Extracting and Encoding: Leveraging Large Language Models and Medical Knowledge to Enhance Radiological Text Representation

Anonymous ACL submission

Abstract

Recent advancements in representation learning face challenges in specialized domains like medicine, where acquiring expert annotations for texts and images is burdensome due to the limited availability of medical professionals. To address this, we propose a two-stage framework aimed at extracting factual statements to improve encoder representations for various downstream tasks. First, we introduce a *Fact Extractor* that leverages the effectiveness of large language models in the medical domain to obtain factual statements from well-curated domain-specific datasets. In the second stage, we present a *Fact Encoder* (CXRFE) based on a BERT model that we fine-tune with objective functions designed to enhance its representations using the gathered data. Additionally, our framework introduces a novel embedding-based metric (CXRFEScore) for evaluating chest X-ray text generation systems, utilizing the two stages of our framework. Through extensive evaluations, we demonstrate that our fact extractor and encoder outperform multiple state-of-the-art methods across tasks such as sentence ranking, natural language inference, and label extraction from radiology reports. Furthermore, we show that our metric is robust, outperforming existing metrics commonly used in the report generation literature.

1 Introduction

In the context of medical image analysis, radiology reports serve as a rich source of information. Radiologists routinely generate these free-text reports, which typically include sections such as *comparison*, *indication*, *findings*, and *impression*, as illustrated in Figure 1. Radiology reports have been employed for various purposes, including label extraction for structured supervision in medical image tasks (Irvin et al., 2019; Wu et al., 2021; Jain et al., 2021a), training data for models in report generation (Messina et al., 2022; Miura et al., 2021; Tanida et al., 2023), summarization tasks (Chen et al., 2023b; Ma et al., 2023), and the development of multimodal models capable of jointly understanding medical images and text (Wang et al., 2022; Boecking et al., 2022; Bannur et al., 2023).

Comparison: Chest radiographs XXXX.
Indication: XXXX-year-old male, chest pain.
Findings: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.
Impression: No acute cardiopulmonary process.

Figure 1: Example image and report from the IU X-ray dataset (Demner-Fushman et al., 2015)

A crucial aspect in addressing such tasks is the accurate comprehension of factual information within the report. Specifically, the *findings* and *impression* sections can be considered as repositories of factual information regarding the imaging examination. These factual statements encompass various elements, including observations (such as abnormalities, diseases, or devices), interpretations derived from observations, references to anatomical locations, discussions on severity or confidence levels, comparisons to previous studies, and more. For instance, in Figure 1, one factual statement indicates no acute bone abnormality (a normal observation), while another describes a stable calcified granuloma within the right upper lung (an abnormality found in a specific anatomical site).

Despite the various aforementioned applications that recent research has explored in the use of radiology reports, the persistent absence of effective methods for precise fact extraction and encoding from medical reports remains a critical challenge. As demonstrated in our experimental evaluation (Section 4), existing encoders and label extraction techniques frequently struggle to capture the nuanced details within free-text radiology reports, resulting in incomplete or inaccurate depictions of clinical information. This is evident in various

aspects. For instance, existing text encoders developed for the medical domain may struggle to generate consistent representations of paraphrased statements (Table 1) or to differentiate between similar sentences conveying contradictory meanings (Table 3), a crucial requirement to prevent encoding erroneous diagnoses. Similarly, current label extraction methods often rely on rigid labeling schemes based on manually crafted rules, leading to incomplete capture of all factual statements within a report (Table 6). Similar limitations are also observed in commonly used evaluation metrics for radiology text generation (Table 5).

In this work, we propose a novel approach that leverages the capabilities exhibited by Large Language Models (LLMs) such as ChatGPT, which have showcased outstanding performance in medical contexts (Liu et al., 2023c; Katz et al., 2023; Liu et al., 2023b; Adams et al., 2023), to improve factual statement representation. Our methodology also takes advantage of the existing knowledge in expert-annotated datasets. These datasets offer indispensable training data and also serve as a benchmark to enhance our model's clinical terminology and context comprehension. Concretely, our contributions are three-fold:

- A fact extractor: a novel and simple approach to extracting facts that leverages ChatGPT and a fine-tuned version of T5 (Raffel et al., 2020) to capture relevant information from Chest X-ray radiology reports without requiring annotations from radiologists.

- A fact encoder: **C**hest **X**-**r**ay **F**act **E**ncoder (CXRFE), a CXR BERT-based model (Boecking et al., 2022) fine-tuned with a multi-task approach that leverages domain expertise from radiologists as well as ChatGPT and T5 generated annotations. CXRFE exhibits significant improvement in fact comprehension, demonstrated on sentence ranking and natural language inference tasks.

- A new evaluation metric for radiology text generation: CXRFEScore, which measures the factual accuracy of a generated text relative to a real text, by extracting and comparing the similarity of fact embeddings.

We release the weights of all our models, as well as the data and code necessary to replicate the results. We also release CXRFEScore as a Python library for ease of use by the research community.

2 Related Work

In this section, we discuss prior work on BERT-based approaches to radiology text representation and label extraction from radiology reports, and relegate discussion of prior work on evaluation of factual correctness in radiology text generation, applications of LLMs to medical text, and knowledge distillation from LLMs to Appendix A.1.

BERT-Based Approaches for Radiology Text Representation. The advent of BERT (Devlin et al., 2019) has sparked notable progress in numerous NLP domains. This has inspired researchers to customize BERT for specific applications, including the medical field. Pioneering works such as BioClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2020), and BioLinkBERT (Yasunaga et al., 2022) have applied the masked language modeling (MLM) objective introduced by BERT to domain-specific corpora, such as PubMed paper abstracts and MIMIC-III (Johnson et al., 2016), an electronic health records dataset.

More recently, specialized variants like CXR-BERT (Boecking et al., 2022) and BioViL-T (Bannur et al., 2023) have been developed, targeting the unique challenges posed by CXR reports. CXR-BERT offers both a general version pretrained with MLM on PubMed abstracts and MIMIC-III documents, and a specialized version fine-tuned with MLM coupled with a radiology section matching loss, specifically tailored for reports from the MIMIC-CXR dataset (Johnson et al., 2019b). BioViL-T adopts the same pretraining strategy of CXR-BERT, but is subsequently fine-tuned with global and local multi-modal contrastive learning and image-informed MLM objectives, by combining reports with temporally sequenced image pairs, leading to an improved understanding of radiological sentences with temporal descriptions.

Drawing inspiration from these works, we adopt BERT as our base model for text encoding. However, unlike prior approaches, we employ a novel domain-specific multi-task learning training protocol that leverages LLMs to generate supervision at scale, alongside annotations provided by domain-specific datasets curated by experts.

Label extraction from Radiology Reports. Our work is also related to the problem of extracting information, usually in the form of labels, from free-text radiology reports. A well-known example in the literature is the CheXpert labeler (Irvin

et al., 2019), which uses a rule-based system to infer the presence or absence of 13 observations (plus the label "No findings"). CheXbert (Smit et al., 2020), and VisualCheXbert (Jain et al., 2021b) are subsequent versions that follow the same labeling standard of CheXpert but are based on BERT.

The Chest ImaGenome (Wu et al., 2021) dataset is another example that used a rule-based NLP system to label reports to construct scene graphs for the corresponding frontal images in the MIMIC-CXR dataset (Johnson et al., 2019a). RadGraph (Jain et al., 2021a) proposed a labeling scheme of entities and relations for radiology reports and trained a variant of BERT, DyGIE++ (Wadden et al., 2019), for entity and relation extraction on examples annotated by radiologists. PadChest (Bustos et al., 2019) followed a similar approach, by labeling reports with an LSTM that was previously trained on examples annotated by physicians.

Our work contributes to this field by introducing a more flexible, open-vocabulary approach to information extraction, focused on extracting the essential factual information contained in the report, without imposing constraints that are too rigid. Specifically, we propose extracting factual statements, referred to as "facts," from reports, by leveraging the proven effectiveness of recent LLMs.

3 Method

We introduce a two-stage method for encoding the information within a CXR report. In the first stage, called fact extraction (Section 3.1), we utilize LLMs to extract facts from the original sentences of the report. In the second stage, termed fact encoding (Section 3.2), we employ a BERT-based text encoder to generate sentence embeddings for each extracted fact. When used in tandem, these two stages form a cohesive system capable of producing vectorial representations of the factual statements found within a CXR report.

3.1 Fact Extraction

Figure 2 outlines our method for extracting facts from radiology reports, with examples taken from the MIMIC-CXR dataset (Johnson et al., 2019b). Initially, we use regular expressions and simple rules to pinpoint relevant radiological sections in MIMIC-CXR reports, mainly *Findings* and *Impression*, but we also handle alternate headings. These sections are then divided into sentences. For simplicity, we use the *sent_tokenize* function from the NLTK library[1]. Next, we proceed to extract concise factual statements from each sentence. The main reason for doing this is that radiologists often write sentences that are noisy or complicated. Figure 2 shows two examples of such sentences. The first example contains multiple factual statements in one sentence, which can be simplified into shorter phrases. The second example is overly verbose, but the essential observation can be summarized in a brief phrase. We provide more examples in Table 7. Given the recent success of LLMs, an effective strategy to achieve this sort of extraction is by directing ChatGPT using a custom prompt.

T5 as an alternative to ChatGPT. In theory, this entire process could be executed using off-the-shelf LLMs. However, the expenses associated with accessing the API to annotate the entire dataset can be prohibitive. Therefore, the alternative approach we adopted was to annotate only a strategically selected subset of sentences and then transfer the acquired knowledge from these annotations to a more cost-effective sequence-to-sequence model, such as T5, through fine-tuning. This approach mirrors the strategy employed by Yang et al. (2023), where a T5 is fine-tuned to condense verbose descriptions from GPT-3 in LLM-assisted image classification.

We provide detailed implementation steps for this fact extraction procedure in Appendix A.2

3.2 Fact Encoding

After we extract facts, we encode to obtain vectors in a latent space via a text encoder model, called CXRFE. In this work, we rely on CXR-BERT (Boecking et al., 2022) to implement our fact encoder. Specifically, we use the CXR-BERT-specialized variant available on the Huggingface hub[2], leveraging its built-in [CLS] token projection, which yields a 128-D vector serving as the final representation of the text.

Building on top of CXR-BERT-specialized, we explore 6 different training tasks to enhance the latent representation of radiological sentences: triplet loss for sentence ranking (T), natural language inference (NLI), quadruplet loss to enforce a separation between entailment and contradiction pairs (EC), entity and relation extraction (ER), sentence classification tasks (C), and sentence decoding (SD). We provide details on the implementation of each task in Appendix A.3.

[1] https://www.nltk.org/

[2] https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized

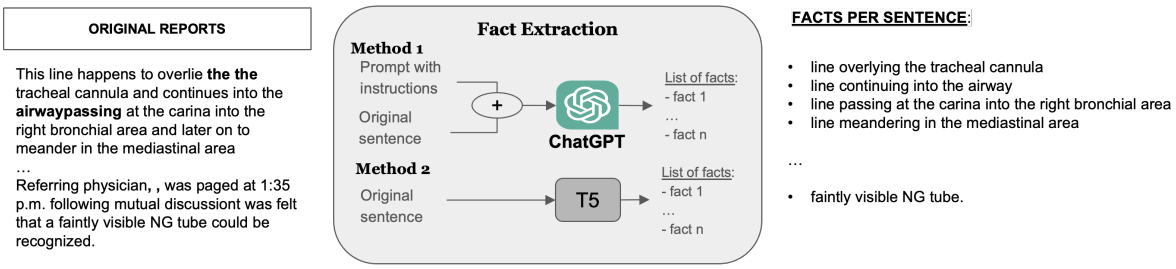

Figure 2: Fact Extraction from Radiology Reports, designed to handle noisy input, including repeated words, incorrect sentence tokenization, typos, and verbose sentences, among other issues. When employing ChatGPT, use context with instructions, whereas T5 is directly applied on the original sentence.

Thus, by combining the two stages, the whole framework can accurately extract and encode facts from CXR reports, thus providing a rich and convenient representation of the factual information for downstream applications.

3.3 CXRFEScore

A potential application of our framework is in evaluating report generation from chest X-rays. We introduce CXRFEScore, an embedding-based metric that leverages our CXRFE model for evaluation.

Given a reference and generated report, we extract facts from each and represent them as embedding vectors, denoting the sets for the original and generated reports as O and G respectively. The cosine similarity matrix M of size $|O| \times |G|$ is formed, where $M_{i,j}$ represents the cosine similarity between the ith vector of O and the jth vector of G. This allows us to calculate S_{row}, S_{col}, and a final score S as follows:

$$S_{\text{col}} = \frac{\sum_j \max_i M_{i,j}}{|G|} \quad S_{\text{row}} = \frac{\sum_i \max_j M_{i,j}}{|O|} \quad S = \frac{S_{\text{row}} + S_{\text{col}}}{2}$$

The equations of this metric resemble those of BERTScore (Zhang et al., 2020a). The key difference lies in CXRFEScore's comparison of fact embeddings rather than token embeddings. This metric illustrates the fusion of the two proposed stages: fact extraction (implemented with T5) and fact encoding (implemented with CXRFE). We provide evidence of the robustness of this metric compared to many existing metrics in Section 4.

3.4 Dataset construction

In our experiments, we primarily utilize the MIMIC-CXR dataset (Johnson et al., 2019b), which comprises 227,827 radiology reports associated with 377,110 chest X-ray images. However, we focus solely on utilizing the reports for our experiments, deferring the exploration of images and multi-modality for future work. Additionally, we incorporate annotations from the Chest ImaGenome dataset (Wu et al., 2021), which provides scene graphs linking observations to anatomical image locations for each frontal view image in MIMIC-CXR. These annotations serve two main purposes: facilitating our creation of a binary multi-label classification task and introducing a radiologist-informed annotation standard covering various observation types and anatomical locations. Similarly, we utilize RadGraph (Jain et al., 2021a), which offers an entity and relation annotation scheme for radiology reports in MIMIC-CXR, and datasets such as MedNLI (Romanov and Shivade, 2018), RadNLI (Miura et al., 2021) and MS-CXR-T (Bannur et al., 2023) for experiments on Natural Language Inference (NLI). It is important to highlight that radiologists or doctors have partially or fully annotated these datasets, which adds significant value to training and evaluation. We direct the reader to Appendix A.3 for further details on these datasets and their utilization in our experiments.

3.4.1 LLM-assisted data augmentations for training supervision

As mentioned in Section 3.2, we use six distinct tasks to train CXRFE. For certain tasks, we leverage ChatGPT to generate additional training data. We elaborate in depth on these aspects in Appendix A.3, so here we only offer a concise overview.

For the triplet loss task (detailed in Section 3.4.2), some triplets incorporate paraphrased facts, which we generate using ChatGPT (an example is depicted in Figure 14). Similarly, we employ ChatGPT to produce challenging triplets, as indicated by the prompt in Figure 20.

In classification-related tasks, each fact is anno-

tated as a JSON metadata object using ChatGPT. This object encompasses fields such as *anatomical location*, *detailed observation*, *short observation*, *category*, *health status*, and *comparison status*. The respective prompt for this process is shown in Figure 16. To refine the "comparison status" field, we use another prompt showcased in Figure 17. Also, we employ ChatGPT to label according to the scheme of the Chest ImaGenome dataset, producing additional observation and anatomical location labels, as shown in Figures 18 and 19.

For the task of natural language inference, we extensively utilize GPT-4 to generate training examples with distinct prompts (refer to Figures 21, 22, 23, 24, and 25).

3.4.2 Triplet Sampling

CXRFE is trained to generate sentence embeddings that cluster semantically similar sentences in the embedding space through a triplet ranking task with binary cross-entropy loss. This approach uses a dataset of triplets, each one with an anchor, a positive sample (akin to the anchor), and a negative one. The difference in similarities is computed as $\Delta\text{sim}(a, p, n) = \text{sim}(a, p) - \text{sim}(a, n)$ from their embeddings' dot product. By minimizing the binary cross-entropy loss, the encoder ensures closely related sentences are nearer and unrelated ones are more distant in the embedding space.

We define six triplet sampling rules to guide the selection process. **Rule 1** prioritizes paraphrases generated with ChatGPT. **Rule 2** involves sampling triplets based on the consensus of BioVIL-T and Levenshtein distance, with the anchor and positive sample sharing the same health status. **Rule 3** ensures proximity between short observations, detailed observations, and original facts, along with their paraphrases. **Rule 4** samples triplets based on Chest ImaGenome labels, ensuring that the anchor and positive sample share at least one label and that BioVIL-T and Levenshtein distance match. **Rule 5** ranks triplets according to the overlap of entities and relations from RadGraph. **Rule 6** includes hard triplets generated by ChatGPT. For each rule, we sampled approximately 3 million training triplets and 1,000 each for validation and testing.

These rules encapsulate specific intuitions and heuristics regarding the ranking of sentence embeddings within the semantic space. The design of these sampling rules and the construction of the triplets dataset involve several technical details, which are elaborated upon in Appendix A.4.

3.4.3 Natural Language Inference

Natural Language Inference (NLI) aims to classify the relationship between a premise and a hypothesis into one of three categories: "entailment", "neutral", or "contradiction". For instance, consider a premise stating *"There are no evident signs of pleural effusion"*, while a hypothesis asserts *"There are evident signs of pleural effusion"*. Despite their structural and lexical similarities, these sentences express contradictory meanings. A robust sentence embedding should possess the capability to distinguish between sentences conveying contradictory diagnoses. Thus, our objective in leveraging NLI during training is to refine sentence embeddings to accurately discern such subtle distinctions.

For training, all MedNLI splits (Romanov and Shivade, 2018) are used, amounting to 14,049 annotated sentence pairs. Radiology-specific datasets include RadNLI (Miura et al., 2021) with 960 pairs and MS-CXR-T (Bannur et al., 2023), an evaluation set with 361 pairs emphasizing condition evolution over time. Given the dearth of training data, we build a custom NLI dataset combining MedNLI (14,049), RadNLI development set (480) and GPT-4 generated examples (154,498), resulting in a total of 169,025 pairs (entailment: 25,175, neutral: 44,729, contradiction: 99,121), while RadNLI test set (480) and MS-CXR-T (361) for evaluation.

4 Experimental Results

In our experiments, we evaluate different versions of CXRFE, each defined by a subset of the six tasks (T, C, NLI, EC, ER, SD) outlined in Section 3.2. This results in a total of 64 potential combinations. However, for the sake of simplicity, we heuristically assess only 11 combinations. Further details on our rationale are provided in Appendix A.3.

Triplet Ranking. We evaluate CXRFE and multiple baselines from the literature on triplet ranking accuracy, using a separate test set of 1000 triplets sampled according to the sampling rules detailed in Section 3.4.2. Table 1 presents these results. Notably, all different versions of CXRFE outperform all the baselines in triplet rules where ChatGPT is heavily involved, namely, paraphrases (R1(o), R1(a), R3) and hard triplets (R6). The hard triplets are especially challenging for the baselines, with BioViL-T only achieving 0.765 accuracy (row 6), while the best performing version of CXRFE achieves 0.952 (row 17). In addition to triplet loss (T), which is key to learn an embedding for these

rules, we notice that sentence decoding (SD) and classification (C) appear to be helpful auxiliary tasks since most of the best scores are achieved by variants that include them (rows 8, 10, 15, 17).

Table 1: Triplet ranking accuracy results. An asterisk (*) denotes perfect scores achieved by BioViL-T due to an unfair advantage from how triplets were heuristically sampled for these rules. Notice that CXR-BERT-specialized also shows a performance boost in these columns, since it is trained in a similar fashion (previous work by the same authors). (o) stands for observations and (a) for anatomical locations.

ID	Text Model	Triplets Test Set (1000 samples per rule)						
		R1(o)	R1(a)	R2	R3	R4	R5	R6
1	BioLinkBERT	0.753	0.725	0.786	0.756	0.644	0.774	0.520
2	PubMedBERT	0.901	0.853	0.905	0.873	0.767	0.834	0.603
3	BioClinicalBERT	**0.922**	**0.864**	0.933	0.912	0.834	**0.948**	0.601
4	CheXbert	0.855	0.771	0.908	0.884	0.760	0.937	0.635
5	CXR-BERT-specialized	0.880	0.804	0.992	0.914	0.904	0.932	0.717
6	BioViL-T	0.910	0.851	**1.000***	**0.938**	**1.000***	0.944	**0.765**
7	CXRFE(T)	0.968	0.955	0.925	0.964	0.798	0.952	0.946
8	CXRFE(T+C)	0.967	0.945	0.967	0.982	**0.926**	**0.988**	0.937
9	CXRFE(T+ER)	0.962	0.946	0.917	0.961	0.798	0.954	0.927
10	CXRFE(T+SD)	**0.981**	**0.966**	0.954	0.977	0.875	0.981	0.898
11	CXRFE(T+EC)	0.957	0.953	0.950	0.971	0.809	0.965	0.943
12	CXRFE(T+NLI)	0.910	0.898	0.958	0.970	0.861	0.968	0.903
13	CXRFE(T+EC+NLI)	0.928	0.950	0.923	0.961	0.777	0.934	0.925
14	CXRFE(T+C+EC+NLI)	0.971	0.932	0.974	0.980	0.892	0.982	0.945
15	CXRFE(T+C+EC+NLI+ER)	0.972	0.944	**0.978**	**0.983**	0.911	0.984	0.936
16	CXRFE(T+C+EC+NLI+SD)	0.977	0.953	0.972	0.980	0.883	0.980	0.951
17	CXRFE(T+C+EC+NLI+ER+SD)	0.976	0.958	0.976	0.982	0.880	0.980	**0.952**

Sentence Ranking. To complement the triplet ranking evaluation, which is based on heuristic sampling rules, we conduct a sentence ranking evaluation using 2412 carefully annotated sentences provided by radiologists from the gold standard of the Chest ImaGenome dataset (Wu et al., 2021). These sentences are annotated with a vocabulary of 70 observations (*yes* (1), *no* (0), or *omitted* (-1)), and 38 anatomical locations (*mentioned* (1) or *unmentioned* (0)), resulting in a discrete vector of size 108 for each sentence.

We report the results of this evaluation in Table 2. In this evaluation, each sentence serves as a query, against which all other sentences are ranked based on the cosine similarity of their embeddings. AUC, acc@k, and cont@k are computed by comparing the labels of the query sentence with those of each ranked sentence. AUC requires defining which sentences are relevant (1) or irrelevant (0) for the query. For two sentences a and b, a is deemed relevant for b if a's labels logically entail b's or vice versa; otherwise, they are considered irrelevant. acc@k represents the mean average accuracy up to the kth sentence in the ranking, while cont@k signifies the mean number of sentences contradicting the query up to the kth sentence, by having contradictory values in at least one observation (1 vs. 0).

Among the baselines, CheXbert (row 4) consistently performs the strongest, likely due to its training in a similar multi-label classification task. CXRFE achieves the best overall performance. Notably, the T+C variant (row 8), combining triplet loss (T) and classification (C), attains the highest AUC and accuracy, while T+EC+NLI (row 13), containing tasks designed for pulling contradictory sentences apart, yields the smallest cont@k scores (where smaller is better). Rows 14-17 represent intermediate points between these two extremes.

Table 2: Sentence ranking results obtained from 2412 sentences annotated by radiologists, sourced from the Chest ImaGenome Gold dataset. Notation: a@k represents the mean average accuracy of the top k ranked sentences (larger is better), while c@k represents the mean number of contradictory sentences among the top k ranked sentences (smaller is better).

ID	Text Model	Chest ImaGenome Gold (2412 sentences)				
		AUC	a@50	a@100	c@50	c@100
1	BioLinkBERT	0.717	0.951	0.945	5.523	12.158
2	PubMedBERT	0.775	0.954	0.947	4.552	10.134
3	BioClinicalBERT	0.830	0.957	0.950	3.823	8.615
4	CheXbert	**0.864**	**0.962**	**0.955**	**1.914**	**4.299**
5	CXR-BERT-specialized	0.837	0.953	0.947	2.905	6.230
6	BioViL-T	0.845	0.956	0.949	3.158	6.903
7	CXRFE(T)	0.896	0.963	0.957	1.668	3.940
8	CXRFE(T+C)	**0.919**	**0.975**	**0.967**	2.955	9.157
9	CXRFE(T+ER)	0.888	0.964	0.957	1.403	3.543
10	CXRFE(T+SD)	0.897	0.961	0.955	2.427	5.465
11	CXRFE(T+EC)	0.840	0.957	0.951	1.237	2.522
12	CXRFE(T+NLI)	0.820	0.951	0.945	1.378	3.240
13	CXRFE(T+EC+NLI)	0.812	0.950	0.943	**0.969**	**2.478**
14	CXRFE(T+C+EC+NLI)	0.890	0.969	0.960	1.437	3.789
15	CXRFE(T+C+EC+NLI+ER)	0.917	0.972	0.964	1.719	4.694
16	CXRFE(T+C+EC+NLI+SD)	0.906	0.971	0.962	1.445	3.603
17	CXRFE(T+C+EC+NLI+ER+SD)	0.907	0.970	0.962	1.502	3.804

NLI. Table 3 presents the NLI results using cosine similarity between sentence vectors, following a methodology akin to Bannur et al. (2023). This methodology specifically focuses on *entailment* and *contradiction* pairs, aiming to assess the efficacy of a text embedding in distinguishing between the two given a similarity threshold. We present results across three datasets: our NLI custom dataset mentioned in Section 3.4.3, the RadNLI test set, and MS-CXR-T. Our reported results are based on thresholds fine-tuned in the NLI custom dataset, alongside upper bounds obtained by tuning thresholds within the same data used for evaluation.

Notably, the employment of the quadruple entailment/contradiction loss (EC) and natural language inference (NLI) (rows 11 to 17) is essential to achieve high performance, significantly outperforming all baselines. In contrast, variants lacking EC and NLI (rows 7 to 10) exhibit a weaker result.

Among the baselines, CheXbert (row 4) demon-

Table 3: NLI results as sentence similarity. a^*_{E+C} denotes the upper bound accuracy with optimal similarity threshold tuning on the same evaluation data. BT stands for best threshold (tuned in NLI custom dataset).

		NLI custom dataset Entailment pairs: 25,175 Contradiction pairs: 99,121				**RadNLI test set** Entailment pairs: 102 Contradiction pairs: 98				**MS-CXR-T** Entailment pairs: 141 Contradiction pairs: 220			
ID	**Text Model**	**BT**	$\mathbf{a_E}$	$\mathbf{a_C}$	$\mathbf{a_{E+C}}$	$\mathbf{a_E}$	$\mathbf{a_C}$	$\mathbf{a_{E+C}}$	$\mathbf{a^*_{E+C}}$	$\mathbf{a_E}$	$\mathbf{a_C}$	$\mathbf{a_{E+C}}$	$\mathbf{a^*_{E+C}}$
1	PubMedBERT	1.000	1.9	99.8	50.8	0.0	100.0	50.0	63.8	0.0	100.0	50.0	56.5
2	BioLinkBERT-large	0.999	1.9	100.0	50.9	0.0	100.0	50.0	65.9	0.0	100.0	50.0	58.1
3	BioClinicalBERT	1.000	1.9	100.0	50.9	0.0	100.0	50.0	69.2	0.0	100.0	50.0	69.5
4	CheXbert	0.556	88.3	33.8	61.0	96.1	83.7	**89.9**	**90.4**	100.0	1.8	50.9	63.3
5	CXR-BERT-specialized	0.619	76.9	44.6	60.8	73.5	85.7	79.6	82.3	100.0	8.6	**54.3**	77.5
6	BioViL-T	0.709	70.9	51.4	**61.2**	58.8	89.8	74.3	78.0	100.0	7.7	53.9	**87.8**
7	CXRFE(T)	0.640	79.8	57.2	68.5	69.6	89.8	79.7	87.3	100.0	18.6	59.2	78.0
8	CXRFE(T+C)	0.934	43.3	69.9	56.6	44.1	95.9	70.0	75.4	97.9	12.3	55.1	62.6
9	CXRFE(T+ER)	0.606	81.6	54.3	67.9	70.6	89.8	80.2	85.9	100.0	16.8	58.4	78.5
10	CXRFE(T+SD)	0.620	68.1	50.5	59.3	81.3	79.6	80.5	81.9	99.3	10.9	55.1	70.3
11	CXRFE(T+EC)	0.308	97.0	93.7	**95.3**	98.0	93.8	96.0	96.4	96.5	69.5	83.0	93.3
12	CXRFE(T+NLI)	0.233	89.3	78.6	84.0	99.0	93.9	**96.4**	**97.0**	100.0	75.9	87.9	95.7
13	CXRFE(T+EC+NLI)	0.267	96.7	93.3	95.0	98.0	93.9	96.0	**97.0**	99.3	84.1	**91.7**	97.6
14	CXRFE(T+C+EC+NLI)	0.288	94.9	82.4	88.6	99.0	91.8	95.4	96.0	100.0	75.9	88.9	97.2
15	CXRFE(T+C+EC+NLI+ER)	0.431	95.2	83.0	89.1	99.0	89.8	94.4	95.5	100.0	59.0	79.5	**98.5**
16	CXRFE(T+C+EC+NLI+SD)	0.480	96.6	87.3	92.0	99.0	89.8	94.4	95.5	100.0	70.5	85.2	93.6
17	CXRFE(T+C+EC+NLI+ER+SD)	0.455	97.1	87.0	92.0	98.0	91.8	94.9	96.0	100.0	72.7	86.4	96.0

Table 4: RadNLI test set accuracy results. Results for CXR-BERT, IFCC, PTUnifier and DoT5 are from the original papers.

ID	**Text Model**	**Acc**
1	CXR-BERT (Boecking et al., 2022)	65.2
2	IFCC (Miura et al., 2021)	77.8
3	PTUnifier (Chen et al., 2023a)	80.0
4	DoT5 (Liu et al., 2023a)	**82.1**
5	GPT-4 with prompt 1 (Fig. 23)	82.3
6	GPT-4 with prompt 2 (Fig. 24)	**89.0**
7	CXRFE(T+NLI)	84.2
8	CXRFE(T+EC+NLI)	81.3
9	CXRFE(T+C+EC+NLI)	86.7
10	CXRFE(T+C+EC+NLI+SD)	85.2
11	CXRFE(T+C+EC+NLI+ER)	88.1
12	CXRFE(T+C+EC+NLI+ER+SD)	88.5
13	CXRFE(NLI fine-tuning)	**89.8**

strates superior performance on RadNLI, while BioVil-T (row 6) is the clear victor on MS-CXR-T. However, all baselines struggle considerably in our NLI custom dataset and are outperformed across all three datasets by variants 11-17 of CXRFE.

Additionally, Table 4 presents the accuracy achieved on the RadNLI test set in the context of the typical 3-class classification task encompassing entailment, contradiction, and neutral classes. In this evaluation, we exclusively assess variants of CXRFE equipped with an NLI classification head (rows 7-13). For insights into the implementation of NLI classification, please consult Figure 7.

Among the existing literature, the strongest baseline identified is DoT5 (Liu et al., 2023a) (82.1), employing a sophisticated sequence-to-sequence approach based on T5. Furthermore, we conducted evaluations on GPT-4 utilizing two distinct prompts (rows 5-6). Figures 23 and 24 in the Appendix show the respective prompts. Intriguingly, the inclusion of chain-of-thought reasoning and two examples in the second prompt led to a significant performance boost (from 82.3 to 89.0). Consequently, GPT-4 with prompt 2 was selected as our "oracle" for generating additional training data (more details on this in Appendix A.3).

Most versions of CXRFE showed superior performance compared to baselines. Surprisingly, a version tuned explicitly for NLI (row 13) even outperforms GPT-4 (row 6) by a narrow margin (89.8).

CXRFEScore vs. existing metrics. To evaluate the quality of our proposed metric, we conduct a meta-evaluation of CXRFEScore alongside multiple metrics from the literature, as shown in Table 5. This assessment encompasses three settings: (1) a sentence ranking evaluation using 2412 sentences, (2) a report ranking evaluation with 500 reports, both sourced from the gold standard of Chest ImaGenome, and (3) a natural language inference evaluation utilizing entailment (336) and contradiction (424) pairs obtained from RadNLI and MS-CXR-T. Note that all these datasets are manually annotated by radiologists, thus serving as a reliable gold standard for metric comparison.

Regarding the baseline metrics, RadGraph F1 (row 11) emerges as one of the most promising, a result that is consistent with the findings of Yu et al. (2022). It is closely followed by BERTScore (row 5). Notably, CheXbert (rows 7 and 9), closely followed by CheXpert (rows 6 and 8), shows the least number of contradictions among the baselines in the gold standard of Chest ImaGenome. However, all baseline metrics are surpassed by CXRFEScore (rows 12-15). Particularly striking is the observation that the baseline metrics struggle significantly to differentiate between entailed and contradictory sentences, as indicated by the AUC results in last column of Table 5. RadGraph F1 achieves only an AUC of 0.610. In contrast, the best version of CXRFEScore (row 15) achieves an AUC of 0.938. This suggests that current metrics are assigning elevated scores to pairs of sentences with contradictory semantics, underscoring the necessity for improved metrics that can discern these subtleties, which is precisely what our metric is designed to accomplish. Additional details and results about these metrics can be found in Appendix A.5.

Fact Extraction Quality. To evaluate the quality of the fact extraction stage, we interpret these facts as *open-vocabulary* labels and compare them

Table 5: Comparative evaluation of text generation metrics. Notation: a@k denotes the mean average accuracy up to the kth sentence, c@k represents the mean number of contradictory sentences up to the kth sentence, and j@k represents the mean average Jaccard index up to the kth sentence.

ID	Metric	Chest ImaGenome Gold Sentences (2412)					Chest ImaGenome Gold Reports (500)			RadNLI + MS-CXR-T
		AUC	a@50	a@100	c@50	c@100	j@20	j@50	j@100	AUC
1	BLEU (Papineni et al., 2002)	0.767	0.951	0.944	4.432	9.148	0.509	0.469	0.442	0.537
2	ROUGE-L (Lin, 2004)	0.773	0.953	0.946	4.228	9.029	0.508	0.466	0.439	0.547
3	METEOR (Banerjee and Lavie, 2005)	0.829	0.954	0.947	4.282	10.100	0.514	0.471	0.444	0.596
4	CIDEr-D (Vedantam et al., 2015)	0.778	0.952	0.946	4.883	10.234	0.482	0.441	0.413	0.556
5	BERTScore (Zhang et al., 2020a)	0.822	**0.962**	**0.955**	5.138	11.018	0.536	0.492	0.461	0.562
6	CheXpert Accuracy (Irvin et al., 2019)	0.764	0.941	0.939	3.263	6.263	0.467	0.446	0.434	0.592
7	CheXbert Accuracy (Smit et al., 2020)	0.778	0.941	0.941	**2.667**	**5.103**	0.485	0.452	0.438	0.592
8	CheXpert F1 (Irvin et al., 2019)	0.742	0.939	0.938	3.658	6.877	0.460	0.441	0.423	0.582
9	CheXbert F1 (Smit et al., 2020)	0.753	0.939	0.939	2.709	5.301	0.472	0.445	0.424	0.583
10	RadGraph Jaccard (Jain et al., 2021a)	**0.833**	**0.962**	0.954	3.204	8.143	**0.546**	**0.500**	**0.466**	**0.610**
11	RadGraph F1 (Jain et al., 2021a)	**0.833**	**0.962**	0.954	3.203	8.133	**0.546**	**0.500**	**0.466**	**0.610**
12	CXRFEScore(T+C+EC+NLI)	0.912	0.967	0.958	1.563	3.953	0.562	0.510	0.472	0.927
13	CXRFEScore(T+C+EC+NLI+ER)	**0.921**	**0.968**	**0.959**	1.575	4.122	**0.563**	**0.511**	**0.474**	0.931
14	CXRFEScore(T+C+EC+NLI+SD)	0.911	0.967	0.958	**1.518**	**3.763**	0.553	0.505	0.470	0.936
15	CXRFEScore(T+C+EC+NLI+ER+SD)	0.911	0.967	0.958	1.606	3.874	0.555	0.507	0.470	**0.938**

Table 6: Template-based report generation metrics on MIMIC-CXR test set for different label extraction methods. CXRFEScore was calculated using the T+C+EC+NLI+SD variant.

ID	Label Extraction Method	CXRFEScore	RadGraph-based metrics				CheXpert		CheXbert		BERTScore			BLEU	CIDEr-D	ROUGE-L	METEOR
			Jacc	F1	P	R	Acc	F1	Acc	F1	F1	P	R				
1	CheXpert labeler	0.661	0.066	0.121	0.159	0.106	**0.999**	**0.990**	0.970	0.854	**0.849**	**0.860**	**0.838**	**0.056**	**0.023**	**0.123**	**0.179**
2	CheXbert	0.664	**0.067**	**0.122**	**0.161**	0.107	0.974	0.921	**0.992**	**0.907**	**0.849**	**0.860**	**0.838**	**0.056**	**0.023**	**0.123**	**0.179**
3	Chest ImaGenome	**0.687**	0.051	0.095	0.065	**0.220**	0.869	0.693	0.874	0.751	0.811	0.801	0.822	0.029	0.002	0.086	0.170
4	Fact extraction (T5)	**0.986**	**0.644**	**0.776**	**0.799**	**0.768**	**0.986**	**0.964**	**0.979**	**0.946**	**0.927**	**0.939**	**0.916**	**0.366**	**0.555**	**0.523**	**0.630**

against three existing radiology report label extraction methods: CheXpert labeler (Irvin et al., 2019), CheXbert (Smit et al., 2020), and Chest ImaGenome (Wu et al., 2021). For Chest ImaGenome, we use the labels from the dataset's scene graphs, as the original NLP algorithm is not publicly available. Our evaluation protocol involves the following: for each MIMIC-CXR test set report and label extraction method, labels are extracted, converted into a report, and evaluated against the original report by several metrics. We adopt the templates suggested by Pino et al. (2021) for CheXpert labeler and CheXbert, while for Chest ImaGenome, we utilize basic templates such as "(no) {observation} in {anatomical location}". For fact extraction, we simply concatenate the facts. These template-based reports are illustrated in Table 8.

Table 6 presents these results. Notably, fact-based reports generated by T5 (row 4) achieve the top scores. This underscores the efficacy of the fact extraction process. Moreover, CXRFEScore suggests that Chest ImaGenome outperforms both the CheXpert labeler and CheXbert, although the improvement is rather marginal. However, it is evident that all three baseline methods fall short of fully capturing the factual information within the report, a conclusion further supported by RadGraph metrics, likely due to their rigid annotation rules.

5 Conclusions & Future Work

In this work, we present a novel two-stage framework for extracting and encoding factual information from radiology reports. The first stage, fact extraction, uses ChatGPT and T5 to extract fact statements. The second stage, fact encoding, introduces CXRFE, a specialized variant of CXR-BERT, fine-tuned through multitask learning incorporating tasks that support representation improvement. Our system's effectiveness is validated through comprehensive evaluations. Additionally, we introduce CXRFEScore, a novel metric for evaluating radiology text generation, leveraging our two-stage system. We anticipate that our work will stimulate further research in enhanced fact extraction and representation, LLM utilization, advanced training methodologies, and improved evaluation metrics. For future work, we aim to explore expanding our framework to integrate visual modality, focusing on image-based fact detection and visual grounding.

6 Limitations

Our study acknowledges several limitations and identifies areas for improvement. Initially, we acknowledge the necessity for more expert evaluation, especially from radiologists, to refine the utilization of large language models (LLMs) in radiology. While we extensively utilized publicly available gold standards, such as those from the Chest ImaGenome dataset, RadNLI, and MS-CXR-T, we believe there is still ample opportunity for improvement. Collaborative prompt engineering and rigorous evaluation protocols are crucial to enhancing prompt quality.

We also see potential for refining triplet sampling heuristics and involving radiologists in their design. Optimizing LLM prompts for triplet sampling and incorporating superior auxiliary embeddings could further improve our approach.

Furthermore, while our study focuses on text-only analysis, we recognize the importance of integrating visual data, such as chest X-ray images, into a multimodal framework. Devising a training protocol that effectively combines supervision from both images and text is an area of potential improvement that can be explored in subsequent work.

In this work, we limited our experiments to the sections "findings", "impression", and similar headings providing factual statements about the imaging exam. However, other sections, such as "comparison", "indication", and "history", were left out of the analysis, and yet they can provide deeper insights into patient information and context. Investigating how this broader information can be extracted and encoded to enhance downstream applications is another avenue for future exploration and potential improvement.

Lastly, we acknowledge that our fact extraction algorithm may be limited due to its reliance on the *sent_tokenize* function of the NLTK library, which we use to obtain a preliminary division of the report into coarse sentences (prior to fact extraction). This method could falter when a fact spans multiple sentences connected through co-reference. While such occurrences are relatively uncommon in our observations, a deeper exploration of this linguistic aspect could guide the development of a more refined fact extraction mechanism that overcomes this challenge.

References

Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bressem. 2023. Leveraging gpt-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4):e230725.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. ACL.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2019. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv:1901.07441*.

Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. 2023a. Towards unifying medical vision-and-language pre-training via soft prompts. *arXiv preprint arXiv:2302.08958*.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023b. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–484, Toronto, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 478–482.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2006–2013. IOS Press.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. 2023. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint arXiv:2307.06439*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021a. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A. Young, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. 2021b. Visualchexbert: Addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 105–115, New York, NY, USA. Association for Computing Machinery.

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(1):317.

Alistair EW 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. *arXiv:1901.07042*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, 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.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

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

Fangyu Liu, Qianchu Liu, Shruthi Bannur, Fernando Pérez-García, Naoto Usuyama, Sheng Zhang, Tristan Naumann, Aditya Nori, Hoifung Poon, Javier Alvarez-Valle, Ozan Oktay, and Stephanie L. Hyland. 2023a. Compositional Zero-Shot Domain Transfer with Text-to-Text Models. *Transactions of the Association for Computational Linguistics*, 11:1097–1113.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, YANG FENG, and Zuozhu Liu. 2023b. A chatGPT aided explainable framework for zero-shot medical image diagnosis. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. 2023c. Exploring the boundaries of GPT-4 in radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14414–14445, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, et al. 2023. Impressiongpt: an iterative optimizing framework for radiology report summarization with chatgpt. *arXiv preprint arXiv:2304.08448*.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andia, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40.

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: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. ACL, ACL.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 654–663. Springer.

Pablo Pino, Denis Parra, Pablo Messina, Cecilia Besa, and Sergio Uribe. 2020. Inspecting state of the art performance and nlp metrics in image-based medical report generation. *arXiv preprint arXiv:2011.09257*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Yucheng Shi, Hehuan Ma, Wenliang Zhong, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. 2023. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. *arXiv preprint arXiv:2305.03513*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *CoRR*, abs/2004.09167.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549.

Joy T Wu, Nkechinyere Agu, Ismini Lourentzou, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest imagenome dataset for clinical reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

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, et al. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, pages 2022–08.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

A Appendix

A.1 Additional Related Work

Evaluation of Factual Correctness in Radiology Text Generation. One important area of application motivating this work is the evaluation of factual correctness in systems that generate radiological text, usually from input medical imaging. Recent research emphasizes enhancing the accuracy of generated facts in applications like report generation (Miura et al., 2021; Delbrouck et al., 2022; Pino et al., 2020, 2021) and summarization (Zhang et al., 2020b; Delbrouck et al., 2023). Pino et al. (2020) conducted an evaluation of several trivial report generation baselines using established metrics such as BLEU (2002), ROUGE-L (2004), and CIDEr-D (2015). They achieved results comparable to state-of-the-art papers at the time. However, when assessed using the CheXpert labeler (Irvin et al., 2019), a domain-specific NLP tool designed to detect 13 findings, the performance was notably poor, underscoring the urgent necessity for standardizing improved evaluation metrics among researchers. More recently, Delbrouck et al. (2022) repurposed RadGraph's entity and relation extraction model (Jain et al., 2021a) to create a factual correctness reward. This reward measures the overlap of entities and relations between real and generated reports, serving as the guiding signal to optimize a report generation model through reinforcement learning. Interestingly, their proposed reward aligns functionally with the RadGraph F1 metric introduced by Yu et al. (2022). Yu et al. conducted a study on metrics for radiology report generation, determining that RadGraph F1 and BLEU show the highest correlation with radiologists' judgement. Recently, the RadSum23 challenge (Delbrouck et al., 2023) evaluated multimodal radiology report summarization quality using BLEU4 (2002), ROUGEL (2004), BERTScore (2020a), CheXbert F1 (2020), and RadGraph F1 (2021a).

Our work holds significant relevance in this domain due to our development of the CXRFEScore metric. This metric combines the strengths of both stages within our framework: fact extraction and encoding. As a result, CXRFEScore is explicitly designed to capture the factual correctness of generated radiological text in comparison to a reference text. Section 4 provides comprehensive evaluations, demonstrating the effectiveness of our two-stage system. This includes a comparison of CXRFEScore against commonly used metrics in the literature, with favorable results.

LLMs in Medicine. Our work falls under the category of applications of LLMs to the medical domain. Specifically, in this work we make use of ChatGPT versions GPT-3.5 and GPT-4 through OpenAI's API[3]. Recent works have shown the effectiveness of ChatGPT applied to medical tasks. Most notably, recent work by Liu et al. (2023c) explored the boundaries of GPT-4 in radiology, evaluated in tasks such as classification, summarization, and natural language inference, with remarkable performance. Liu et al. (2023b) employed ChatGPT to generate short sentences with plausible symptoms of medical conditions for interpretable zero-shot medical image diagnosis. Adams et al. (2023) used GPT-4 to transform free-text radiology reports into structured templates, with remarkable results. GPT-4 is also known for having passed the bar exam (Katz et al., 2023). Inspired by these results, we make extensive use of ChatGPT to produce abundant annotations through diverse prompts.

Knowledge Distillation from LLMs. Our approach can be also viewed as a form of LLM knowledge distillation, where a LLM ("teacher") is queried to generate annotations for training a more compact model ("student"). Shi et al. (2023) illustrated this idea by using ChatGPT to extract knowledge graphs from text to train a smaller model for text classification. Similarly, Gu et al. (2023) applied this concept in the biomedical field, distilling knowledge from GPT-3.5 for adverse drug event

[3] https://platform.openai.com/

extraction, with student models like PubMedBERT and BioGPT. Like these works, our work presents a form of knowledge distillation in the fact that we fine-tune T5 (student) using annotations generated with ChatGPT (teacher), essentially training an imitator that is cheaper and easier to run in our machines. We do this in the task of fact extraction from reports, and in the task of generating metadata and labels for each fact. Additionally, we make extensive use of GPT-4 to produce high quality NLI labels, but in this case we don't use T5. However, these labels are used as part of the training supervision of CXRFE, which can be regarded as a form of knowledge distillation as well.

A.2 Fact Extraction Implementation Details

In our experiments, we analyzed the 227,827 radiology reports provided by the MIMIC-CXR dataset (Johnson et al., 2019b). To pinpoint relevant radiological sections in the MIMIC-CXR reports, such as *Findings*, *Impression*, and various other headings, we employed a combination of regular expressions and simple rules. These sections were then segmented into sentences using NLTK's *sent_tokenize* function, resulting in 677,694 unique sentences after processing the entire dataset. Subsequently, we extracted facts from each sentence.

Extracting factual statements from a sentence of a free-text radiology report using traditional approaches, such as regular expressions, hand-designed rules, and similar heuristics, poses significant challenges due to the complexity and diversity of vocabulary used by radiologists. A more promising alternative is to leverage the capabilities of powerful LLMs like ChatGPT to tackle this task. Table 7 presents several examples of facts extracted by GPT-4 from challenging sentences. This is achieved by providing the model with a specific set of instructions: (Refer to Figure 13 for a screenshot of OpenAI's web interface displaying the same prompt.)

> Relevant facts:
>
> 1. observations of abnormalities 2. observations of diseases 3. observations of strange visual patterns 4. observations of devices 5. observations of foreign bodies 6. observations of specific anatomical regions that look normal or healthy 7. absences of abnormalities (usually expressed with a negation) 8. comparisons with respect to a previous study (something changed or remained the same)
>
> Task:
>
> Given a sentence taken from a chest x-ray report, generate a JSON list of relevant facts. Each fact should be about one observation. If a sentence mentions multiple observations, each observation should be extracted as a separate fact. Each fact should include the anatomical location where it was observed. If multiple facts occur in the same location, repeat the location in each fact.
>
> If no relevant facts are mentioned, return [] (an empty array).
>
> Examples:
>
> Opacity and density in the right lobe
>
> ["opacity in the right lobe", "density in the right lobe"]
>
> Lungs are well inflated without evidence of focal airspace consolidation to suggest pneumonia.
>
> ["well inflated lungs", "lungs without evidence of focal airspace consolidation", "lungs without evidence of pneumonia"]
>
> Taken together, compared with less than 1 hr earlier, the findings are suggestive of worsening of CHF, with new or significantly increased left greater right pleural effusions and underlying bibasilar collapse and/or consolidation, particularly on the left.
>
> ["worsening of CHF", "new or significantly increased left pleural effusions", "new or significantly increased right pleural effusions", "underlying bibasilar collapse on the left", "underlying consolidation on the left",]
>
> No acute cardiopulmonary abnormality
>
> ["no acute cardiopulmonary abnormality"]

Given the relatively high cost of both GPT-4 and GPT-3.5 models, we opted to annotate a subset of the MIMIC-CXR sentences with these LLMs. To select this subset, we employed a two-pronged approach. Firstly, we ranked sentences based on the sum of the inverse frequency of their tokens, prioritizing longer and more complex sentences, which often include infrequent abnormalities, typos, and symbols. Secondly, to ensure diversity of topics, we clustered the sentences into 200 groups using K-Means, utilizing embedding representations obtained with BioViL-T, and sampled equally from each cluster in order of difficulty. This combined strategy allowed us to curate a subset that is both diverse and challenging. GPT-4-0613 extracted facts from 14,999 sentences, while GPT-3.5-turbo-0613 processed 69,936. Subsequently, we trained T5-small for fact extraction using a total of 84,735 training examples, reserving 200 examples for validation. Once trained, T5-small annotated the remaining 592,759 sentences, yielding a total of 1,323,687 facts, including duplicates, of which 583,202 were unique after duplicate removal.

Table 7: Examples of facts extracted by GPT-4, with the prompt shown in Figure 13

Original sentence	Extracted facts	Comment
2.Atelectasis of the left lower lobe 3.Stable left lower lobe laceration.	• Atelectasis of the left lower lobe • Stable left lower lobe laceration	In this example, NLTK's sent_tokenize failed to split the text into two sentences, but GPT-4 has no problems identifying the facts.
A CT of the abdomen showing the lower chest performed on adds the detection of acinar-sized nodules, some of which are confluent at the lung periphery and confirms the presence of small bilateral pleural effusions.	• detection of acinar-sized nodules at the lung periphery • some nodules are confluent at the lung periphery • presence of small bilateral pleural effusions	A long complicated sentence broken down into 3 shorter factual statements.
NO chf, focal inifiltrate, or gross effusionis identified	• no CHF identified • no focal infiltrate identified • no gross effusion identified	A negated sentence becomes 3 short facts, each one preserving the negation. GPT-4 also correctly fixes a typo.
Severe cardiomegaly is again noted as well as aortic valve replacement including stent material at the aortic root compatible with appearance of Carevalve Aortic bioprothesis.	• Severe cardiomegaly • aortic valve replacement • stent material at the aortic root • appearance of Carevalve Aortic bioprothesis at the aortic root	Complex sentence becomes 4 shorter facts.
A right-sided internal jugular central venous catheter is seen, which appears slightly high in position, terminating in the region of the right innominate vein/right internal jugular/innominate vein junction.	• right-sided internal jugular central venous catheter is seen • catheter appears slightly high in position' • catheter terminating in the region of the right innominate vein • catheter terminating in the region of the right internal jugular • catheter terminating in the region of the innominate vein junction	Complex sentence becomes 5 shorter facts. Note GPT-4's ability to handle forward slashes (/).
THEY REPORT TEXT FOLLOWS: The heart and great vessels are normal.	• normal heart • normal great vessels	GPT-4 successfully ignores a distracting header (in uppercase) and correctly distributes the word "normal" to each extracted fact.

A.3 CXRFE's Tasks Details

CXRFE is a fine-tuned version of CXR-BERT-specialized, accessible for download from https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized. This fine-tuning process entails multi-task learning, incorporating six distinct tasks: triplet loss for sentence ranking (T), sentence classification (C), sentence decoding (SD), natural language inference (NLI), quadruplet loss for enforcing separation between entailment and contradiction pairs (EC), and entity and relation extraction (ER). While there exist a total of 64 possible task combinations, we focus on 11 combinations in our experiments, namely: T, T+C, T+ER, T+SD, T+EC, T+NLI, T+EC+NLI, T+C+EC+NLI, T+C+EC+NLI+ER, T+C+EC+NLI+SD, T+C+EC+NLI+ER+SD. This selection is guided by the following heuristics:

- Triplet loss (T) is always included, given our belief, as discussed in Appendix A.4, that the developed triplets dataset captures many desirable properties for learned sentence embeddings.
- We explore each combination of T with the other tasks individually: T+C, T+ER, T+SD, T+EC, T+NLI.
- Recognizing the complementary nature of EC and NLI in exploiting natural language inference data, we explore T+EC+NLI.
- According to our experimental results, both C and EC+NLI serve as effective auxiliary tasks; thus, we keep them fixed while varying other combinations of SD and ER: T+C+EC+NLI, T+C+EC+NLI+ER, T+C+EC+NLI+SD, T+C+EC+NLI+ER+SD.

Next, we delve into the implementation details of each task in our experiments.

Triplet loss for sentence ranking (T). One of the tasks we explore for model fine-tuning is sentence ranking via triplet loss. Figure 3 illustrates this task. Concretely, we forward 3 sentences (anchor, positive, negative) through CXR-BERT-specialized with weight sharing, obtaining three vectors a, b, and c each of dimension 128, and compute $\Delta\text{sim}(a, p, n) = a \cdot p - a \cdot n$. This is serves as the input logit for a binary cross-entropy loss.

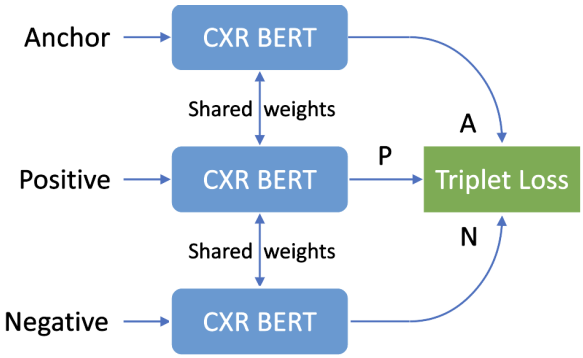

Figure 3: Triplet loss (T)

Sentence classification tasks (C). A second task group is classification tasks (Figure 4). These include category (5 classes: *anatomical finding*, *disease*, *technical assessment*, *tubes and lines* and *device*), health status (4 classes: i.e., *normal*, *abnormal*, *ambiguous*, *unknown*), comparison status (15 classes, see Figure 17), Chest ImaGenome observations (74 classes, see Figure 18) and anatomical locations (38 classes, see Figure 19). Category, Health Status and Comparison Status are single-label multi-class classification tasks, whereas Chest ImaGenome observations and anatomical locations are multi-label multi-class classification tasks. Implementing these tasks require attaching fully connected heads on top of CXR-BERT-specialized's built-in projection layer in order to perform the classification.

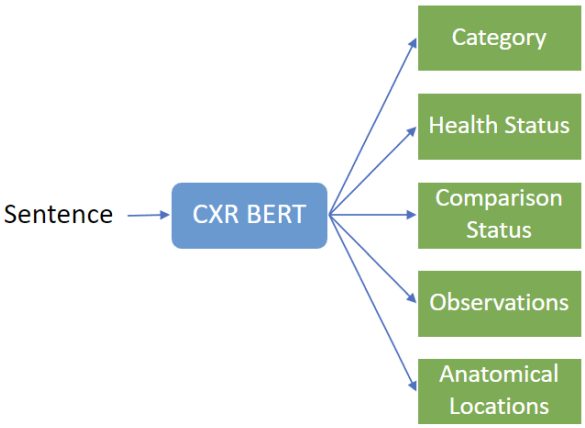

Figure 4: Sentence classification (C)

Sentence decoding (SD). Another task is sentence decoding (Figure 5). We attach a lightweight, shallow Transformer Decoder to CXR-BERT-specialized's projection layer in order to generate back the original sentence. This can be viewed a sort of text autoenconder, forcing the projection layer to capture as much information as possible of the input sentence to facilitate the reconstruction of the sentence by the Transformer Decoder. We use a Transfomer Decoder with embedding, hidden and feedforward dimension 256, only one self-attention head and only one layer.

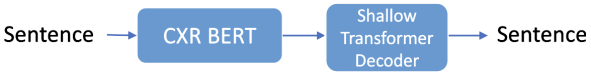

Figure 5: Sentence decoding (SD)

Quadruplet loss: enforcing separation between entailment and contradiction pairs (EC). The next task is what we refer to as entailment/-contradiction quadruplet loss (Figure 6). The goal of this task is to promote a generalized separation of entailment and contradiction sentence pairs in the latent space, by randomly sampling entailment and contradiction pairs and requiring that the entailment pair have greater similarity than the contradiction pair. This loss was crucial to achieve state-of-the-art results in Table 3.

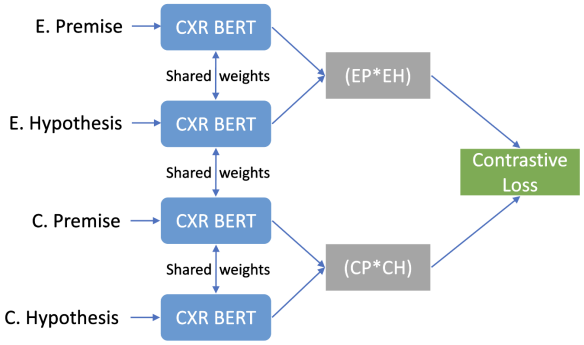

Figure 6: Entailment/contradiction quadruplet loss (EC)

Natural language inference (NLI). For NLI, we adopt an approach similar to that of SBERT (Reimers and Gurevych, 2019), by concatenating the embeddings of the premise, hypothesis and their element-wise multiplication, followed by a fully connected layer and a softmax layer for NLI classification (see Figure 7).

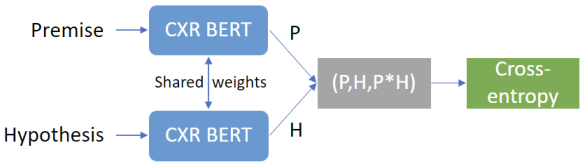

Figure 7: Natural language inference (NLI)

Entity and relation extraction (ER). Lastly, for

entity and relation extraction we augment CXR-BERT-specialized with the layers proposed by SpERT (Eberts and Ulges, 2020). This adaptation was relatively straightforward, since the authors of SpERT released an implementation (https://github.com/lavis-nlp/spert/) that is compatible with Huggingface models like CXR-BERT-specialized.

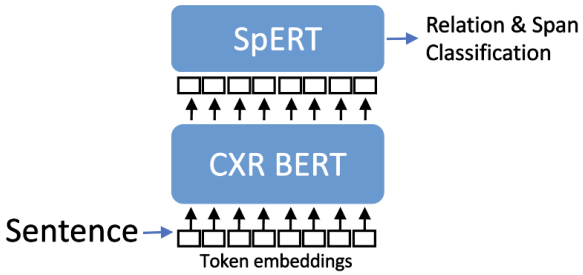

Figure 8: Entity and relation extraction (ER) with SpERT

A.4 Triplets Dataset Construction Details

CXRFE is trained to generate sentence embeddings that organize semantically similar sentences into clusters within the embedding space via a triplet ranking task. This task involves a dataset comprising triplets (a, p, n), where a represents an anchor sentence, p is a positive sentence, and n denotes a negative sentence. The objective is to train a text encoder in a manner that ensures $\text{sim}(a, p) > \text{sim}(a, n)$ holds true for each triplet, with $\text{sim}(x, y)$ computed as the cosine similarity between sentence embeddings (or dot product if they are already normalized).

The rationale behind selecting triplet loss lies in its versatility, offering flexibility in terms of desired properties for the learned embedding function. For instance, if a and p denote identical medical observations using different vocabulary, while n presents an unrelated observation, it's logical to enforce $\text{sim}(a, p) > \text{sim}(a, n)$. Similarly, if there's substantial topic overlap between a and p, whereas n is largely unrelated or contradictory to a, it's reasonable to aim for $\text{sim}(a, p) > \text{sim}(a, n)$ to hold true.

To guide a text encoder towards learning sentence embeddings consistent with triplets, we can define the difference in similarities as $\Delta\text{sim}(a, p, n) = \text{sim}(a, p) - \text{sim}(a, n)$. Then, a straightforward approach is to utilize $\text{sim}(a, p, n)$ as the input for a binary cross-entropy loss, where the ground-truth label is consistently set to 1.

Notation. To define our triplet sampling heuristics, we use the notation $\text{E}(x)$ for the embedding of sentence x, $\cos(\text{E}(x), \text{E}(y))$ for the cosine similarity between embeddings of x and y, $\text{lev}(x, y)$ for the Levenshtein string distance between them, and $\text{levsim}(x, y) = 1 - \text{lev}(x, y) / \max(\text{len}(x), \text{len}(y))$. $\text{c}(x)$ indicates the cluster id for sentence x after running a clustering algorithm like K-Means on the sentence embeddings. This requires having an auxiliary text encoder capable of producing these auxiliary embeddings and clusters. Specifically, we use BioViL-T (Bannur et al., 2023), a state-of-the-art BERT-based model for radiological text, available on Huggingface[4].

With this, we sample triplets based on the following heuristics:

Rule 1: Rank paraphrases highly. $\Delta\text{sim}(a, p, n) > 0$ if p is a paraphrase of a generated by ChatGPT and n is any other sentence (unless $\cos(\text{E}(a), \text{E}(p)) < \cos(\text{E}(a), \text{E}(n))$ and $\text{lev}(a, p) > \text{lev}(a, n)$). To generate paraphrases, we employ the prompts shown in Figures 14 and 15, one for paraphrasing facts and another for paraphrasing anatomical locations. We decided to paraphrase anatomical locations too in order to strengthen the model's understanding of their vocabulary. As a reminder, the anatomical locations are obtained from facts as part of the metadata generated with the prompt shown in Figure 16.

Rule 2: Sample triplets according to the consensus of E and lev, while anchor and positive share the same health status. $\Delta\text{sim}(a, p, n) > 0$ if $\text{HS}(a) = \text{HS}(p)$, $\text{c}(p) = \text{c}(a)$, $\text{c}(p) \neq \text{c}(n)$, $\cos(\text{E}(a), \text{E}(p)) > \cos(\text{E}(a), \text{E}(n)) + \text{margin}_{\cos}$ and $\text{levsim}(a, p) > \text{levsim}(a, n) + \text{margin}_{\text{lev}}$. This rule is very heuristic, and can be interpreted as a sort of mixture of experts. Essentially, we are looking for multiple signals to agree, increasing the likelihood that the resulting triplet passing those tests is in fact a high quality triplet. Levenshtein distance is useful to capture the notion of string similarity, while BioViL-T captures more the semantics (to some extent). The health status condition is adding an additional constraint. As a reminder, the health status is also obtained with the prompt of Figure 16.

Rule 3: Short observation, detailed observation and the original fact (and their paraphrases) should be close to each other. Given

[4] https://huggingface.co/microsoft/BiomedVLP-BioViL-T

Table 8: Examples of template-based generated reports for different label extraction algorithms.

Ground-truth Report	Fact extraction	CheXbert	CheXpert labeler	Chest ImaGenome
New PICC line on the right is projecting with its tip somewhere in the mediastinum. Appears to cross the midline, there is concern for potential arterial location. The initial line concerns were communicated over the telephone at the time of the wet read. Repeat PA and lateral radiograph, taken approximately an hour after the radiograph demonstrated the PICC line in the mid SVC. Potential small right pleural effusion. Stable moderate cardiomegaly.	new PICC line on the right. tip of the PICC line in the mediastinum. potential arterial location crossing the midline. PICC line in the mid SVC. potential small right pleural effusion. stable moderate cardiomegaly **CXRFEScore:** 1.000 **RadGraph F1:** 0.796 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	the heart is enlarged. the cardiomediastinal silhouette is enlarged. no focal consolidation. the lungs are free of focal airspace disease. no atelectasis. a device is seen. pleural effusion is seen. no fibrosis. no pneumonia. no pneumothorax is seen. no pulmonary edema. no pulmonary nodules or mass lesions identified. no fracture is seen **CXRFEScore:** 0.517 **RadGraph F1:** 0.021 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	the heart is enlarged. the cardiomediastinal silhouette is enlarged. no focal consolidation. the lungs are free of focal airspace disease. no atelectasis. a device is seen. pleural effusion is seen. no fibrosis. no pneumonia. no pneumothorax is seen. no pulmonary edema. no pulmonary nodules or mass lesions identified. no fracture is seen **CXRFEScore:** 0.517 **RadGraph F1:** 0.021 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	enlarged cardiac silhouette in cardiac silhouette. abnormal cardiac silhouette. picc in left shoulder. picc in mediastinum. lung opacity in right costophrenic angle. pleural effusion in right costophrenic angle. abnormal right costophrenic angle. lung opacity in right lung. pleural effusion in right lung. abnormal right lung. picc in right shoulder. picc in svc. enlarged cardiac silhouette. lung opacity. pleural effusion. picc **CXRFEScore:** 0.647 **RadGraph F1:** 0.103 **CheXpert Acc:** 0.929 **CheXbert Acc:** 0.929
AP chest compared to :. Moderately severe pulmonary edema is worsening, accompanied by increasing small-to-moderate bilateral pleural effusions and progressive moderate-to-severe cardiomegaly. Right jugular line ends in the upper right atrium. ET tube is in standard placement. No pneumothorax. Thoracic aorta is heavily calcified and at least tortuous if not dilated, but probably not acutely changed.	moderately severe pulmonary edema. worsening pulmonary edema. increasing small-to-moderate bilateral pleural effusions. progressive moderate-to-severe cardiomegaly. right jugular line ends in the upper right atrium. ET tube in standard placement. no pneumothorax. heavily calcified thoracic aorta. tortuous thoracic aorta. thoracic aorta not acutely changed **CXRFEScore:** 1.000 **RadGraph F1:** 0.754 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	the heart is enlarged. the mediastinal contour is normal. no focal consolidation. the lungs are free of focal airspace disease. no atelectasis. a device is seen. pleural effusion is seen. no fibrosis. no pneumonia. no pneumothorax is seen. pulmonary edema is seen. no pulmonary nodules or mass lesions identified. no fracture is seen **CXRFEScore:** 0.676 **RadGraph F1:** 0.103 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	the heart is enlarged. the mediastinal contour is normal. no focal consolidation. the lungs are free of focal airspace disease. no atelectasis. a device is seen. pleural effusion is seen. no fibrosis. no pneumonia. no pneumothorax is seen. pulmonary edema is seen. no pulmonary nodules or mass lesions identified. no fracture is seen **CXRFEScore:** 0.676 **RadGraph F1:** 0.103 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	tortuous aorta in aortic arch. vascular calcification in aortic arch. abnormal aortic arch. enlarged cardiac silhouette in cardiac silhouette. abnormal cardiac silhouette. ij line in cardiac silhouette. lung opacity in left costophrenic angle. pleural effusion in left costophrenic angle. pulmonary edema or hazy opacity in left costophrenic angle. abnormal left costophrenic angle. lung opacity in left hilar structures. pulmonary edema or hazy opacity in left hilar structures. abnormal left hilar structures. lung opacity in left lung. pleural effusion in left lung. no pneumothorax in left lung. pulmonary edema or hazy opacity in left lung. abnormal left lung. tortuous aorta in mediastinum. vascular calcification in mediastinum. abnormal mediastinum. ij line in mediastinum. endotracheal tube in neck. ij line in neck. ij line in right atrium. lung opacity in right costophrenic angle. pleural effusion in right costophrenic angle. pulmonary edema or hazy opacity in right costophrenic angle. abnormal right costophrenic angle. lung opacity in right hilar structures. pulmonary edema or hazy opacity in right hilar structures. abnormal right hilar structures. lung opacity in right lung. pleural effusion in right lung. no pneumothorax in right lung. pulmonary edema or hazy opacity in right lung. abnormal right lung. endotracheal tube in trachea. enlarged cardiac silhouette. lung opacity. pleural effusion. no pneumothorax. pulmonary edema or hazy opacity. tortuous aorta. vascular calcification. endotracheal tube. ij line **CXRFEScore:** 0.670 **RadGraph F1:** 0.061 **CheXpert Acc:** 0.857 **CheXbert Acc:** 0.857

a fact f, $\Delta\text{sim}(a, p, n) > 0$ if a and $p \in \text{S}(f)$, $n \notin \text{S}(f)$ and $\text{c}(a) \neq \text{c}(n)$ (unless $\cos(\text{E}(a), \text{E}(p)) < \cos(\text{E}(a), \text{E}(n))$ and $\text{lev}(a, p) > \text{lev}(a, n)$). Here, $\text{S}(f)$ stands for the union of f, its detailed observation, its short observation and all the paraphrases (if any) generated for all of them with ChatGPT. The intuition here is that they are all closely related semantically, as they are all derived from the same fact (see Figure 16).

Rule 4: Sample triplets according to Chest ImaGenome labels. $\Delta\text{sim}(a, p, n) > 0$ if $\text{CIGL}(a) \cap \text{CIGL}(p) \neq \emptyset$, $\text{CIGL}(a) \cap \text{CIGL}(n) = \emptyset$, $\text{CIGL}(p) \cap \text{CIGL}(n) = \emptyset$, and if $(\cos(\text{E}(a), \text{E}(p)) > \cos(\text{E}(a), \text{E}(n))$ AND $\text{levsim}(a, p) > \text{levsim}(a, n))$. Here, $\text{CIGL}(x)$ stands for the set of Chest ImaGenome labels of the sentence x. With this rule, our aim is to heuristically utilize the Chest ImaGenome labels to identify semantically similar sentences that warrant clustering together. However, we enhance this approach by incorporating signals from Levenshtein distance and BioViL-T to bolster our confidence in the triplet quality.

Rule 5: Rank triplets according to the overlap of entities and relations from RadGraph. $\Delta\text{sim}(a, p, n) > 0$ if $\text{c}(a) = \text{c}(p)$, $\text{c}(a) \neq \text{c}(n)$, and $\text{J}(\text{RG}(a), \text{RG}(p)) > \text{J}(\text{RG}(a), \text{RG}(n)) + \text{margin}_{\text{RG}}$. Here, $\text{RG}(x)$ stands for the set of RadGraph entities and relations for the sentence x, and J for Jaccard similarity. Following the same idea of the previous rule, we seek to utilize the entities and relations provided by the RadGraph dataset as valuable cues for identifying semantically similar sentences that ought to be clustered together in the embedding space.

Rule 6: Hard triplets generated by ChatGPT. $\Delta\text{sim}(a, p, n) > 0$ if (a, p, n) is a hard triplet generated by ChatGPT. The intuition behind this rule is very simple: we want to leverage ChatGPT's remarkable skills to produce challenging triplets, requiring a good understanding of the text to be ranked correctly. Figure 20 shows the prompt used to generate these triplets along with an example.

For each rule, we create approximately 3 million training triplets, 1,000 for validation, and 1,000 for testing.

A.5 Additional Metric Evaluation Details and Results

We conduct a thorough comparison involving CXR-FEScore against several established metrics, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr-D (Vedantam et al., 2015), BERTScore (Zhang et al., 2020a), CheXpert labeler (Irvin et al., 2019), CheXbert (Smit et al., 2020), and RadGraph (Jain et al., 2021a). The last three are considered domain-specific metrics, tailored to the radiology domain, while the others serve as general-purpose evaluation metrics.

Both the CheXpert labeler and CheXbert output a 14-dimensional discrete vector representing 13 observations along with a label indicating "no findings". The values in this vector denote presence (1), absence (0), uncertainty (-1), and unknown (-2). We binarize this vector, treating both presence and uncertainty as 1, and absence and unknown as 0. This enables the computation of CheXpert F1, CheXbert F1, CheXpert Accuracy, and CheXbert Accuracy.

For RadGraph, inspired by the methodology of Delbrouck et al. (2022), we utilize the pretrained entity and relation extraction model provided with the dataset. This model is applied to a text, extracting its entities and relations, both with and without their specific relation types. Subsequently, we construct a "bag" or set comprising these entities and relations. Thus, we compute RadGraph Jaccard and RadGraph F1 by directly comparing the sets obtained from real and generated texts.

In addition to the results reported in Table 5 (Section 4), in this appendix we include the plots shown in Figures 9, 10, 11, and 12.

A.6 Hardware and Other Implementation Details

All of our experiments are implemented using Python 3.10.10 with PyTorch version 1.13.1+cu117 (Paszke et al., 2017). All experiments are conducted on a computing node equipped with a 20-core Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz, three NVIDIA GPUs - two GeForce RTX 2080 Ti with 11GB memory and one GeForce RTX 3090 with 24GB memory. The system is complemented by 125GB of RAM.

We implement multitask learning for CXRFE through gradient accumulation. This is achieved by multiple model forwards, each fed by interleaved dataloaders for different tasks. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a cyclic exponential learning rate that varies from 8e-5 to 1e-6 over 8 epochs. Here, an epoch consists of roughly 800 batches. Typically, our experiments run for 12-15 hours, after which we observe no significant gains in validation metrics.

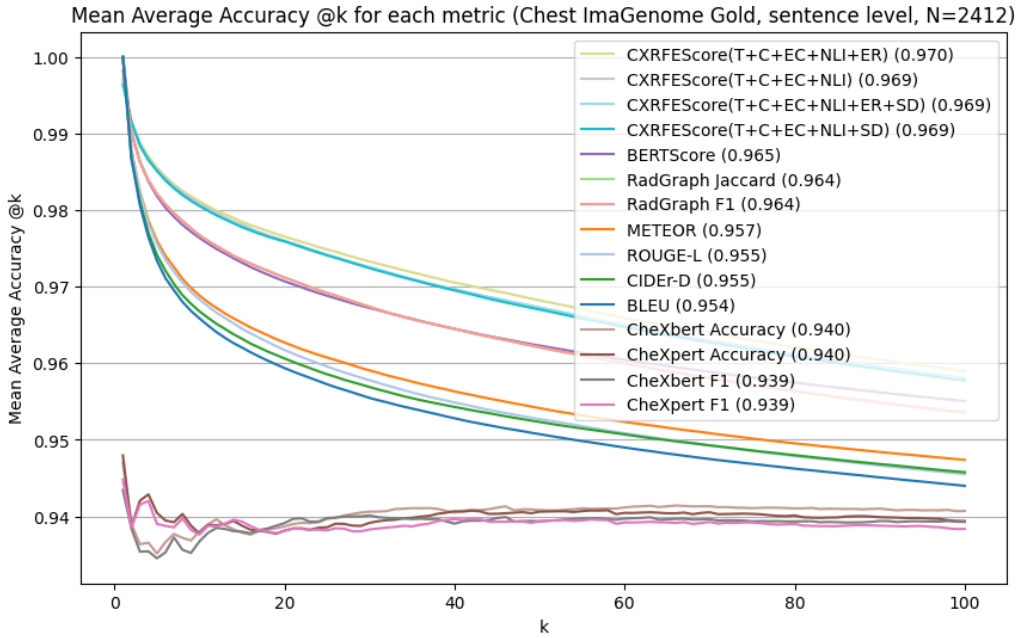

Figure 9: Mean average accuracy at k, for 2412 sentences in the gold dataset of Chest ImaGenome. Larger accuracy is better.

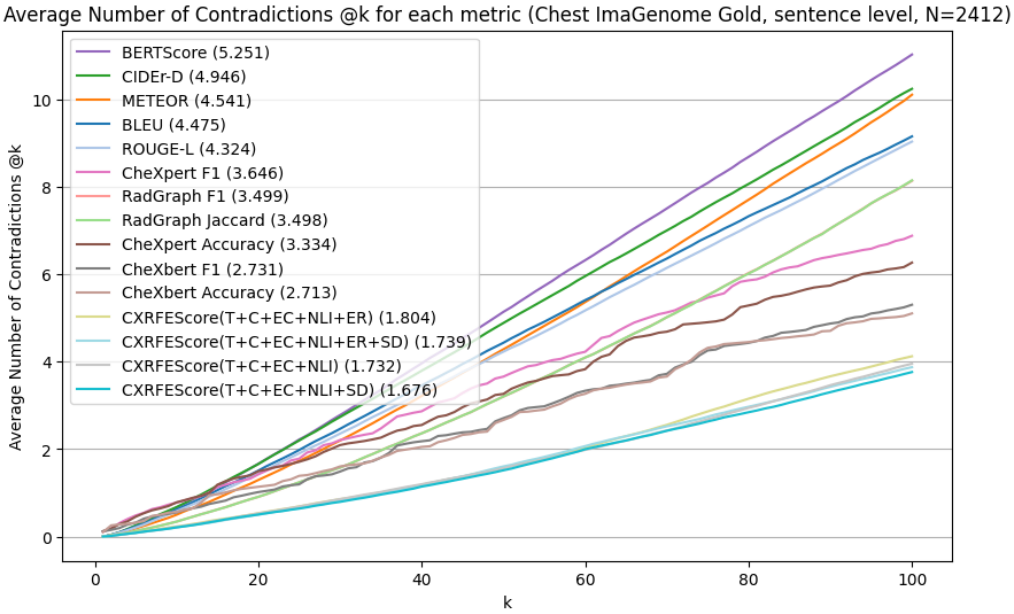

Figure 10: Average number of contradictions at k, for 2412 sentences in the gold dataset of Chest ImaGenome. Smaller number of contradictions is better.

A.7 ChatGPT prompts

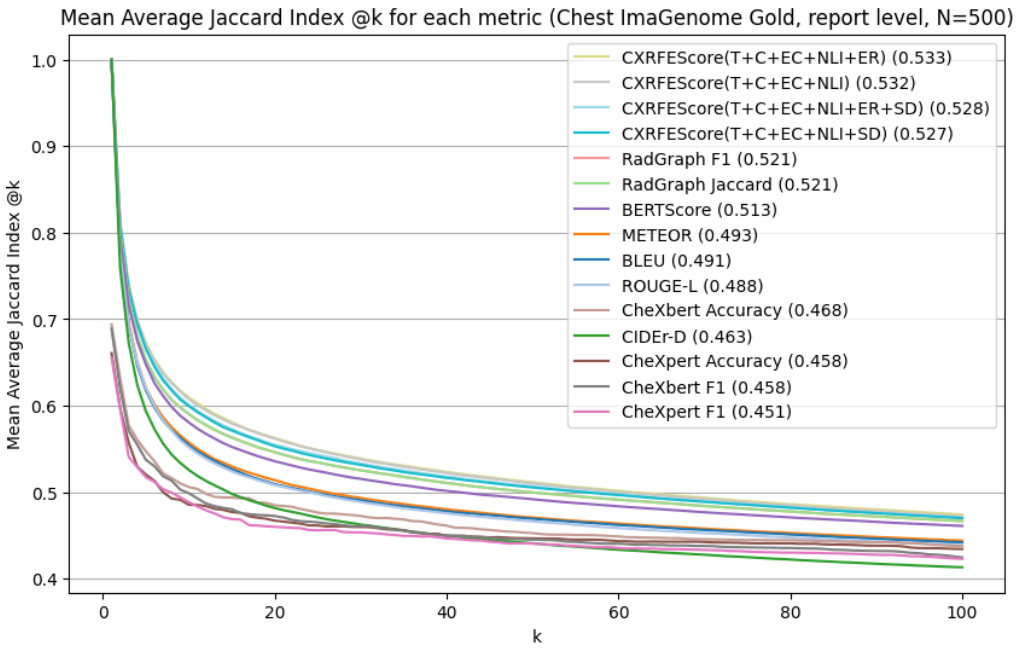

Figure 11: Mean average Jaccard Index at k, for the 500 reports in the gold dataset of Chest ImaGenome. Larger Jaccard Index is better.

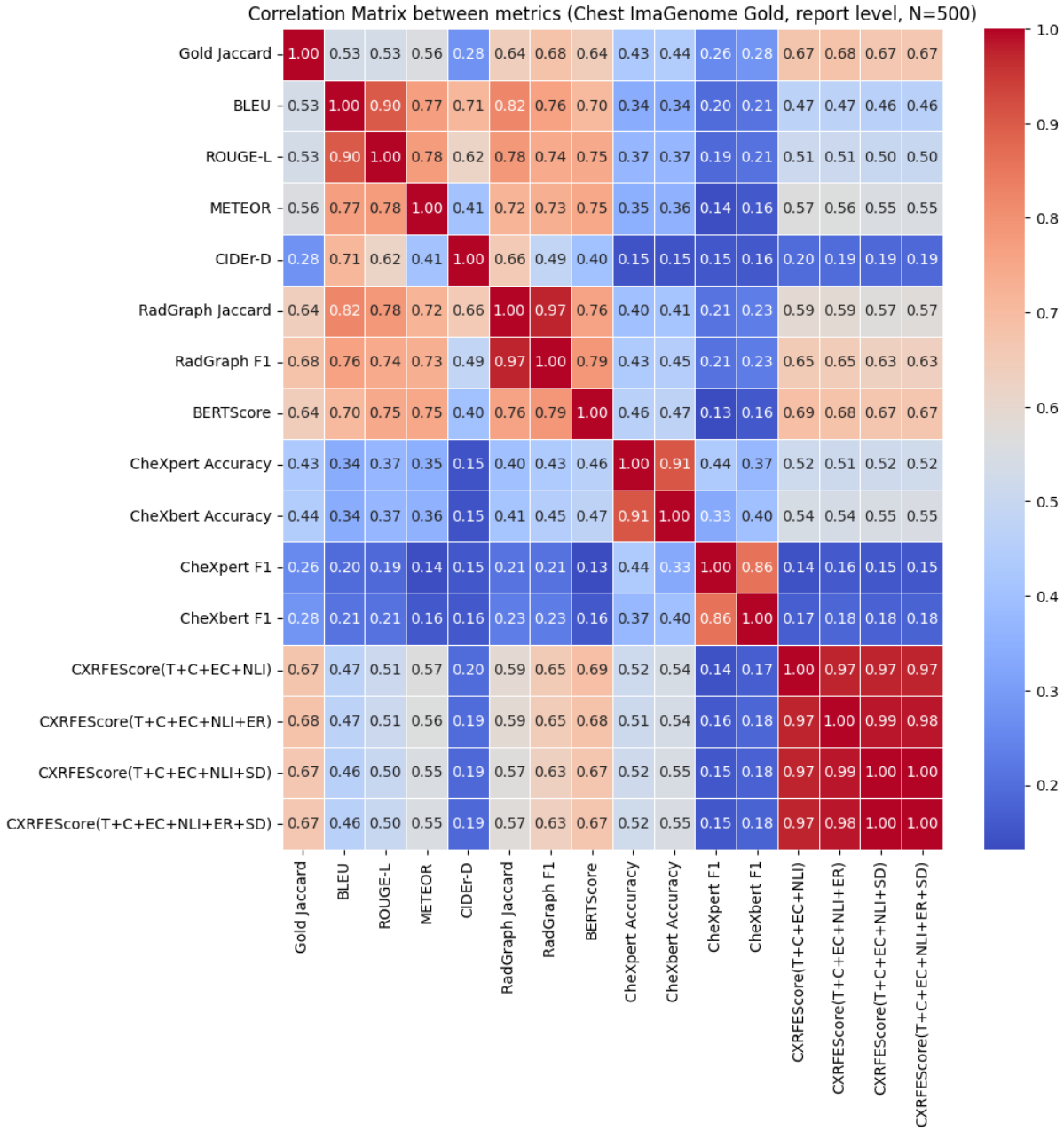

Figure 12: Metrics correlation matrix, using Pearson correlation, for the 500 reports in the gold dataset of Chest ImaGenome. The 'Gold Jaccard' metric is computed as the Jaccard Index between the gold labels of report x and report y, ignoring unmentioned observations and unmentioned anatomical locations. We do this to focus on the overlap of those elements that are explicitly mentioned by the labels.

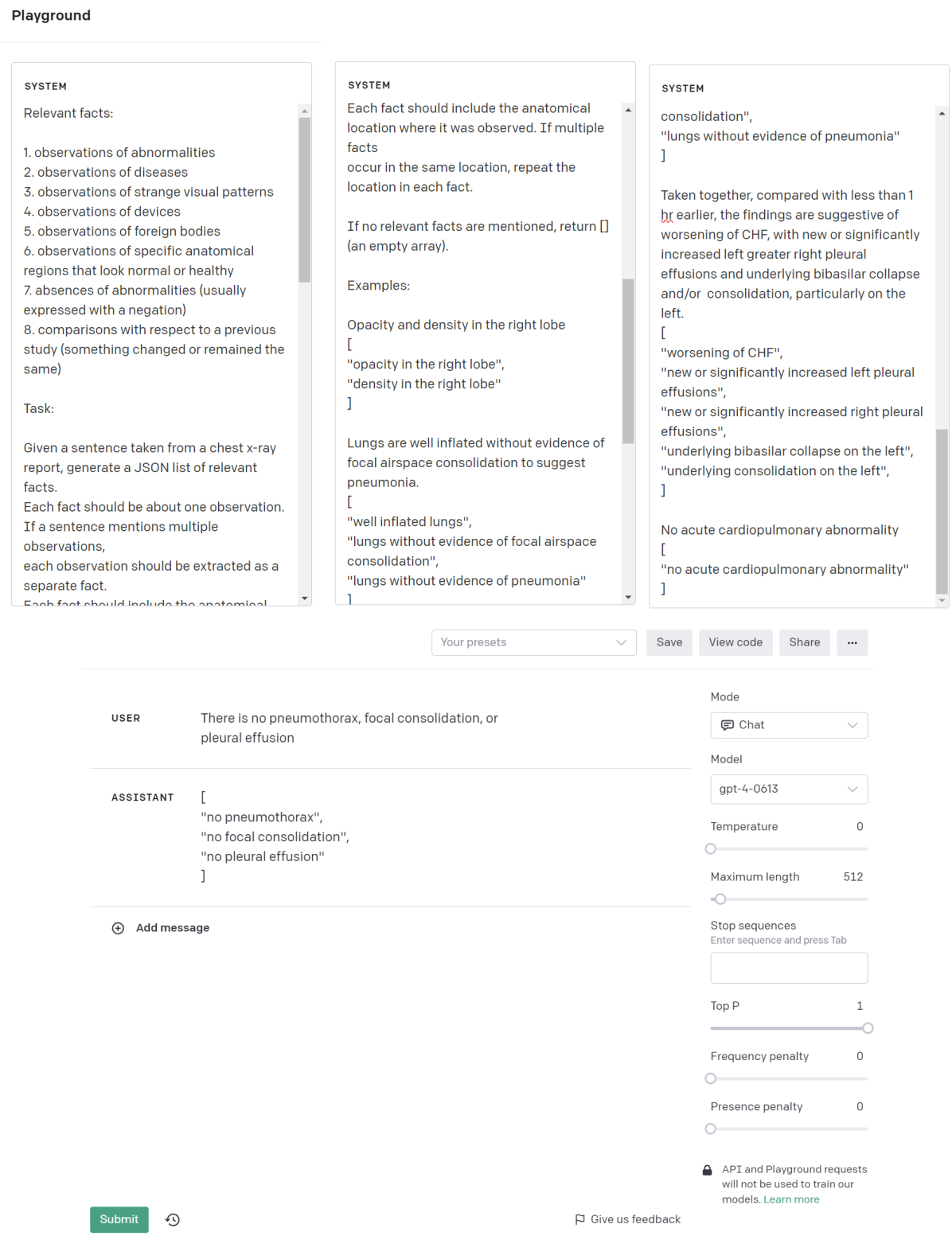

Figure 13: Screenshot of OpenAI's playground web interface showing the prompt used to extract facts from chest X-ray report sentences. The screenshot was manipulated in order to display the full instructions given to ChatGPT.

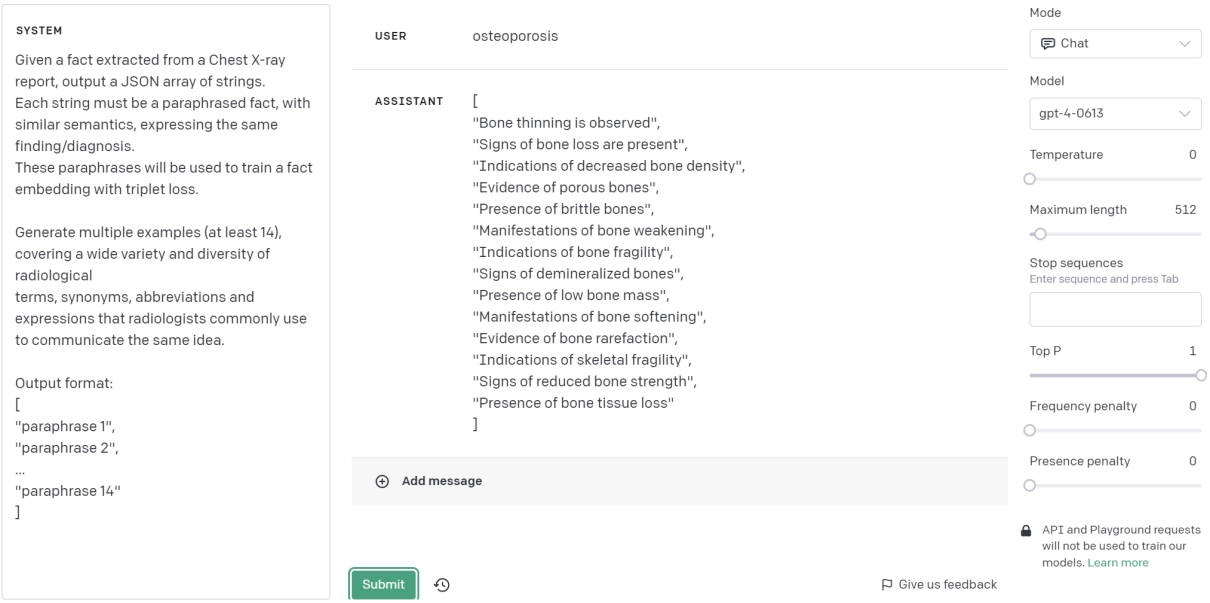

Figure 14: ChatGPT prompt. Fact to paraphrases

SYSTEM

Given a sentence referring to an anatomical location in the context of a chest X-ray, output a JSON array of strings paraphrasing it, covering a wide diverse range of terminology, synonyms and abbreviations that radiologists commonly use to express the same idea in a chest X-ray report.

Examples:

gastroesophageal junction
[
"cardia",
"cardia of the stomach",
"cardiac orifice",
"esophagogastric junction",
"GE junction",
"esophagogastro junction",
"esophagogastric junctional region",
"gastroesophageal interface",
"gastroesophageal transition zone",
"lower esophageal junction",
"esophagogastric transition",
"stomach-esophagus interface",

SYSTEM

"junction between stomach and esophagus",
"cardioesophageal junction",
"junction of the esophagus and stomach",
"transition between the esophagus and stomach",
"point where the esophagus meets the stomach",
"Z-line",
"distal end of the esophagus",
"proximal start of the gastric cardia"
]

costodiaphragmatic recess
[
"costophrenic angle",
"costophrenic recess",
"costophrenic sulcus",
"costophrenic notch",
"costophrenic sinus",
"phrenicocostal sinus",
"diaphragm-chest wall interface",
"junction of diaphragm and chest wall",
"area where the diaphragm meets the ribs",

SYSTEM

"costal margin of the diaphragm",
"angle formed by diaphragm and rib cage"
]

R>L
[
"right greater than left"
]

base
[
"lung base",
"base of the lung",
"lower lung area",
"basal lung region",
"basal portion of the lung",
"lung's basal part",
"bottom part of the lung",
"lower region of the lung",
"pulmonary base",
"base of the pulmonary structure",
"lower zone of the lung",
"lower pulmonary region",
"lower segment of the lung"
]

USER cardiomediastinal silhouette

ASSISTANT
[
"heart and mediastinal contour",
"cardiac and mediastinal silhouette",
"outline of the heart and mediastinum",
"cardiac and mediastinal shadow",
"cardiopulmonary silhouette",
"heart and mediastinal structure",
"cardiac and mediastinal shape",
"cardiac and mediastinal border",
"cardiac and mediastinal profile",
"cardiac and mediastinal image",
"cardiac and mediastinal area",
"cardiac and mediastinal configuration",
"cardiac and mediastinal pattern",
"cardiac and mediastinal appearance",
"cardiac and mediastinal contouring",
"cardiac and mediastinal form",
"cardiac and mediastinal outline",
"cardiac and mediastinal figure",
"cardiac and mediastinal boundary"
]

Mode
Chat

Model
gpt-3.5-turbo-0613

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. Learn more

Submit

Figure 15: ChatGPT prompt. Anatomy location to paraphrases

SYSTEM

Given a medical fact, output a JSON object with 7 fields:

1. "anatomical location"
2. "detailed observation"
3. "short observation"
4. "category"
5. "health status"
6. "prev_study_comparison?"
7. "comparison status"

Definitions:

1. "anatomical location" means the anatomical location of the observation. If not given, it should be "".

2. "detailed observation" means what was observed, with all the details provided (but excluding location, unless the location itself is the main observation).

3. "short observation" should be a brief summary of the previous field, without extra details, but logically consistent.

4. "category" can be "anatomical finding", "disease", "technical assessment", "tubes and lines" or "device".

SYSTEM

5. "health status" can be "normal", "abnormal", "ambiguous" or "unknown".
"normal" means healthy.
"abnormal" means unhealthy.
"ambiguous" means it is not clear whether it is healthy or unhealthy, leaving room for interpretation.
"unknown" means the observation provides no information about abnormality/normality.

6. "prev_study_comparison?" is a boolean field. Can be either "yes" or "no". If the fact is comparing the current study with a previous study (something improved, worsened, changed or remained the same), return "yes". Otherwise, return "no".

7. "comparison status" can be "resolved", "new", "improved", "worsened", "larger", "smaller", "displaced", or "". If no comparison with a previous study is made, the default is "" (empty).

Examples:

small to moderate size left pleural effusion
{
"anatomical location": "left",
"detailed observation": "small to moderate size pleural effusion",
"short observation": "pleural effusion",

SYSTEM

"category": "anatomical finding",
"health status": "abnormal",
"prev_study_comparison?": "no",
"comparison status": ""
}

new finding of mass in the abdomen
{
"anatomical location": "abdomen",
"detailed observation": "mass",
"short observation": "mass",
"category": "anatomical finding",
"health status": "abnormal",
"prev_study_comparison?": "yes",
"comparison status": "new"
}

clear lung fields
{
"anatomical location": "lung fields",
"detailed observation": "clear lung fields",
"short observation": "clear lung fields",
"category": "anatomical finding",
"health status": "normal",
"prev_study_comparison?": "no",
"comparison status": ""
}

no ill-definition of right hemidiaphragm

USER bilateral effusions equivocally getting worse

ASSISTANT
{
"anatomical location": "bilateral",
"detailed observation": "effusions equivocally getting worse",
"short observation": "worsening effusions",
"category": "anatomical finding",
"health status": "abnormal",
"prev_study_comparison?": "yes",
"comparison status": "worsened"
}

Mode: Chat
Model: gpt-4-0613
Temperature 0
Maximum length 512
Stop sequences
Enter sequence and press Tab

Figure 16: ChatGPT prompt. Fact to metadata

SYSTEM

Given a statement extracted from a chest X-ray report, output a comparison category. The category must be one item of the following list:

0. no comparison
1. new finding
2. resolved
3. improved
4. worsened
5. progressed
6. reappeared
7. larger
8. smaller
9. increase
10. decrease
11. position changed
12. stable/unchanged
13. unclear comparison
14. other

Examples:

right pleural effusion similar in size
12. stable/unchanged

pleural density on the left is somewhat increased
9. increase

new bibasal consolidations on the left

SYSTEM

6. reappeared
7. larger
8. smaller
9. increase
10. decrease
11. position changed
12. stable/unchanged
13. unclear comparison
14. other

Examples:

right pleural effusion similar in size
12. stable/unchanged

pleural density on the left is somewhat increased
9. increase

new bibasal consolidations on the left
1. new finding

small foci of opacity in the left mid lung
0. no comparison

no ill-definition of right hemidiaphragm
0. no comparison

heart within upper limits of normal in size
0. no comparison

USER bilateral effusions equivocally getting worse

ASSISTANT 4. worsened

Add message

Mode

Chat

Model

gpt-4-0613

Temperature 0

Figure 17: ChatGPT prompt. Fact to comparison status

SYSTEM

Instructions:

Apply an annotation standard of observations to raw phrases extracted from chest X-ray reports.
The standard considers the following observations:

airspace opacity
atelectasis
bone lesion
bronchiectasis
calcified nodule
clavicle fracture
consolidation
costophrenic angle blunting
cyst/bullae
diaphragmatic eventration (benign)
elevated hemidiaphragm
enlarged cardiac silhouette
enlarged hilum
hernia
hydropneumothorax
hyperaeration
increased reticular markings/ild pattern
infiltration
linear/patchy atelectasis
lobar/segmental collapse
lung lesion
lung opacity
mass/nodule (not otherwise specified)

SYSTEM

mediastinal displacement
mediastinal widening
multiple masses/nodules
pleural effusion
pleural/parenchymal scarring
pneumomediastinum
pneumothorax
pulmonary edema/hazy opacity
rib fracture
scoliosis
shoulder osteoarthritis
spinal degenerative changes
spinal fracture
sub-diaphragmatic air
subcutaneous air
superior mediastinal mass/enlargement
tortuous aorta
vascular calcification
vascular congestion
vascular redistribution
aortic graft/repair
cabg grafts
cardiac pacer and wires
prosthetic valve
alveolar hemorrhage
aspiration
copd/emphysema
fluid overload/heart failure
goiter

SYSTEM

granulomatous disease
interstitial lung disease
lung cancer
pericardial effusion
pneumonia
artifact
breast/nipple shadows
low lung volumes
rotated
skin fold
chest port
chest tube
endotracheal tube
enteric tube
ij line
intra-aortic balloon pump
mediastinal drain
picc
pigtail catheter
subclavian line
swan-ganz catheter
tracheostomy tube

Output format:

JSON array of strings

Rules:

SYSTEM

intra-aortic balloon pump
mediastinal drain
picc
pigtail catheter
subclavian line
swan-ganz catheter
tracheostomy tube

Output format:

JSON array of strings

Rules:

Only output the observations from the standard that best match or are supported by the phrase. If none match the phrase, output an empty array ([]).

Example:
Increased bibasilar opacities likely representing atelectasis, but cannot exclude aspiration or pneumonia in the
correct clinical setting
[
"lung opacity",
"atelectasis",
"aspiration",
"pneumonia"
]

USER dual chamber pacemaker

ASSISTANT [
"cardiac pacer and wires"
]

Add message

Mode: Chat
Model: gpt-4-0613
Temperature 0
Maximum length 512

Figure 18: ChatGPT prompt. Fact to Chest ImaGenome observations

SYSTEM

Instructions:

Apply an annotation standard of anatomical locations to raw phrases extracted from chest X-ray reports.
The standard considers the following anatomical locations:

abdomen
aortic arch
cardiac silhouette
carina
cavoatrial junction
left apical zone
left arm
left breast
left chest wall
left clavicle
left costophrenic angle
left hemidiaphragm
left hilar structures
left lower lung zone
left lung
left mid lung zone
left shoulder
left upper lung zone
mediastinum
neck
right apical zone

SYSTEM

right arm
right atrium
right breast
right chest wall
right clavicle
right costophrenic angle
right hemidiaphragm
right hilar structures
right lower lung zone
right lung
right mid lung zone
right shoulder
right upper lung zone
spine
svc
trachea
upper mediastinum

Output format:

JSON array of strings

Rules:

Only output the anatomical locations from the standard that best match the given phrase.
If none match the phrase, output an empty array ([]).

SYSTEM

Only output the anatomical locations from the standard that best match the given phrase.
If none match the phrase, output an empty array ([]).

Examples:

cardiomediastinal silhouette
[
"cardiac silhouette",
"mediastinum"
]

A tracheostomy tube is in place
[
"neck",
"trachea"
]

bilateral pulmonary nodules
[
"left lung",
"right lung"
]

left IJV
[
"neck"
]

USER dual chamber pacemaker

ASSISTANT [
"right atrium",
"right ventricle"
]

Add message

Mode: Chat
Model: gpt-4-0613
Temperature 0
Maximum length 512

Figure 19: ChatGPT prompt. Fact to Chest ImaGenome anatomical locations

SYSTEM

Given an anchor, output a list of positives and a list of negatives.
The anchor will come from a chest X-ray report. The positives must be 6 paraphrases of the anchor, expressing the same idea with different terms and synonyms (cover a wide range of medical terminology).
The negatives must be 10 hard negatives, that is, similar to the anchor but semantically different, for example, by expressing a different diagnosis (if applicable), by referring to a different anatomical location (if applicable), etc.

Output format: a JSON object as follows
{
"positives": [...],
"negatives": [...]
}

USER removal of endotracheal tube

ASSISTANT
```
{
"positives": [
"Endotracheal tube has been removed",
"Extraction of the endotracheal tube",
"Endotracheal tube has been taken out",
"Endotracheal tube has been extracted",
"Endotracheal tube removal has been performed",
"Discontinuation of the endotracheal tube"
],
"negatives": [
"Insertion of endotracheal tube",
"Removal of nasogastric tube",
"Removal of chest tube",
"Endotracheal tube is in place",
"Endotracheal tube repositioned",
"Removal of central venous catheter",
"Endotracheal tube is unchanged",
"Removal of tracheostomy tube",
"Endotracheal tube is malpositioned",
"Removal of urinary catheter"
]
}
```

Mode: Chat
Model: gpt-4-0613
Temperature 0
Maximum length 512
Stop sequences
Enter sequence and press Tab
Top P 1
Frequency penalty 0
Presence penalty 0

API and Playground requests will not be used to train our models. Learn more

Figure 20: ChatGPT prompt. Hard triplets

SYSTEM

Given a premise from a Chest X-ray report, output an entailment, a contradiction and a neutral sentence.
The entailment must strictly follow logically from the premise. The neutral can optionally be about a different chest X-ray observation.
The output format must be a JSON as follows:
{
"e": ...,
"c": ...,
"n": ...
}

USER removal of endotracheal tube

ASSISTANT
```
{
"e": "The patient no longer has an endotracheal tube.",
"c": "The endotracheal tube is still in place.",
"n": "The patient's lungs appear clear."
}
```

Add message

Mode: Chat
Model: gpt-4-0613
Temperature 0
Maximum length 512
Stop sequences
Enter sequence and press Tab

Figure 21: NLI Prompt 1. Given a premise, output entailment, contradiction, and neutral hypotheses.

SYSTEM

You will receive a NLI example from a chest X-ray dataset with premise, hypothesis and label ("entailment", "contradiction" or "neutral"). Generate 10 new NLI examples with the same label and following a similar writing style as the example provided. They have to be about chest X-ray reports. Output the examples as a JSON array of objects. If the premise or the hypothesis (or both) are complex (i.e. they say several things), create challenging examples following a similar structure as the provided example that require a good understanding of logic in order to deduce the label.
Avoid verbosity and words such as "patient" or "X-ray". State observations succinctly.

USER {"P": "No pleural effusions or pneumothorax.", "H": "Previously seen pneumothorax is no longer visualized.", "L": "entailment"}

ASSISTANT
```
[
 {"P": "No evidence of pulmonary edema or pneumonia.", "H": "The previously noted pneumonia is not present.", "L": "entailment"},
 {"P": "No signs of lung nodules or masses.", "H": "The lung mass seen before is not visible now.", "L": "entailment"},
 {"P": "No visible signs of cardiomegaly or mediastinal widening.", "H": "The previously observed cardiomegaly is not seen.", "L": "entailment"},
 {"P": "No signs of pulmonary embolism or pleural effusion.", "H": "The earlier detected pleural effusion is not visible.", "L": "entailment"},
 {"P": "No evidence of pneumothorax or lung collapse.", "H": "The previously seen lung collapse is not present.", "L": "entailment"},
 {"P": "No signs of rib fractures or lung contusions.", "H": "The rib fracture noted before is not visible now.", "L": "entailment"},
 {"P": "No visible signs of bronchiectasis or lung fibrosis.", "H": "The previously observed bronchiectasis is not seen.", "L": "entailment"},
 {"P": "No signs of lung cancer or tuberculosis.", "H": "The earlier detected tuberculosis is not visible.", "L": "entailment"},
 {"P": "No evidence of aortic dissection or aneurysm.", "H": "The previously seen aortic aneurysm is not present.", "L": "entailment"},
 {"P": "No signs of pericardial effusion or heart failure.", "H": "The
```

Mode: Chat
Model: gpt-4-0613
Temperature 0
Maximum length 512
Stop sequences
Enter sequence and press Tab
Top P 1
Frequency penalty 0
Presence penalty 0

API and Playground requests will not be used to train our models. Learn more

Figure 22: NLI Prompt 2. Given a ground-truth NLI example, generate multiple similar examples.

SYSTEM

Context: natural language inference.

Given a premise and a hypothesis, output "entailment", "contradiction", or "neutral".

Use "entailment" when the facts stated by the premise necessarily entail the truth of the hypothesis.

Use "contradiction" when premise and hypothesis are mutually exclusive/contradictory (both cannot be true at the same time).

Use "neutral", if there is no contradiction (premise and hypothesis are compatible), but the premise does not entail the hypothesis (it's possible for the premise to be true and the hypothesis still be false). In other words, use "neutral" when neither "entailment" nor "contradiction" adequately fit.

USER Premise: Lungs are well-expanded and clear without focal consolidation concerning for pneumonia. | Hypothesis: Lungs are hyperinflated but clear.

ASSISTANT Neutral

USER Enter a user message here.

Add message

Mode: Chat

Model: gpt-4-0613

Temperature 0

Maximum length 10

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Figure 23: NLI Prompt 3. Given a premise and a hypothesis, output the correct label. In this version, GPT-4 has to output the label immediately.

SYSTEM

Context: natural language inference.

Given a premise (#P) and a hypothesis (#H), output "Reason: {reason}. Label: {label}" where {reason} is a short sentence and {label} is one of "entailment," "contradiction," or "neutral."

Use "entailment" when the premise necessarily entails the truth of the hypothesis.

Use "contradiction" when premise and hypothesis are mutually exclusive/contradictory. Pay attention to logical inconsistencies, such as expressions suggesting presence vs. absence, etc.

Use "neutral" when there's no contradiction, but the premise doesn't necessarily entail the hypothesis.

Examples:

1. #P: increased pulmonary edema. | #H: worsened pulmonary edema.
Label: entailment

2. #P: No pulmonary edema, consolidation, or pneumothorax. | #H: No focal consolidation, pleural effusion, or pneumothorax is present.
Label: neutral

USER #P: Increased opacity is seen in the left peritracheal region and left lung base | #H: There are signs of abnormality in the left lung base and left peritracheal region

ASSISTANT Reason: The premise states that there is increased opacity in the left peritracheal region and left lung base, which is a sign of abnormality. Therefore, the hypothesis is a rephrase of the premise. Label: entailment.

Add message

Model: gpt-4-0613

Temperature 0

Maximum length 256

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. Learn more

Submit

Give us feedback

Figure 24: NLI Prompt 4. Given a premise and a hypothesis, output the correct label. In this version, GPT-4 is given the chance to reason first and then output the label. It is also shown a couple of examples with the expected labels. These simple changes to the prompt significantly boosted GPT-4's performance.

SYSTEM

Given a premise from a Chest X-ray report, output 5 statements that explicitly logically contradict the premise, that is, the premise and any generated statement cannot both be true at the same time. Include at least one sentence that is very similar to the premise but with a very slight difference, thus creating a logical contradiction.
For example:
- "There is mid thoracic dextroscoliosis" vs. "There is no mid thoracic dextroscoliosis"
- "Left basal consolidation has slightly improved" vs. "Left basal consolidation has slightly increased"
Output format: a JSON array of strings.

USER Lungs are well-expanded and clear without focal consolidation concerning for pneumonia.

ASSISTANT ["Lungs are not well-expanded and clear without focal consolidation concerning for pneumonia.", "Lungs are well-expanded but show focal consolidation concerning for pneumonia.", "Lungs are collapsed and clear without focal consolidation concerning for pneumonia.", "Lungs are well-expanded and clear with focal consolidation not concerning for pneumonia.", "Lungs are well-expanded and clear, but there is focal consolidation concerning for pneumonia."]

Add message

Mode: Chat

Model: gpt-4-0613

Temperature 0

Maximum length 500

Stop sequences
Enter sequence and press Tab

Top P 1

Figure 25: NLI Prompt 5. Given a premise, generate diverse hypotheses that contradict it.