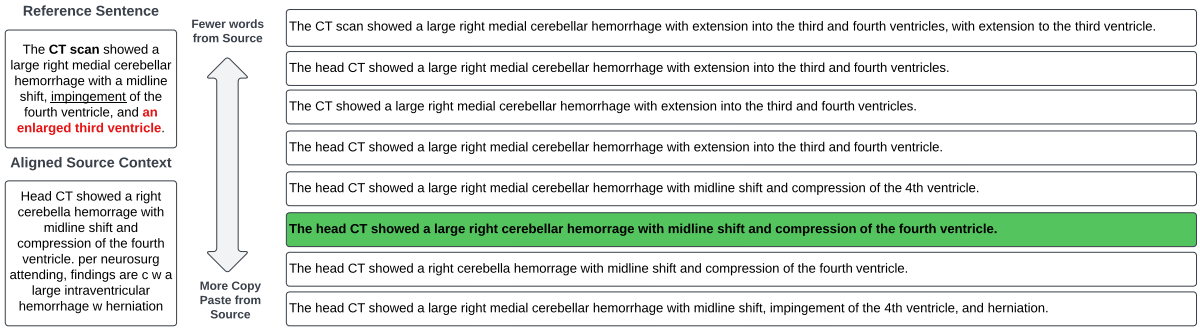

Figure 10: Over-Generated Revisions from varying revision style codes. The highlighted sentence is the one ultimately selected as the revision.

ventricle, but ultimately, the reviser edits them out, and, its place, introduces something that is only mentioned in the aligned source context: herniation. The green highlight is the sentence ultimately chosen by the More Abstractive strategy for its blend of abstraction and faithfulness (groundedness).